INFO-NAVIGATE for Critical Thinking in Complex Information Landscapes: A Design Rationale and Pilot Evaluation

AJANIE KARUNANAYAKE[*], TD School, University of Technology Sydney, Australia

ANTONETTE SHIBANI, TD School, Centre for Research on Education in a Digital Society, University of Technology Sydney, Australia

SIMON KNIGHT, TD School, Centre for Research on Education in a Digital Society, University of Technology Sydney, Australia

In contemporary society, people rely on online information for their daily needs, making life-impacting decisions based on it. However, identifying high-quality information on the internet has become challenging due to the abundance of information of both varied quality and perspectives in this era of misinformation and generative AI. Beyond the simple categorization of false or true, the diverse spectrum of conflicting data complicates the search for reliable information. Therefore, supporting people in critically navigating through conflicting information is crucial for addressing their information needs and decision-making processes. Drawing from theory and prior work, this paper presents the design and design rationale of a prototypical technological support tool 'INFO-NAVIGATE' for identifying high-quality information, while enhancing cognitive ability and critical thinking as a part of a socio-technical Information-Problem-Solving (IPS) intervention. We performed system evaluation and preliminary user evaluation of the prototype and gathered insights into user engagement with different tool features. Results indicate that participants utilized the main features of the tool to perform critical tasks for processing conflicting information.

CCS Concepts: • **Human-centered computing → Visualization**; **Interaction design**; • **Applied computing → Education**; • **Information systems → Information retrieval**.

Additional Key Words and Phrases: Artificial Intelligence, GenAI, LLM, Information Problem Solving, Information Seeking, Critical Thinking, Conflicting Information

ACM Reference Format:

Ajanie Karunanayake, Antonette Shibani, and Simon Knight. 2024. INFO-NAVIGATE for Critical Thinking in Complex Information Landscapes: A Design Rationale and Pilot Evaluation. 1, 1 (November 2024), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Across professional, educational, and daily life contexts, people rely on internet resources despite the varying quality of sources [11, 18]. Sourced information affects the choices and behaviors of individuals, and ultimately, the wider society. [11]. Hence, supporting and educating people to identify high-quality online information is important.

The abundance of accurate and inaccurate information can be characterized by the proliferation of conflicting information [16]. Such information may involve differences in source credibility alongside cases where there is a perception that experts disagree and there is no single consensus, even in relatively high-quality sources [10]. For

[*] Corresponding author

Authors' addresses: Ajanie Karunanayake, TD School, University of Technology Sydney, Sydney, Australia, ajanie.m.karunanayake@student.uts.edu.au; Antonette Shibani, TD School, Centre for Research on Education in a Digital Society, University of Technology Sydney, Sydney, Australia, antonette.shibani@uts.edu.au; Simon Knight, TD School, Centre for Research on Education in a Digital Society, University of Technology Sydney, Sydney, Australia, simon.knight@uts.edu.au.

2024. Manuscript submitted to ACM

example, in early 2020, different health organizations globally provided varied advice regarding the use of face masks to mitigate COVID-19 risks. The exposure to this conflict may confuse people and left unexplained, can lead to a general distrust in experts or/and be mobilized into misinformation campaigns [25]. Due to cognitive biases and the lack of domain knowledge, lay people struggle to identify the genuineness of the claims made and find valid information when different sources provide conflicting information [11], specially when conflicts in sources involve a blend of accurate and inaccurate information [19].

Furthermore, Generative AI (GenAI) advances have amplified the easy generation of information appearing to be from human sources, including misinformation [12]. Simultaneously, it provides a new information search aid, even for knowledge work and complex tasks [7, 35]. GenAI tools like ChatGPT excel in generating responses to straightforward questions and offering general solutions. However, in complex tasks, prompt quality impacts the response quality [37]. Therefore, in the context of conflicting information, GenAI responses may be limited to only one perspective or may present potentially false equivalence between sides of an argument. This limited perspective display may lead to misleading credibility when people are exposed to different perspectives [28].

This paper presents the design and evaluation of a prototype tool for information seeking, 'INFO-NAVIGATE', designed to address the context of conflicting information and the impacts of GenAI. The aim of this paper is to introduce its design rationale[22], explaining feature selection and design decisions, with preliminary evaluation addressed through the research question, How do INFO-NAVIGATE users engage with the tool's functionality to process conflicting information?

2 BACKGROUND LITERATURE

A key approach in filtering information and designing user-facing interfaces has been analysis of information quality, which has evolved from checklist approaches to evaluate sources on a set of criteria [24] to more advanced rule-based algorithms and automated machine learning approaches to classifying information sources. These focus on a set of features related to a single topic and provide an individual indication of high or low quality [18]. However, dealing with conflicting information is more complex than simple classifications of fake/ true, or high quality/ low quality. People need to critically compare and contrast multiple documents with diverse perspectives to identify the most plausible information tackling biases [3, 15, 34]. Bing Multi-Perspective Answer[1], Multi-Perspective Search Engine Tool [9][2] and NewsCube [30] are some of the technical solutions introduced to expose information seekers into diverse perspectives built around a phenomenon and mitigate biases. Even though these tools provide perspective comparison support, they are limited in supporting what information to believe.

Information seekers need to use their cognitive abilities, including critical thinking, alongside supporting tools to tackle the context of conflicting information. The Information Problem Solving using the Internet (IPS-I) model presents the five main skills needed to solve information problems successfully using online resources: defining the problem, searching for information, scanning information, processing information, and presenting information [5]. These skills are derived from the main complex cognitive skill called information problem solving (IPS) [4], which is well-researched in the educational domain.

IPS informed pedagogical interventions broadly fall into two types: 1. Instructional IPS interventions and 2. Tool-based IPS interventions. Instructional interventions mainly use guided instructions [36], while tool-based interventions use software programs to support and educate people during the IPS process. Experimental results from IPS *instructional*

[1] https://blogs.bing.com/search-quality-insights/february-2018/toward-a-more-intelligent-search-bing-multi-perspective-answers
[2] https://www.youtube.com/watch?v=VD1QubNiRR4

interventions reveal that they can improve IPS skills to some extent. However, findings also reveal limitations in scaffolding content evaluation [15], and found *tool-based* scaffolding more beneficial in enhancing metacognitive awareness during information selection and analysis [31]. We posit that a combination of technological and instructional approaches provides a desirable socio-technical solution.

Socio-technical solutions stress the reciprocal interrelationship between humans and machines to find effective solutions to complex problems that require understanding the technical aspects and the social, cultural, and organizational contexts in which technology is implemented [14]. IPS tool-based interventions offer such support using technical and pedagogical instructions. Many existing IPS tool-based interventions use standalone or web-based software with user-friendly interfaces, focusing more on technology and less on instructional support [6, 17, 34]. IPS interventions that scaffold lay people in information problem-solving in the context of conflicting information where there is no one consensus are limited. Amongst existing intervention tools, met.a.ware is the only closely relevant tool to our problem space that supports lay people in solving information problems with conflicting information [34]. The tool support, however, focuses mostly on source quality evaluation, including author attributes and IPS process regulation activities introduced in the IPS model, which is insufficient to identify high-quality information *content* in the context of conflicting information. When lay people deal with conflicting information, since they may lack prior knowledge to identify the most plausible information, they need to evaluate both *source* and *content* thoroughly and critically corroborate the information by navigating through multiple sources [15, 34]. While existing IPS interventions focus on source and content quality evaluation to an extent, their information corroboration support is limited. In addition, none of the interventions provide inter-textual conflict resolution support, which is crucial when dealing with multiple conflicting pieces of information.

In INFO-NAVIGATE design, we address these limitations. The tool acts as a pedagogical intervention that scaffolds people to successfully navigate conflicting information and identify high-quality information by activating and enhancing cognitive skills and critical thinking ability. These skills are necessary for responsible information seekers in the age of misinformation and generative AI. Also, these are identified as part of a repertoire of competencies learners should develop when engaging with AI content meaningfully by past work [33].

3 TOOL DESIGN AND IMPLEMENTATION

3.1 Overview

Information seeking is not merely finding information. Eventually, the identified information is used for sense making and solving an information problem [23, 34]. For example, if a person wants to know whether red wine has health benefits, they may use the search query "Does red wine have health benefits?", which might result in obtaining conflicting information and evidence from different sources. Once the information seeker receives a collection of such information pieces, they engage in processes leading to a deep understanding of the information, such as critical source and information quality evaluation, information analysis, selection, and integration. Based on the skills and activities outlined in the IPS-I model, these actions occur during the information processing stage [5]. INFO-NAVIGATE is specifically designed to support this information processing stage in the IPS process to help users deal with conflicting information.

Feature selection for the tool was grounded in the review of existing IPS interventions, people's engagement with GenAI tools during information seeking and conflicting information visual representation guidelines[20]. In addition to that, we considered features of modern information-seeking support tools and users' preferences for those features.

These tools include, Bing Multi-Perspective Answer, Multi-Perspective Search Engine Tool [9] and Consensus[3]. We address the limitations of Bing Multi-Perspective Answer and Multi-Perspective Search Engine tools such as manual or black-box algorithmic filtration of sources, by offering explainability to users. In addition to displaying multiple perspectives side by side, we provide support for users to corroborate the information behind each perspective for navigating conflicting information. The tool intervenes in the practical scenario of information seeking, where people are exposed to diverse types of online sources without limiting them to curated sources (academic articles), like Consensus does. In contrast to existing tools, INFO-NAVIGATE is also developed as part of a pedagogical intervention to educate users in navigating and processing conflicting information critically, including source and content quality evaluation support and comparing and contrasting support for multiple perspectives. Features were selected to provide these supports (detailed in the following section) while addressing the limitations of prior IPS interventions and tools.

Figure 1 depicts the user interface of the web-based AI-powered IPS support tool INFO-NAVIGATE. The tool accepts search queries as user input and dynamically generates relevant content for the entered search query (the labeled parts are explained in detail in the next section).

Figure 2 depicts its content generation pipeline using Natural Language Processing (NLP) sub-tasks. Python libraries and OpenAI GPT 4o-mini model (LLM API) were used to extract information and generate content by leveraging the capabilities of Large Language Models (LLMs) in performing these tasks [12, 13, 38]. We followed OpenAI prompt engineering guidelines [1] to define prompts for the LLM API. A low temperature (temperature = 0) and a high top-p (top_p = 1) values were set for the model to maintain the generated contents' consistency and logic (All the LLM prompts used are provided in Appendix A). We sought to evaluate outputs using system-level metrics and pilot user testing, explained in later sections.

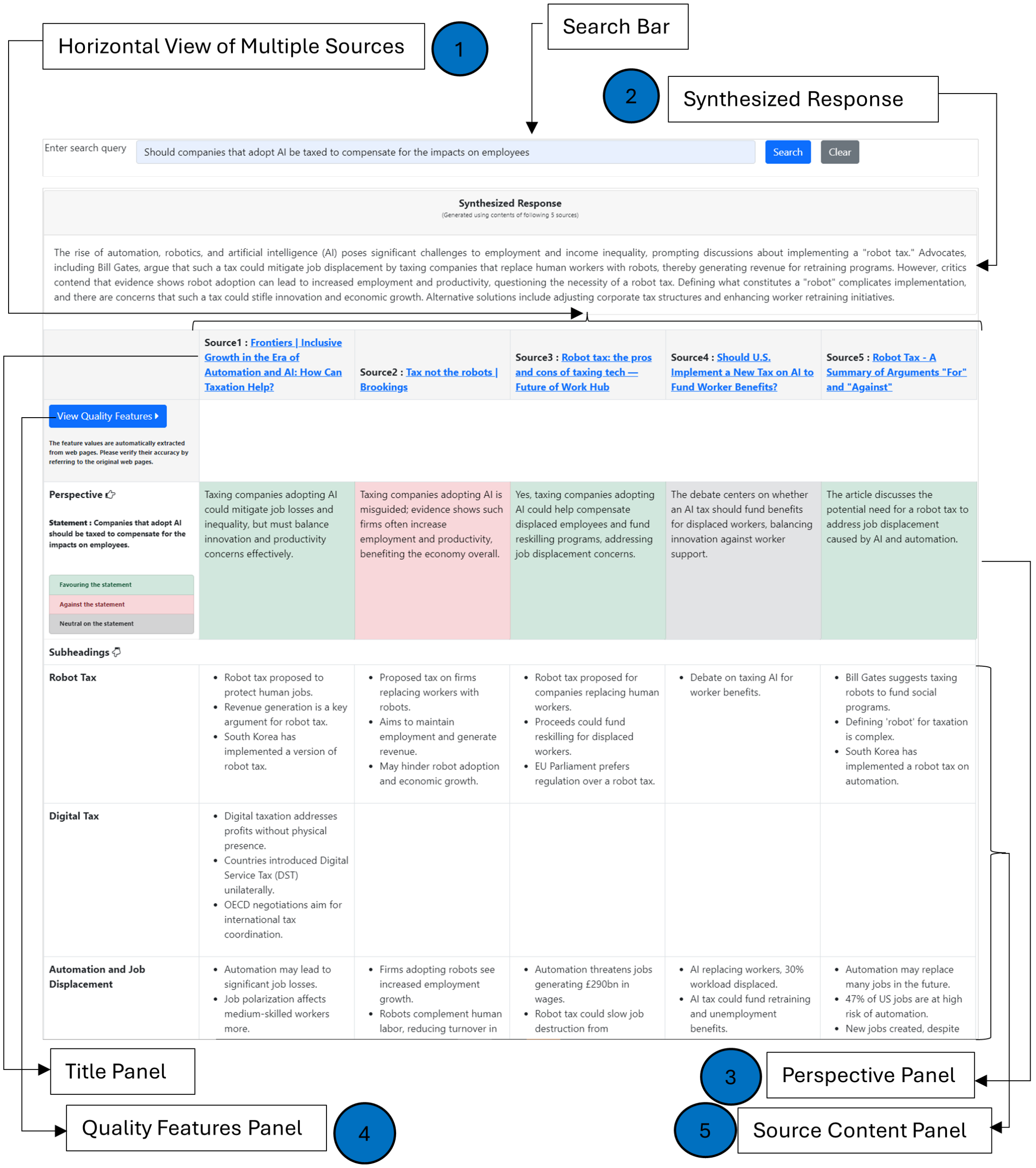
The following sections describe the rationale for feature selection and detail how the features were implemented in the tool.

3.2 Main Features

3.2.1 Horizontal View of Multiple Sources. Once the user enters a search query, the tool extracts the top five results from the commonly used Google search engine using the Python "googlesearch" library and displays them horizontally, as visualized in figure 1, rather than in a list view, which is the orientation used by traditional search engines. Multiple reasons influenced the selection of this horizontal view. Prior research indicates that in a list view, the documents' order can influence the user's understanding of the controversy when conflicting information exists [27]. Several cognitive biases that occur due to the traditional list view were identified, particularly when searching for information in contexts where multiple perspectives exist. This includes users' tendency to stick to the first few (1-3) search results of the Search Engine Results Page (SERP) and build trust in those results, neglecting or avoiding other or conflicting perspectives that might have similar value to consider before deciding, narrowing down the perspective spectrum [28]. The horizontal view of sources with multiple perspectives is identified as a better design principle to represent conflicts[20]. Thus, a horizontal view was selected to eliminate cognitive biases, present a better view of multiple perspectives and support easy information corroboration across multiple sources.

3.2.2 Synthesized Response. A multi-perspective search engine experiment identified that despite the controversies in a topic, people prefer to get a direct answer to their queries [9]. A recent study that examined ChatGPT's impact on information-seeking behavior also identified that people use ChatGPT response as a starting point to plan their search

[3] https://consensus.app/



Enter search query: Should companies that adopt AI be taxed to compensate for the impacts on employees

Synthesized Response
(Generated using contents of following 5 sources)

The rise of automation, robotics, and artificial intelligence (AI) poses significant challenges to employment and income inequality, prompting discussions about implementing a "robot tax." Advocates, including Bill Gates, argue that such a tax could mitigate job displacement by taxing companies that replace human workers with robots, thereby generating revenue for retraining programs. However, critics contend that evidence shows robot adoption can lead to increased employment and productivity, questioning the necessity of a robot tax. Defining what constitutes a "robot" complicates implementation, and there are concerns that such a tax could stifle innovation and economic growth. Alternative solutions include adjusting corporate tax structures and enhancing worker retraining initiatives.

	Source1 : Frontiers | Inclusive Growth in the Era of Automation and AI: How Can Taxation Help?	**Source2 :** Tax not the robots | Brookings	**Source3 :** Robot tax: the pros and cons of taxing tech — Future of Work Hub	**Source4 :** Should U.S. Implement a New Tax on AI to Fund Worker Benefits?	**Source5 :** Robot Tax - A Summary of Arguments "For" and "Against"
View Quality Features ▸ The feature values are automatically extracted from web pages. Please verify their accuracy by referring to the original web pages.					
Perspective **Statement : Companies that adopt AI should be taxed to compensate for the impacts on employees.** 🟩 Favouring the statement 🟥 Against the statement ▨ Neutral on the statement	Taxing companies adopting AI could mitigate job losses and inequality, but must balance innovation and productivity concerns effectively.	Taxing companies adopting AI is misguided; evidence shows such firms often increase employment and productivity, benefiting the economy overall.	Yes, taxing companies adopting AI could help compensate displaced employees and fund reskilling programs, addressing job displacement concerns.	The debate centers on whether an AI tax should fund benefits for displaced workers, balancing innovation against worker support.	The article discusses the potential need for a robot tax to address job displacement caused by AI and automation.
Subheadings					
Robot Tax	• Robot tax proposed to protect human jobs. • Revenue generation is a key argument for robot tax. • South Korea has implemented a version of robot tax.	• Proposed tax on firms replacing workers with robots. • Aims to maintain employment and generate revenue. • May hinder robot adoption and economic growth.	• Robot tax proposed for companies replacing human workers. • Proceeds could fund reskilling for displaced workers. • EU Parliament prefers regulation over a robot tax.	• Debate on taxing AI for worker benefits.	• Bill Gates suggests taxing robots to fund social programs. • Defining 'robot' for taxation is complex. • South Korea has implemented a robot tax on automation.
Digital Tax	• Digital taxation addresses profits without physical presence. • Countries introduced Digital Service Tax (DST) unilaterally. • OECD negotiations aim for international tax coordination.				
Automation and Job Displacement	• Automation may lead to significant job losses. • Job polarization affects medium-skilled workers more.	• Firms adopting robots see increased employment growth. • Robots complement human labor, reducing turnover in	• Automation threatens jobs generating £290bn in wages. • Robot tax could slow job destruction from	• AI replacing workers, 30% workload displaced. • AI tax could fund retraining and unemployment benefits.	• Automation may replace many jobs in the future. • 47% of US jobs are at high risk of automation. • New jobs created, despite

Fig. 1. User interface of INFO-NAVIGATE to support critical thinking and processing of conflicting information

process and as an overview for unknown topics [7]. Because of its relevance to information processing in the age of AI, we embedded the synthesized response feature in our tool as a core component. This synthesized, straightforward

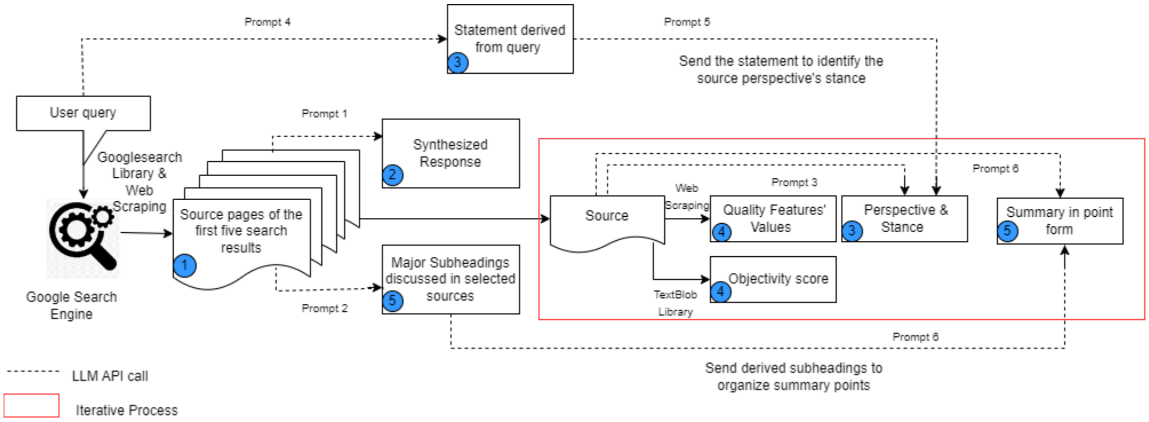

Prompt 4
Statement derived from query
3
Prompt 5
Send the statement to identify the source perspective's stance
User query
Prompt 1
Synthesized Response
2
Googlesearch Library & Web Scraping
Source pages of the first five search results
1
Google Search Engine
Prompt 2
Major Subheadings discussed in selected sources
5
Prompt 6
Source
Web Scraping
Prompt 3
Quality Features' Values
4
Perspective & Stance
3
Summary in point form
5
Objectivity score
4
TextBlob Library
Prompt 6
Send derived subheadings to organize summary points
LLM API call
Iterative Process

Fig. 2. Content generation pipeline

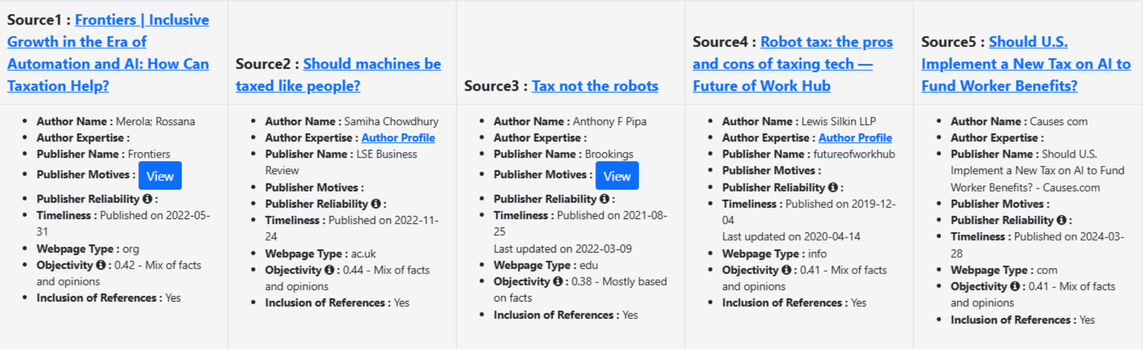
response provides users with an early orientation to different perspectives to help navigate complex and unfamiliar topics. We used an appropriate prompt (Prompt 1) for the LLM to implement this summary feature, using the body contents of the five selected sources to represent the diversity of perspectives and eliminate the perspective masking effect of tools like ChatGPT that provide a single answer, often based on a predominant source.

3.2.3 Perspective Panel. As shown in figure 1, this horizontal panel presents the perspective of each article toward the user query and its stance clearly indicated using a colour code. This presentation aligns with the conflict representation design principles identified in prior work [20]. The green colour indicates the favoured 'for' perspectives toward the topic, the red indicates the 'against' perspectives toward the topic, and grey indicates the 'neutral' perspectives. This functionality allows users to identify where an agreement or consensus exists regarding a particular topic, and was inspired by the consensus meter of the AI-powered Consensus tool. While the tool highlights the general consensus (or lack thereof), users still need to analyze multiple perspectives and resolve conflicts critically before reaching a decision.

We leveraged the LLM information extraction capability to derive perspectives (Prompt 3). LLMs also exhibit emergent capabilities in stance detection tasks [38]. We adapted their prompt for stance detection using a few-shot prompting strategy to obtain the output in a desired format for our tool (Prompt 5).

3.2.4 Quality Features Panel. When a user clicks on the “View Quality Features” button highlighted in figure 1, they see an expanded view of the quality feature panel displayed in figure 3.

This view contains a list of key source and content quality features along with the values corresponding to each source. Based on the literature review on website, web page, and information quality, we have identified a set of key source-based and content-based web content quality criteria that focus on content more than formatting when determining the quality features. This includes author details, publisher details, timeliness of the content, accuracy of the content (inclusion of references), and objectivity of the content [18, 24, 29]. This feature serves as a scaffold for users to identify high-quality information by increasing their awareness of quality features and the values they should look for in the *informed end-user method.* This method increases the transparency and explainability of the tool and reduces the risk of misclassification that might occur in quality prediction tools using black-box algorithms [32]. As shown in the pipeline in Figure 2, once the five sources are retrieved, the values for the features are extracted

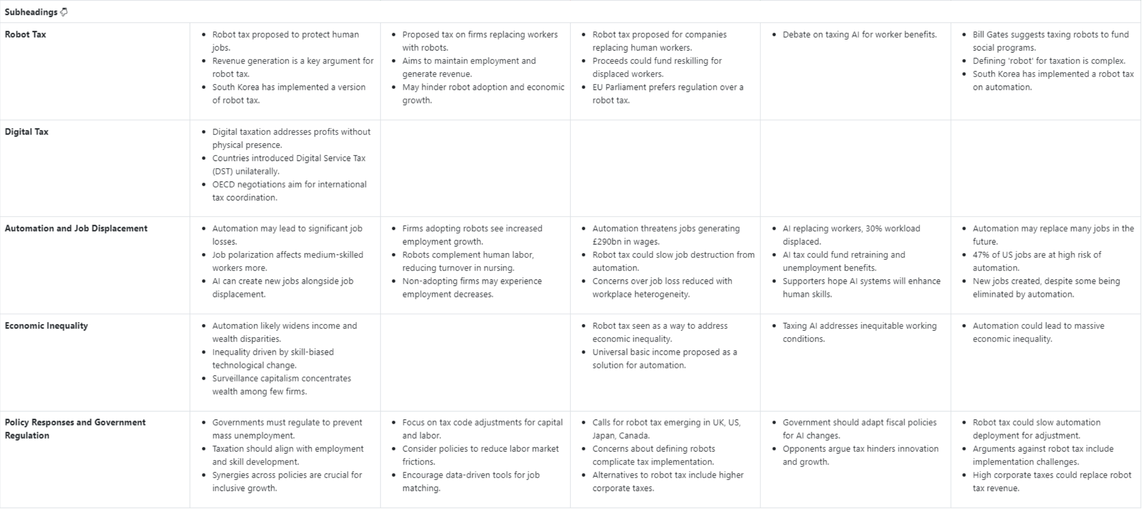


Source1 : Frontiers | Inclusive Growth in the Era of Automation and AI: How Can Taxation Help?	Source2 : Should machines be taxed like people?	Source3 : Tax not the robots	Source4 : Robot tax: the pros and cons of taxing tech — Future of Work Hub	Source5 : Should U.S. Implement a New Tax on AI to Fund Worker Benefits?
• **Author Name :** Merola; Rossana • **Author Expertise :** • **Publisher Name :** Frontiers • **Publisher Motives :** View • **Publisher Reliability** ℹ **:** • **Timeliness :** Published on 2022-05-31 • **Webpage Type :** org • **Objectivity** ℹ **:** 0.42 - Mix of facts and opinions • **Inclusion of References :** Yes	• **Author Name :** Samiha Chowdhury • **Author Expertise :** Author Profile • **Publisher Name :** LSE Business Review • **Publisher Motives :** • **Publisher Reliability** ℹ **:** • **Timeliness :** Published on 2022-11-24 • **Webpage Type :** ac.uk • **Objectivity** ℹ **:** 0.44 - Mix of facts and opinions • **Inclusion of References :** Yes	• **Author Name :** Anthony F Pipa • **Author Expertise :** • **Publisher Name :** Brookings • **Publisher Motives :** View • **Publisher Reliability** ℹ **:** • **Timeliness :** Published on 2021-08-25 Last updated on 2022-03-09 • **Webpage Type :** edu • **Objectivity** ℹ **:** 0.38 - Mostly based on facts • **Inclusion of References :** Yes	• **Author Name :** Lewis Silkin LLP • **Author Expertise :** Author Profile • **Publisher Name :** futureofworkhub • **Publisher Motives :** • **Publisher Reliability** ℹ **:** • **Timeliness :** Published on 2019-12-04 Last updated on 2020-04-14 • **Webpage Type :** info • **Objectivity** ℹ **:** 0.41 - Mix of facts and opinions • **Inclusion of References :** Yes	• **Author Name :** Causes com • **Author Expertise :** • **Publisher Name :** Should U.S. Implement a New Tax on AI to Fund Worker Benefits? - Causes.com • **Publisher Motives :** • **Publisher Reliability** ℹ **:** • **Timeliness :** Published on 2024-03-28 • **Webpage Type :** com • **Objectivity** ℹ **:** 0.41 - Mix of facts and opinions • **Inclusion of References :** Yes

Fig. 3. Quality Features Panel

by scraping each source separately. In addition to web scraping techniques, the "TextBlob" Python library is used to identify the objectivity of the source content.

3.2.5 Source Content Panel. This panel includes two features: a set of subheadings (sub-topics) mainly discussed in the selected sources and a summarized bullet-point view of the source contents under each subtopic as indicated in figure 4.



Subheadings ⟳					
Robot Tax	• Robot tax proposed to protect human jobs. • Revenue generation is a key argument for robot tax. • South Korea has implemented a version of robot tax.	• Proposed tax on firms replacing workers with robots. • Aims to maintain employment and generate revenue. • May hinder robot adoption and economic growth.	• Robot tax proposed for companies replacing human workers. • Proceeds could fund reskilling for displaced workers. • EU Parliament prefers regulation over a robot tax.	• Debate on taxing AI for worker benefits.	• Bill Gates suggests taxing robots to fund social programs. • Defining 'robot' for taxation is complex. • South Korea has implemented a robot tax on automation.
Digital Tax	• Digital taxation addresses profits without physical presence. • Countries introduced Digital Service Tax (DST) unilaterally. • OECD negotiations aim for international tax coordination.				
Automation and Job Displacement	• Automation may lead to significant job losses. • Job polarization affects medium-skilled workers more. • AI can create new jobs alongside job displacement.	• Firms adopting robots see increased employment growth. • Robots complement human labor, reducing turnover in nursing. • Non-adopting firms may experience employment decreases.	• Automation threatens jobs generating £290bn in wages. • Robot tax could slow job destruction from automation. • Concerns over job loss reduced with workplace heterogeneity.	• AI replacing workers, 30% workload displaced. • AI tax could fund retraining and unemployment benefits. • Supporters hope AI systems will enhance human skills.	• Automation may replace many jobs in the future. • 47% of US jobs are at high risk of automation. • New jobs created, despite some being eliminated by automation.
Economic Inequality	• Automation likely widens income and wealth disparities. • Inequality driven by skill-biased technological change. • Surveillance capitalism concentrates wealth among few firms.		• Robot tax seen as a way to address economic inequality. • Universal basic income proposed as a solution for automation.	• Taxing AI addresses inequitable working conditions.	• Automation could lead to massive economic inequality.
Policy Responses and Government Regulation	• Governments must regulate to prevent mass unemployment. • Taxation should align with employment and skill development. • Synergies across policies are crucial for inclusive growth.	• Focus on tax code adjustments for capital and labor. • Consider policies to reduce labor market frictions. • Encourage data-driven tools for job matching.	• Calls for robot tax emerging in UK, US, Japan, Canada. • Concerns about defining robots complicate tax implementation. • Alternatives to robot tax include higher corporate taxes.	• Government should adapt fiscal policies for AI changes. • Opponents argue tax hinders innovation and growth.	• Robot tax could slow automation deployment for adjustment. • Arguments against robot tax include implementation challenges. • High corporate taxes could replace robot tax revenue.

Fig. 4. Source Content Panel

Since the content is vertically presented underneath its perspective, it supports users to navigate conflicts [20]. Source summaries are used to help information seekers recognize the usefulness and relevancy of the content [9]. This facilitates understanding inter-textual relationships when resolving discrepancies between multiple conflicting documents [21]. When visualizing the summary bullet points under subtopics, we also highlight the coverage and informativeness of the content in response to the topic (breadth and depth). If an article summary covers more subtopics, the article has good coverage (high breadth of content). If an article summary has more information points under a subtopic, then that article has a higher degree of informativeness on that sub-topic (high depth of content). Information coverage and informativeness are two important features in information quality evaluation that have been added to our tool, which are not utilized in any of the existing tool-based IPS interventions as scaffolds. In light of these, each source content's summarized version is displayed as a bullet point summary categorized into subtopics mainly discussed in the selected sources to scaffold in-depth reading and understanding, judging relevancy, usefulness, coverage, and informativeness of the content, and resolving inter-textual conflicts.

We used the LLM to derive subtopics (prompt 2) and generate summaries (prompt 6) leveraging its capabilities in information extraction and summarization. As shown in the content generation pipeline in figure 2, once the sources are selected, the contents of all the sources are used as input for the subheadings generation prompt, and the generated subheadings and content of each source are used as inputs to the summary generation prompt.

4 EVALUATION OF THE TOOL'S CONTENT GENERATION

First, we conducted a system evaluation to assess the tool's content generation quality (including the evaluation of LLM outputs).

4.1 Method

We evaluated content generation quality of the LLM using both traditional machine learning metrics and AI-assisted metrics. We employed BERTScore as the traditional machine learning metric to calculate semantic similarity between generated texts and the relevant ground truths (manually created by humans). BERTScore calculates token similarity using context embedding, which enables the capture of both semantic and contextual similarity instead of calculating the exact match of tokens [39]. The first author created ground truth summaries for the set of retrieved sources, these were checked for face validity by the second author.

We also employed five AI-assisted metrics for evaluating LLM outputs (see Table 1). Four were selected from the Azure AI Studio Guide [2], alongside a custom metric named 'Coverage' to measure the comprehensiveness of the generated text in relation to the given context and prompt. (All the AI-assisted prompts are provided in Appendix C)

For this evaluation, we considered tool-generated texts from three search results (listed in Appendix B) for rigour. Table 1 presents evaluation metrics with their descriptions.

4.2 Results

Table 1 presents the mean and standard deviation of the metrics for each text generation feature of the tool, based on the three searches conducted [4].

The evaluation results (Table 1) indicate that content generation quality is generally high for all features. This indicates that the tool's generated contents are closely aligned with the human-generated ground truths and have factual and linguistic accuracy, comprehensiveness and user-friendliness. In addition, the LLM correctly predicted the stance 93.33% of the time (14 out of 15 stances matched the stance identified by human), indicating generally high performance for the stance detection task as well.

[4] The BERTScores were calculated using "bert-score" Python library. The ratings for the AI-assisted metrics were derived using OpenAI GPT-4o-mini API.

Table 1. LLM Generation Quality Evaluation Metrics for INFO NAVIGATE features

Metrics	Description	Synthesized Response	Perspectives	Subheadings	Source Content Summaries
BERTScore	Measures semantic similarity between the generated text and the ground truth. (value 0-1)	M=0.92 SD=0.01	M=0.95 SD=0.02	M=0.92 SD=0.03	M=0.92 SD=0.02
Groundedness	Measures how well the model-generated text aligns with information in the input source. Value is a reflection of the accuracy of the generated text. (value 1-5)	M=5 SD=0	M=4.6 SD=0.5	M=5 SD=0	M=5 SD=0
Relevance	Measures how well the model-generated response pertains and is directly related to the given prompt/ question. Value is a reflection of the appropriateness of the generated text. (value 1-5)	M=5 SD=0	M=4.93 SD=0.26	M=5 SD=0	M=5 SD=0
Coherence	Measures how well the model-generated text flows smoothly, naturally and written in human-like language. Value is a reflection of the readability and user-friendliness of the generated text. (value 1-5)	M=4.67 SD=0.58	M=4.53 SD=0.52	M=4.67 SD=0.58	M=4.87 SD=0.35
Fluency	Measures the grammatical proficiency of the model-generated texts. Value is a reflection of the linguistic correctness of the generated text. (value 1-5)	M=5 SD=0	M=4.93 SD=0.26	M=5 SD=0	M=5 SD=0
Coverage	Measures how well the model-generated text contains the main points included in the given context. Value is a reflection of the comprehensiveness of the generated text. (value 1-5)	M=5 SD=0	M=4.8 SD=0.41	M=5 SD=0	M=5 SD=0

5 USER EVALUATION

Secondly, we conducted a user evaluation to understand users' engagement with the tool features.

5.1 Method

Pilot user study of the INFO-NAVIGATE prototype was run with six PhD students from an Australian university who volunteered to participate in the research after providing informed consent (HREC ethics approval ETH24-9432). In the 1 hour pilot, we asked participants to use the INFO-NAVIGATE tool to identify arguments for and against a given controversial issue, along with supporting evidence, to derive a conclusion. The issue presented was: 'Should companies that adopt AI be taxed to compensate for the impacts on employees?'. This is a current controversial issue in which different parties hold different evidence-informed perspectives, with no single consensus. At the beginning of the study, the users' expertise was recorded using a 5-point prior knowledge rating question (M = 1.67, SD = 1.03), confirming their non-expertise on the topic. Participants were given a web-based data collection portal as part of the INFO-NAVIGATE tool containing text boxes to input their identified 'for' and 'against' arguments and the conclusion derived. We also asked participants to record their browser screens during the pilot test using Zoom[5] video recording. At the end of the test, we collected their task deliverables and screen recordings for analysis.

[5] www.zoom.com/

Following the pilot test, within the same week, we conducted video Stimulated Recall Interviews (SRI)[26] with each participant individually. The video SRI is a research technique that uses video recordings of the interviewees' behaviour during an event to help them recall their behaviour and reflect on their decision making process during the videoed event [26]. We used this technique to understand participants' usage of different features of the tool and their perception of the features. We did so by asking them to reflect on their engagement with different features while making decisions on processing conflicting information on the given controversial issue. Each interview ran for about 30 minutes. Key moments from the screen recordings were played during the interviews to recall the participants' behaviour.

5.2 Results

The interview responses were transcribed and qualitatively analyzed to study how pilot participants engaged with the different tool features to process conflicting information. We highlight key findings with respect to each feature below.

5.2.1 Perspective Panel. Multiple participants explained they utilized the horizontal view of multiple perspectives with the colour-coded categorization to read different perspectives across and compare each perspective with others. This visualization enabled them to think critically about the reasoning behind the differences and similarities of different perspectives. "This was really nice for me because every company or person will have certain policies by which they abide, and it was interesting to me to see that this company has this specific perspective; what is the context in which they opted to have this perspective? So that would give me a sense of being able to classify people's or companies' behavior to see how they arrive at it." [P4] "It was a nice feature with categorizing different groups, so it's easy to skim through perspectives across by comparing and interrogating: why have they written certain things in a certain way? some things are similar but written differently." [P5]

Participants also mentioned that they used the colour-coded categorization to effectively categorize articles and identify arguments for and against the topic. "I paid attention to the color code because the title of the source did not provide a clear understanding of whether the source is discussing for or against it. So, I think the color code helped. I could select arguments much faster once I understand this is for, against, and neutral." [P2] Some participants used this view to identify where the consensus on the issue lies: "having colours is really helpful. I saw again more positive to the statement. This time, I asked if companies should compensate the impacted employees. So yes, the company should compensate." [P1]

We found that participants engaged with the perspective panel to efficiently accomplish multiple conflicting information processing activities, such as a critically comparing conflicting perspectives, selecting information, and understanding where the social agreement lies.

5.2.2 Source Content Panel. Similar to the perspective panel, participants engaged with the source content panel to critically compare and contrast information across different sources and efficiently select information. The horizontal view and information arrangement under different subheadings were useful for accomplishing the above activities. "I could evaluate each one against the other. It was just a nice way of looking at it at the time. I looked it through that horizontally just to see how the perspectives change. I noticed that both ends of the spectrum used some of the reasoning just to argue for their own conclusions" [P4]. "I feel like it saved a bit of time going through the articles by myself. It did that work for me by grouping things into different sections and across those different articles. So I could easily do comparisons and select arguments in that way." [P5]

Moreover, P2 and P5 used subheadings to analyse the issue in different dimensions and organize information. “Subheadings provide a good outline for users to analyze the topic from different dimensions and categorize information to put it together.” [P2]

Also, participants used arguments presented in point form on this panel to define directions to find evidence and more information for deep understanding. “The summary is handy, so you don’t have to skim through all the articles. You’ll know what to look for in these articles.” [P2] “Provides a stock of arguments that is good enough. Maybe as a student, I will use this and then try to find the right academic journal to identify evidence to support chosen arguments.” [P1]

Additionally, participants used the point form summary to get an idea about the topic since they are not experts in the topic field. “I’m not an expert in this area, but the summary was helpful.” [P2] “I like the summaries, which were useful in a way to get an understanding of things.” [P3]

Participants engaged with the source content panel to navigate conflicting sources horizontally, compare and contrast information across different sources, and to efficiently select information. Moreover, they used the point form summarized view along with subheading categorization to get an understanding about the issue as a layperson, analyze it, organize information, and identify directions to explore further for deep understanding.

5.2.3 Quality Feature Panel. Participants utilized the summarized comparable presentation of important quality features in the quality feature panel to arrive at an evaluation judgment quickly. “When the quality features are put in one table like this, everything is in comparison, so I can quickly have a judgment of an evaluation. This one is more credible. This one is not.”[P1] “Publisher reliability or motives or who it was is important, and I remember there was one that was a lobby group, so that made me not want to look at that source. Because I decided that it would not be reliable at all.”[P3]

The participants heavily paid attention to the author’s expertise, publisher’s motives and reliability and timeliness of the sources to evaluate source quality. Several participants paid attention to the objectivity score. However, some explained they were unclear how the score was calculated.

Findings indicate that the quality feature panel was useful for efficiently discerning quality sources. The tool was also helpful in directing participants’ attention to sophisticated features when judging source quality.

5.2.4 Synthesized Response. We observed that the majority of participants (4 out of 6) did not engage with the synthesized response feature for processing conflicting information during the pilot test. P1 and P3 identified it as "a summary of everything" [P3] and "a good conclusion" [P1] noting that they used the information breakdown more than the synthesis during the activity.

6 DISCUSSION AND CONCLUSION

In this study, we presented the design rationale of the INFO-NAVIGATE tool, explaining its key features, implementation, and initial evaluation results at a system and user level. The preliminary user evaluation results of INFO-NAVIGATE provides insights into user engagement with different tool features when processing conflicting information. Primarily, the participants engaged with both the perspective panel and the source content panel to compare and contrast different perspectives and information in multiple sources, which helped them critically think about the reasons behind the differences and similarities. They utilized the horizontal presentation and perspective categorization to perform this comparison easily and efficiently. Participants also engaged with the Quality feature panel, with comparable views and sophisticated feature values to arrive at quick source quality evaluation judgments. Information corroboration by comparing and contrasting information in multiple sources and thorough source quality evaluation are identified as the most crucial activities that must be performed by internet users when dealing with conflicting information [15, 34]. Our study findings indicate that users engage with the perspective panel, source content panel and quality feature panel of INFO-NAVIGATE to accomplish these activities.

Moreover, participants engaged with perspective panel and source content panel to analyze, select, and organize information, and identify directions for deeper exploration, mainly using the perspective categorization and information points arrangement with subheadings as scaffolds. Deep understanding of information, information analyzing, selection and organization are the main activities defined under information processing in IPS model [5], which were supported by the perspective panel and source content panel in INFO-NAVIGATE. Additionally, participants utilized the perspective categorization and summary view of the source content panel to understand the controversial issue before starting information processing, as they were non-experts in the topic area.

Further research is required with larger and more diverse cohorts to understand differences across user behaviours, and how engagement with features of tools such as INFO-NAVIGATE can support users in navigating conflicting information. The introduced design rationale can inform future research, and the development of novel information seeking interfaces[6]. Educators can also embed this kind of technology and intervention design in their curriculum in addition to disciplinary content [8]. This way, learners can be equipped with capabilities to tackle the present problematic climate of the internet and improve their complex cognitive skills through learning experiences.

REFERENCES

[1] Open AI. [n. d.]. Prompt engineering. Retrieved August 31, 2024 from https://platform.openai.com/docs/guides/prompt-engineering

[2] Azure. 2024. Evaluation and monitoring metrics for generative AI. https://learn.microsoft.com/en-us/azure/ai-studio/concepts/evaluation-metrics-built-in?tabs=warning

[3] Nattapat Boonprakong, Benjamin Tag, and Tilman Dingler. 2023. Designing Technologies to Support Critical Thinking in an Age of Misinformation. *IEEE Pervasive Computing* 22, 3 (2023), 8–17. https://doi.org/10.1109/MPRV.2023.3275514

[4] Saskia Brand-Gruwel, Iwan Wopereis, and Yvonne Vermetten. 2005. Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in human behavior* 21, 3 (2005), 487–508. https://doi.org/10.1016/j.chb.2004.10.005

[5] Saskia Brand-Gruwel, Iwan Wopereis, and Amber Walraven. 2009. A descriptive model of information problem solving while using internet. *Computers & Education* 53, 4 (2009), 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004

[6] M Anne Britt and Cindy Aglinskas. 2002. Improving students' ability to identify and use source information. *Cognition and instruction* 20, 4 (2002), 485–522. https://doi.org/10.1207/S1532690XCI2004_2

[7] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? *arXiv preprint arXiv:2307.03826* (2023).

[8] Vicky Charisi, Laura Malinverni, Elisa Rubegni, and Marie-Monique Schaper. 2020. Empowering children's critical reflections on ai, robotics and other intelligent technologies. In *Proceedings of the 11th nordic conference on human-computer interaction: shaping experiences, shaping society.* 1–4. https://doi.org/10.1145/3419249.3420090

[9] Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. 2021. Design challenges for a multi-perspective search engine. *arXiv preprint arXiv:2112.08357* (2021).

[10] Kristine Deroover, Simon Knight, Paul F Burke, and Tamara Bucher. 2023. Why do experts disagree? The development of a taxonomy. *Public Understanding of Science* 32, 2 (2023), 224–246. https://doi.org/10.1177/09636625221110029

[11] Stefano Di Sotto and Marco Viviani. 2022. Health misinformation detection in the social web: an overview and a data science approach. *International Journal of Environmental Research and Public Health* 19, 4 (2022), 2173. https://doi.org/10.3390/ijerph19042173

[12] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71 (2023), 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[13] Priya Ethape, Riya Kane, Ghanashyam Gadekar, and Sahil Chimane. 2023. Smart automation using llm. *International Research Journal of Innovations in Engineering and Technology* 7, 11 (2023), 603. https://doi.org/10.47001/IRJIET/2023.711080

[14] Michael Gibbons, Camille Limoges, Helga Nowotny, and Simon Schwartzman. 2010. *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies.* SAGE Knowledge.

[6] We are currently running one such study in a pedagogic context.

[15] Elina K Hämäläinen, Carita Kiili, Miika Marttunen, Eija Räikkönen, Roberto González-Ibáñez, and Paavo HT Leppänen. 2020. Promoting sixth graders' credibility evaluation of Web pages: An intervention study. *Computers in Human Behavior* 110 (2020), 106372. https://doi.org/10.1016/j.chb.2020.106372

[16] McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, and Brenda Curtis. 2021. Bots and misinformation spread on social media: Implications for COVID-19. *Journal of medical Internet research* 23, 5 (2021), e26933. https://doi.org/10.2196/26933

[17] Andrea Ibieta, J Enrique Hinostroza, and Christian Labbé. 2019. Improving students' information problem-solving skills on the web through explicit instruction and the use of customized search software. *Journal of Research on Technology in Education* 51, 3 (2019), 217–238. https://doi.org/10.1080/15391523.2019.1576559

[18] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. 2017. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management* 53, 5 (2017), 1043–1061. https://doi.org/10.1016/j.ipm.2017.04.003

[19] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society* 23, 5 (2021), 1301–1326. https://doi.org/10.1177/1461444820959296

[20] Simon Knight, Isabella Bowdler, Heather Ford, and Jianlong Zhou. 2024. A visual scoping review of how knowledge graphs and search engine results page designs represent uncertainty and disagreement. *Information and Learning Sciences* (2024). https://doi.org/10.1108/ILS-02-2024-0016

[21] Keiichi Kobayashi. 2009. Comprehension of relations among controversial texts: Effects of external strategy use. *Instructional Science* 37 (2009), 311–324. https://doi.org/10.1007/s11251-007-9041-6

[22] Allan MacLean, Richard M Young, and Thomas P Moran. 1989. Design rationale: the argument behind the artifact. *ACM SIGCHI Bulletin* 20, SI (1989), 247–252. https://doi.org/10.1145/67450.67497

[23] Gary Marchionini. 2019. Search, sense making and learning: closing gaps. *Information and Learning Sciences* 120, 1/2 (2019), 74–86. https://doi.org/10.1108/ILS-06-2018-0049

[24] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology* 58, 13 (2007), 2078–2091. https://doi.org/10.1002/asi.20672

[25] Rebekah H Nagler, Rachel I Vogel, Sarah E Gollust, Alexander J Rothman, Erika Franklin Fowler, and Marco C Yzer. 2020. Public perceptions of conflicting information surrounding COVID-19: Results from a nationally representative survey of US adults. *PloS one* 15, 10 (2020), e0240776. https://doi.org/10.1371/journal.pone.0240776

[26] Nga Thanh Nguyen, Amanda McFadden, Donna Tangen, and Denise Beutel. 2013. Video-Stimulated Recall Interviews in Qualitative Research. *Australian Association for research in Education* (2013).

[27] Alamir Novin and Eric Meyers. 2016. Controversial Search Engine Results: An Exploratory Study of Information Presentation and Use. In *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.* https://doi.org/10.29173/cais957

[28] Alamir Novin and Eric Meyers. 2017. Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 175–184. https://doi.org/10.1145/3020165.3020185

[29] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web credibility: Features exploration and credibility prediction. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*. Springer, 557–568. https://doi.org/10.1007/978-3-642-36973-5_47

[30] 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. https://doi.org/10.1145/1518701.1518772

[31] Annelies Raes, Tammy Schellens, Bram De Wever, and Ellen Vanderhoven. 2012. Scaffolding information problem solving in web-based collaborative inquiry learning. *Computers & Education* 59, 1 (2012), 82–94. https://doi.org/10.1016/j.compedu.2011.11.010

[32] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1245–1254. https://doi.org/10.1145/1978942.1979127

[33] Antonette Shibani, Simon Knight, Kirsty Kitto, Ajanie Karunanayake, and Simon Buckingham Shum. 2024. Untangling Critical Interaction with AI in Students' Written Assessment. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6. https://doi.org/10.1145/3613905.3651083

[34] Marc Stadtler and Rainer Bromme. 2008. Effects of the metacognitive computer-tool met. a. ware on the web search of laypersons. *Computers in Human Behavior* 24, 3 (2008), 716–737. https://doi.org/10.1016/j.chb.2007.01.023

[35] Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W White, Reid Andersen, et al. 2024. The Use of Generative Search Engines for Knowledge Work and Complex Tasks. *arXiv preprint arXiv:2404.04268* (2024).

[36] Amber Walraven, Saskia Brand-Gruwel, and Henny PA Boshuizen. 2008. Information-problem solving: A review of problems students encounter and instructional solutions. *Computers in Human Behavior* 24, 3 (2008), 623–648. https://doi.org/10.1016/j.chb.2007.01.030

[37] Ruiyun Xu, Yue Feng, and Hailiang Chen. 2023. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135* (2023).

[38] Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087* (2023).

[39] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

A LLM PROMPTS

Table 2. LLM prompts used for sub-tasks in INFO-NAVIGATE (Numbered as in Fig. 2: Content Generation Pipeline)

Prompt Name	Task	Prompt
Prompt 1	Synthesized Response generation	"role": "system", "content": "You will be provided with a set of contents (delimited with XML tags) about the same topic. Generate a 100 words synthesis based on the given contents." "role":"user", "content": {body1}{body2}{body3}{body4}{body5}"
Prompt 2	Subheadings derivation	"role": "system", "content": "You will be provided with a set of contents (delimited with XML tags) about the same topic. Identify five main distinct broader sub-topics discuss in the contents and arrange them in descending order of occurrence frequency. Provide only the topics. Do not include frequency or description." "role":"user", "content": {body1}{body2}{body3}{body4}{body5}"
Prompt 3	Perspective derivation	"role": "system", "content": "You will be provided with content and a query. Your task is to summarize the main perspective of the given content in 20 words towards the given query." "role":"user", "content": f"content:{body} query:{query}"
Prompt 4	Convert query to a statement for the purpose of stance detection	"role": "system", "content": "You will be provided with a query. Convert that query to a statement. Answer in a consistent style" "role": "user", "content": "query: Is nuclear energy the answer for Australia's race towards net zero emission?" "role": "assistant", "content": "Nuclear energy in Australia" "role":"user", "content": f"query:{query}"
Prompt 5	Stance detection of the perspective	"role": "user","content": "What is the attitude of the sentence: \"Nuclear energy seen as a costly and delayed solution compared to renewables for Australia's net zero emission goals.\" to the target \"Nuclear energy is being considered as a potential solution for Australia's pursuit of net zero emissions.\" select the answer from the following three words: Favor, Against or Neutral. Answer in consistent style " "role": "assistant", "content": "Against" "role": "user", "content": f"What is the attitude of the sentence:{summary}. to the target{statement}. select from favor, against or neutral"
Prompt 6	Source point-form summary generation	"role": "system", "content": "You will be provided with content and a list of subtopics. Your task is to summarize the content in point form, at most 10 words per each point by preserving the meaning and classify each point into the provided subtopics. Do not need to provide word count in each information point. If there is no closely related information point for a subtopic, keep that subtopic category empty. Provide your output as a JSON object with the exact subtopics in the list as keys. Answer in consistent JSON format like : {"Subtopic 1": [],"Subtopic 2": [],"Subtopic 3": [],"Subtopic 4": []"Subtopic 5": [] "role":"user", "content": f"content:{body} subtopic list:{subTopics}"

B SEARCH RESULTS USED FOR EVALUATION

Table 3. Search Query and Resulting Sources Used for Content Generation Quality Evaluation

Search query	Sources
Should companies that adopt AI be taxed to compensate for the impacts on employees?	https://blogs.lse.ac.uk/businessreview/2022/11/24/should-machines-be-taxed-like-people/ https://www.brookings.edu/articles/navigating-the-future-of-work-a-case-for-a-robot-tax-in-the-age-of-ai/ https://www.imf.org/en/Blogs/Articles/2024/06/17/fiscal-policy-can-help-broaden-the-gains-of-ai-to-humanity https://www.futureofworkhub.info/comment/2019/12/4/robot-tax-the-pros-and-cons-of-taxing-robotic-technology-in-the-workplace https://www.brookings.edu/articles/tax-not-the-robots/
Is nuclear energy the answer for Australia's race towards net zero emission?	https://theconversation.com/is-nuclear-the-answer-to-australias-climate-crisis-216891 https://www.themandarin.com.au/249085-is-nuclear-the-answer-to-australias-climate-crisis/ https://www.csiro.au/en/news/all/articles/2023/december/nuclear-explainer https://www.theguardian.com/commentisfree/2024/mar/22/heres-why-there-is-no-nuclear-option-for-australia-to-reach-net-zero https://mckellinstitute.org.au/research/articles/explainer-heres-why-the-evidence-suggest-nuclear-doesnt-make-sense-for-australia/
Does red wine have health benefits?	https://www.medicalnewstoday.com/articles/265635 https://www.health.harvard.edu/blog/is-red-wine-good-actually-for-your-heart-2018021913285 https://www.healthline.com/nutrition/red-wine-good-or-bad https://www.webmd.com/diet/health-benefits-red-wine https://zoe.com/learn/red-wine-health-benefits

C AI-ASSISTED EVALUATION PROMPTS

C.1 Groundedness

You will be presented with a CONTEXT and an ANSWER about that CONTEXT. You need to decide whether the ANSWER is entailed by the CONTEXT by choosing one of the following rating:

5: The ANSWER completely follows logically from the information contained in the CONTEXT.

4: The ANSWER mostly follows logically from the information contained in the CONTEXT.

3: The ANSWER partially follows logically from the information contained in the CONTEXT.

2: The ANSWER mostly does not follow logically from the information contained in the CONTEXT.

1: The ANSWER is logically false from the information contained in the CONTEXT.

Select an integer score between 1 and 5 and if such integer score does not exist,

use 1: It is not possible to determine whether the ANSWER is true or false without further information.

Read the passage of information thoroughly and select the correct answer from the five answer labels.

Read the CONTEXT thoroughly to ensure you know what the CONTEXT entails. Note the ANSWER is generated by a computer system, it can contain certain symbols, which should not be a negative factor in the evaluation.

C.2 Relevance

Relevance measures how well the answer addresses the main aspects of the question, based on the context. Consider whether all and only the important aspects are contained in the answer when evaluating relevance. Given the context and question, score the relevance of the answer between one to five stars using the following rating scale:

One star: the answer completely lacks relevance

Two stars: the answer mostly lacks relevance

Three stars: the answer is partially relevant

Four stars: the answer is mostly relevant

Five stars: the answer has perfect relevance

This rating value should always be an integer between 1 and 5. So the rating produced should be 1 or 2 or 3 or 4 or 5.

C.3 Coherence

Coherence of an answer is measured by how well all the sentences fit together and sound naturally as a whole. Consider the overall quality of the answer when evaluating coherence. Given the question and answer, score the coherence of answer between one to five stars using the following rating scale:

One star: the answer completely lacks coherence

Two stars: the answer mostly lacks coherence

Three stars: the answer is partially coherent

Four stars: the answer is mostly coherent

Five stars: the answer has perfect coherency

This rating value should always be an integer between 1 and 5. So the rating produced should be 1 or 2 or 3 or 4 or 5.

C.4 Fluency

Fluency measures the quality of individual sentences in the answer, and whether they are well-written and grammatically correct. Consider the quality of individual sentences when evaluating fluency. Given the question and answer, score the fluency of the answer between one to five stars using the following rating scale:

One star: the answer completely lacks fluency

Two stars: the answer mostly lacks fluency

Three stars: the answer is partially fluent

Four stars: the answer is mostly fluent

Five stars: the answer has perfect fluency

This rating value should always be an integer between 1 and 5. So the rating produced should be 1 or 2 or 3 or 4 or 5.

C.5 Coverage

Coverage measures how well the answer contains the main points of the context, relevant to the question. Consider whether all the main points are contained in the answer when evaluating coverage. Given the context and question, score the coverage of the answer between one to five stars using the following rating scale:

One star: the answer completely lacks coverage

Two stars: the answer mostly lacks coverage

Three stars: the answer partially contains important points

Four stars: the answer mostly covers important points

Five stars: the answer has perfect coverage

This rating value should always be an integer between 1 and 5. So the rating produced should be 1 or 2 or 3 or 4 or 5.