Which questions should I answer?
Salience Prediction of Inquisitive Questions

Yating Wu*[1], Ritika Mangla*[2], Alexandros G. Dimakis [1,4], Greg Durrett[2], Junyi Jessy Li[3]
[1]Electrical and Computer Engineering, [2]Computer Science, [3]Linguistics
The University of Texas at Austin
[4]BespokeLabs.ai
{yating.wu, jessy}@utexas.edu

Abstract

Inquisitive questions — open-ended, curiosity-driven questions people ask as they read — are an integral part of discourse processing (Van Kuppevelt, 1995; Onea, 2016; Kehler and Rohde, 2017) and comprehension (Prince, 2004). Recent work in NLP has taken advantage of question generation capabilities of LLMs to enhance a wide range of applications. But the space of inquisitive questions is vast: many potential questions can be evoked from a given context. So which of those should be prioritized to find answers? Linguistic theories, unfortunately, have not yet provided an answer. This paper presents QSALIENCE, a salience predictor of inquisitive questions. QSALIENCE is instruction-tuned over our dataset of linguist-annotated salience scores of 1,766 (context, question) pairs. A question scores high on salience if answering it would greatly enhance the understanding of the text (Van Rooy, 2003). We show that highly salient questions are empirically more likely to be answered in the same article, bridging potential questions (Onea, 2016) with Questions Under Discussion (Roberts, 2012). We further validate our findings by showing that answering salient questions is an indicator of summarization quality in news.

1 Introduction

Asking questions is the natural language manifestation of human inquisitiveness: we insist on getting answers for what we are curious about since childhood (Chouinard et al., 2007). Acquired strategies of question generation have a profound impact on education (Davey and McBride, 1986; Prince, 2004). In linguistics, both theoretical and psycholinguistic work argued that readers generate inquisitive questions, seeking information in a conversation or as they read (Van Kuppevelt, 1995; Ginzburg, 1996; Onea, 2016; Kehler and Rohde, 2017). In NLP, pre-trained models have enabled the

generation of these *open-ended, curiosity-driven, information-seeking* questions, leading to a flourish of recent work: identifying information loss between two texts (Trienes et al., 2024; Cole et al., 2023), analyzing the diversity of news perspectives (Laban et al., 2022), generating elaborations or explanations (Wu et al., 2023b; Fok et al., 2023), evaluating summaries (Pratapa et al., 2023), asking follow-up questions (Meng et al., 2023), decontextualization (Newman et al., 2023), and planning (Narayan et al., 2023).

However, the space of possible inquisitive questions is vast. Prior work (Ko et al., 2020; Westera et al., 2020) showed that many distinct questions can be evoked from a given context, yet not all questions are equally good for an application. In theoretical linguistics, this also brings up a long-standing gap in understanding how discourse progresses (Warstadt, 2020): some of such inquisitive "potential questions" (as named in Onea (2016)) are likely more pertinent than others. Some of these questions may be answered (by the writer) later in the article and thus become Questions Under Discussion (QUDs) (Roberts, 2012). Evidence in psycholinguistics indicate that readers form expectations how a discourse would progress (Kehler and Rohde, 2017), providing a foundation for the predictability of QUDs (Westera et al., 2020). Van Rooy (2003) argues that a question is important if answering it provides high utility. However, there is so far no computational work to predict whether or not those questions should be answered or how salient they are.

This work (Figure 1) seeks to answer these questions by training a salience prediction model for inquisitive questions, using a new linguist-annotated corpus. In line with Van Rooy (2003), a question scores high on salience if answering it would greatly enhance the understanding of the text. First, we collected validity and salience ratings of 1,766 inquisitive questions over English news articles
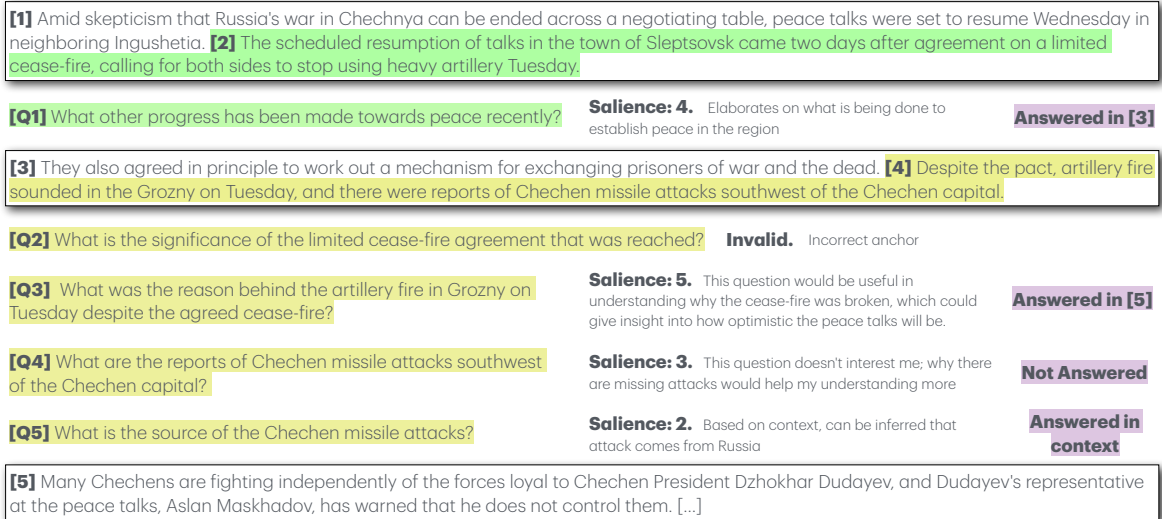
---

*Yating and Ritika contributed equally.

Figure 1: Examples of inquisitive questions and their annotated salience (with rationales). Each question is evoked by an anchor sentence (shown in the same highlight color). Whether the question is answered is shown on the right. Q1 is taken from human-annotated QUDs in DCQA (Ko et al., 2022); Q2-5 are GPT-4 generated questions.

(Ko et al., 2022; Huang et al., 2024) and TED talks (Westera et al., 2020). A subset of these questions were also annotated in terms of "answerability", i.e., how well they were answered by the same article. Not only do our annotators largely agree with each other on their salience ratings, these ratings also correlate with answerability. Furthermore, human-generated QUDs from Ko et al. (2022), whose answers were guaranteed to be present in the article, also received high salience ratings. These empirical findings support the hypothesis that there is, to some degree, a "common" notion of question salience—capturing reader expectations—that connects to answer utility (Van Rooy, 2003).

We then present QSALIENCE, instruction-tuned from open-sourced LLMs that predict salience ratings given an article context. QSALIENCE is the first application-agnostic question salience model, and outperforms GPT-4 under zero-shot, few-shot, and Chain-of-Thought prompting (Wei et al., 2024) when evaluated on our dataset. Encouragingly, even much smaller models such as Flan-T5 (Chung et al., 2024) achieves significant correlations with human annotated salience. Our experiments show the utility of instruction-tuning on long-context discourse tasks that capture implicit information arising from the cognitive processing of text.

Finally, we take a first step to investigate the value of question salience prediction in a downstream task, where a short TL;DR of a news article is expanded to a 240-word summary. We show that summaries that answer more salient inquisitive

questions from the TL;DRs are also ranked higher by human readers.

QSALIENCE is available at https://github.com/ritikamangla/QSalience.

2 Background

Theory: Potential Questions and QUDs There are two concepts in linguistics that are relevant to inquisitive questions discussed in this work. First, Onea (2016) define the semantics of "potential questions"; informally, a question Q evoked (or licensed) by an utterance u in a given context c such that the answer space of Q is entailed by the common ground of $\{c, u\}$, and that c alone does not license Q. Second, the pragmatics theory of Question Under Discussion (QUD) views discourse as progressively answering questions that were explicitly or implicitly generated in prior context (Van Kuppevelt, 1995; Roberts, 2012; Benz and Jasinskaja, 2017). Under the QUD view, a potential question Q' is a QUD if it is answered by an utterance u' where u' is not part of the common ground but is entailed by it (e.g., an upcoming utterance later in an article). In Figure 1, Q1 evoked by sentence 2 is the QUD of sentence 3, and Q3 evoked by sentence 4 is the QUD of sentence 5. Some QUDs evoked earlier in an article can be answered much later (Ko et al., 2022).

The salience of potential questions and its connection with QUDs, however, is understudied. Onea (2016) listed several hypotheses for the ordering of potential questions, but acknowledged

that they are limited and presented no formal or empirical validation. Kehler and Rohde (2017)'s psycholinguistic experiments showed that people form *expectations* of what QUDs are upcoming using contextual cues, and that such expectations affect their interpretation of linguistic forms. This is compelling evidence for the incremental processing of discourse (Altmann and Steedman, 1988; Cristea and Webber, 1997) and why people ask questions. Westera et al. (2020) later studied how likely a potential question is answered, using the TED-Q corpus that annotates both questions and answers in a (limited) moving window of context. Yet this work focuses on the predictability of QUDs, rather than a reader-centric view of salience as in ours. Salience defined in our work is in line with Van Rooy (2003)'s information-theoretic argument that questions are salient when information utility of the answer is high; yet empirical evidence at-scale is yet to be seen.

Applications: Generating Inquisitive Questions Prior work developed datasets and models for generating inquisitive questions (defined as open-ended high-level questions targeting discourse understanding) (Ko et al., 2020; Gao et al., 2022), which was later used in a range of applications (Laban et al., 2022; Wu et al., 2023b; Fok et al., 2023; Newman et al., 2023; Trienes et al., 2024). However, this existing work does not explicitly define or model question salience. In QUD parsing, prior work focuses on what makes questions linguistically felicitous QUDs (Riester et al., 2018; Wu et al., 2023a).

A question salience model, however, is often necessary in downstream applications. For instance, in elaborative simplification (Wu et al., 2023b), the lack of a salience model means that existing approaches cannot predict which concepts to insert explanations for. Indeed, the over-generation of valid, fine-grained inquisitive questions is undesirable and can easily overwhelm the readers (Trienes et al., 2024). In goal-oriented forums, Rao and Daumé III (2018) calculated information utility from the answers to rank clarification questions; however this presupposes an explicit discourse goal to solve a specific task. While domain-specific notions of salience can sometimes be implicitly captured in end-to-end training with a downstream gold-standard (e.g., in summarization planning (Narayan et al., 2023)), it does not apply to most prior work mentioned above, as they are more open-ended.

3 Task Definition

A (human or machine) reader is reading a document, with established context C_p (preceding context) consisting of sentences $\{1, ..., S_{k-1}\}$. The reader generates a potential question (Section 2; Onea (2016)) Q evoked at sentence S_k (also called the "anchor sentence" (Wu et al., 2023a)).

The salience of Q is the measure of the extent to which it is important for a question Q to be answered, in order to gain a fuller understanding of the situation described, after its invocation at sentence S_k (Van Rooy, 2003). Specifically, for all valid questions, we define a Likert scale of 1-5 (full definitions found in Appendix A):

- Score = 1: Q is not related to C_p
- Score = 2: Q is related to C_p but seems to be a stretch to ask, and answering it is not useful.
- Score = 3: Q is related to C_p but whether it is answered does not matter much to the reader.
- Score = 4: Q is related to C_p and answering it might clarify some newly introduced concepts, or might expand on an idea already introduced.
- Score = 5: Q is related to C_p and it should definitely be answered as it clarifies a concept introduced in S_k or asks for more information about newly introduced humans (or animated) individuals into the discourse.

A question is *invalid* if it contains grammatical or factual errors, or is not anchored in S_k. The last criteria follows linguistic constraints in Wu et al. (2023a) reflecting that the content of Q is not grounded in S_k, hence should not be evoked at S_k.

A note on subjectivity. The salience values are, to some degree, subjective. However, prior work has shown compelling evidence that certain QUDs are more predictable than others (Westera et al., 2020) and that linguistic cues in the text play a significant role in readers' expectation (Kehler and Rohde, 2017). Under the assumption that there isn't too much divergence between the authors' intended audience and the background of the actual readers, our work sets out to capture such expectations through a question salience score.

4 Data Collection

We first present QSALIENCE-*data*, a corpus of 1,766 inquisitive questions annotated with salience,

plus natural language rationales for their judgments. Although question generation has been used widely, application-independent datasets consisting of human-generated inquisitive questions are scarce. Thus, we generate questions with LLMs, both to obtain a sizable amount of data and also to understand inquisitive question generation capabilities of LLMs. We supplement these questions with a smaller number of human-generated questions from prior work, which allows us to perform deeper analysis (Section 5).

4.1 Source Texts and Questions

Table 1 summarizes the number of source articles and questions in QSALIENCE-*data*. We draw these from different existing corpora to support different facets of our experimental analysis. They are from:

(1) News texts from DCQA (Ko et al., 2022).[1] We generate questions from DCQA articles, with gradually increasing size of C_p. Additionally, the annotated QUDs in DCQA allows us to study the salience of QUDs compared to inquisitive questions in general (Section 5).

(2) TED talks from TED-Q (Westera et al., 2020). In addition to LLM-generated questions, we also annotate the salience of one of the 6 excerpts with human generated questions in the TED-Q dataset. This provides data for further analysis on question salience vs. how answerable they are (Section 5).

(3) DiverseSumm (Huang et al., 2024) contains a newer set of news articles for which we annotate salience of LLM-generated questions. For convenience, we denote this subset as *Div. Article*. These are source articles for Section 7, our downstream task. To ensure fair evaluation for the task, the articles we selected were all roughly 1,500 words. Additionally, we annotate question salience on a set of GPT-4 generated short TL;DRs for these articles. We denote this subset *Div. TL;DR*.

Machine Generated Questions Given the preceding context C_p along with the anchor sentence S_k, we prompt LLMs to generate 5 questions about a part of the sentence that a reader may be curious about (settings and prompt in Appendix C.1). Multiple LLMs were used to cover a more diverse set of question styles in the dataset. Specifically, 250 questions were generated from Llama-2-7B-chat,

249 from Mistral-7B-instruct, 100 from GPT-3.5-turbo, and 1,106 from GPT-4-turbo.

For full articles, we begin the question generation process from the 4th sentence until the 16th sentence, maintaining a gap of two sentences between consecutive question generation probes, similar to Westera et al. (2020). For the DiverseSumm TL;DRs which are typically 3 sentences long (50 words), we generate questions per sentence.

Human Generated Questions We annotate the salience of 61 human generated questions from the above sources, to perform analyses in Section 5. Among those, 36 of them are derived from 2 articles of DCQA and 25 of them are from one article of TED-Q.

4.2 Salience Annotation

QSALIENCE-*data* is annotated by three linguistics undergraduate students at our institution who are native English speakers. They have previously been involved in multiple linguistic annotation tasks and have been trained on our specific annotation guideline on 50 questions (25 questions × 2 articles). The annotation guideline can be found in Appendix A.[2] In addition to the labels, annotators also provide natural language rationales which we release with QSALIENCE-*data*. These rationales are used in few-shot Chain-of-Thought prompting (Wei et al., 2024) in Section 6. The annotators were paid at least $15/hr.

Agreement The inter-annotator agreement (IAA) as measured by the Krippendorff's alpha (Krippendorff, 2011) (ordinal, with the "invalid" label set to 0) is 0.719 for the DCQA articles, 0.632 for TED-Q, 0.751 for Div.Article and 0.649 for Div.TL;DR. These values indicate substantial agreement (Artstein and Poesio, 2008), providing evidence to the predictability of reader expectations manifested as inquisitive questions.

Aggregation For label aggregation, we take the average salience of all annotations, then round it to the closest integer.

Analysis Examples of the annotated data are shown in Figure 1 and Appendix Table 13. Table 2 provides the label distribution for QSALIENCE-*data*. Notably, more than 90.8% of the questions generated from LLMs are valid, making them

---

[1] Ko et al. (2020) contains human-generated questions, but all the questions are from only the first five sentences of news texts. Thus, annotating them will provide only signals when C_p is very small.

[2] 1,150 questions were annotated by all three annotators; the rest, which is a subset of DiverseSumm articles and TL;DRs, were annotated by two of the three annotators.

dataset	#articles	#questions	average length	standard deviation
DCQA	4	260	11.97	4.76
TED-Q	1	100	11.07	3.9
Div. Article	27	957	14.8	4.44
Div. TL;DR	34	449	16.32	3.78
All	66	1766	13.99	4.57

Table 1: Count of articles and questions, average length and standard deviation of human and machine-generated questions per dataset.

dataset	Invalid	1	2	3	4	5
DCQA	13.4	2.3	19.2	36.9	20.0	8.0
TED-Q	14.0	0	6.0	29.0	35.0	16.0
Div. Article	9.9	0.8	17.2	22.6	31.9	17.6
Div. TL;DR	3.6	0.2	4.7	12.0	47.0	32.5
All	9.1	0.8	13.7	22.4	34.1	19.9

Table 2: Validity and salience distribution (in %) of human-annotated labels for the questions in QSALIENCE-*data*.

promising tools for inquisitive question generation. Our qualitative analysis of annotator rationales for invalid questions show that many of them does not have the right anchor sentence (i.e., Q not anchored in S_k); this was also found in Wu et al. (2023a). A few invalid questions also contain non-factual presuppositions. Among valid questions, those with the lowest score of 1 (question was irrelevant to C_p) is rare. However, the salience of the questions varies, indicating the potential usefulness of a salience predictor for LLM-generated questions in downstream tasks. We further analyze salience scores stratified by the LLMs that generate the questions in Appendix C.2.

In Appendix H, we show that question types that are more likely to associate with high salience ratings are *Consequence, Example*, and *Procedural. Disjunctive, Concept* and *Judgmental*.

5 Salience vs. Answerability

A valid potential question evoked at S_k can be deemed a QUD anchored at S_k if the subsequent discourse answers it. In order to study the relationship between potential questions and QUDs, we annotate a subset of the questions in QSALIENCE-*data* in terms of their *answerability*, i.e., the extent to which Q is answered in the subsequent context C_s.[3] We annotate answerability on a Lik-

dataset	0	1	2	3
DCQA	0.28	0.35	0.16	0.21
TED-Q	0.07	0.22	0.23	0.48

Table 3: Distribution (in %) of human-annotated answerability labels for 311 questions stratified by data source.

Random Questions	Human Ratings		
	All	DCQA	TED-Q
-0.02*	0.65	0.59	0.74

Table 4: Spearman rank correlation between salience and answerability annotated by humans and a random baseline. The correlation values that are not statistically significant ($p < 0.05$) are marked with a *.

ert scale: fully answered (3), partially answered (2), unanswered by C_s (1), and already answered in $C_p + S_k$ (0). The same annotators as in Section 4.2 annotated 225 and 86 valid questions from the DCQA and TED-Q subsets, respectively. Krippendorff's alpha (ordinal) is 0.799, indicating substantial agreement. Table 3 shows the distribution of answerability scores. Appendix C.3 provides more analysis on answerability scores per question generation model.

Do salience and answerability correlate? Table 4 presents the Spearman rank correlation coefficients between annotated salience and answerability. As comparison, we also compute a random correlation baseline between salience and the answerability of a random question, averaged across 10 trials. The annotated salience and answerability values have a significant Spearman's ρ of 0.65 (compared to the random baseline of -0.02). **This is evidence suggesting that salient potential questions are also likely to be answered later in the discourse,** even though the writers do not observe reader questions. This suggests that reader and writer expectations align to a certain degree.

Are QUDs salient potential questions? Table 5 presents the salience distribution of 36 DCQA questions that are annotated QUDs. Similar to the previous analysis, we also take a random subset of 36 potential questions, averaging their scores over 10 trials and present their salience distribution. We see that QUDs, which are potential questions answered in later context, are overall much more

---

[3] Although Westera et al. (2020) annotated this (they call this QUD predictability), we did not use their labels because they annotated whether a question was answered within a window of 4 sentences after S_k rather than the full texts.

	1	2	3	4	5
Random Questions	0.01	0.20	0.39	0.28	0.12
Annotated QUDs	0	0.11	0.25	0.47	0.17

Table 5: Salience distribution for 36 human annotated QUDs from DCQA, compared to a random set of inquisitive questions of the same size.

salient compared to a random set of potential questions sampled from QSALIENCE-*data*.

6 Salience Prediction

We experiment with a range of models for the prediction of question salience, given valid questions. Our finding is that salience prediction is a discourse task that recovers implicit information not readily grasped by LLMs, while our best instruction-tuned model, QSALIENCE, can achieve moderate agreement with humans. We further present question validity classifiers in Appendix B, which can be used with QSALIENCE in a pipeline fashion.

6.1 Models

Instruction Tuning We instruction fine-tune several open-source LLMs with QLoRA (Dettmers et al., 2023): **Mistral** (Jiang et al., 2023) (Mistral-7B-Instruct), **Llama 2** (Touvron et al., 2023) (Llama-2-7b-chat), **TinyLlama** (Zhang et al., 2024a) (TinyLlama-1.1B-chat), and **Flan-T5** (Chung et al., 2024) (flan-t5-base). AdamW (Loshchilov and Hutter, 2018) is used for optimization. Hyperparameters can be found under Table 11 in the Apendix.

The training data is formulated as (*input, output*) pairs where *input* consists of C_p, S_k, Q, and *output* is the salience score.[4] Appendix E shows the instructions for these models. For Flan-T5, since the context span is 512 tokens, we also experiment without using instructions, and truncate the context in the reverse sentence order from C_p and S_k until the context length is filled.

LLM Zero-/Few-shot Baselines We perform extensive experiments with various zero-shot and in-context learning scenarios with **GPT-4**-turbo. We show prompts in Appendix D.

(1) **Zero-shot (vanilla)**, where the model is given an instruction similar to that of the annotators.

---
[4]We also tried fine-tuning a classification head; however performance is inferior.

(2) **Few-shot (vanilla)**, where 15 in-context learning examples (3 per label) of $((C_p, C_k, Q), scr)$ are given, where scr denotes the salience score. We utilize LLMs' recency bias (Liu et al., 2024) to nudge its prediction to better align with our label distribution. Thus we altered the order of in-context demonstrations such that the examples at the end have labels more frequent within our train set.

(3) **Few-shot (kNN).** Performance of LLMs is often sensitive to the selection of the in-context examples (Rubin et al., 2022). Hence we use a kNN-based approach (Liu et al., 2022) to find the closest in-context examples to the current test instance. We encode C_p and S_k separately with RoBERTa-large (Liu et al., 2019) and take the average of the CLS tokens of each. We use Euclidean distance and retrieve one closest example for each salience label. These examples are put in-context following a similar ordering as the few-shot (vanilla) setting.

(4) **Chain-of-Thought (CoT).** We experimented with Chain-Of-Thought prompting (Wei et al., 2024). For **few-shot CoT**, we use 5 in-context examples, with the reasoning taken from the natural language rationales that the annotators gave during salience annotation.

6.2 Evaluation

Data We create a test set of 235 valid questions. The rest of the dataset is split into training (1,228 valid questions) and validation (143 valid questions). The data splits are stratified by articles. We upsample the training data to balance the label distribution. Our final training set consists of 2,355 examples, where each label has its 471 examples. We do not upsample validation or test sets.

Evaluation Metrics We measure the performance of salience prediction using four metrics: **(1)** Mean Absolute Error (MAE) between the predicted salience scores and the aggregated human scores; **(2)** Spearman's ρ between the two; **(3)** macro-averaged F1. These are standard metrics for ordinal classification or regression. In addition, we report **(4)** Krippendorff's α that measures agreement, also used in Section 4.2 between annotators.

Results Table 6 shows that the fine-tuned models clearly outperform zero- or few-shot LLMs, even with stronger prompting techniques such as kNN-based in-context example retrieval and Chain-of-Thought. On the contrary, among the fine-tuned smaller models, the best performing Mistral-based

Model	MAE ↓	Spearman ↑	Macro F1 ↑	krippendorff's α ↑
GPT4 zero-shot (vanilla)	1.314	0.229	0.193	-0.141
GPT4 few-shot (vanilla)	0.910	0.417	0.316	0.358
GPT4 few-shot (kNN)	1.063	0.359	0.245	0.215
GPT4 CoT zero-shot	1.144	0.366	0.197	0.058
GPT4 CoT few-shot	1.034	0.327	0.292	0.165
QSALIENCE (Mistral-7B-instruct)	**0.579**	**0.623**	**0.417**	**0.615**
Llama-2-7B-chat	**0.626**	**0.566**	**0.413**	**0.557**
Flan-t5-base	0.706	0.542	0.370	0.526
TinyLlama-1.1B-chat	0.664	0.522	0.402	0.496

Table 6: Model performance on the salience prediction task, for GPT-4 zero/few-shot baselines (top) and instruction-tuned LLMs (bottom). **Bold**: top-2 performance; blue shades: best performance for baselines and for fine-tuned models.

model achieves moderate agreement with human annotation with a substantial correlation. Even `flan-T5-base` with only 250M parameters and a small context window can be fine-tuned for this task to achieve competitive performance. These conclusions indicate that question salience is difficult to elicit from LLM prompting or in-context learning, and that explicit training can successfully capture this notion.

Appendix Figure 3 shows the confusion matrix for zero-shot GPT-4, the best-performing GPT-4 setting (few-shot vanilla), and our fine-tuned models.[5] Compared to fine-tuned models, GPT-4 tends to give a high score for the question by predicting many 4s and 5s, indicating its inability to distinguish good vs. bad questions irrespective of the Likert scale. By comparison, predictions from fine-tuned models tend to confuse primarily labels closer to each other. This also shows our fine tuned models understand the tasks better than in-context learning with GPT-4.

7 Use Case: Do better expanded summaries answer more salient questions?

We demonstrate the usefulness of question salience prediction in a downstream task: summary expansion. Given a document D and a short TL;DR S_s, the summary expansion task aims to generate a longer summary S_l that captures a fuller picture of the article, shown in Figure 2. This task captures the situation where after reading the TL;DR, a curious reader often wants to know more in order to decide if they want to read the entire article. A similar task is deployed in Semantic Scholar, though in a very different domain than ours.

Given findings in Section 4.2, our hypothesis is that reader expectation aligns with QUDs in the expanded summary; namely, a higher-quality summary answers more salient questions.

7.1 Data

Articles and TL;DRs We sample 34 articles from DiverseSumm as source articles; these articles are roughly 1,500 words long. Given the strong performance of GPT-* in summarization (Goyal et al., 2022; Zhang et al., 2024b), we prompt GPT-4-turbo to produce a 50-word summary TL;DR of the article. All prompts needed in this section are shown in Appendix F.

Expanded Summaries For each TL;DR, we generate three expanded summaries while controlling their lengths to be between 230–250 words:

(1) GPT-4: given (D, S_s), we prompt GPT-4-turbo for S_l.

(2) Flan-T5: we use `flan-t5-large` to produce an elaboration in a similar manner as above.[6] Since this model can produce summaries with obvious errors such as repeating sentences and violating the length control, we refine the Flan-T5 outputs to fix these obvious errors, while preserving the summary as much as possible, using GPT-4-turbo.

(3) GPT-4-Corrupted: We synthetically generate "bad" summaries to serve as a baseline that is missing the most prevalent topics. First, we prompt GPT-4-turbo with (D, S_s) and ask it to identify important topics from S_s. Using these topics and D, we then prompt the model to generate a long summary within our expected word count that does not include these relevant topics. Since this response can sometimes be incoherent and disobey

---

[5]Note that the label 1 is extremely rare (Table 2) and is not present in the test set.

[6]We prompt the model with the instruction, TL;DR and the article in that order. Owing to the short context size of `flan-t5-large`, the entire article is not used for the task.
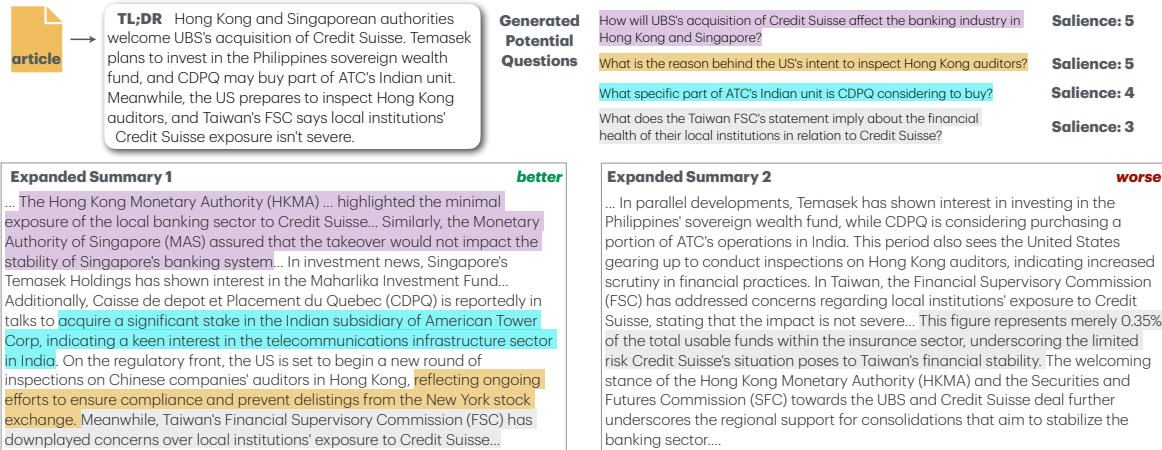
Figure 2: This example illustrates the expanded summarization task and how question salience is used in evaluation. The summaries are generated by expanding on a short TL;DR of an article. We show the inquisitive questions generated from the TL;DR and their salience scores. A better summary (left) answers more salient questions than the worse summary (right), where only one medium-salient question is answered.

the length control, so we again use GPT-4-turbo to refine the output with the same refinement prompt as Flan-T5.

Question Salience Each sentence in every TL;DR is associated with 5 LLM generated inquisitive questions (Section 4.1), annotated with salience (Section 4.2). Additionally, we run fine-tuned models from Section 6 for all valid questions in this set to obtain predicted salience values.

7.2 Evaluation

Human Ranking of Summaries We recruited 5 experienced annotators to rank the expanded summaries, focusing on their content rather than style.[7] Each of the three summaries is given a score from 1 (lowest) to 3 (highest) based on their rankings. Our annotation interface can be seen in Figure 4 in the Appendix. Examples are shown in Figure 2 and Table 14.

The ranking results show an expected order of quality: GPT-4 (2.91 average ranking), Flan-T5 (1.73), GPT-4-Corrupted (1.35). This is the oracle ordering that we aim to reproduce by utilizing the salience of QUDs in the expanded summary.

Measuring QUD Salience To score a summary S_l, we measure the salience scr of the questions evoked in each TL;DR S_s that become QUDs (i.e., answered in S_l). First, we filter all questions that are not answered by the article D itself. Next, with the remaining questions and S_l, we prompt GPT-

---

[7]Prior work showed that GPT3-generated summaries are faithful (Zhang et al., 2024b).

	GPT-4	Flan-T5	GPT-4-Crpt
Human Salience	26.9	18.3	17.8
Mistral-7B-instruct	29.0	20.2	19.2
Llama-2-7B-chat	29.4	21.0	19.8
Flan-T5-base	30.8	21.8	20.8
TinyLlama-1.1B-chat	29.0	20.6	18.8

Table 7: SummScr of human and model salience for 5 TL;DR expansion instances

4-turbo to return all the questions that S_l answers. Prompts in this section are shown in Appendix G.

The salience scores $scr(q)$ of these QUDs are then aggregated into SummScr $= \frac{1}{n} \sum_{i=1}^{n} \sum_{q \in Q_i} scr(q)$ where Q_i denotes all answered questions in the generated expanded summary i.

Results The SummScr (calculated from human salience) for all 34 articles is GPT-4 (21.93), Flan-T5 (16.25) and GPT-4-Corrupted (8.81). The Kendall's τ rank correlation between the majority summary ranking and SummScr (human) for the full set is 0.529, a moderately high correlation.

Table 7 shows the SummScr on the test set portion of DiverseSumm. All SummScr values derived from *predicted* salience values, using fine-tuned systems, reproduce the same system-level ranking as humans: GPT-4, Flan-T5, then GPT-4-Corrupted. We consider this evidence that better expansions answer more salient questions.

8 Conclusion

In this paper, we explored predicting salience for inquisitive questions. Our work connects two ideas:

a theoretical idea of which questions are useful for understanding and likely to be answered later in a text, and an empirical notion of what questions are useful. We showed that predicting salience is possible with fine-tuned models, and these approaches outperform GPT-4. Furthermore, we showed in a pilot use case that notions of summary quality align with how many salient questions were answered.

Limitations

While this work takes the first step at empirically connecting prior discourse literature and developing a salience model for inquisitive questions, we have not engaged in the formal semantics of potential questions as in Onea (2016). An additional limitation is that we have not explicitly measured information utility (in information-theoretic terms) given the open-ended nature of the questions, although our notion of salience is consistent with Van Rooy (2003).

This work has considered only English text, sourcing articles from existing datasets that provided groundwork for various analyses in this paper, both theoretical ones and empirical experiments. Thus even though our notion of question salience is application-agnostic, we believe an exciting future direction is to explore question salience in other domains and languages.

Finally, when considering the notions of salience for our texts, we assume that the reader backgrounds are not too divergent from what the writer has intended. A large discrepancy between the two could lead to readers having very different salient questions; e.g., when the reading level of the reader is much lower than that of the intended audience (Wu et al., 2023b). Thus our tool and dataset should not be used when reader backgrounds are too different from the writer expectations or among themselves.

Acknowledgments

References

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse. *Discourse Processes*, 54(3):177–186.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. 2007. Children's questions: A mechanism for cognitive development. *Monographs of the society for research in child development*, pages i–129.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jeremy R. Cole, Palak Jain, Julian Martin Eisenschlos, Michael J.Q. Zhang, Eunsol Choi, and Bhuwan Dhingra. 2023. DiffQG: Generating questions to summarize factual changes. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3088–3101, Dubrovnik, Croatia. Association for Computational Linguistics.

Dan Cristea and Bonnie Webber. 1997. Expectations in incremental discourse processing. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 88–95, Madrid, Spain. Association for Computational Linguistics.

Beth Davey and Susan McBride. 1986. Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, 78(4):256.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with

recursively expandable summaries. *arXiv preprint arXiv:2310.07581*.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2022. "what makes a question inquisitive?" a study on type-controlled inquisitive question generation. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 240–257, Seattle, Washington. Association for Computational Linguistics.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. *Logic, Language and Computation*, 1:221–237.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

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, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Andrew Kehler and Hannah Rohde. 2017. Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*, 54(3):219–238.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555, Online. Association for Computational Linguistics.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. Discourse comprehension: A question answering framework to represent sentence connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FollowupQG: Towards information-seeking follow-up question generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–271, Nusa Dua, Bali. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

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, Singapore. Association for Computational Linguistics.

Edgar Onea. 2016. Potential questions at the semantics-pragmatics interface. In *Potential Questions at the Semantics-Pragmatics Interface*. Brill.

Adithya Pratapa, Kevin Small, and Markus Dreyer. 2023. Background summarization of event timelines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8111–8136, Singapore. Association for Computational Linguistics.

Michael Prince. 2004. Does active learning work? a review of the research. *Journal of engineering education*, 93(3):223–231.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for questions under discussion and information structure. *Information structure in lesser-described languages. Studies in prosody and syntax*, pages 403–443.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv 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, Bangkok, Thailand. Association for Computational Linguistics.

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

Robert Van Rooy. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy*, 26:727–763.

Alex Warstadt. 2020. " just" don't ask: Exclusives and potential questions. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 373–390.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

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, Singapore. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

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

A Annotation Guidelines for Question Salience

Motivation As one reads an article, it is natural to ask curiosity-driven questions to enhance one's understanding of the article. Amongst different potential questions that one might ask while reading the article, to what extent is it important for it to be answered later in the article? Can we perhaps rank these questions? We develop an evaluation schema to do just that!

Task Given the prior context, anchor sentence, and a list of potential questions, score the questions on the basis of the following schema.

Score=0: These are questions which satisfy atleast one of the following criterion:

1. Question has grammar errors

2. Question is not anchored in the given anchor sentence

3. Question contains multiple sub-questions

4. Question misinterprets the context

Score=1: The question is not very related to the topic (basically to weed out any odd questions)

Score=2: The question is related to the concepts introduced in the prior context and the anchor sentence but asking the question seems like a stretch. Answering the question doesn't seem useful in making the article feel complete. Typically questions that also seem to be completely answered by the prior context and the anchor sentence are given this score.

Score=3: The question is related to the prior context and anchor sentence but answering it doesn't matter to me. Answering it may provide additional information which may/may not enhance my understanding of the article.

Score=4: Answering the question is somewhat useful because, for example, it might clarify some newly introduced concepts, or might expand on an idea already introduced. It is useful to answer the question because it might influence the narrative. There is a degree of uncertainty here as compared to when you would score a question 5.

Score=5: This question should definitely be answered in the subsequent context. Some reasons why the question should definitely be answered:

1. It clarifies a concept introduced in the anchor sentence

2. It asks about surprising or mysterious events/objects

3. It asks for more information about newly introduced humans (or animated) individuals into the discourse

4. Answering this question is essential to understanding the narrative.

Do keep in mind that one shouldn't make an inference about other people. For instance, if the question is about defining or explaining a concept, and you don't need that explanation, don't say that answering the question may still be helpful just because you think some other people will find the answer useful.

B Validity Classification

Per Table 2, invalid questions accounted for 9.1% of the annotated data. Thus we also experiment with question validity classification, which can be used in a pipeline to first find invalid questions to exclude, before scoring their salience.

LLM Zero-/Few-shot Baselines Since many invalid questions are caused by anchor issues (Section 4.2), we use the anchor relevance prompt in QUDEval (Wu et al., 2023a) for few-shot prompting using GPT-4 and Mistral-7B-instruct to classify question validity.[8].

Fine-Tuning We also fine tune `flan-t5-base` and `TinyLlama-1.1B-chat` on this task. Prompt B.1 list the instruction for TinyLlama-chat.[9] AdamW (Loshchilov and Hutter, 2018) was used as optimizer with a learning rate of 3e-4, trained for 2 epochs. We perform downsampling to balance the data distribution.

Results Table 8 shows that both prompting and fine-tuned models perform decently well on question validity classification. The fine-tuned models are on-par with prompting LLMs.

Model	Macro F1 ↑
GPT-4 few-shot	0.689
Mistral-7B-instruct few-shot	0.538
Flan-t5-base fine tuned	0.662
TinyLlama-1.1B-chat fine tuned	0.693

Table 8: Question Validity Performance

---

[8]We merge "fully grounded" and "partially grounded" in their label set as a single valid label

[9]Due to the small context window of Flan-T5, we do not use instructions.

C LLM Question Generation and Additional Analyses

C.1 Settings

We show the prompt for LLM inquisitive question generation in Prompt C.1. A temperature of 0 is used so that the questions generated are grounded within the context provided. Greedy decoding is used due to its computational efficiency and deterministic behaviour. We also use a frequency penalty of 0.5 to make the model more conservative in generating repetitive tokens.

C.1 Prompt for Question Generation

Context:
{{article context C_p}}

After reading the sentence {{anchor sentence S_k}}, ask 5 questions about a part of this sentence that you are curious about which you don't have an answer for.

C.2 Salience Analysis Per Model

Table 9 shows the average salience scores of questions produced by each model. Qualitative analysis of annotator rationales reveals that both Mistral-7B-instruct and Llama-2-7B-chat struggled to generate questions anchored in S_k, resulting in many invalid questions; Mistral-7B-instruct was worse than Llama-2-7B-chat. On the other hand, GPT-* models produced more interesting, curiosity driven questions.

	#questions	avg salience
human	61	3.49
mistral-7B-instruct	249	2.26
llama-2-7B-chat	250	2.6
GPT-3.5-turbo	100	3.18
GPT-4-turbo	1106	3.75

Table 9: Count and average human salience of questions stratified by the question generation model used

	#questions	avg answerability
human	58	2.25
mistral-7B-instruct	73	1.32
llama-2-7B-chat	88	1.15
GPT-3.5-turbo	92	1.52

Table 10: Distribution and average human answerability of questions stratified by question generation model

C.3 Answerability Analysis Per Model

Table 10 shows the count and average answerability of valid potential questions stratified by the question generating model. Qualitative analyses of annotator rationales reveals that Mistral-7B-instruct and Llama-2-7B-chat often produced questions which were already answered in $C_p + S_k$ or were about specific parts of S_k, not relevant to the article as a whole; llama was generally worse than mistral. While GPT-3.5-turbo is good at generating salient questions, they sometimes becomes too diverse to be actually answered by the article. Human questions are QUDs and thus are mostly answered.

D Prompts for GPT-4 salience prediction

GPT-4 zero-shot salience prompts are shown in Prompt D.1. For few-shot, 3 examples are taken for each of the 5 labels using the same format. Similarly, zero-shot Chain-of-Thought prompt is shown in Prompt D.2. Few-shot Chain-of-Thought uses the same format with 3 examples for each of the 5 labels.

D.1 Zero-shot (vanilla) prompt for salience prediction.

article: {{article context C_p + anchor sentence S_k}}

question: {{question}}

system: Imagine you are a curious reader who is reading the article. You come across a question and you need to determine if it should be answered in the following article or not. You have to give a score for this input. Score = 1 means the question is completely unrelated to the topic of the article. Score = 2 means the question is related to the article but it has already mostly been answered by

the article. Score = 3 means the question is related to the article but answering it is not useful as it might expand of an idea that is not very important or central to the context of the article. Score = 4 means the question is related to the article and answering it is somewhat useful in enhancing the understanding of the article. Score = 5 means the question is related to the article and should definitely be answered because it expands on some ideas which are central to the article. Note that the score is given according to the information utility of its answer. If a question is related to the article but doesn't need to be answered or is not central to the article, do NOT give it a high score of 4 or 5, instead give a score of 3 if the question is unanswered by the article and 2 if it has already been answered by the article. To differentiate between a score of 4 vs 5, think of how the article would look like if you don't answer the question - if the article would not feel complete without the answer to the question, give a score of 5, else a 4. A score of 4 is usually given if answering the question will be useful but there might be other questions that are more important to answer as compared to this. A score of 5 is only given to the best and most important questions that MUST be answered so use it carefully and sparingly. Do not be biased towards giving a high score and follow the above instructions carefully. The score should strictly be an integer from 1 to 5.

score:

D.2 Zero-shot CoT prompt for salience prediction.

article: {{article context C_p + anchor sentence S_k}}

question: {{question}}

system: Imagine you are a curious reader who is reading the article. You come across a question and you need to determine if it should be answered in the following article or not. You have to give a reason and a score for this input. Score = 1 means the question is completely unrelated to the topic of the article or misinterprets the context of the article. Score = 2 means the question is related to the article but it has already mostly been answered by the article. Score = 3 means the question is related to the article but answering it is not useful as it might expand of an idea that is not very important or central to the context of the article. Score = 4 means the question is related to the article and answering it is somewhat useful in enhancing the understanding of the article. Score = 5 means the question is related to the article and should definitely be answered because it expands on some ideas which are central to the article. Note that the score is given according to the information utility of its answer. If a question is related to the article but doesn't need to be answered or is not central to the article, do NOT give it a high score of 4 or 5, instead give a score of 3 if the question is unanswered by the article and 2 if it has already been answered by the article. To differentiate between a score of 4 vs 5, think of how the article would look like if you don't answer the question - if the article would not feel complete without the answer to the question, give a score of 5, else a 4. A score of 4 is usually given if answering the question will be useful but there might be other questions that are more important to answer as compared to this. A score of 5 is only given to the best and most important questions that MUST be answered so use it carefully and sparingly. Do not be biased towards giving a high score and follow the above instructions carefully. First provide a reasoning for your response and then the score. Now let's think step by step.

---

reason:

E Setups for instruction fine-tuning

Model	Seq Len	Learn. Rate	Epoch
Mistral-7B-instruct	4096	0.0003	3
Llama-2-7B-chat	4096	0.0001	5
Flan-t5-base	512	0.0003	3
TinyLlama-1.1B-chat	4096	0.0003	4

Table 11: Parameters for fine-tuned Models in salience scoring

E.1 Instruction for fine-tuned models for salience prediction.

System: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: Give a score from 1 to 5 for how important it is for the question to be answered later in the article.
Score = 1 means the question is completely unrelated to the topic of the article.
Score = 2 means the question is related to the article but answering it is not useful in making the article feel complete.
Score = 3 means the question is related to the article but answering it might not enhance the understanding of the article.
Score = 4 means the question is related to the article and answering it is somewhat useful in enhancing the understanding of the article.
Score = 5 means the question is related to the article and should definitely be answered because it might provide explanation for some new concepts.

Input:
article: {{article context C_p + anchor sentence S_k}}
question:{{question}}

Response::
{{score}}

F Prompts for Expanded Summary Generation

F.1 Prompt for generating a short TL;DR.

Context: {{article}}

Generate a short 50-word summary for the above article. Remember, do not exceed 50 words.

Summary:

F.2 Prompt for GPT-4 summary expansion.

article: {{article}}

short summary: {{tl;dr}}

Produce an elaboration of the short summary by including relevant details from the article within a word count range of 230 to 250 words. Strive for conciseness and clarity in

delivering a comprehensive expansion within the specified word limit. The response MUST NOT exceed 250 words at any cost. Produce outputs less than 250 words.

elaboration:

F.3 Prompt for GPT-4-Corrupted expansion (Step 1).

article: {{article}}

short summary: {{tl;dr}}

Read the article and the short summary. Provide a list of all the important topics from the short summary and related to it which are spoken about in the article. Your response should be a comma separated list.

response:

F.4 Prompt for GPT-4-Corrupted expansion (Step 2).

article: {{article}}

irrelevant topic: {{irrelevant topic}}

In 230 to 250 words, produce an elaboration of the article by omitting as many topics included or related to the 'irrelevant topic' field as possible. Your response MUST be strictly more than 230 words and under 250 words. Remember, you MUST produce ATLEAST 230 word count responses.

response:

F.5 Prompt for improving the expansion style while imposing minimal content changes.

paragraph: {{original expanded summary}}

Make minor alterations to the paragraph above such that its narrative style is similar to a usual summary. Do not use very flowery language and stick to the contents in the paragraph ONLY. Your response should NOT include any new content. Your response should be over 230 words but not exceed 250 words. Remember, do not produce responses below 230 words. Don't start the sentences like the 'article talks about this' or 'the article sheds light on..'. Remember, you MUST produce ATLEAST 230 word count responses.

response: :

G Prompts for Expanded Summary Evaluation

G.1 Prompt for identifying questions that are unanswered in the article.

article: {{article}}

Which sentences from the article completely answer the question {{question}} Include only the relevant sentences extracted from the article that are answers to the question and NOT just vaguely related to the topic introduced in the question. Be concise. Your response should not exceed 3 lines. If the article doesn't provide a SPECIFIC answer to the question, respond with 'No Answer'.

response:

G.2 Prompt for finding questions that an expanded summary answers.

article: {{expanded summary}}

questions: {{list of questions}}

Read the above article and find the questions from the 'questions' list provided above which are answered in the article. Your response should be a comma separated list of only questions that are completely or partially answered by the article.

response:

H Question Type vs Salience

We computed PMI (pointwise mutual information) between question type and salience scores to see if they relate. We divided salience into three levels: low, mid, high. We classified questions into 13 types defined in (Cao and Wang, 2021). The ranked PMI by the highest salience level can be found under Table 12.

I Compute

We used 3 NVIDIA A40 (48GB) and 1 A100 (80GB) for fine-tuning models. For A40, each training took less than 30 minutes. For A100, each training took under 10 minutes. The training process for all models finished within 4 hours.

J License and Copyright

We will release our annotations under the Creative Commons CC-BY license; the original texts (DCQA (Ko et al., 2022), TED-Q (Westera et al., 2020) and DiverseSumm (Huang et al., 2024)) come with their original licenses.

Question Type	PMI
CONSEQUENCE	0.413
EXAMPLE	0.279
PROCEDURAL	0.146
COMPARISON	0.127
CAUSE	0.019
EXTENT	-0.060
VERIFICATION	-0.112
DISJUNCTIVE	-0.401
CONCEPT	-0.451
JUDGMENTAL	-0.648

Table 12: Question types ranked by PMI with high salience scores.

article: [1] The desperate actions by governments, regulatory authorities, and banks in both the US and Europe have not only failed to stem the growing financial crisis but in some ways are making it worse.[2] In the US, following the failure of the Silvergate bank, Silicon Valley Bank and Signature over the past two weeks, the latter two recording the second- and third-largest banking failures in US history respectively, attention has turned to the travails of the First Republic Bank with growing concerns that it could be the next to go.[3] Last week, a consortium of 11 major banks, under the leadership of JPMorgan Chase CEO Jamie Dimon, with the collaboration of Treasury Secretary Janet Yellen, deposited $30 billion with the struggling bank.[4] It was hoped this show of confidence would stop the outflow of depositors' money, ease the pressure on its share price and stabilise it.

questions:
1) Who initiated the act of depositing $30 billion into the struggling First Republic Bank? **Salience = 2**
2) Why was it thought that this deposit would stem the outflow of depositors' money? **Salience = 5**
3) What role did Treasury Secretary Janet Yellen play in this financial effort? **Salience = 0**

---

article: [5] In just a few days, the operation has been revealed as a complete failure.[6] While the outflows are reported to have slowed somewhat, First Republic has lost $70 billion out of the total of $176 billion it held at the start of the year.[7] And despite the injection of cash, the company's shares have continued to plummet.

questions:
1) How has the injection of cash affected the overall financial health of the company? **Salience = 3**
2) What are the strategies that the company intends to use to stabilize its shares amid the injection of cash? **Salience = 4**

---

article: [8] Its share price has fallen by 90 percent since the beginning of the month, closing 47 percent down yesterday.[9] Long-term bonds that mature in 2046 were trading at 55 cents on the dollar, down from 75 cents in early March.[10] First Republic took another hit before trading opened yesterday, when the ratings agency S&P Global downgraded its credit rating for the second time in a week.

questions:
1) What was First Republic's credit rating prior to these two downgrades by S&P Global? **Salience = 3**
2) Why did the ratings agency S&P Global downgrade First Republic's credit rating for the second time in a week? **Salience = 5**

Table 13: An example of an article with potential questions it evokes and their corresponding human-annotated salience scores.
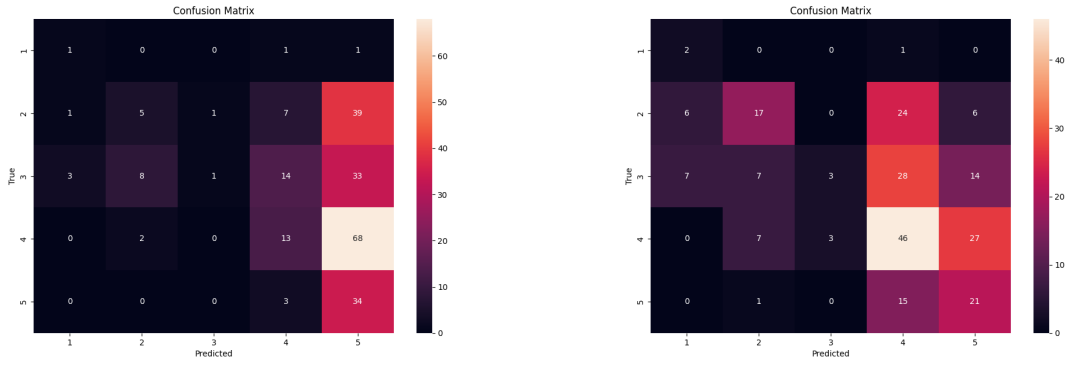
(a) GPT-4-turbo zero-shot vanilla (left), GPT-4-turbo few-shot vanilla (right)
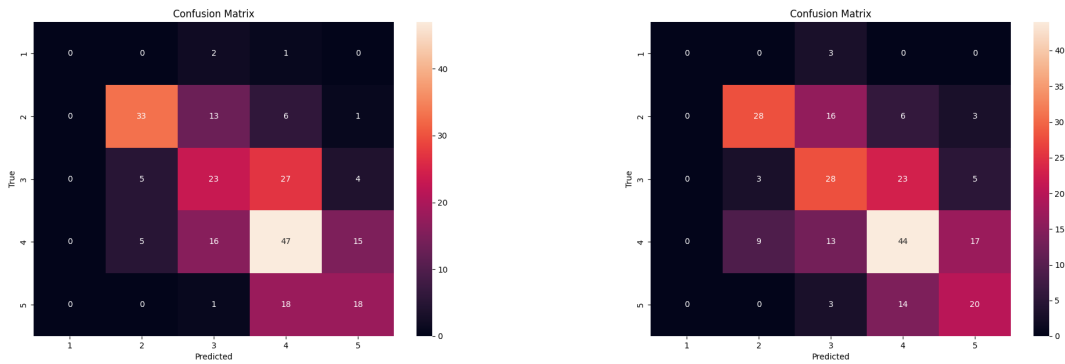
(b) Mistral-7B-Instruct (left), Llama-2-7B-chat (right)
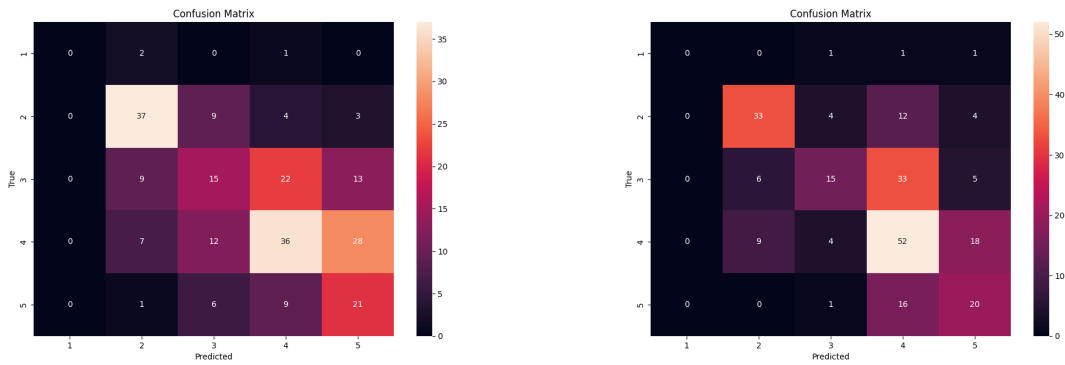
(c) Flan-t5-base (left), TinyLlama-1.1B-chat (right)

Figure 3: confusion matrices for salience prediction across different models.

Figure 4: Annotation interface for the summary expansion task; the three candidates are ordered via drag-and-drop.

TL;DR: The IMF reached a staff-level agreement to offer Ukraine a $15.6bn financial package, marking the first loan to a country at war. The two-phase agreement involves enhancing fiscal and financial stability, followed by more extensive reforms. The loan follows recent amendments allowing the IMF to fund countries facing 'exceptionally high uncertainty'.

best expansion: The International Monetary Fund (IMF) has agreed to provide Ukraine with a significant financial package worth $15.6 billion, a groundbreaking move as it marks the first instance of the organization offering a loan to a nation currently engaged in war. This decision comes after the IMF recently amended its policies to permit funding under conditions of "exceptionally high uncertainty," enabling support for countries grappling with extraordinary challenges such as war or severe natural disasters. The financial aid is structured into a two-phase program spanning four years. Initially, Ukraine will focus on enhancing fiscal, external, price, and financial stability, including the cessation of monetary financing, over a period of 12-18 months. The subsequent phase will pivot towards broader reforms aimed at securing macroeconomic stability, supporting recovery and early reconstruction efforts, and fostering resilience and higher long-term growth. These reforms are particularly geared towards Ukraine's aspirations for European Union accession, with an emphasis on reverting to pre-war policy frameworks like a flexible exchange rate and inflation targeting regime. This loan package is not only the largest the IMF has proposed for a country amidst conflict but also a testament to the international community's support for Ukraine during its ongoing war with Russia. The IMF's decision reflects a recognition of the severe economic impact of the conflict on Ukraine, including a 30% contraction in economic activity in 2022, significant destruction of capital stock, and rising poverty levels. By securing this financial assistance, Ukraine aims to stabilize its economy, mobilize further support from international partners, and lay the groundwork for post-war reconstruction and recovery.

good expansion: The International Monetary Fund (IMF) has agreed to provide Ukraine with a $15.6 billion financial package, a significant move as it's the first loan offered to a country engaged in war. This agreement, structured in two phases, aims to first enhance Ukraine's fiscal and financial stability, with a subsequent focus on broader reforms. This decision comes after recent changes in the IMF's policies, now allowing funding for countries experiencing 'exceptionally high uncertainty'. The journey to this agreement began last June when Ukraine first initiated discussions with the IMF, leading to a final approval by the IMF executive board on March 20, after an initial round of talks failed and a second was launched in August. Meanwhile, Pakistan faces challenges in securing a vital IMF bailout package. The government revealed that the IMF is seeking financial support for Pakistan from 'friendly' nations before proceeding with a $6.5 billion bailout program. This stance was reiterated with a slight variation in the required amount, mentioning a $6.1 billion tranche. Criticism has been directed at the IMF's approach to assisting poor countries. Nobel Prize-winning economist Joseph Stiglitz, in his 2002 book "Globalisation and Its Discontents," criticized the IMF's economic reform conditions. He argued that these often counterproductive measures have had devastating effects on the populations of the target countries, highlighting a broader debate on the IMF's role and policies in global financial stability.

worst expansion: The International Monetary Fund (IMF) has updated its policies to support countries experiencing "exceptionally high uncertainty," aiming to assist those affected by crises like wars or natural disasters. This policy revision, revealed on March 17, diverges from the IMF's earlier approach that restricted its capacity to offer non-emergency loans under conditions of significant uncertainties. The modification in the financing assurances policy intends to cover scenarios where external shocks, beyond a country's economic policy influence, require financial aid. The IMF has been instrumental in aiding countries during financial crises by providing loans to address challenges such as foreign debt payments and reducing foreign exchange reserves. Funding for these loans is sourced from the capital subscriptions or quotas contributed by member countries, reflecting their economic standing globally. With 190 member countries, the United States is the largest contributor to the IMF. Before approving a loan, the IMF conducts discussions with the applicant country to assess its financial situation and requirements. The country typically needs to agree to undertake certain economic policy measures. Following an agreement, the IMF's executive board reviews and approves the loan, which is then disbursed in phases. Throughout this period, the IMF monitors the country's adherence to the agreed economic policies. This procedure highlights the IMF's commitment to promoting economic stability and reform among its member countries facing financial challenges.

Table 14: An example of GPT-4, Flan-T5 and GPT-4-Corrupted expansions from best to worst.