



Designing and Evaluating Interfaces that Highlight News Coverage Diversity Using Discord Questions

Philippe Laban
plaban@salesforce.com
Salesforce AI Research
United States

Chien-Sheng Wu
Salesforce AI Research
United States

Lidiya Murakhov'ska
Salesforce AI Research
United States

Xiang 'Anthony' Chen
UCLA
United States

Caiming Xiong
Salesforce AI Research
United States

ABSTRACT

Modern news aggregators do the hard work of organizing a large news stream, creating collections for a given news story with tens of source options. This paper shows that navigating large source collections for a news story can be challenging without further guidance. In this work, we design three interfaces – the Annotated Article, the Recomposed Article, and the Question Grid – aimed at accompanying news readers in discovering coverage diversity while they read. A first usability study with 10 journalism experts confirms the designed interfaces all reveal coverage diversity and determine each interface's potential use cases and audiences. In a second usability study, we developed and implemented a reading exercise with 95 novice news readers to measure exposure to coverage diversity. Results show that Annotated Article users are able to answer questions 34% more completely than with two existing interfaces while finding the interface equally easy to use.

ACM Reference Format:

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhov'ska, Xiang 'Anthony' Chen, and Caiming Xiong. 2023. Designing and Evaluating Interfaces that Highlight News Coverage Diversity Using Discord Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3544548.3581569>

1 INTRODUCTION

In order to complement their understanding of political and world affairs, readers increasingly access the news through multiple channels, with 64% of Americans relying on articles shared on social media [56], and 62% using smart speakers to listen to the news [50]. In this multi-channel setting, it becomes important for news readers to have easy ways to compare and contrast opinions of varying sources [53], as a lack of transparency risks reader exposure to bias on critical societal issues such as elections or international affairs [7].

Modern news aggregators play a crucial role in giving readers access to broad coverage diversity on any given topic, with features

like “See Full Coverage” on Google News that put thousands of sources at the fingertips of news readers. In practice, however, news aggregator users must invest more time and effort to obtain broad coverage for a news story [37], reading through several sources and sifting through overlapping content to build a more complete story understanding.

Some news aggregators provide additional guidance to simplify the source selection process, such as the political alignment of a source or information on the organization that owns each source [2], assisting readers in anticipating potential bias. However, news aggregators treat news articles as *atomic units* and do not typically help users compare details in coverage differences within news articles. The burden of aligning article text to compare source differences on story-specific issues is therefore left to the reader.

In this work, we explore an extended role of the news aggregator, in which details in coverage that sources diverge on are explicitly revealed, accompanying the reader in understanding the subtlety of the story. Elements of disagreement between sources are extracted using the Discord Questions framework [35], an automatic pipeline based on Natural Language Processing (NLP). For multi-source news stories, the framework defines discord questions as questions answered by a large proportion of the sources, with a level of diversity or disagreement in the sources' answers. Authors of the framework hypothesize that discord questions can be a powerful tool in helping news readers navigate source differences, with each question revealing how the sources side on a specific issue within the story.

As an illustration, Figure 1 gives an example of two generated discord questions for a story on inflation in Summer 2022 in the US. The questions highlight source disagreements on the role of the Federal Reserve and predictions of its future actions, revealing elements of the story-specific debate. Although promising, it is not evident how to integrate Discord Questions into news reading interfaces that can accompany novice users in realistic reading settings. In this work, we design three interfaces that leverage discord questions – Annotated Article, Recomposed Article, and Question Grid – and evaluate them in two usability studies. The three interfaces we propose vary in levels of complexity. With the Annotated Article, we select an existing news article and augment its contents with discord questions and selected multi-source answers. In the Recomposed Article, a QA-format article is algorithmically composed using the content of multiple sources. Finally, the Question Grid achieves higher information density by laying out information within a

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
<https://doi.org/10.1145/3544548.3581569>

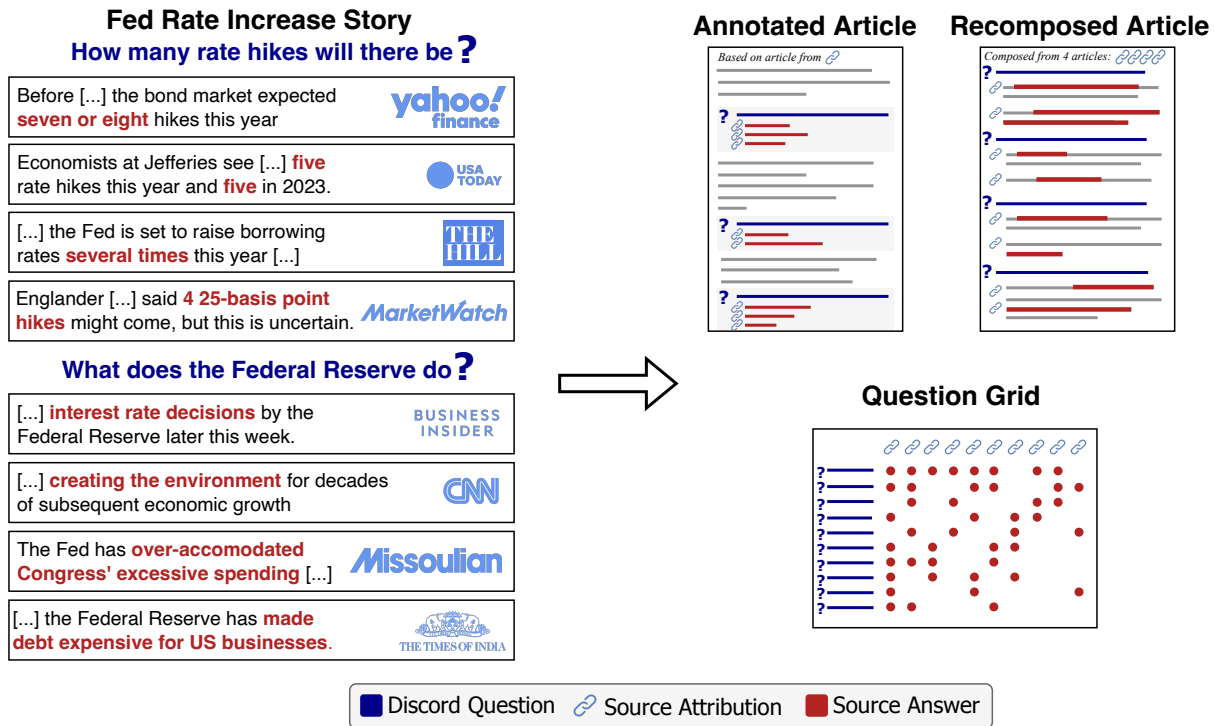


Figure 1: We design and evaluate news reading interfaces that incorporate discord questions to reveal coverage diversity. The Annotated Article is based on a single source article annotated with questions and answers from multiple sources. The Recomposed Article is a QA-format article composed from the ground up using multiple source articles. The Question Grid presents content on a (question, source)-matrix, specifying the answer each source provides to a list of questions.

(question, source-grid, specifying the answer each source provides to a list of questions.

We conduct the first usability study with 10 journalism experts. We ask participants to compare the proposed interfaces with two baselines (i.e., a plain News Article and a Headline List), with an aim to narrow down use cases and target audiences for each interface. Experts find that only the two baseline interfaces and the Annotated Article are suitable for novice news readers, with the Annotated Article giving the most complete overview of the story. The Question Grid is found to highlight the most coverage diversity overall and be relevant for advanced use cases such as helping newsrooms find unique perspectives to differentiate from the competition.

In a second usability study, we design a news-reading exercise to measure exposure to coverage diversity in a realistic news-reading scenario for novice readers, which we define as readers with no professional experience in journalism. We implement the exercise with 95 novice users who complete the exercise using three interfaces: the two baseline interfaces, and the Annotated Article. We find that the Annotated Article leads to a significantly broader understanding of news stories while being preferred in terms of ease of use. Our work makes the following contributions:

- The design and implementation of three news interfaces¹ – the Annotated Article, the Recomposed Article, and the Question Grid – based on the Discord Questions framework, aimed at highlighting news coverage diversity in multi-source news stories,
- The design of a news-reading exercise that measures reader exposure to news coverage diversity, in which readers are tasked to answer open-ended questions as thoroughly as possible, followed by manual scoring of answer completeness,
- An evaluation through two usability studies of proposed designs, confirming that integrating discord questions into news interfaces effectively highlights coverage diversity, and provides readers with a more complete story understanding in a realistic news reading situation, compared to existing baseline interfaces.

2 RELATED WORK

2.1 Exposing Bias To News Readers.

Prior work has proposed interfaces to accompany news readers in explicitly seeking diverse opinions. We classify approaches as

¹A live demonstration of the interfaces is available at <http://assembly.salesforceresearch.ai/>

supplying source indicators, annotating single articles, or multi-article interfaces.

2.1.1 Source Position Indicators. Source Position indicators typically depict the valence (left/right) and magnitude (moderate/extreme) of each source of information in a news interface [41]. These indicators have become widespread, and are now integrated into popular aggregators such as the Wall Street Journal’s Blue Feed Red Feed², and AllSides [2]. Some studies have shown the effectiveness of source position indicators, for example, Munson et al. [47] built a system that sends a weekly report to readers summarizing the source positions they accessed, and showed that sending such reports increases reader visits to opposite or moderate news sites. One limitation of source indicators is their tendency to be overly generic, as source bias can vary based on the topic, and does not accompany the reader beyond article recommendation.

2.1.2 Article Annotation. Prior work has proposed annotations that augment and modify articles once they are selected by a reader. Munson and Resnick [48] propose to highlight passages in articles that are in agreement with user opinion and to reorder article content such that agreeable content comes first, finding limited improvement in reader satisfaction. Hamborg et al. [21] more directly go after educating readers, automatically highlighting terminology likely to be biased (e.g., freedom fighters vs. terrorists). Beyond annotating the content, prior work has also inserted information best practices in the form of a checklist to encourage readers to detect misinformation[23]. The single-source setting is limited, as the user cannot get exposed to content not present in the chosen source. We implement an interface in which we annotate a single article with contents from other sources, providing multi-source information within a single-source framing.

2.1.3 Multi-Article Interface. Some work has proposed interfaces to contrast and compare biases of multiple sources on a common story. Newscube[51, 52] proposes a clustering-based interface that organizes story sources by viewpoints. Hamborg et al. [20] propose NewsBird, a matrix-based news aggregator that organizes sources based on geopolitical origin, with a follow-up study showing the limited impact on user awareness of content bias[58]. Another approach to tackling news coverage diversity is to diversify the output during the user’s search process, by modifying the recommender engine’s results, for example, 3bij3[43] proposed a common framework to evaluate recommender systems, and Heitz et al. [22] find that diversifying the recommendation of a news search engine can have depolarizing effects on news readers. Prior work most commonly treats individual articles as atomic units, limiting the scope to recommending a more diverse set of articles for a given topic to the user.

In this work, we propose one article annotation interface, and two multi-article interfaces, all leveraging the common Discord Questions framework to extract annotations. Unlike prior work, the annotations are not solely focused on highlighting bias, but more generally the diversity of opinion, by bringing to light questions that receive diverse answers.

2.2 Surfacing Novel and Personalized Content.

Another vein of work focuses on an information retrieval problem formulation, with an objective to assist a reader in discovering novel information, reducing a user’s requirements to search through information surplus [45]. Iacobelli et al. [25] build a system that recommends novel content to news readers, categorizing each recommendation as providing a novel quote, actor, or figure. NewsJunkie [16] leverages user history to tailor a news feed that prioritizes information novelty, while Yom-Tov et al. [63] integrates information diversity directly into search engine results.

Other work has leveraged recent progress in automated question generation [49] to personalize the presentation of news content, such as the question-driven news chatbot [33], or the NewsPod [36] project for automating Q&A-based news podcasts.

In this work, we surface novel answers centered on questions for given stories and do not use personal user information or history to tailor the highlighted content.

2.3 Surfacing Neutral Content.

News bias can be argued to be detrimental to news readers, and some prior work has put forward methods to surface objective content and mute unwanted biases. Babaei et al. [3] leverage social media to find consensus articles, and create a “Purple Feed” that focuses on such articles that are well received by both sides of the political spectrum. Text generation, either through the form of full article generation [38] or multi-document summarization [14, 39] has also been proposed as a method to mute bias in news. Prior work has however shown the negative effects of content moderation [53], finding that opinion heterogeneity provides users with a feeling of fairness. In this work, we view opinion diversity not as a danger but as a necessity, believing in readers’ ability to construct informed opinions.

2.4 Background: Discord Questions

We introduce the terminology used in this work to describe news content production, based on the Discord Question framework [35]. A *news story* represents an event that happened at a specific point in time (e.g., Sweden and Finland applying to join NATO). The news story receives coverage from *news sources* represented by a domain name (e.g., cnn.com). News sources publish *news articles* that typically cover a single news story. A news article is at a minimum composed of a headline and main content, which can be decomposed into paragraphs. We focus on the news stories that receive coverage from at least 10 distinct news sources, in which exhaustively reading all news articles is prohibitively time-consuming.

In the Discord Question framework, news coverage diversity is defined as the formulation of a question, accompanied by diverse – and sometimes contradicting – answers from the sources. The premise of the framework is that each question can serve as an analysis tool to reveal how sources position themselves on a specific aspect of a news story (see Figure 1 for example discord questions). An automatic pipeline is used to generate candidate discord questions and filter them down to a final set of discord questions for a given news story. Each question is paired with a consolidated list of answers from news sources. According to the

²<https://graphics.wsj.com/blue-feed-red-feed/>

framework, a question qualifies as a discord question if it qualifies several properties: it must (a) be answered by at least 30% of the sources of a news story, (b) receive a diverse set of answers (i.e., the majority of answers should not be semantically equivalent), and (c) be specific to the story (i.e., unanswered in other news stories). Property (a) filters overly specific questions (e.g., In what year did the factory first open?), (b) limits consensus factoid questions (e.g., Who is the president of the US?), and (c) filters vague or generic questions (e.g., What did they say?).

Authors of the Discord Questions pipeline experiment with varying models to optimize the pipeline and enable the generation of discord questions for any given news story. The final pipeline is composed of three Transformer-based models: (1) a BART-large[40] a question generation model trained on a combination of the InquisitiveQG [27] and NarrativeQA [29] datasets, (2) a RoBERTa-large [42] extractive question answering model trained on the NewsQA [61] and SQuAD 2.0 [55] datasets, and (3) a RoBERTa-large answer consolidation model trained on the MOCHA dataset [9].

Evaluation of the final pipeline finds that although the automatic generation lags human ability at generating discord questions, the pipeline is able to produce multiple discord questions for any news story containing at least 10 news articles. Manual analysis of the produced questions reveals that discord questions can surface four main types of coverage diversity: (1) differences in the level of detail, (2) differences in the aspect discussed (e.g. a political vs. economics perspective on a question), (3) differences in sentiment, and (4) differences in reasoning.

However, Laban et al. [35] did not integrate the questions into news-reading interfaces or evaluate their usefulness to accompanying readers while they read the news. In this work, we use the existing implementation of the pipeline and focus on evaluating its integration into practical news-reading interfaces.

3 ASSEMBLY INTERFACES

We designed three novel interfaces that incorporate discord questions in varying ways. We attempted to select designs of different novelty and complexity to explore the space of possible designs: the Annotated Article introduces minimal changes to existing news articles and could be implemented as an add-on to a news website, the Recomposed Article remains similar to a news article in appearance but is generated from scratch, and the Question Grid proposes a new information-dense layout. We also reproduce two baseline interfaces – the Headline List and the News Article – that do not present discord questions but represent existing news reading interfaces.

3.1 Annotated Article

In the Annotated Article, illustrated in Figure 2, a *basis article* is selected and its contents are reproduced unaltered. Annotations based on discord questions are inserted between paragraphs of the basis article.

3.1.1 Design Rationale. By leveraging a high-quality article as a starting point and adding further annotation, the interface is likely to follow a coherent reading order, and introduce discord questions when they become relevant to the story. The annotations act as an

add-on to the basis article, each representing optional additional content, which a user can opt into reading.

3.1.2 Implementation. The headline of the basis article is at the top of the interface, followed by a by-line detailing the basis article source, and the sequence of paragraphs of the article. If a paragraph contains the answer to a discord question (i.e., it belongs to the answer group of a discord question), an *annotation* is appended directly after the paragraph. In the example of Figure 2, the second paragraph mentioned that the readers' budget could be impacted by inflation, and the discord question "Who does inflation affect?" is inserted as an annotation.

Annotations are rendered as a collapsible rectangular box, toggleable through a user's click. When collapsed only the discord question is visible, and once opened a list of all answers to the question from other sources' becomes visible. Each answer is inserted on a separate line, with a clickable link to the original source that supplied the answer. All annotations are collapsed initially. In Figure 2, the first annotation is expanded, and the second collapsed.

For a given news story, we automatically select the basis article by counting the number of annotations each news article would have, picking the article with the most annotations. In the extreme, each paragraph of the basis article is associated with a discord question, and the Annotated Article alternates between paragraphs and annotations.

3.1.3 Expected Benefits. Familiarity. An expected benefit of the interface is the ease of use, as a user can choose to not expand the annotations, reducing the interface to reading the basis article, an interface most news readers should be familiar with.

Coherence of Text Order. Because the interface adopts the order of the human-written basis article, which likely follows journalistic guidelines, introducing the required background as needed for an average reader. Because the annotations are placed following the content that they are most related to within the article, they are likely to minimally affect content coherence.

3.1.4 Expected Drawbacks. Limited Coverage. The Annotated Article is limited in the number of discord questions presented to the user, as only questions addressed within the basis article are appended as annotations. For example, the framework might produce a total of 20 discord questions, but the best basis article might only accommodate eight questions, and the user will not be exposed to an additional 12 discord questions.

Low Diversity by Default. Because all annotations are collapsed by default, a second drawback is that without additional user effort, the majority of content originates from a single source – the basis article – reducing coverage diversity. Although it is possible to expand the annotations by default, initial feedback from users showed that could be too disruptive, and collapsing the annotations by default gives the user more control on alternating between reading the basis article and the annotations.

3.2 Recomposed Article

In the Recomposed Article, illustrated in Figure 3, an article is created de novo using the content of several news sources. First, a summary is extracted from one of the source articles, intended to present the basic facts and context of the news story. The second

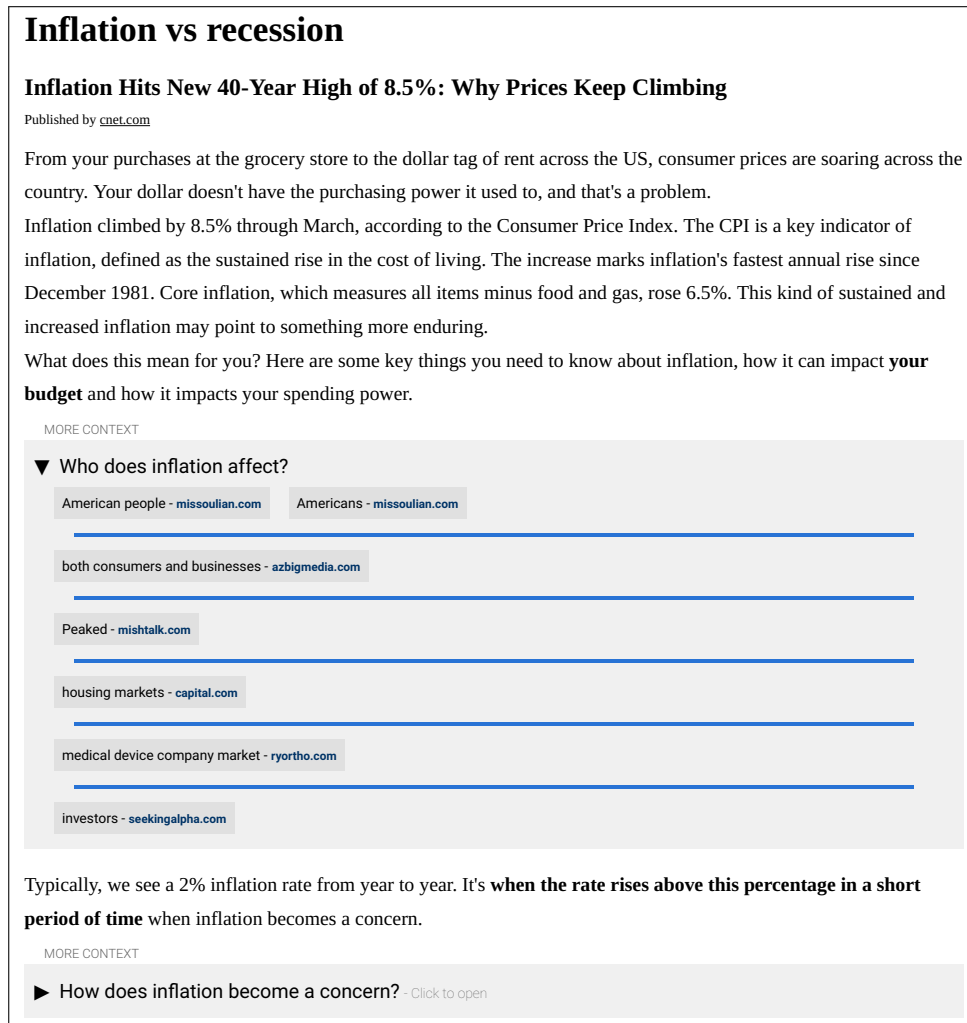


Figure 2: Annotated Article interface. Two discord question-based annotations are inserted into a CNet news article. The first annotation is expanded and the second is collapsed.

portion consists of a sequence of discord questions, each composed of the question itself followed by a list of paragraphs containing answers to the question from the story's sources.

3.2.1 Design Rationale. The Recomposed Article keeps the overall layout of a textual document meant to be read from top to bottom to resemble a standard news article. The two-step reading process, first introducing necessary minimal context followed by in-depth content, follows the inverted pyramid style common in journalism [54].

3.2.2 Implementation. The upper portion of the interface introduces the story name, followed by a by-line of all source articles used in the composition.

Summary Selection: News articles often include a manually written summary within the page's metadata [17]. We extract summaries from all sources and select one closest in length to 60 words.

In practice, we find that this simple heuristic yields summaries that give an appropriate introduction to the story.

Sequence of Discord Questions: A composition algorithm (pseudo-code in Appendix 8) is used to select and order discord questions. At a high level, the algorithm iterates over discord questions, selecting questions that introduce the largest amount of unseen content. With this algorithm, questions that are addressed by more sources tend to come earlier in the sequence, and more specific questions discussed by fewer sources appear later. This ordering mirrors the inverted pyramid writing style.

Visually, each selected discord question is represented by a rectangular unit in the interface. The question is followed by two-paragraph answers from distinct sources, in bullet-point format. For each paragraph, the answer span is bolded. In cases where more answers are available, they are added as a horizontal carousel, allowing the user to dig deeper when interested. Source attribution

Inflation vs recession

Story assembled with content from: [cnet.com](#), [theweek.com](#), [marketwatch.com](#), [mishstalk.com](#), [komonews.com](#), and 41 other sources · Open Grid View

Story Summary:
To combat the highest inflation rate during the height of "stagflation" in 1981, the Federal Reserve on Wednesday announced yet another increase to the central bank's overnight interest rate, hiking it 75 basis points. Overall, the Fed has gradually increased interest rates from near 0 to between 2.25 and 2.50 percent in just four months, the fastest tightening of monetary policy since the Fed's attempt to battle stagflation in the early 1980s.

Q. How does inflation affect the economy?

- With the economy facing headwinds thanks in part to soaring inflation, **both consumers and businesses will see an erosion of their finances and ability to plan for the future.** [azbigmedia.com](#)
- Simply put, inflation is **a sustained increase in consumer prices. It means a dollar bill doesn't get you as much as it did before**, whether you're at the grocery store or a used car lot. [cnet.com](#)

Q. What does the Federal Reserve do?

- To combat the highest inflation rate during the height of "stagflation" in 1981, the Federal Reserve on Wednesday **announced yet another increase to the central bank's overnight interest rate**, hiking it 75 basis points. Overall, the Fed has gradually increased interest rates from near 0 to between 2.25 and 2.50 percent in just four months, the fastest tightening of monetary policy since the Fed's attempt to battle stagflation in the early 1980s. [lobserveur.com](#)
- The Federal Reserve also **plays a significant role in slowing down inflation.** It should continue to raise the Federal Funds Rate and pay interest on reserves while simultaneously reducing the assets on its balance sheet. The Fed will need to pursue a more neutral plan focused mainly on curbing inflation. Unfortunately, the Fed has placed itself in a challenging position to do these things. It procrastinated tightening for too long. [missoulian.com](#)

Figure 3: Recomposed Article interface. In the upper portion, the sources used are listed, followed by an introductory summary, and a sequence of discord questions with corresponding answer paragraphs.

is appended to each paragraph as a blue link, allowing the user to easily access the source of an answer of interest.

3.2.3 Expected Benefits. Article Appearance. The Recomposed Article maintains the appearance of a news article, with a top-to-bottom reading direction, and an inverted pyramid writing style, which should be familiar to news readers.

Full Paragraphs. The Recomposed Article is the only Assembly interface to present full-paragraph answers to discord questions, providing additional context in cases when spans alone can be hard to interpret.

3.2.4 Expected Drawbacks. Lack of Coherence. Since the order of Discord Questions is chosen algorithmically, the overall article will likely lack thematic coherence. **Excessive Length.** The length of the Recomposed Article could be excessive, particularly for stories with a large number of discord questions. The unabridged version of the interface shown in Figure 3 contains 39 discord question units and approximately 3,900 words.

3.3 Question Grid

In the Question Grid, illustrated in Figure 4, the news story is rendered as a two-dimensional grid. Each row of the grid represents a question and each column a source. Each (i, j) -element in the grid is either empty if source j did not answer question i , or a colored shape when an answer is found.

3.3.1 Design Rationale. Inspired by prior work [19] leveraging grid-based visualization of news stories, we adapt the discord questions data to the grid format. The information-dense visualization is intended for advanced users, to help compare and contrast sources, and inspect the framing choices of newsrooms.

3.3.2 Implementation. The grid representation relies on choosing an order for the questions and the sources. Question ordering – similarly to the Recomposed Article – is based on the number of source answers to each question, such that most answered questions are in the upper portion of the grid. Sources are ordered based on the number of questions they answer, with the sources that answer

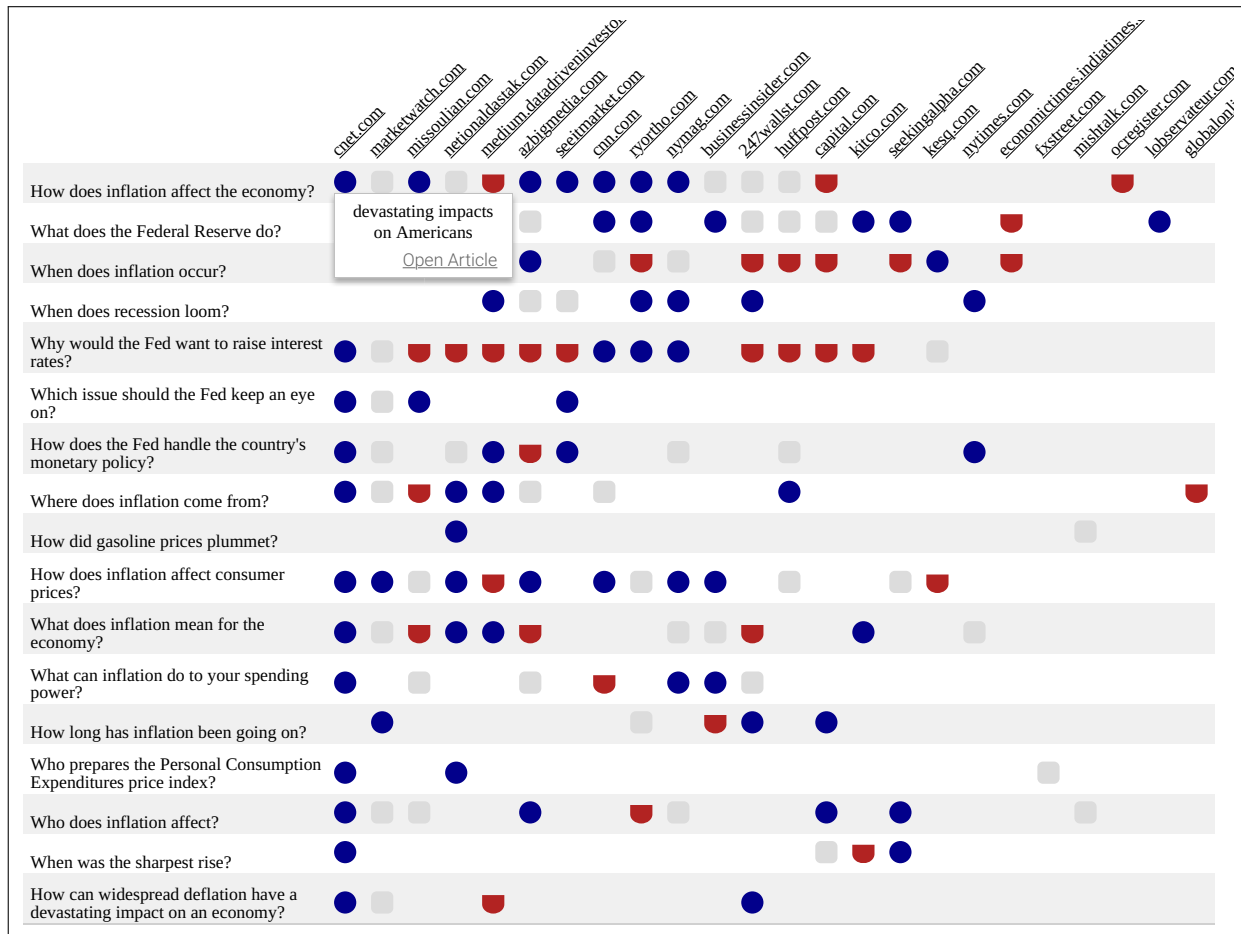


Figure 4: Question Grid interface for a news story on “Inflation / Recession in 2022”. Each row lists a question, each column represents a source. For each element in the grid, the presence of a square indicates a source answering the row’s question, and the color indicates answer similarity. Hovering over a square displays the source’s answer.

more questions in the left portion of the grid. The combined orders result in the upper-left corner of the grid being the most populated, and other areas of the grid gradually losing answer density.

When source j answers question i , a colored shape is inserted in element (i,j) of the matrix. A hover window appears when the user moves the mouse over answer shapes, containing the answer span of the source j for question i . The user can click on the hover window to open the source’s article in a new browser tab.

Color and shape indicate semantic similarity between answers. The Discord Questions framework organizes answers to a question into groups, such that all answers within a group relay similar answers to the question. In the grid, we assign each answer group to a distinct color and shape. For example, the first row of the Question Grid in Figure 4 corresponds to the question: “How does inflation affect the economy?”. The seven blue shapes relay that inflation leads to negative effects on consumers, the three red shapes that it leads to an overheating of the economy, and the five grey that it affects the housing market.

In order to reduce redundancy in the grid, questions that overlap in their answering groups (i.e., the paragraphs that answer the

question) by more than 80% are deduplicated, keeping only the question with the larger number of answers. Deduplication ensures that each question brings on a unique axis of diversity within the story. We note that it is likely that the composition algorithm of the Recommended Article and the deduplication process of the Question Grid select similar questions, leading to the underlying content of the two interfaces to be alike, and the differences exposed in the Usability Studies arising from the presentation of the content.

3.3.3 *Expected Benefits. Information Density.* The Question Grid is the most efficient interface of the three Assembly interfaces in terms of information density. This can be beneficial for users looking to scan for questions of interest, or looking to analyze the coverage of a particular source, as the rectangular layout enables horizontal and vertical inspection. **Source Comparisons.** The matrix-based interface lends itself to the pairwise comparison of sources. For example using Figure 4, it can be visually deduced that BusinessInsider and CNet provide similar coverage of the story, as they answer eleven common questions, whereas MarketWatch and BusinessInsider are more dissimilar, with only five questions

in common. **Answer Omission.** The Question Grid is the only interface to explicitly visualize a source not answering a question, represented by a blank entry in the matrix. As pointed out by prior work[10, 13], surfacing omissions in coverage is an important signal in analyzing source bias.

3.3.4 Expected Drawbacks. Lack of Context. The Question Grid interface does not provide a high-level introduction to the story, potentially causing confusion in users unfamiliar with the story. **Information Overload.** The benefits of information density come at a cost, and it is likely that a densely packed Question Grid might be challenging to some users, due to the overwhelming choice of questions and sources. The requirement to hover over individual answer elements could prove tedious to some users as well.

3.4 Baseline Interfaces

In order to expand the field of comparison during our usability studies, we implement two baseline interfaces: the News Article and the Headline List.

3.4.1 News Article. The News Article corresponds to the unannotated content of a single news article. This interface is equivalent to the Annotated Article with zero annotations. This basic interface is intended to simulate a user reading content from a single source, offering a baseline of coverage diversity from a single source. We take the article of median length amongst available sources, with an objective to represent the average level of coverage of a single news article (rather than the extrema of longest and shortest articles). We do not include a screenshot of the News Article, as it simply corresponds to the Annotated Article in Figure 2 without any annotations boxes.

3.4.2 Headline List. The Headline List iterates over each source presenting solely the headline of the article. A user can then click on the headline to open the full news article associated with the source and headline. This interface is common to news aggregators such as Google News or Yahoo News, with prior studies showing that headline-based interfaces reduce the fraction of participants that delve deeper into news stories beyond the headline [56]. Headline List is intended to simulate users of standard news aggregators. Appendix B provides a screenshot of the Headline List.

4 SYSTEM IMPLEMENTATION

We introduce the data source and computational resources to build the live version of the Assembly interfaces.

4.1 Data Source

When exposing readers to diverse opinions, we have a responsibility to limit the visibility of harmful sources that introduce misrepresentations of important events, which can lead to manipulation of public opinion[1, 30].

There is an engineering challenge in maintaining a list of trusted news sources. Google News – the most popular news aggregator in the US according to Pew Research[5] – bases its news recommendation on content from more than 50,000 sources [15].

Google recently released a document describing the principles behind the source selection for Google News[11], outlining the

manual review process and the editorial expectations for sources within its collection.

In our prototypes, we rely on Google's source selection process, acquiring groups of diverse sources covering a common story directly from the live Google News website. We programmatically visit the Google News pages for World, Finance, Politics, Business, and Science sections, extracting each news story with at least 10 distinct news sources.

For each story, we then directly access each of the source articles and use the newspaper³ library to extract the plain text article. In some cases, an article can only be accessed with a paid account, in which case we only extract basic metadata such as the headline and summary when available.

Although we depend on Google's source selection process, it is not a gold standard and is known to have Western bias [62], for instance with the aggregator recently removing major Russian sources from its platform⁴. Striking the right balance between increasing access to opinion diversity and minimizing harmful content is fundamentally challenging.

4.2 Computational Resources

The Discord Question pipeline processes Google News stories as they are published. On average, the stories contain 37 news articles. QGen takes as input individual news articles and generates questions, producing on average around 987 candidate discord questions per story. QA takes as input each candidate's discord question paired with each news article and extracts an answer when one is found. A third and final process confirms or discards each candidate question, based on whether it receives answers from enough sources and whether the answer set is diverse. On average, the pipeline produces 16 discord questions.

We run the Discord Questions pipeline on a single server equipped with 4 Nvidia V100 GPUs, one allocated to QGen, two to QA, and one to filtering. With the described resources, we are able to process the incoming stream of stories from Google News, on average processing 403 stories per day.

5 USABILITY STUDY A: JOURNALISM EXPERTS

We conducted a usability study with 10 journalism professionals. The objective were to:

- (1) Understand potential use-cases of each Assembly interface,
- (2) Determine how the Assembly interfaces compare to baseline interfaces in terms of ease of use and coverage diversity presentation,
- (3) Assess which of the interfaces are suitable for use by experts and/or novice news readers.

5.1 Participants

We recruited 10 participants (5 women, 4 men, and 1 non-binary, aged between 23 and 71, all living in the US) on a user research recruiting platform⁵. Participants were recruited based on a screener survey, with an objective to recruit journalism professionals with

³github.com/codelucas/newspaper

⁴reuters.com/technology/google-drops-rt-other-russian

⁵<https://www.userinterviews.com>

diverse experience. In the screener survey, participants reported having 1-35 years of experience in journalism (mean 10.4 years), and working at both local news organizations (e.g., Arizona Pbs, Laredo Morning Times) and national news organizations (e.g., CNN, NBC News). In terms of roles within the newsroom, 9 participants listed reporting/writing, 4 listed editing, and 2 producing. Finally, participants listed different specializations, from Local Politics to Crime & Courts, and Technology.

5.2 Study Procedure

The study was conducted in 1-hour sessions online via video-conferencing software, with participants receiving \$80 upon completion. The interface and study procedure were approved by the ethics review agent of our organization.

Each participant first watched a 3-minute introductory video introducing terminology and features of the five interfaces (2 baselines, 3 Assembly). Participants then completed two 25-minute story-reading sessions each involving all interfaces, and finally completed an open-ended feedback form.

The two story-reading sessions were centered on major US news stories at the time of the study: the Baby Formula Shortage, and Potential Recession. Each story-reading session consisted of interface browsing and a comparison questionnaire. During interface browsing, participants spent 4 minutes viewing the story with each interface, going in the following order: News Article, Headline List, Annotated Article, Recomposed Article, Question Grid. Participants were instructed to browse the interface with the objective to learn about the story and were permitted to click on external links when available.

In the comparison questionnaire, participants provide 5-pt Likert scale ratings of the interfaces. Rather than requesting ratings after using each interface, we waited until participants had seen all interfaces and used a grid of Likert scales to allow participants to rate interfaces relative to each other. Before they provided ratings, we confirmed with each participant that they were familiar with each interface's name, and allowed participants to go back to individual interfaces as they answered. Figure 5 summarizes comparison questionnaire results.

5.3 Study Results

5.3.1 Revealing Coverage Diversity (Figure 5c). A larger proportion of experts found that the Assembly interfaces revealed coverage diversity in the story than the baseline interfaces. The Question Grid achieved the highest ratings, followed by the Recomposed Article and the Annotated Article. The single News Article was rated largely lower, confirming that content from a single source does not provide satisfactory coverage diversity for complex stories. The Headline List – a popular interface in existing news aggregators – obtains ratings between the News Article and the Assembly interfaces, confirming that it is partly successful at conveying coverage diversity.

The Discord Questions-based annotations are the only difference between the News Article baseline and the Annotated Article, yet lead to a large rise in ratings. This difference highlights that a news article can provide support to introduce a news story, and annotations can be inserted to increase coverage diversity.

The Recomposed Article and Question Grid achieve the highest ratings, with the latter slightly better rated. Although both interfaces relay largely the same information (see Section 3.3), we hypothesize that the increased information density of the grid gave experts a perception of better access to coverage diversity.

5.3.2 Good Overview & Coherent Order (Figure 5a-b). The Annotated Article achieves the highest ratings in terms of both coherence and providing an accessible overview of the story. This confirms that the method of adding annotations directly after the related paragraph in a human-written news article is the least disruptive way to introduce discord questions content without reducing story accessibility.

The two more advanced Assembly interfaces – Recomposed Article and Question Grid – obtain worse ratings overall, with the Question Grid lagging other interfaces in terms of giving a good overview of the story. These ratings validate that the pair achieves higher coverage diversity at the cost of accessibility.

The two baseline interfaces are rated between the advanced Assembly interfaces and the Annotated Article, and are overall seen as providing good introductory material to news stories.

5.3.3 Ease of Use (Figure 5d-e). Expert participants were asked to estimate whether the interface would be easy to operate for other media professionals, as well as for novice news readers without expertise in journalism.

With respect to an expert population, all interfaces received high ratings, with 80% of experts agreeing that fellow experts should be able to use any of the interfaces. The most information-dense interface – the Question Grid – received slightly lower ratings, coming from two participants that found the grid hard to understand (more detail in qualitative feedback §5.4).

With respect to a novice reader population, rating distributions showed a larger variance. Only three interfaces were considered suitable by a large majority of participants: the two baselines and the Annotated Article. The Recomposed Article and Question Grid both received lower ratings, with no participant strongly agreeing that the Question Grid is suited for novice users. We hypothesize that lower ratings are due to the two interfaces not being anchored on a news article and as such not providing the introductory context that novice readers might require.

5.4 Qualitative Feedback

Participants were given time to provide open-ended feedback and were asked to reflect on: (1) the limitations or benefits of any interface, (2) potential use cases of any interface, and (3) any aspect that did or did not work well. We employed a thematic analysis [8] to organize the feedback, and we discuss themes brought up by three of the ten experts or more.

5.4.1 Preference for the Annotated Article – 8 participants. Eight of the ten participants explicitly expressed an overall preference for the Annotated Article, mirroring the interface's high ratings in Figure 5.

"I really thought the annotated article was well done." – (P10)

We observed during the interviews that participants interacted differently with the annotations, with some choosing to first read

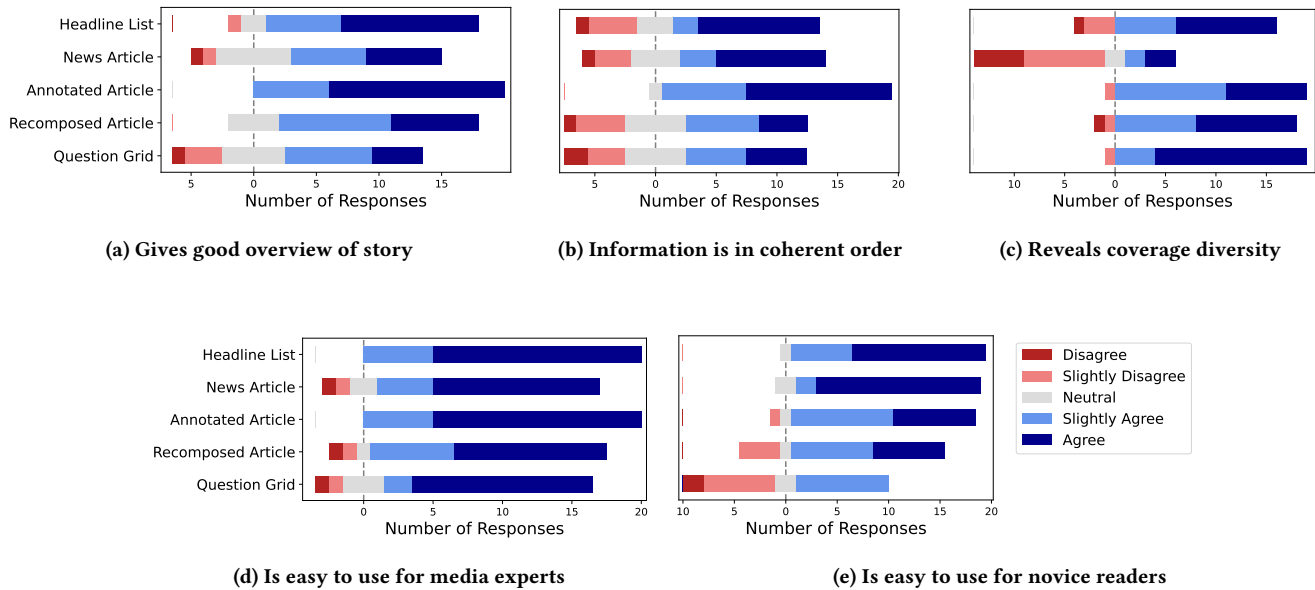


Figure 5: Expert Comparison Questionnaires Results. Participants rated each interface on whether it: (a) provides a good overview of the story, (b) presents information in a coherent order, (c) reveals coverage diversity, (d) is easy to use by experts, and (e) easy to use by novice users.

the article, followed by opening some annotations, and others choosing to open each annotation as they appeared in the article. Overall, participants often expanded 50-80% of the annotations of the article to see multi-source answers.

5.4.2 News Article does not provide coverage diversity – 5 participants. Half the participants noted the lack of coverage diversity in the single news article baseline interface. This finding serves as a justification for the other four interfaces which integrate multi-source content to achieve higher coverage diversity.

“I did not see a lot of use for the single article – I am not interested in having my reading tailored that specifically.” – (P5)

5.4.3 Question Grid Use Cases – 5 participants. First, several participants saw an opportunity in the blank spaces of the Question Grid. By seeing the parts of the Grid that lack coverage, a newsroom could decide on story angles to produce stories that are differentiated from the competition.

“I could see journalists using the question grid as a way to inspire different angles in stories, as well as start researchers down different paths of questioning.” – (P3)

Second, some participants proposed processing the grid itself to inspect story coverage. By analyzing which questions are more or less answered, a journalist could computationally discover the frames of a news story.

“I can also see the question grid being used as a kind of data source to make an argument about news diversity itself” – (P1)

However, two participants expressed difficulty in using the Question Grid, finding the interface sometimes overwhelming, which would require better training to operate.

“The grid was a bit difficult to understand at first. It’s a good tool but needs a bit more of explanation.” – (P8)

5.4.4 Discord questions can be noisy – 4 participants. Although participants were not told the questions and answers were automatically generated, some spotted noisy elements in the questions or answers.

“Some questions that were visible in the interface had inaccurate or imprecise “answers” listed” – (P9)

To counter the imperfections, one participant suggested manual editing would be required before publishing.

5.4.5 Trustworthiness of Sources – 2 participants. We further spotlight two insights from individual participants that are relevant to future work and the limitations of our work. First, during the study, P5 was visibly irritated by the opinion of certain sources that were being highlighted in some interfaces. In their feedback, P5 mentions the importance of the user’s trust in the sources included in the interface:

“I need assurance that the search algorithm that produces the stories I see is not slanted toward any particular results and has a reasonably broad scope. [...] Again, that is only useful if I trust the sources.” – (P5)

By leveraging the source selection process of Google News, we ensure some level of quality of included sources. Source trust is however user-specific, and an empowering solution for future work would be to give the user control over source inclusion, allowing them to filter out undesired sources.

On a related note, participant P10 pointed out the danger of plurality of answers at the center of the Discord Questions framework when a single answer is correct, again hinting at trust in sources being important:

“The downside to that [discord questions] might be all the different answers you get. It could be confusing for a reader to know which source is the trusted one with several different answers.” – (P10)

Diversity in answers is valuable in some cases but can be detrimental when it is used to spread misinformation. Providing the user information about included sources and control to remove unwanted sources could be implemented in future work, reducing reliance on Google’s source selection process.

5.5 Summary of Results

In summary, only three of the five interfaces were estimated to be operable by novice news readers. The only Assembly interface to make the cut – the Annotated Article – was particularly well received, as it was rated to provide the best overview of the story while highlighting almost as much coverage diversity as more advanced interfaces.

The two advanced Assembly interfaces – Recomposed Article and Question Grid – maximize exposure to coverage diversity at the cost of ease of use and are limited to subject-matter experts looking to deepen their understanding of a story. Surprisingly, although the Question Grid was rated as slightly more challenging to use, participants were more vocal about its potential use cases in their open-ended feedback. The Recomposed Article – which can be thought of as a middle-ground between the other two Assembly interfaces – did not find its preferred audience, as the Annotated Article was a favorite for introductory use cases and the Question Grid for advanced analysis.

Following the recommendation of the participating experts, we exclude the Recomposed Article and the Question Grid interfaces from the usability study involving novice news readers, focusing on the three interfaces that were predominantly rated as easy to use by novice news readers.

6 USABILITY STUDY B: NOVICE NEWS READERS

In a second usability study with novice news readers, the objectives were to:

- (1) Assess whether the Annotated Article leads readers to gain a broader understanding of a news story compared to baseline interfaces,
- (2) Verify whether the Annotated Article is as straightforward to use as baseline interfaces,
- (3) Understand user pain points in using the News Article, the Headline List, and the Annotated Article.

Measuring reader exposure to coverage diversity is challenging [58], prompting us to design an active reading exercise to testbed interfaces (§6.1). We then detail the study protocol (§6.2), report on recruited participants (§6.3), detail the manual analysis we conducted (§6.4), and analyze quantitative and qualitative outcomes (§6.5-6.6).

6.1 Reading Exercise Design

We pose three requirements for the study design. Requirement I: the study should simulate a realistic reading scenario for a novice reader (i.e., it should not require participants to use an interface for several hours, or require advanced training). Requirement II: the study should strive to be topically interesting to the user to maximize genuine interaction from the user while being relatively novel to minimize the effects of prior knowledge from participants. Requirement III: the study should be reproducible and not provide an unfair advantage to one of the settings. We introduce the proposed study design, explaining choices made with respect to the target requirements.

Overall exercise. The exercise is a time-limit reading comprehension exercise consisting of four open-ended questions. During the exercise, a participant has six minutes to answer the questions using a single interface, assigned at random. The participant is urged to answer the questions in the bullet-point form and as thoroughly as possible, listing any answer elements they read.

Concurrent reading and answering. A major design choice is whether the participant completes the comprehension questions during or after the reading session. We choose a concurrent design using a two-column interface shown in Appendix C. The advantage of the concurrent design is that it does not rely on participants recalling answers, and allows participants to actively use the interface with an objective, avoiding passive navigation (requirement II).

Story Selection. Participants are given story choices and are prompted to select a story they are interested in but not up to date on. It is important that the participant is both interested to generate genuine interaction (requirement II), without having participants with too much prior knowledge that would bias results (requirement III).

Unbiased Question Selection. Questions selection should be interface-independent. For example, the questions should not be taken from the discord questions in the Annotated Article, as this would bias results and reduce reproducibility (requirement I). As further detailed in Section 6.2, we hired external experts to write questions based on their review of the study’s selected news stories.

Time-limited Exercise. The average reading time for long-form news articles in 2016 was a little over 2 minutes [46]. By setting the exercise duration at 6 minutes, we account for the additional time needed to write down answers and give participants roughly the reading time of a realistic news-reading session (requirement I). Once the time has elapsed, a message is shown and participants are given up to three additional minutes to finalize their answers.

Open-ended Short-form Questions. We select comprehension questions that require reasoning, prediction, and generally short-form answers, rather than factoid questions that would focus the exercise on narrow fact-finding. We did not impose within the interface the use of bullet points, as we performed manual scoring of the answers (see more in §6.2) and can in this way detect low-quality participants.

Allowing no-ans. Participants were explicitly told it is acceptable to leave an answer blank or answer it with “No answer”. Although we expected most questions to be answerable with the three interfaces, we believe that participant perception of a lack of

answers is an interesting signal. Some participants might abuse the ability to leave answers empty, and we discuss how we filtered out such participants in §6.3.

6.2 Study Procedure

6.2.1 Study Content. We selected the two expert study stories – Baby Formula Shortage, and Potential Recession – and three new stories: New Iran Nuclear Deal, Taiwan/China Diplomatic Row, and Sweden/Finland Joining NATO.

To generate comprehension questions, we re-contacted three experts from the first usability study. Each expert spent at least 10 minutes getting informed on each story, and was tasked with writing 4-6 questions (without answers). Question-writers were asked to favor core questions with multiple answer aspects or opinions. Through this process, we obtained roughly 12 questions per story and selected a final four by deduplicating and prioritizing questions suggested by several experts. As an example, comprehension questions for Taiwan/China Diplomatic Row are: (1) “How is China responding to separatists in China?”, (2) “How is Taiwan responding to Chinese threats?”, (3) “Why was Nancy Pelosi visiting Taiwan?”, (4) “What did the Chinese government say in a white paper?”.

We manually verified that no selected question matched verbatim to any of the discord questions within the Annotated Article, confirming no interface provides an unfair advantage. We estimated that for roughly 30% of comprehension questions, the answers to a discord question in the Annotated Article could be directly useful in finding answers, however, we judge this to be a confirmation of the ability of discord questions to surface salient content, rather than an unfair advantage since the surface question formulations were different.

We studied the answerability of each comprehension question based on the information in each interface. We found that at least one answer element is provided for each question in the Headline List and Annotated Article interfaces, but two of the twenty questions were unanswered in the News Article interface, affecting the completability of the questionnaire in that setting. We discuss question choice further in the Limitations section.

6.2.2 Study Protocol. The study had a target duration of 25 minutes, split between an introduction, three reading exercises, and a feedback form. Participants first viewed a timed slideshow presenting the reading exercise and filled out a news reading habits form.

Afterward, participants completed three 6-minute reading exercises. For each exercise, they selected a new story from the five available and were assigned to a random interface, in such a way that each participant completed one exercise with each interface – the News Article, the Headline List, and the Annotated Article – in random order.

Finally, participants filled out a completion survey reviewing their experience, both prompting for interface ratings and open-ended feedback.

6.3 Participants

Participants were recruited on the Amazon Mechanical Turk crowdsourcing platform⁶, completing the 25-minute unmoderated task for

compensation of \$7 (\$16/hour). We set requirements for recruited participants: be residents of the US, have completed 5000+ tasks on the platform, and have a 98%+ acceptance rate. Participant results were manually reviewed to further improve result quality.

In total 110 crowd-workers completed our task. Participants were removed from the analysis for one of three reasons: (1) they completed the study in 20 minutes or less (impossible without cheating disabled buttons), or (2) they navigated to other tabs more than three times during the study (we used JavaScript to detect tab switching and warned participants to remain focused on the task), or (3) their feedback hinted at spam participation (e.g., in the open feedback: “LIFE IS GOOD”). Fifteen participants were filtered-out, and we see the added overhead as a cost of ensuring reproducible results. The results presented are based on task completion from 95 participants we believe completed the task to the best of their ability.

News reading habit questionnaire results are summarized in Figure 6. Roughly 80% of the participants access the news daily, with the five most common platforms being: social media, web news, mobile application, TV, and newspapers. About 85% of participants at least slightly agreed that the news they read can be biased and that it is difficult to get the complete picture of some news stories. Most participants agreed that they try to obtain their news from multiple sources.

The results confirm that ordinary news readers struggle to obtain broad coverage diversity of some news stories, and put additional effort into reading from multiple sources, justifying the goals of the Assembly prototypes we built.

6.4 Manual Scoring of Answers

Once the study was completed, we manually reviewed the answers. For each question, we extracted all unique answer aspects, considering any aspect that could be a direct answer to the question. For example, eight answer aspects were extracted for the question “What are the reasons for the Baby Formula shortage?": (1) the Abbott plant closure, (2) supply chain issues, (3) low retailer stocks, (4) hard-to-find specialized products, (5) a recall due to child sickness, (6) hoarding by some consumers, (7) government inaction, and (8) high prices. Each answer was then scored based on the number of aspects it mentioned. Additional details on the annotation procedure are included in Appendix D.

We then aggregated the scores based on the reading interface used and computed four metrics: (1) the average score (**Score**), (2) the percentage of answers left blank intentionally (**%No Ans**), (3) the percentage of answers with a score of 0 (**%S0**), and (4) the percentage of answers with a score of two or more (**%S2+**). We hypothesize that an interface that exposes its user to more coverage diversity will lead to higher average scores and multi-aspect answers (**Score**, **%S2+**), and lower unattempted and zero-score answers (**%No Ans**, **%S0**).

6.5 Study Results

6.5.1 Quantitative Results. Table 1 summarizes the main quantitative results, based on the manual score analysis as well as statistics collected during the study. In terms of statistics, we measured how frequently (**%Any L**) and how many (**#Links**) links users opened during each exercise (apart from the News Article which does not

⁶<https://www.mturk.com>

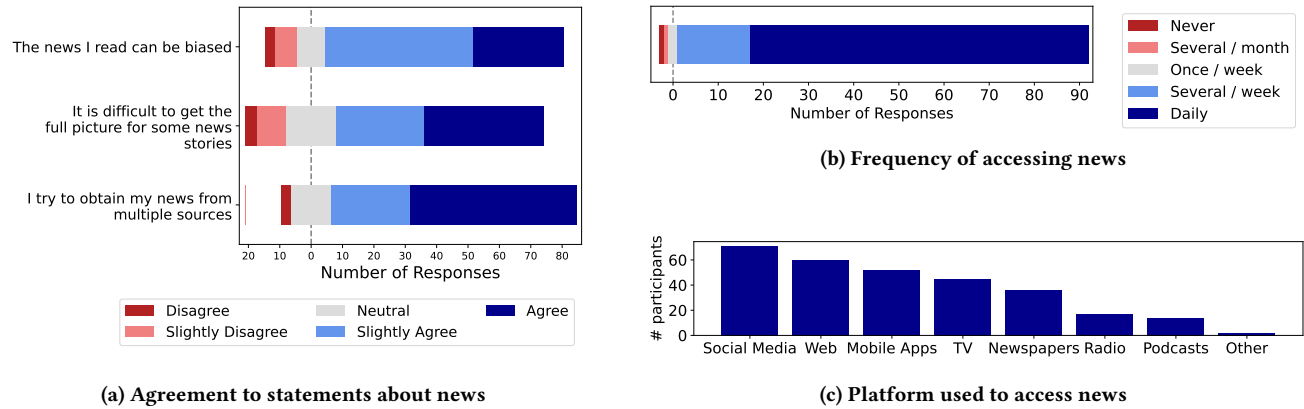


Figure 6: Responses from novice participants on news reading habits. Participants were asked: (a) whether they agree with three statements about news consumption, (b) how frequently they access the news, and (c) which platform they use.

Interface	Answer Scores				Statistics			
	Score \uparrow	%No Ans \downarrow	%S0 \downarrow	%S2+ \uparrow	#Links	%Any L	#Words \downarrow	#Min \downarrow
News Article \clubsuit	0.96	22.1	36.5	22.9	–	–	457	6m13s $\diamond\heartsuit$
Headline List \diamond	1.37 \clubsuit	6.0 \clubsuit	20.3 \clubsuit	39.4 \clubsuit	4.3	83.1	2490	8m46s
Annotated Article \heartsuit	1.61 $\clubsuit\diamond$	7.6 \clubsuit	15.0 $\clubsuit\diamond$	48.4 $\clubsuit\diamond$	1.3	42.1	1570	7m05s
Upper-Bound	7.58	–	–	100	46.8	–	21100	–

Table 1: Quantitative results of the reading exercise. Participants used three interfaces: News article, Headline List, and Annotated Article. Each answer was manually assigned a score based on the number of answer aspects it provides, %No Ans, %S0, and %S2+ are the percentage of non-attempted answers, answers with a score of 0, and answers with a score of two or more. #Links, %Any L, correspond to the number of links and percentage of exercises where 1+ link was opened. #Words is the number of words shown to the users, and #Min average time spent. Column arrows indicate whether higher (\uparrow) or lower (\downarrow) values are better. We provide upper-bound values for certain elements. Each interface is assigned a symbol ($\clubsuit\diamond\heartsuit$), which is used to signify pair-wise statistical significant difference ($p < 0.05$).

provide external links) and used link openings to extrapolate how many words were presented to the user (#Words). We could not measure how many words users actually read, as we did not use eye-tracking technology. To compare quantitative results across settings, we use a t-test to measure pairwise statistical differences.

We find that exercises completed with the Annotated Article lead to the highest overall comprehension scores (mean 1.61), statistically higher than the Headline List (mean 1.37) and the News Article (mean 0.96), confirming that discord question annotations succeed at exposing users to coverage diversity in a realistic reading scenario.

The percentage of questions left blank is much larger for the News Article interface (22.1%) than the other two (6–8%), reflecting the lack of exposure from reading a single news article. News Article users’ answers receive a score of zero 36.5% of the time, more than twice as much as the users of the Annotated Article (15.0%), showing that annotations in the article help participants locate more answer elements and answer a larger fraction of the questions.

The Annotated Article leads participants to produce significantly more multi-aspect answers (%S2+), with almost half the answers

achieving such a score, compared to 39% for the Headline List and 22.9% for the news article.

Headline List users predominantly needed to open external links (83% of the time) showing that simply reading headlines is not sufficient to answer news reading comprehension questions. Because Headline List users opened an average of 4.3 news articles, they were expected to read or skim through much more content (on average 2,490 words), which would take 16 minutes to read at a pace of 150 words per minute. This content overload explains why Headline List participants remained in the interface longer than the two others, staying more than two additional minutes to complete the exercise compared to the delimited time.

In comparison, Annotated Article users completed in roughly 7 minutes, and opened an external link 42% of the time, with an of average 1.3 links in each session. This is noteworthy, as Annotated Article users were able to achieve higher scores while viewing less textual content (1,570 words on average). This finding underlines the efficiency of discord questions in highlighting coverage diversity, which allowed participants to understand more with less effort.

Interface	%1-side ↓	%Hypo ↑	%2-Side ↑
News Article ♣	26.7	53.3	20.0
Headline List ◇	34.6	46.2	19.2
Annotated Article ♥	14.3	23.8	61.9 ♣◇

Table 2: Breakdown of answer categories for prediction question. When asked to predict the future likelihood of a recession, participants could answer with a one-sided assurance (%1-side), with a hypothetical (%Hypo), or with a two-sided answer (%2-Side). Each interface is assigned a symbol (♣◇♥), which is used to signify pair-wise statistical significant difference ($p < 0.05$).

News Article users were only given a single news article with no external links, and they completed the task mostly on time at the cost of low scores, leaving 22% of the questions blank. However, they were still able to provide multi-aspect answers for 22.9% of questions, providing evidence that single news articles occasionally provide multi-perspective news coverage. The quality of the chosen news article has a direct effect on performance, and this study attempts to establish average performance by selecting a news article of median length. Future work can expand on this by studying the extremes as well, the average score received using the longest and shortest news article in the collection.

When looking at upper-bound values, the stories in the study had an average of 46.8 sources, totaling on average 21,100 words per story, which would take more than eleven hours to read in its entirety (at a rate of 150 wpm) for the five stories, underscoring the impossibility of exhaustive reading. The maximum achievable score (7.58) largely surpasses average interface scores (0.9-1.6) highlighting that the designs we propose are the first steps in improving access to news coverage diversity, and there is a large room for improvement.

6.5.2 From Exposure To Persuasion. Overall, the exercise we design focuses on exposure and does not evaluate persuasiveness. Out of the twenty comprehension questions, one of them asked participants to make a prediction about future events, allowing us to investigate participant perception of future events. More specifically, in the “Potential Recession” exercise the third comprehension question asked for a prediction: “Is there going to be a recession?”. We isolate answers to this question for detailed inspection.

In total, 62 participants selected this story. By manual analysis, answers were assigned to three categories: (1) **one-sided**, when participants expressed certainty on the outcome (e.g., *We are already in it!*, P31), (2) **hypothetical** when participants express some level of uncertainty (e.g., *Likely. Might be avoided after consumers return to normal spending patterns*, P47), or (3) **two-sided** when participants express two opinions or more (e.g., *Yes, if recession proves to be the only way to get inflation under control. No, if the Fed can engineer a “soft landing,”*, P9).

We argue that a successful interface should accompany readers in discovering the uncertainty, reducing the proportion of one-sided answers, and increasing two-sided and hypothetical answers. Results are summarized in Table 2.

Only 14.3% of Annotated Article users gave one-sided answers, compared to 26-34% of baseline interfaces, but the difference is not statistically significant. However, an increase in two-sided answers with the Annotated Article is statistically significant, showing that the Annotated Article encouraged participants to consider several potential alternatives. This result is limited yet encouraging, and we encourage future related work to increase the proportion of open-ended prediction questions, as they are a useful tool in measuring participant persuasion.

6.5.3 Completion Questionnaire. In the completion questionnaire, participants were asked to rate each interface on two features: whether it is easy to use, and whether it highlights coverage diversity. We labeled each interface with the name of the story they read in it, to avoid interface name confusion.

Regarding ease of use, the News Article was largely preferred, followed by the Annotated Article and finally the Headline List. The pronounced gap was not predicted by experts, which had estimated all three interfaces to be usable by novice users. We hypothesize this gap is due to the difference in study objective: experts browsed the interface freely, while novice users were completing a comprehension exercise, which might affect impressions of ease of use.

Regarding coverage diversity highlighting, the News Article obtained the lowest ratings, and the Headline List and Annotated Article were virtually tied. Surprisingly, although Annotated Article users scored significantly higher on the comprehension questions, they did not rate the interface higher in terms of coverage diversity. This shows the difficulty for users of news-reading interfaces to directly assess coverage diversity exposure, similar to findings from prior work [22], and the usefulness of comprehension questionnaires in measuring user exposure.

6.6 Qualitative Feedback

In their open-ended feedback, participants were asked to reflect on: (1) their favorite and least favorite interface, and (2) any aspect that did or did not work well. We employed a thematic analysis [8] with the feedback, grouping by the interface, and filtering to themes mentioned by at least five participants.

6.6.1 News Article. The News Article was the favorite interface of 10 participants, and the least favorite of 35. Twelve participants mentioned ease of use as the main benefit (e.g., *“Of course the single story is very easy to read”* - P66). The central disadvantage discussed by 30 participants is insufficient information preventing participants from answering all comprehension questions (e.g., *The basic news article was my least favorite as it did not provide enough information.* - P8).

6.6.2 Headline List. The Headline List was mentioned as a favorite by 20 participants, and least favorite by 20 participants, showing a polarity among participants. Two positive features stood out, first, 13 participants thought the interface gave access to the most information (e.g., *because I was able to see so many more options and ‘sides’ to the story* - P26), and second, seven appreciated doing their “own research” (e.g., *The headline list was more like a traditional search engine result page and so it was the most known interface.* - P71).

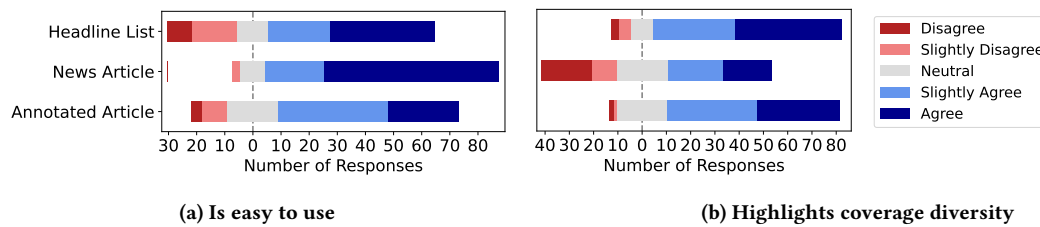


Figure 7: Results of Novice Completion Questionnaires. Upon completing the three reading exercises, participants were asked to rate each interface on whether it is easy to use, and if it highlights coverage diversity (a term definition was provided).

Negative features were more diverse: 16 participants complained of a lack of time (e.g., *it would be the most useful with a lot more time to get used to it* – P5), 15 struggled with opening articles that were later behind a paywall, deleted, or with too many ads (e.g., *but there were so many with subscription blocks and ad-littered articles.* – P59), 11 disliked the back-and-forth required between picking a headline and reading the full article (e.g., *The headline list was way too broad [...] since you had to go to the articles and read them to see if they even answered questions.* – P67), 7 did not like the sources listed (e.g., *The headline list seemed to only be 'western' or 'western' friendly perspectives, nothing from China media/propaganda.* – P89), and 7 expressed distrust of headlines due to their potential click-bait and vague nature (e.g., *It is harder to sift through it all with just headlines which in my experience[sic] are sometimes misleading.* – P41).

6.6.3 Annotated Article. The Annotated Article was the favorite interface of 35 participants, and the least favorite of 12. There were three negative aspects discussed: 12 participants found the layout of the interface and particularly the annotations challenging to navigate (e.g., *but the dropdown lists were sort of jarring* – P84), 11 participants found the annotations could provide incomplete information (e.g., *and wasn't always the most accurate in the information it provided.* – P71), and 5 participants found that annotations were not placed in a pertinent position (e.g., *[did not] connect in an organic way to their placement within the larger article* – P9).

In terms of positive features, 17 participants found the positioning of annotations to be adequate, sometimes coming as reflections occurred while reading the article (e.g., *Gave the story and added links to how you can learn more information without overtaking the article.* – P86).

6.7 Summary of Results

In conclusion, the execution of the reading exercise with 95 participants to compare three interfaces – News Article, Headline List, and Annotated Article – reveals significant differences in access to coverage diversity between interfaces. Annotated Article users are more successful at answering comprehension questions with multiple aspects while finding the interface slightly easier to use than the Headline List.

The Annotated Article is however still more challenging to use than an (unannotated) news article, with imperfections in the automatically generated annotations and dropdown design of the annotations causing some user dissatisfaction.

The Headline List falls between the other two interfaces, exposing users to more diversity than a single News Article but less

than the Annotated Article. The interface received more polarized feedback, with some users liking the manual aspect of doing their own research, while others were frustrated with having to go back and forth due to uninformative news articles. We hypothesize that the Headline List is most adequate in longer reading sessions when a user has ample time to navigate between articles (e.g., 1 hour to do research on a topic), but less adapted to shorter reading sessions, explaining the frustration of some of its users in our 6-minute exercises.

In Appendix E, we assess the reproducibility of the study design through bootstrap re-sampling [12] and find that statistical significance mostly holds when varying the selection of stories and that result reliability requires recruiting at least 60 participants.

7 DISCUSSION & LIMITATIONS

Imperfect NLP Methodology. By choosing to work with an existing NLP framework, we evaluate automatically generated interfaces, measuring whether the value added by discord questions outweighs the noise introduced by imperfect NLP. For example, in Figure 2, the framework extracted the answer “Peaked” to the question “Who does inflation affect?”, which is invalid. The findings are therefore tied to model quality and are likely to change as NLP methods mature. We considered manually post-editing discord questions in the study interfaces to obtain feedback on an ideal version of the interfaces. Still, we preferred the realistic scenario of the fully automatic interface. As suggested by experts in §5, post-editing in a newsroom could ensure the quality of discord questions for production settings.

Amazon Mechanical Turk population bias. Even though recruiting through a crowd-sourcing platform typically leads to more representative participants than on-site recruitment (typically, undergraduate students at a university) [6], the crowd-worker population differs from the U.S. population on many metrics (e.g., age, gender, income level) [26] which would likely have an impact on the results of our second usability study. All participants also received a monetary reward for completing the study, which could affect the authenticity of the interactions we base our results on. We aim to release a public version of prototyped interfaces, which would allow us to observe users interacting with our tools in a more genuine setting.

Completeness of the Comprehension Questionnaire. We relied on experts to select comprehension questions to evaluate news readers on. In order to avoid biasing the question towards information in one interface, in particular, the experts were given

access to all the content of the news event and would propose comprehension questions irrespective of the studied interfaces. Because of this choice, some of the selected questions are not answerable for some interface-story combinations, reducing the completeness of the comprehension questionnaire in some settings. We believe that enforcing the answerability of the questions would be detrimental, as it would lower the difficulty of the comprehension questions which could artificially favor simpler interfaces. We did not however analyze the effect of unanswerability on our results.

Realistic Reading Setting. The participants in the reading exercise were instructed to complete the comprehension questionnaire while making use of the reading interface in a limited time, making for a more active and focused reading setting which is not representative of more passive news reading. Some participants noted in their feedback that the exercise felt like a research assignment rather than news reading. It is possible that the effect of the discord questions would vary in a more passive news reading scenario, and future work could focus on passive reading scenarios.

Recomposed Article Limitations. The Recomposed Article design can be seen as a middle-ground between the more straightforward Annotated Article and the information-dense Question Grid. The interface received the least enthusiastic feedback from interviewed experts, which found it more confusing than the Annotated Article, while not offering as many analysis opportunities as the Question Grid. We hypothesize that the sorting algorithm in Figure 8 used to compose the article is overly simplistic and leads to a lack of story coherence. The idea of composing novel articles that efficiently present points of discord remains promising, and perhaps generative question-answering models [31, 59] will offer new avenues for improvements in this domain.

Headline List Order. Some participants found the order of headlines in the interface to lack meaning. We simply reproduced the order in which headlines appeared on Google News' original page, and did not further deduplicate or re-order headlines. Prior work could be integrated to organize the headlines into groups [4, 32], or provide information on relations between headlines [18].

Discord Questions for non-news data. In this paper, we focus the interface design and usability study on the news domain, and interview journalists with expertise in news production. However, the Discord Questions framework [35] is not restricted to the news domain. It could prove useful in other multi-document exploration settings, such as helping shoppers navigate 100+ reviews of a product, or instructors navigate the end-of-course student feedback. Although some design components in our work might transfer to new domains, others require domain-specific adaptations.

English-Only Data Source. Our current prototype is inherently limited due to our focus on English-written news sources, as coverage diversity on international topics is likely to come from non-English news sources [44]. However, improvements in automatic news translation [60], as well as multi-lingual models [24] draw a path towards a multi-lingual version of our prototype.

Realizability of Advanced Use Cases. Several of the interviewed experts were enthusiastic about the Question Grid, proposing several scenarios within a newsroom that could benefit from a Question Grid interface. In this paper, we did not follow up with

these ideas, focusing instead on evaluating novice-compatible interfaces. Future work can however collaborate with members of newsrooms to tailor a Question Grid-like interface to journalism applications.

8 FUTURE WORK

Long-Term Effects. Participants in our studies spent at most 20 minutes with a given interface: enough time for an initial opinion, but not enough for prolonged use judgment. Repeated access to diverse coverage for an ongoing story might have complex effects on readers, either gradually assisting them in diversifying their understanding, or causing them to distrust certain sources over time [28]. Extending the reading exercise to a longitudinal study could provide insights into the interface's long-term effects, but would likely prove challenging to implement. Prior work has built interfaces specific to long-ranging news stories [34, 57], and their adaptation with the Discord Question framework is a promising research direction.

Design and Deployment Recommendations. The findings from our usability can provide insights to designers of future news-reading interfaces eyeing to facilitate access to diversity in news coverage. First, questions – generated or manually curated – can serve as a tool for alignment of source stances and can trigger curiosity in the reader. Second, as the user reads a single source content, the in-context positioning of the additional source's opinion or analysis (such as in the Annotated Article) can effectively indicate to the user which parts of the news story are complex, and lower the barrier to access broad coverage for curious readers. Finally, in multi-source presentations, transparently indicating the origin of each content element, and the use of hyperlinking to allow access to original source presentation is crucial to maintain user trust.

Reproducibility of Results. We designed the reading exercise with the objective to maximize reproducibility and minimize potential bias toward any of the conditions in the study, for example by recruiting external experts to select the comprehension questions, or through an anonymized scoring of participants' answers (see Appendix D). Statistical testing described in Appendix E confirm that the study design should be robust to a different selection of news stories, and a different group of participants, as long as 60 participants or more are recruited. We hope future work can use our study design as a template to measure the efficacy of reading interfaces at surfacing coverage diversity. Some adaptations will likely be necessary, such as selecting newer, more relevant news stories, or updating the selection of NLP models to the latest generation.

Measuring Reader Persuasion. We did not explicitly plan to measure user persuasion at the inception of the project and were focused on reader exposure to content diversity. However, the selection by experts of a prediction question in one of the story's comprehension questionnaires offered an opportunity for the analysis in Section 6.5.2, revealing minor evidence that participants might have been influenced in their perception of the likelihood of a future event. We encourage future work to use prediction questions in future usability studies to expand on our preliminary findings.

9 CONCLUSION

This paper introduced three news reading interfaces – the Annotated Article, the Recomposed Article, and the Question Grid – that leverage the Discord Questions automated pipeline to add context to news stories and highlight coverage diversity. We first conducted a usability study with journalism experts, gaining insights into potential use cases of the different interfaces. Experts find that discord questions help highlight coverage diversity in all Assembly interfaces, and judge that the Annotated Article is generally accessible to a wide audience, while the other two are suitable for advanced use within newsrooms. In a second study with 95 novice news readers, we assess the usability of the Annotated Article compared to existing news interfaces – a single News Article, and a Headline List. Novice readers found the Annotated Article roughly as easy to use as existing interfaces while scoring significantly higher on a story understanding questionnaire that measured user exposure to coverage diversity. The findings demonstrate that NLP technology can be integrated into news reading interfaces to assist readers in gaining diverse views on complex news stories, reducing the barrier to educating informed citizens.

ACKNOWLEDGMENTS

We thank Jesse Vig, Marti Hearst, Alex Fabbri, and the CHI reviewers for their helpful feedback on the manuscript.

REFERENCES

- [1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.
- [2] AllSides. 2021. How AllSides rates media bias: Out methods. (2021).
- [3] Mahmoudreza Babaei, Jui Kulshrestha, Abhijnan Chakraborty, Fabricio Benvenuto, Krishna P Gummadi, and Adrian Weller. 2018. Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 10–16.
- [4] Joshua Bambrick, Minjie Xu, Andy Almonte, Igor Malioutov, Guim Perarnau, Vittorio Selo, and Iat Chong Chan. 2020. NSTM: Real-Time Query-Driven News Overview Composition at Bloomberg. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 350–361.
- [5] Michael Barthel, Amy Mitchell, Dorene Asare-Marfo, Courtney Kennedy, and Kirsten Worden. 2020. Measuring News Consumption in a Digital Era. *Pew Research* (8 December 2020).
- [6] Adam J. Berinsky, G. Huber, and G. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk. *Political Analysis* 20 (2012), 351–368.
- [7] Dan Bernhardt, Stefan Krasa, and Mattias Polborn. 2008. Political polarization and the electoral effects of media bias. *Journal of Public Economics* 92, 5-6 (2008), 1092–1104.
- [8] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- [9] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A Dataset for Training and Evaluating Generative Reading Comprehension Metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6521–6532.
- [10] Barry Cooper. 1994. *Sins of omission: Shaping the news at CBC TV*. University of Toronto Press Toronto.
- [11] Google Developers. 2021. Understanding the sources behind Google News. *Google Search Central Blog* (1 June 2021).
- [12] Bradley Efron. 1982. The Jackknife, the Bootstrap and other resampling plans. In *CBMS-NSF Regional Conference Series in Applied Mathematics*.
- [13] Jonas Ehrhardt, Timo Spinde, Ali Vardasbi, and Felix Hamborg. 2021. Omission of information: Identifying political slant via an analysis of co-occurring entities. (2021).
- [14] Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1074–1084.
- [15] Felix Filloux. 2013. Google News: the secret sauce. *The Guardian* (Feb 2013).
- [16] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th international conference on World Wide Web*. 482–490.
- [17] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 708–719.
- [18] Ilya Gusev and Alexey Tikhonov. 2021. HeadlineCause: A Dataset of News Headlines for Detecting Causalities. *arXiv preprint arXiv:2108.12626* (2021).
- [19] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2017. Matrix-Based News Aggregation: Exploring Different News Perspectives. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2017), 1–10.
- [20] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2018. Bias-aware news analysis using matrix-based news aggregation. *International Journal on Digital Libraries* 21 (2018), 129–147.
- [21] Felix Hamborg, Anastasia Zhukova, Karsten Donnay, and Bela Gipp. 2020. Newsalyze: enabling news consumers to understand media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 455–456.
- [22] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism* (2022), 1–21.
- [23] Hendrik Heuer and Elena Leah Glassman. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *CHI Conference on Human Factors in Computing Systems*. 1–21.
- [24] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*. PMLR, 4411–4421.
- [25] Francisco Iacobelli, Larry Birnbaum, and Kristian J Hammond. 2010. Tell me more, not just” more of the same”. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 81–90.
- [26] Panagiotis G. Ipeirotis. 2010. Demographics of Mechanical Turk. *Labor: Supply & Demand eJournal* (2010).
- [27] Wei-Jen Ko, Te-Yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. *arXiv preprint arXiv:2010.01657* (2020).
- [28] Thomas Koch and Thomas Zerback. 2013. Helpful or harmful? How frequent repetition affects perceived statement credibility. *Journal of Communication* 63, 6 (2013), 993–1010.
- [29] Tomáš Kočický, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics* 6 (2018), 317–328.
- [30] Jim A Kuypers. 2006. *Bush’s war: Media bias and justifications for war in a terrorist age*. Rowman & Littlefield Publishers.
- [31] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.
- [32] Philippe Laban, Lucas Bandarkar, and Marti A Hearst. 2021. News headline grouping as a challenging nlu task. *arXiv preprint arXiv:2105.05391* (2021).
- [33] Philippe Laban, John Canny, and Marti A Hearst. 2020. What’s The Latest? A Question-driven News Chatbot. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 380–387.
- [34] Philippe Laban and Marti A Hearst. 2017. newsLens: building and visualizing long-ranging news stories. In *Proceedings of the Events and Stories in the News Workshop*. 1–9.
- [35] Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs’ ka, Xiang’Anthony’ Chen, and Caiming Xiong. 2022. Discord Questions: A Computational Approach To Diversity Analysis in News Coverage. *arXiv preprint arXiv:2211.05007* (2022).
- [36] Philippe Laban, Elicia Ye, Srulay Korklunka, John F. Canny, and Marti A. Hearst. 2022. NewsPod: Automatic and Interactive News Podcasts. *27th International Conference on Intelligent User Interfaces* (2022).
- [37] Angela M Lee and Hsiang Iris Chyi. 2015. The rise of online news aggregators: Consumption and competition. *International Journal on Media Management* 17, 1 (2015), 3–24.
- [38] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Mitigating Media Bias through Neutral Article Generation. *arXiv preprint arXiv:2104.00336* (2021).
- [39] Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias. *ArXiv abs/2204.04902* (2022).
- [40] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

- [41] Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now? Mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 184–196.
- [42] 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* (2019).
- [43] Felicia Loecherbach and Damian Trilling. 2020. 3bij3—Developing a framework for researching recommender systems and their effects. *Computational Communication Research* 2, 1 (2020), 53–79.
- [44] Scott R Maier. 2020. The world view (ed) through the English-speaking media lens: foreign news coverage steadfastly narrow and uniform. *The Journal of International Communication* 26, 2 (2020), 155–170.
- [45] Brian McNair. 2006. *Cultural chaos: journalism and power in a globalised world*. Routledge.
- [46] Amy Mitchell, Galen Stocking, and Matsa Katerina Eva. 2016. Long-Form Reading Shows Signs of Life in Our Mobile News World. *Pew Research* (5 May 2016).
- [47] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of The International AAAI Conference on Web and Social Media*, Vol. 7. 419–428.
- [48] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.
- [49] Lidiya Murakhov'ska, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural Question Generation with Mixed Answer Types. *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022* (2022).
- [50] NPR and Edison Research. 2020. *The Smart Audio Report*. Technical Report. National Public Media. <https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/>
- [51] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 443–452.
- [52] Souneil Park, Minsam Ko, Jungwoo Kim, Ho-Jin Choi, and Junehwa Song. 2011. NewsCube 2.0: an exploratory design of a social news website for media bias mitigation. In *Workshop on Social Recommender Systems*.
- [53] Simon T Perrault and Weiyu Zhang. 2019. Effects of moderation and opinion heterogeneity on attitude towards the online deliberation experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
- [54] Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies* 4, 4 (2003), 501–511.
- [55] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 784–789.
- [56] Tom Rosenstiel, Jeff Sonderman, Kevin Loker, Millie Tran, Trevor Tompson, Jennifer Benz, Nicole Wilcoxon, Rebecca Reimer, Emily Alvarez, Dan Malato, and et al. 2014. How Americans get their news. *American Press Institute* (Mar 2014). <https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/>
- [57] Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 623–632.
- [58] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 2020. Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (2020).
- [59] Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with macaw. *arXiv preprint arXiv:2109.02593* (2021).
- [60] Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 News Translation Task Submission. In *Proceedings of the Sixth Conference on Machine Translation*. 205–215.
- [61] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. 191–200.
- [62] Kohei Watanabe. 2013. The western perspective in Yahoo! News and Google News. *International Communication Gazette* 75 (2013), 141–156.
- [63] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting civil discourse through search engine diversity. *Social Science Computer Review* 32, 2 (2014), 145–154.

A RECOMPOSED ARTICLE CONTENT ALGORITHM

Figure 8 introduces a pseudo-code version of the algorithm used to sort through and select the discord questions that are presented in the Recomposed Article.

B BASELINE INTERFACES

Figure 9 presents the Headline List baseline interface, which is simply a list of the source's headline. Each headline is clickable and opens the original website in a new browser tab.

C READING EXERCISE INTERFACE

Figure 10 displays the interface used during the reading exercise study, using a two-column layout for concurrent reading and question answering.

D MANUAL SCORING PROCEDURE

The scoring of answers was performed manually by the authors of the paper, and we took steps to ensure the impartiality of the scoring. All answers to a question were loaded into a Google Sheet, with an anonymized identifier that could not readily identify which user wrote each answer, or which interface was used.

In the first step, one reader read through all the answers and identified the set of all answer elements. We required that an answer element to be a plausible answer to the question be added as one possible answer element to be considered in the score. In some cases, participants provided answers that were unrelated to the question, or could not be interpreted as a direct answer, and these were discarded. We did not attempt to assess the veracity of each answer element, or whether each answer was explicitly stated in at least one of the provided articles, as some of the questions required the user's interpretation. For example, to the question "Is there going to be a recession?", some participants answered with the answer element: "we are already in one", which was factually incorrect during the study dates (according to the definition of a recession requiring two consecutive quarters of economic contraction). In order to maximize our assessment of access to coverage diversity, any plausible answer element was added to the answer element set.

In the second step, the same reader shuffled all the answer elements and tagged each answer with all the answer elements it mentioned. For all answers, we generously assigned the presence of an answer element, for example in cases when the element is partially mentioned or strongly implied.

Once the tagging was completed for all answers, answers were then processed automatically and scores were aggregated. Although we did not perform inter-annotator agreement evaluation for the manual scoring process, and it is likely that there exists some variance in the methodology we used, we believe the process did not favor any particular interface, and this variance should not negatively impact our results or their interpretation.

E STUDY REPRODUCIBILITY EXPERIMENTS

In order to determine the level of reproducibility of the reading exercise we designed, we perform two reproducibility experiments, leveraging bootstrap re-sampling [12] to simulate a change in study

```
def select_discord_qs(questions):
    # Initialize output and progress word count
    selected_qs, seen_paragraphs = [], set([])
    while True:
        # Score each discord question
        for q in questions:
            # A question's score is the number of unseen answering paragraphs
            q.score = len(q.paragraph_set - seen_paragraphs)
        # Select the highest_scoring question
        selected_q, score = questions.select_best_q()
        if score == 0:
            # Exit if no question introduces unseen paragraphs
            break
        # Add question to final list
        selected_qs.append(selected_q)
        # Mark question's paragraph as seen
        seen_paragraphs = seen_paragraphs | selected_q.paragraph_set
    return selected_qs
```

Figure 8: Python code for the Composition Algorithm used for sequence selection in the Recomposed Article interface.

results and verify whether the results remain statistically significant. We experiment with two re-sampling methods to verify whether the results remain significant with a different story selection, or a different participant population.

Varying the stories. We consider subsets of the data where only four of the five stories are kept, recompute the average reading scores under each interface and perform a paired t-test to verify whether results remain significant. In 80% of cases, the score differences are significant ($p < 0.05$), confirming that no individual story is responsible for the result, and giving evidence that the results would extend to other stories. When decreasing subset sizes to only 3 out of the five stories, only 66% of the cases are significant, showing that results are more stable when using at least 4 distinct stories, as variances amongst individual stories exist.


Varying the participants. We simulate a lower number of participants, testing each value of participants from 5 to 95 in increments of 5. For each number of participants, we sample 40 random sets of participants of that size, compute results for that subset and test whether there are statistically significant differences in answer scores. The results are summarized in Figure 11. We find that the results remain largely significant with 60 participants or more, and then the results become less significant as the number of participants decreases. This can serve as a rule-of-thumb for similar studies, encouraging a participant population of at least 60 participants to increase the likelihood of statistical significance in the results.

Inflation vs recession

Inflation Hits New 40-Year High of 8.5%: Why Prices Keep Climbing
 cnet.com


Economy: As inflation booms, does recession loom?
 theweek.com


'Peak inflation' trade is driving financial markets as investors brace for U.S. economic slowdown
 marketwatch.com

Unleaded Gasoline Futures Declined 26 Percent, Has Inflation Peaked This Economic Cycle?
 mishtalk.com

LISTEN: A recession may be the only way to stop inflation
 komonews.com

The Inflation Picture Darkens
 forbes.com

Jesse Ramos: The solution to the inflation crisis
 missoulian.com

Here's why inflation may have peaked -- but a recession could still loom
 businessinsider.com

The 2024 Recession
 247wallst.com

Editorial: Fed has no good choice between inflation and recession
 herald-dispatch.com

Hiltzik: Inflation is starting to come down
 netionaldastak.com

A Fed-induced recession is a medicine worse than the disease
 ft.com

Will hiking interest rates tackle inflation and recession? Even the Fed is confused.
 economictimes.indiatimes.com

Figure 9: Headline List Interface. We purposefully imitate the style of the “More Coverage” view in Google News, listing all the source’s headlines within a news story.

Story 1: US Inflation / Recession

Inflation vs recession

Inflation Hits New 40-Year High of 8.5%: Why Prices Keep Climbing

Published by [cnet.com](#)

From your purchases at the grocery store to the dollar tag of rent across the US, consumer prices are soaring across the country. Your dollar doesn't have the purchasing power it used to, and that's a problem. Inflation climbed by 8.5% through March, according to the Consumer Price Index. The CPI is a key indicator of inflation, defined as the sustained rise in the cost of living. The increase marks inflation's fastest annual rise since December 1981. Core inflation, which measures all items minus food and gas, rose 6.5%. This kind of sustained and increased inflation may point to something more enduring.

What does this mean for you? Here are some key things you need to know about inflation, how it can impact **your budget** and how it impacts your spending power.

MORE CONTEXT

▼ Who does inflation affect?

American people - [missoulain.com](#)
Americans - [missoulain.com](#)

both consumers and businesses - [azbigmedia.com](#)

Questions

*Use the interface on the left to answer questions as thoroughly as possible. Use **bullet-point** format. If you don't find any answer, write down: "No answer".*

What are the reasons for the inflation?

- Answer 1

- Answer 2

...

What products have been affected by inflation?

- Answer 1

- Answer 2

...

Is there going to be a recession?

- Answer 1

- Answer 2

...

What does the government do to reduce inflation?

- Answer 1

- Answer 2

...

Figure 10: Two-column interface used during reading exercise. Participants completed the comprehension questions in the right column while using the reading interface in the left column. In this case, the exercise was assigned with the Annotated Article reading interface.

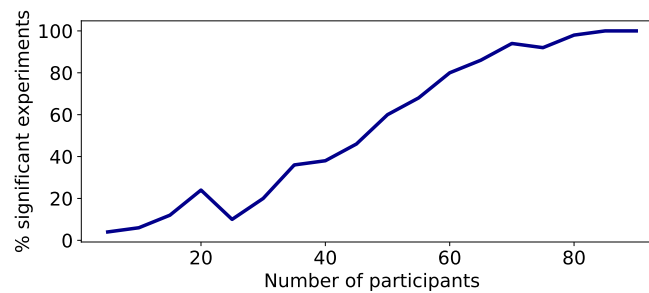


Figure 11: Reproducibility experiments for the subset of participants. As the number of participants decreases, the results are less likely to be significant, showing the benefit of increasing the population size when running the study.