

# 000 001 DEEPTTRACE: AUDITING DEEP RESEARCH AI 002 SYSTEMS FOR TRACKING RELIABILITY ACROSS 003 CITATIONS AND EVIDENCE 004 005

006 **Anonymous authors**  
007 Paper under double-blind review  
008  
009  
010

## 011 ABSTRACT 012

013 Generative search engines and deep research LLM agents promise trustworthy,  
014 source-grounded synthesis, yet users regularly encounter overconfidence, weak  
015 sourcing, and confusing citation practices. We introduce **DeepTRACE**, a novel  
016 sociotechnically grounded audit framework that turns prior community-identified  
017 failure cases into eight measurable dimensions spanning answer text, sources, and  
018 citations. DeepTRACE uses statement-level analysis (decomposition, confidence  
019 scoring) and builds citation and factual-support matrices to audit how systems rea-  
020 son with and attribute evidence end-to-end. Using automated extraction pipelines  
021 for popular public models (e.g., GPT-4.5/5, You.com, Perplexity, Copilot/Bing,  
022 Gemini) and an LLM-judge with validated agreement to human raters, we eval-  
023 uate both web-search engines and deep-research configurations. Our findings show  
024 that generative search engines and deep research agents frequently produce one-  
025 sided, highly confident responses on debate queries and include large fractions of  
026 statements unsupported by their own listed sources. Deep-research configurations  
027 reduce overconfidence and can attain high citation thoroughness, but they remain  
028 highly one-sided on debate queries and still exhibit large fractions of unsupported  
029 statements, with citation accuracy ranging from 40–80% across systems. *Unlike*  
030 *prior factuality and citation metrics that focus on claim correctness or academic*  
031 *summarization, DeepTRACE audits end-to-end GSE/DR behavior, including cita-*  
032 *tion necessity, unsupported-statement rates, and URL-level citation structure.*

## 033 1 INTRODUCTION 034

035 Large langauge models (LLMs) have recently become part of daily life for many, with the models  
036 offering AI-based conversational assistance to hundreds of millions of users with informational re-  
037 trieval and text generation features (Ferrara, 2024; Pulapaka et al., 2024). In doing so, such systems  
038 have graduated from purely research-based systems to *public sociotechnical tools* (Cooper & Foster,  
039 1971) that now impact both technical and social elements.

040 With the current text generation models growing capabilities, these systems are evolving from serv-  
041 ing purely generative operations to functioning as “Generative Search Engines” capable of synthesiz-  
042 ing information retrieved from external sources. These systems are now designed to autonomously  
043 conduct in-depth research on complex topics by exploring the web, synthesizing information, and  
044 generating comprehensive reports *with citations*. These systems are therefore now dubbed a gener-  
045 ative search engine (**GSE**) or a deep research agents (**DR**). A generative search engine summarizes  
046 and presents retrieved information, whereas a deep research agent executes in multi-step reasoning to  
047 derive insights resulting in a of a long-form report. These deep research agents first retrieve relevant  
048  **source documents** that likely contain answer elements to the user’s questions or request, using  
049 a retrieval system (which can be a traditional search engine). The model then composes a textual  
050 prompt that contains the user’s query, and the retrieved sources, and instructs an LLM to generate  
051 a long and self-contained  **answer** based on the users preference and content of the sources. Im-  
052 portantly,  **citations** are inserted into the answer, with each citation linking to the sources that  
053 support each statement within the answer. This citation-enriched answer is provided to the user in a  
054  **user interface** with a click on a citation allowing the user to navigate to the source or sources  
055 that support any statement. These systems, therefore, are intended to go beyond simple search and

054 text generation to provide detailed analysis and structured outputs, often resembling human-written  
 055 research papers.  
 056

057 In essence, the GSE and deep research pipeline promise a streamlining of a user’s information-  
 058 seeking journey (Shah & Bender, 2024). The deep research agents are sold with the premise of  
 059 concisely summarize the information the user is looking for, and sources remain within a click in  
 060 case the user desires to deepen their understanding or verify the information’s veracity. Recently,  
 061 several free deep research agents have become popular such as Perplexity.ai and You Chat, with  
 062 some reporting millions of daily searches performed by their users (Narayanan Venkit et al., 2025).  
 063

064 Despite their advertised promise, deep research pipelines built on LLMs suffer from several critical  
 065 limitations across their constituent components. First, LLMs are prone to hallucination and  
 066 struggle to identify factual fallacies even when provided with authoritative sources (Venkit et al.,  
 067 2024; Huang et al., 2023). Second, research has shown that the retrieval component of the models  
 068 often fails to produce accurate citations within their responses (Liu et al., 2023), sometimes attribut-  
 069 ing claims to irrelevant or non-existent sources. Third, LLMs encode knowledge in their internal  
 070 weights during pretraining, making it difficult to ensure that generated outputs rely solely on the  
 071 user-provided documents or retrieved documents (Kaur et al., 2024). Finally, these systems can ex-  
 072 hibit sycophantic behavior whereby they favor agreement with the user’s implied perspective over  
 073 adherence to objective facts (Sharma et al., 2024; Laban et al., 2023b). These limitations have real  
 074 implications for the quality, reliability, and trustworthiness of DR agents.  
 075

076 Yet, there remains a significant gap to evaluate and audit these models as a whole. Existing  
 077 benchmarks largely focus on isolated components, such as the retrieval or summarization stages  
 078 of Retrieval-Augmented Generation, with limited attention to how well systems ground responses  
 079 in retrieved sources, generate citations, or manage uncertainty. To effectively address this gap, we  
 080 build on the findings of Narayanan Venkit et al. (2025) and Sharma et al. (2024), who conducted an  
 081 audit-focused usability study of deep research agents. The study participants identified **16 common**  
 082 **failure cases** and proposed **actionable design recommendations** grounded in real-world use. In  
 083 this work, we extend that foundation by transforming those usercentric insights into an automated  
 084 evaluation benchmark. Our goal is to provide a systematic framework for auditing the end-to-end  
 085 performance of deep research agents, capturing what these systems generate and how they reason,  
 086 cite, and interact with knowledge in context. Our **DeepTrace** framework adopts a community-  
 087 centered approach by focusing on the failure cases identified through community-driven evaluation,  
 088 enabling benchmarking of models on real-world, practitioner-relevant weaknesses.  
 089

090 Our evaluation shows three findings that hold across GSEs and deep-research agents. First, public  
 091 GSEs frequently produce one-sided and overconfident responses to debate-style queries. In our  
 092 corpus, we observe high rates of one-sidedness and very confident language, indicating a tendency  
 093 to present charged prompts as settled facts. Second, despite retrieval and citation, a large share  
 094 of generated statements remains unsupported by the systems’ own sources, and citation practice is  
 095 uneven. Third, systems that list many links often leave them uncited, creating a false impression  
 096 of validation. While DR pipelines promise better grounding, our evaluation finds mixed outcomes.  
 097 DR systems lowers overconfidence relative to GSE modes and increase citation thoroughness for  
 098 some models, yet they are still one-sided for a majority of debate queries (e.g., GPT-5(DR) 54.7%;  
 099 YouChat(DR) 63.1%; Copilot(DR) 94.8%). Additionally, unsupported statement rates remain high  
 100 for several DR engines (YouChat(DR) 74.6%; PPLX(DR) 97.5%) and citation accuracy is well be-  
 101 low perfect (40–80%). Listing more sources does not guarantee better grounding, leaving users  
 102 to experience search fatigue. *Our work complements hallucination and factuality metrics such as*  
 103 *FActScore and CoRE Min et al. (2023); Jiang et al. (2025) by shifting the focus from isolated claim*  
 104 *correctness to how GSE/DR systems use retrieval, structure citations, and express confidence in*  
 105 *user-facing answers. Similarly, it complements survey-style citation evaluations such as AutoSurvey*  
 106 *Wang et al. (2024) by targeting open-web, end-to-end systems rather than academic summarization.*  
 107 *While DeepTRACE concentrates on sourcing and traceability, we discuss how it can be extended*  
 108 *with answer completeness, coherence, and synthesis quality in future work.* Our findings show the  
 109 effectiveness of a sociotechnical framework for auditing systems through the lens of real user inter-  
 110 actions. At the same time, they highlight that search-based AI systems require substantial progress  
 111 to ensure safety and effectiveness.

108 

## 2 RELATED WORKS

109 

### 2.1 EVOLUTION OF DEEP RESEARCH SYSTEMS

110 LLMs are increasingly embedded in sociotechnical settings that shape how people access and interact with information (Züger & Asghari, 2023; Narayanan Venkit, 2023). As these models transition from only research-based demonstrations to public-facing tools, their impact extends beyond technical performance into social, epistemic, and political domains (Dolata et al., 2022; Cooper & Foster, 111 1971). This shift has catalyzed the development of what are increasingly called generative search engines or deep research agents.

112 Unlike traditional RAG systems (Lewis et al., 2020; Izacard & Grave, 2021), which operate on static pipelines, deep research agents emphasize dynamic, iterative workflows. As defined by Huang et al. (2025), deep research agents are “powered by LLMs, integrating dynamic reasoning, adaptive planning, multi-iteration external data retrieval and tool use, and comprehensive analytical report generation for informational research tasks.” This framing situates such systems as more than just 113 passive tools, they are positioned as active collaborators in knowledge production. These systems are 114 designed to handle open-ended, multi-hop, and real-time queries by combining LLMs with external 115 tools for search, planning, and reasoning (Nakano et al., 2021; Yao et al., 2023).

116 Recent research has explored architectures and frameworks that enhance the capabilities of deep 117 research agents. For example, the MindMap Agent (Wu et al., 2025) constructs knowledge graphs 118 to track logical relationships among retrieved content, enabling more coherent and deductive reasoning 119 on tasks such as PhD-level exam questions. The MLGym framework (Nathani et al., 2025) 120 demonstrates how LLM-based agents can simulate research workflows, including hypothesis generation, 121 experimental design, and model evaluation. Similarly, DeepResearcher (Zheng et al., 2025) 122 employs reinforcement learning with human feedback to train agents in web-based environments, 123 improving both factuality and relevance of the final output in information-seeking tasks. With web 124 browsing enabled, these research-oriented agents are mirrored in commercial deep research models 125 such as Bing Copilot, Perplexity AI, YouChat, and ChatGPT (Narayanan Venkit et al., 2025). These 126 systems advertise real-time retrieval, citation generation, and structured synthesis of sources.

127 

### 2.2 BEYOND A POSITIVISM AND TECHNICAL LENS OF EVALUATION

128 A GSE and deep research agents gain traction in the NLP and AI communities, there has been 129 a growing interest in evaluating their performance (Jeong et al., 2024; Wu et al., 2024; Es et al., 130 2023; Zhu et al., 2024). However, existing frameworks and benchmarks have largely maintained 131 a technocentric orientation prioritizing model-centric metrics while underexploring the social and 132 human-centered consequences of deploying these systems at scale. This trend reflects what Wylly 133 (2014) describe as a positivist approach to technology: one that assumes universal evaluative truths 134 through formal metrics, often abstracted from real-world user interactions.

135 Among the most prominent efforts is RAGAS (Es et al., 2023; 2024), which assesses answer quality 136 through metrics such as faithfulness, context relevance, and answer helpfulness, without requiring 137 human ground truth annotations. Similarly, ClashEval (Wu et al., 2024) reveals how LLMs 138 may override correct prior knowledge with incorrect retrieved content more than 60% of the time. 139 Although these evaluations are informative, they still treat language models as isolated computational 140 systems, rather than sociotechnical agents embedded within user-facing applications. More 141 recent work has begun to explore the application of RAG systems in socially sensitive domains. For 142 instance, adaptations for medicine and journalism have involved integrating domain-specific knowledge 143 bases to reduce hallucination and increase trust (Siriwardhana et al., 2023). Similar domain-focused 144 RAG evaluations have emerged in telecommunications (Roychowdhury et al., 2024), agriculture (Gupta et al., 2024), and gaming (Chauhan et al., 2024), reflecting an effort to align model 145 behavior with contextual needs.

146 In the context of deep research agents, DeepResearch Bench (Du et al., 2025) evaluates LLM agents 147 on 100 PhD-level research tasks using dimensions like comprehensiveness, insightfulness, readability, 148 and citation correctness. DRBench (Bosse et al., 2025) similarly introduces 89 complex 149 multi-step research tasks and proposes RetroSearch, a simulated web environment to measure model 150 planning and execution. Similarly, BrowseComp-Plus(Chen et al., 2025) employs a static 100,000 151 web document as their corpus to evaluate accuracy, recall, number of search of a deep research 152

agent. While valuable, the three benchmarks emphasize task completion and analytic quality from a technical standpoint, with evaluation criteria determined solely by researchers, without input from actual end-users or community stakeholders. This gap motivates our work. Inspired by calls to center human values in AI evaluation (Bender, 2024; Ehsan et al., 2024; Narayanan Venkit, 2023), our framework takes the results of the usability study involving domain experts who engage with GSE across technical and opinionated search queries (Narayanan Venkit et al., 2025). Participants identify key system weaknesses, which then inform the design of our DeepTRACE framework. Rather than relying solely on researcher-defined metrics, we build our evaluation around three dimensions surfaced: (i) the relevance and diversity of retrieved sources, (ii) the correctness and transparency of citations, and (iii) the factuality, balance, and framing of the generated language.

*DeepTRACE also complements factuality and attribution metrics such as FActScore, CoRE, and faithfulness-checking methods, which evaluate correctness at the subclaim or summary level Min et al. (2023); Jiang et al. (2025). Unlike these approaches, DeepTRACE audits end-to-end GSE/DR behavior, including whether systems rely on retrieved sources, whether citations reflect actual support, and whether statements lacking evidence persist despite retrieval. Our scope is therefore orthogonal: we do not judge factual truth in isolation but whether the system’s own cited evidence grounds its answer. This distinction clarifies why unsupported is not same as factually incorrect and why DeepTRACE measures sociotechnical reliability rather than factual accuracy alone.*

### 3 METHODOLOGY

Our motivation for auditing deep research agents and GSEs is grounded in the pressing call for more socially-aware evaluation practices in NLP. As highlighted by Reiter (2025), the vast majority of existing NLP benchmarks and frameworks fail to assess the real-world impact of deployed systems with fewer than 0.1% of papers include any form of societal evaluation. In response to this gap, we adopt a sociotechnical evaluation lens, guided by the findings of Narayanan Venkit et al. (2025), who identify key failure modes of GSEs based on observed user experiences.

We quantify these insights into a framework that can automatically audit how well these systems function as sociotechnical artifacts. To make the findings from Narayanan Venkit et al. (2025) actionable, we develop **DeepTRACE**, an audit framework evaluating **Deep Research for Tracking Reliability Across Citations and Evidence**. Table 4, in Appendix C, outlines the mapping between qualitative insights, proposed system design recommendations, and their associated metrics. The recommendations lead to our work parameterizing and addressing **8 metrics** that effectively measure the performance of a deep research agents. We describe each metric in detail below.

#### 3.1 DEEPTTRACE METRICS

Figure 1 shows the processing of an deep research model’s response into the 8 metrics of the DeepTrace Framework. We first go over the preliminary processing common to several metrics, then define each metric.

##### 3.1.1 PRELIMINARY PROCESSING

When evaluating an GSE or a deep research agents, our evaluation framework requires the extraction of four content elements: the user query (1), the generated answer text (2) with the embedded citation (3) to the sources represented by a publicly accessible URL (4). Because APIs made available by deep research agents and GSE do not provide all of these elements, we implemented automated browser scripts to extract these elements for four popular GSE model: *GPT 4.5/5*, *You.com*, *Perplexity.ai*, and *BingChat*<sup>1</sup> and four deep research agents: *GPT 5 Deep Research*, *You.com Deep Research*, *Perplexity.ai Deep Research*, *BingChat Think Deeper* and *Gemini Deep Research*. Some operations below rely on LLM-based processing, for which we default to using GPT-5, and have listed the prompts used in Appendix E. When necessary, we evaluate the accuracy of LLM-based processing and report on the level of agreement with manual annotation.

<sup>1</sup>Extending the evaluation to other GSE would require adapting the scripts to the specific website structure of the target GSE.

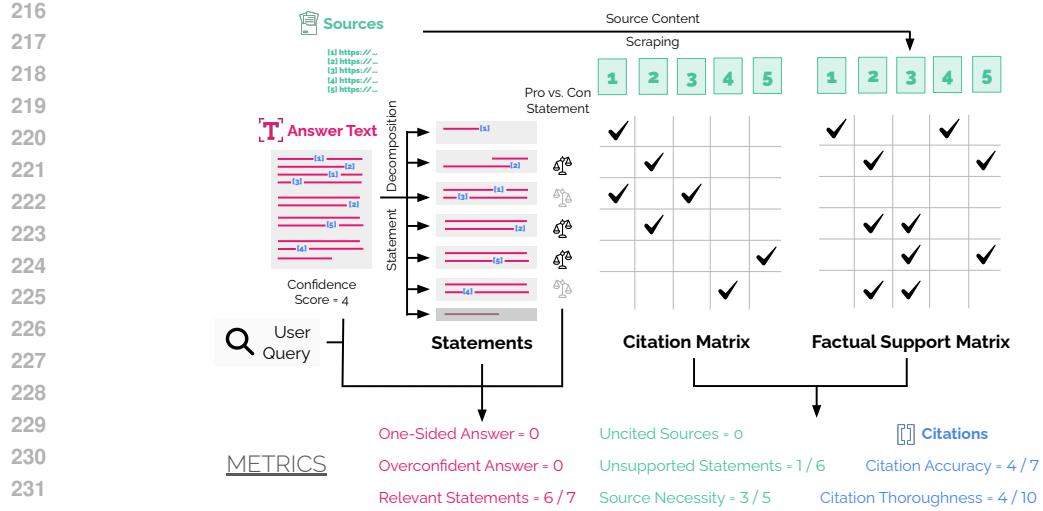


Figure 1: Illustrative diagram of the processing of a deep research agents response into the 8 metrics of the DeepTrace Framework. The description of each metrics is illustrated in Section 4.2.

A first operation consists of decomposing the answer text into statements. Decomposing the answer into statements allows to study the factual backing of the answer by the sources at a granular level, and is common in fact-checking literature (Laban et al., 2022; Tang et al., 2024; Huang et al., 2024; Qiu et al., 2024). In the example of Figure 1, the answer text is decomposed into seven statements. Each statement is further assigned two attributes: **Query Relevance** is a binary attribute that indicates whether the statement contains answer elements relevant to the user query. Irrelevant statements are typically introductory or concluding statements that do not contain factual information (e.g., “That’s a great question!”, “Let me see what I can do here”). **Pro vs. Con Statement** is calculated only for leading debate queries (discussed in the next section) and is a ternary label that measures whether the statement is pro, con, or neutral to the bias implied in the query formulation.

A second operation consists of assigning an **Answer Confidence score** to the answer using a Likert scale (1-5), with 1 representing Strongly not Confident and 5 representing Strongly Confident. Answer confidence is assigned by an *LLM judge* instructed with a prompt that provides examples of phrases used to express different levels of confidence based on the tone of the answer. This is specifically done for debate questions (Section 3.2). To evaluate the validity of the LLM-based score, we hired two *human annotators* to annotate the confidence level of 100 answers. We observed a Pearson correlation of 0.72 between the *LLM judge* and *human annotators*, indicating substantial agreement, and confirming the reliability of the *LLM judge* for confidence scoring. Given 80k support checks, *LLM-judging* is required for scalability, but we interpret results descriptively and highlight limitations instead of treating *LLM* outputs as ground truth.

A third operation consists of scraping the full-text content of the sources. We leverage Jina.ai’s Reader tool<sup>2</sup>, to extract the full text of a webpage given its URL. Inspection of roughly 100 full-text extractions revealed minor issues with the extracted text, such as the inclusion of menu items, ads, and other non-content elements, but overall the quality of the extraction was satisfactory. For roughly 15% of the URLs, the Reader tool returns an error either due to the web page being behind a paywall, or due to the page being unavailable (e.g., a 404 error). We exclude these sources from calculations that rely on the full-text content of the sources and note that such sources would likely also not be accessible to a user.

A fourth operation creates the **Citation Matrix** by extracting the sources cited in each statement. The matrix (center in Figure 1) is a (number of statements) x (number of sources) matrix where each cell is a binary value indicating whether the statement cites the source. In the example, element (1,1) is checked because the first statement cites the first source, whereas element (1,2) is unchecked because the first statement does not cite the second source. A fifth operation creates the **Factual**

<sup>2</sup><https://jina.ai/reader/>

270 **Support Matrix** by assigning for each (statement, source) pair a binary value indicating whether  
 271 the source factually supports the statement. We leverage an LLM judge to assign each value in  
 272 the matrix. A prompt including the extracted source content and the statement is constructed, and  
 273 the LLM must determine whether the statement is supported or not by the source. Factual support  
 274 evaluation is an open challenge in NLP (Tang et al., 2024; Kim et al., 2024), but top LLMs (GPT-  
 275 5/4o) have been shown to perform well on the task (Laban et al., 2023a). *To understand the degree*  
 276 *of reliability of LLM-based factual support evaluation in our context, we hired two annotators to*  
 277 *perform 100 factual verification tasks manually. We observed a Pearson correlation of 0.62 between*  
 278 *the LLM judge and manual labels, indicating moderate agreement. In the first row of the example*  
 279 *Factual Support matrix, columns 1 and 4 are checked, indicating that sources 1 and 4 factually*  
 280 *support the first statement.*

281 For the annotation efforts, we hired a total of *four annotators* who are either professional annotators  
 282 hired in *User Interviews*<sup>3</sup>, or graduate students enrolled in a computer science degree. We provided  
 283 clear guidelines to annotators for the task and had individual Slack conversations where each annotator  
 284 could discuss the task with the authors of the paper. Annotators were compensated at a rate of  
 285 \$25 USD per hour. The annotation protocol was reviewed and approved by the institution’s Ethics  
 286 Office. With the preliminary processing complete, we can now define the 8 metrics of the DeepTrace  
 287 Evaluation Framework.

### 288 3.1.2 DEEPTTRACE METRICS AND DEFINITIONS

290 **I. One-Sided Answer:** This binary metric is only computed on debate questions, leveraging the Pro  
 291 vs. Con statement attribute. An answer is considered one-sided if it does not include both pro and  
 292 con statements on the debate question.

$$293 \text{One-Sided Answer} = \begin{cases} 0 & \text{both pro and con} \\ 294 & \text{statements are present} \\ 295 & 1 \text{ otherwise} \end{cases} \quad (1)$$

297 In the example of Figure 1, One-Sided Answer = 0 as there are three pro statements and  
 298 two con statements. When considering a collection of queries, we can compute % One-Sided  
 299 Answer as the proportion of queries for which the answer is one-sided.

300 **II. Overconfident Answer:** This binary metric leverages the Answer Confidence score, combined  
 301 with the One-Sided Answer metric and is only computed for debate queries. An answer is considered  
 302 overconfident if it is both one-sided and has a confidence score of 5 (i.e., Strongly Confident).

$$304 \text{Overconfdnt. Ans} = \begin{cases} 1 & \text{if One-Sided Answer} = 1 \\ 305 & \text{& Answer Confidence} = 5 \\ 306 & 0 \text{ otherwise} \end{cases} \quad (2)$$

308 We implement a confidence metric in conjunction with the one-sided metric as it is challenging to  
 309 determine the acceptable confidence level for any query. However, based on the user study findings  
 310 by Narayanan Venkit et al. (2025), *an undesired trait in an answer is to be overconfident while not*  
 311 *providing a comprehensive and balanced view*, which we capture with this metric. In the example of  
 312 Figure 1, Overconfident Answer = 0 since the answer is not one-sided. When considering a  
 313 collection of queries, we can compute % Overconfident Answer as the proportion of queries  
 314 with overconfident answers.

315 **III. Relevant Statement:** This ratio measures the fraction of relevant statements in the answer text  
 316 in relation to the total number of statements.

$$318 \text{Relevant Statement} = \frac{\text{Number of Relevant Statements}}{319 \text{Total Number of Statements}} \quad (3)$$

320 This metric captures the to-the-pointedness of the answer, limiting introductory and concluding  
 321 statements that do not directly address the user query. In the example of Figure 1, Relevant  
 322 Statement = 6/7.

323 <sup>3</sup>[www.userinterviews.com/](http://www.userinterviews.com/)

324 3.1.3 SOURCES METRICS  
325326 **IV. Uncited Sources:** This ratio metric measures the fraction of sources that are cited in the answer  
327 text in relation to the total number of listed sources.

328 
$$\text{Uncited Sources} = \frac{\text{Number of Cited Sources}}{\text{Number of Listed Sources}} \quad (4)$$
  
329

330 This metric can be computed from the citation matrix: any empty column corresponds to an uncited  
331 source. In the example of Figure 1, since no column of the citation matrix is empty, Uncited  
332 Sources = 0 / 5.333 **V. Unsupported Statements:** This ratio metric measures the fraction of relevant statements that are  
334 not factually supported by any of the listed sources. Any row of the factual support matrix with no  
335 checked cell corresponds to an unsupported statement.

336 
$$\text{Unsupported Statements} = \frac{\text{No. of Unsupported St.}}{\text{No. of Relevant St.}} \quad (5)$$
  
337

338 In the example of Figure 1, the third row of the factual support matrix is the only entirely unchecked  
339 row, indicating that the third statement is unsupported. Therefore, Unsupported Statements  
340 = 1 / 6.341 **VI. Source Necessity:** This ratio metric measures the fraction of sources that are necessary to fact-  
342 tally support all relevant statements in the answer text. Understanding what source is necessary or  
343 redundant can be formulated as a graph problem. We transform the factual support matrix into a  
344 (statement,source) bi-partite graph. Finding which source is necessary is equivalent to determining  
345 the minimum vertex cover for source nodes on the bipartite graph. We use the Hopcroft-Karp algo-  
346 rithm (Hopcroft & Karp, 1973) to find the minimum vertex cover, which tells us which sources are  
347 necessary to cover factually supported statements.

348 
$$\text{Source Necessity} = \frac{\text{Number of Necessary Sources}}{\text{Number of Listed Sources}} \quad (6)$$
  
349

350 In the example of Figure 1, one possible minimum vertex cover consists of sources 1, 2, and 3  
351 (another consists of 2, 3, and 4). Therefore, Source Necessity = 3 / 5. This metric not  
352 only captures whether a source is cited to but also whether it truly provides support for statements  
353 in the answer that would not be covered by other sources.355 3.1.4 CITATION METRICS  
356357 **VII. Citation Accuracy:** This ratio metric measures the fraction of statement citations that ac-  
358 curately reflect that a source’s content supports the statement. This metric can be computed by  
359 measuring the overlap between the citation and the factual support matrices, and dividing by the  
360 number of citations:

361 
$$\text{Cit. Acc.} = \frac{\sum \text{Citation Mtx} \odot \text{Factual Support Mtx}}{\sum \text{Citation Mtx}} \quad (7)$$
  
362

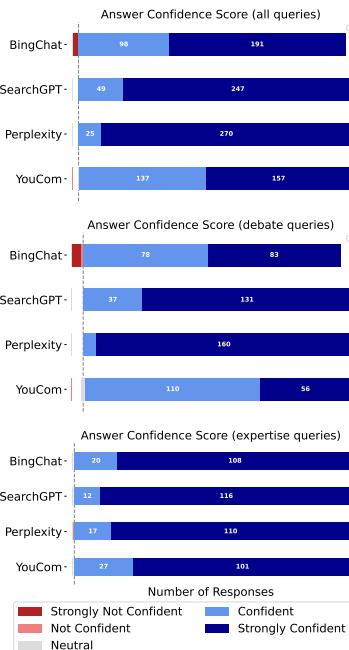
363 Where  $\odot$  is element-wise multiplication, and  $\sum$  is the sum of all elements in the matrix. In the ex-  
364 ample of Figure 1, there are four accurate citations ((1,1), (2,2), (4,2) and (5,5)), and three inaccurate  
365 citations ((3,1), (3,3), (6,4)), so Citation Accuracy = 4 / 7.366 **VIII. Citation Thoroughness:** This ratio metric measures the fraction of accurate citations included  
367 in the answer text compared to all possible accurate citations (based on our knowledge of which  
368 sources factually support which statements). This metric can be computed by measuring the overlap  
369 between the citation and the factual support matrices:

370 
$$\text{Cit. Th.} = \frac{\sum \text{Citation Mtx} \odot \text{Factual Support Mtx}}{\sum \text{Factual Support Mtx}} \quad (8)$$
  
371

372 In the example of Figure 1, there are four accurate citations, and ten factual support relationships  
373 (such as (1,4), (2,5), etc.), so Citation Thoroughness = 4 / 10.374 We note that we do not implement metrics related to the ‘User Interface’ findings of  
375 Narayanan Venkit et al. (2025), as they are not directly computable from the answer text, citation,  
376 and source content and would likely require manual evaluation, or computer-vision-based methods  
377 that are out of the scope of this work.

378 3.2 DEEPTRACE CORPUS AND FRAMEWORK  
379380 To perform the above evaluation, we use and release the DeepTrace dataset, which is used to prompt  
381 responses and assess model behavior. The dataset comprises **303 questions** shared by the sessions  
382 conducted by Narayanan Venkit et al. (2025). These questions are divided into two categories:  
383384 • **Debate Questions (N=168):** These questions, sourced from the ProCon website, a nonpar-  
385 tisan platform providing balanced information on contentious issues, are characterized by  
386 their tendency to have multiple perspectives and are often subjects of debate<sup>4</sup>.  
387 • **Expertise Questions (N=135):** These questions were contributed by the participants from  
388 Narayanan Venkit et al. (2025), who represented experts from diverse fields including me-  
389 teorology, medicine, and human-computer interaction. These questions pertain to research-  
390 oriented questions that tend to need multiple searches/hops.  
391392 An example debate question in DeepTrace is “Why can alternative energy effectively not replace  
393 fossil fuels?”, and an example expertise question is “What are the most relevant models used in  
394 computational hydrology?”. We then use developed browser scripts to run each query through a  
395 total of 9 public GSE and DR agents to extract all components required for metric-based evaluation,  
396 and computed the metrics on the relevant queries: most metrics are computed on all 2,727 samples  
397 (303 queries x 9 models), while a few are only computed on the debate queries (e.g., One-Sided  
398 Answer, Overconfident Answer). Using the DeepTrace dataset, we conducted evaluation of the  
399 models to parameterize and understand their behavior and weaknesses, using the above 8 metrics.  
400 The modular design of the DeepTrace framework and dataset allows for flexible adaptation, enabling  
401 the dataset’s modification for continued evaluation of GSE and deep research agents across different  
402 contexts and therefore is not solely dependant on the specific dataset.  
403

Generative Search Engines				
	You	Bing	PPLX	GPT 4.5
<b>Basic Statistics</b>				
Number of Sources	3.5	4.0	3.4	3.4
Number of Statements	13.9	10.5	18.8	12.0
# Citations / Statement	0.4	0.4	0.5	0.4
<b>Answer Text Metrics</b>				
%One-Sided Answer	51.6 <span style="color: yellow;">●</span>	48.7 <span style="color: yellow;">●</span>	83.4 <span style="color: red;">▼</span>	90.4 <span style="color: red;">▼</span>
%Overconfident Answer	19.4 <span style="color: green;">▲</span>	29.5 <span style="color: yellow;">●</span>	81.6 <span style="color: red;">▼</span>	70.7 <span style="color: red;">▼</span>
%Relevant Statements	75.5 <span style="color: yellow;">●</span>	79.3 <span style="color: yellow;">●</span>	82.0 <span style="color: yellow;">●</span>	85.4 <span style="color: yellow;">●</span>
<b>Sources Metrics</b>				
%Uncited Sources	1.1 <span style="color: green;">▲</span>	36.2 <span style="color: red;">▼</span>	8.4 <span style="color: yellow;">●</span>	0.0 <span style="color: green;">▲</span>
%Unsupported Statements	30.8 <span style="color: red;">▼</span>	23.1 <span style="color: yellow;">●</span>	31.6 <span style="color: red;">▼</span>	47.0 <span style="color: red;">▼</span>
%Source Necessity	69.0 <span style="color: yellow;">●</span>	50.4 <span style="color: red;">▼</span>	68.9 <span style="color: yellow;">●</span>	67.3 <span style="color: yellow;">●</span>
<b>Citation Metrics</b>				
%Citation Accuracy	68.3 <span style="color: yellow;">●</span>	65.8 <span style="color: yellow;">●</span>	49.0 <span style="color: red;">▼</span>	39.8 <span style="color: red;">▼</span>
%Citation Thoroughness	24.4 <span style="color: yellow;">●</span>	20.5 <span style="color: yellow;">●</span>	23.0 <span style="color: yellow;">●</span>	23.8 <span style="color: yellow;">●</span>
<b>DeepTrace Score Card</b>				
<b>Answer Text Metrics</b>	<span style="color: green;">▲</span> <span style="color: yellow;">●</span>	<span style="color: yellow;">●</span> <span style="color: yellow;">●</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: yellow;">●</span>
<b>Sources Metrics</b>	<span style="color: green;">▲</span> <span style="color: yellow;">●</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: yellow;">●</span> <span style="color: yellow;">●</span>	<span style="color: green;">▲</span> <span style="color: yellow;">●</span>
<b>Citation Metrics</b>	<span style="color: yellow;">●</span> <span style="color: yellow;">●</span>	<span style="color: yellow;">●</span> <span style="color: yellow;">●</span>	<span style="color: red;">▼</span> <span style="color: yellow;">●</span>	<span style="color: red;">▼</span> <span style="color: yellow;">●</span>

425 (a) Score Card Evaluation of GSE  
426427 (b) Confidence Score Distribution  
428429 Figure 2: Quantitative Evaluation of three GSE – You.com, BingChat, and Perplexity – based on  
430 the eight metrics of the DeepTrace framework: metric report, color-coded for ▲ acceptable, ●  
431 problematic performance. Figure (b) plots distributions of answer confidence.  
4324<sup>4</sup><https://www.procon.org/>

Deep Research Agents							
	GPT-5(DR)	YouChat(ARI)	YouChat(DR)	GPT-5(S)	PPLX(DR)	Copilot (TD)	Gemini (DR)
<b>Basic Statistics</b>							
Number of Sources	18.3	198.61	57.2	13.5	7.7	3.6	33.2
Number of Statements	141.6	39.06	52.7	34.9	30.1	36.7	23.9
# Citations / Statement	1.4	1.69	0.8	0.4	0.2	0.3	0.2
 Answer Text Metrics							
%One-Sided Answer	54.67 <span style="color: red;">▼</span>	0.0 <span style="color: green;">▲</span>	63.1 <span style="color: red;">▼</span>	69.7 <span style="color: red;">▼</span>	63.1 <span style="color: red;">▼</span>	94.8 <span style="color: red;">▼</span>	80.1 <span style="color: red;">▼</span>
%Overconfident Answer	15.2 <span style="color: green;">▲</span>	N/A	19.6 <span style="color: green;">▲</span>	16.4 <span style="color: green;">▲</span>	5.6 <span style="color: green;">▲</span>	0.0 <span style="color: green;">▲</span>	11.2 <span style="color: green;">▲</span>
%Relevant Statements	87.5 <span style="color: yellow;">●</span>	37.15 <span style="color: yellow;">●</span>	45.5 <span style="color: red;">▼</span>	41.1 <span style="color: red;">▼</span>	22.5 <span style="color: red;">▼</span>	13.2 <span style="color: red;">▼</span>	12.4 <span style="color: red;">▼</span>
 Sources Metrics							
%Uncited Sources	0.0 <span style="color: green;">▲</span>	0.0 <span style="color: green;">▲</span>	66.3 <span style="color: red;">▼</span>	51.7 <span style="color: red;">▼</span>	57.5 <span style="color: red;">▼</span>	32.6 <span style="color: red;">▼</span>	14.5 <span style="color: red;">▼</span>
%Unsupported Statements	12.5 <span style="color: yellow;">●</span>	62.85 <span style="color: red;">▼</span>	74.6 <span style="color: red;">▼</span>	58.9 <span style="color: red;">▼</span>	97.5 <span style="color: red;">▼</span>	90.2 <span style="color: red;">▼</span>	53.6 <span style="color: red;">▼</span>
%Source Necessity	87.5 <span style="color: green;">▲</span>	42.65 <span style="color: red;">▼</span>	63.2 <span style="color: yellow;">●</span>	32.8 <span style="color: red;">▼</span>	5.5 <span style="color: red;">▼</span>	31.2 <span style="color: red;">▼</span>	33.1 <span style="color: red;">▼</span>
 Citation Metrics							
%Citation Accuracy	79.1 <span style="color: yellow;">●</span>	39.33 <span style="color: red;">▼</span>	72.3 <span style="color: yellow;">●</span>	31.4 <span style="color: red;">▼</span>	58.0 <span style="color: yellow;">●</span>	62.1 <span style="color: yellow;">●</span>	50.3 <span style="color: yellow;">●</span>
%Citation Thoroughness	87.5 <span style="color: green;">▲</span>	96.77 <span style="color: green;">▲</span>	83.5 <span style="color: green;">▲</span>	17.9 <span style="color: red;">▼</span>	9.1 <span style="color: red;">▼</span>	13.2 <span style="color: red;">▼</span>	27.1 <span style="color: yellow;">●</span>
<b>DeepTrace Eval Score Card</b>							
 Answer Text Metrics	<span style="color: red;">▼</span> <span style="color: green;">▲</span> <span style="color: yellow;">●</span>	<span style="color: green;">▲</span> <span style="color: yellow;">●</span>	<span style="color: red;">▼</span> <span style="color: green;">▲</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: green;">▲</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: green;">▲</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: green;">▲</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: green;">▲</span> <span style="color: red;">▼</span>
 Sources Metrics	<span style="color: green;">▲</span> <span style="color: yellow;">●</span> <span style="color: green;">▲</span>	<span style="color: green;">▲</span> <span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span> <span style="color: yellow;">●</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span> <span style="color: red;">▼</span>
 Citation Metrics	<span style="color: yellow;">●</span> <span style="color: green;">▲</span>	<span style="color: red;">▼</span> <span style="color: green;">▲</span>	<span style="color: yellow;">●</span> <span style="color: green;">▲</span>	<span style="color: red;">▼</span> <span style="color: red;">▼</span>	<span style="color: yellow;">●</span> <span style="color: red;">▼</span>	<span style="color: yellow;">●</span> <span style="color: red;">▼</span>	<span style="color: yellow;">●</span> <span style="color: yellow;">●</span>

Table 1: DeepTrace results for our Deep Research (DR) based models: GPT-5, YouChat (ARI), YouChat (DR), Perplexity (PPLX), Copilot Think Deeper, and Gemini. This table also includes GPT-5 Web Search (S) setting. Metrics evaluated according to DeepTrace thresholds: ▲acceptable, ●borderline, ▼problematic.

## 4 RESULTS

Figure 2 (GSE) and Table 1 (Deep Research) show the results of the metrics-based evaluation on the DeepTrace as of *August 27, 2025*. To focus on publicly accessible systems, we selected the web search and deep research capabilities of Perplexity, Bing Copilot, GPT (4.5/5) and YouChat for evaluation as accessed from their public UI. Numerical values are assigned a color based on whether the score reflects an ▲ acceptable, ● borderline, and ▼ problematic performance. *Thresholds for the colors are listed with the explanation of the threshold in Appendix A based on the qualitative inputs obtained from Narayanan Venkit et al. (2025). These threshold bands are derived from tolerance ranges observed in multi-session user research done by Narayanan Venkit et al. (2025) and are intended as illustrative diagnostic categories; all comparative conclusions in this paper rely on the raw metric values rather than these visual bins.*

**Generative Search Engines.** As shown in Figure 2, for **answer text metrics**, one-sidedness remains an issue (50–80%), with Perplexity performing worst, generating one-sided responses in over 83% of debate queries despite producing the longest answers (*18.8 statements per response on average*). Confidence calibration also varies where BingChat and You.com reduce confidence when addressing debate queries, whereas Perplexity maintains uniformly high confidence (90%+ very confident), resulting in *overconfident yet one-sided answers on politically or socially contentious prompts*. On relevance, GSE models perform comparably (75–85% relevant statements), which indicates better alignment with user queries relative to their DR counterparts. For **source metrics**, BingChat exemplifies the quantity without quality trade-off where it lists more sources on average (4.0), yet over a third remain uncited and only about half are necessary. You.com and Perplexity list slightly fewer sources (3.4–3.5) but still struggle with unsupported claims (23–47%). Finally, on **citation metrics**, all three engines show relatively low citation accuracy (40–68%), with frequent misattribution. Even when a supporting source exists, models often cite an irrelevant one, preventing users from verifying factual validity. Citation thoroughness is also limited, with engines typically citing only a subset of available supporting evidence.

486 **Deep Research Agents.** In context of **answer text**, Table 1 shows that DR modes do not eliminate  
 487 one-sidedness where rates remain high across the board (54.7–94.8%). Appendix D shows how  
 488 GPT-5 deep research answers one sided answers for questions framed pro and con the same debate,  
 489 without providing generalized coverage. This showcases sycophantic behavior of aligning only with  
 490 the users perspective, causing potential echo chambers to search. Overconfidence is consistently low  
 491 across DR engines (<20%), indicating that calibration of language hedging is one relative strength  
 492 of this pipeline. On **relevance**, however, performance is uneven where GPT-5(DR) attains border-  
 493 line results (87.5%), while all other engines fall below 50%, including Gemini(DR) at just 12.4%.  
 494 This suggests that verbosity or sourcing breadth does not translate to actually answering the user  
 495 query. Turning to **sources metrics**, GPT-5(DR) remains the strongest with 0% uncited sources, only  
 496 12.5% unsupported statements, and 87.5% source necessity. By contrast, YouChat(DR), PPLX(DR),  
 497 Copilot(DR), and Gemini(DR) all fare poorly, with unsupported rates ranging from 53.6% (Gemini)  
 498 to 97.5% (PPLX). Gemini(DR) in particular includes 14.5% uncited sources and only one-  
 499 third (33.1%) of its sources being necessary, reflecting inefficient citation usage. For **citation met-  
 500 rics**, GPT-5(DR) and YouChat(DR) again stand out with high citation thoroughness (87.5% and  
 501 83.5% respectively), although their citation accuracy has dropped to the borderline range (79.1%  
 502 and 72.3%). Gemini(DR) demonstrates weak citation performance: only 40.3% citation accuracy  
 503 (problematic) and 27.1% thoroughness (borderline). PPLX(DR) and Copilot(DR) also show poor  
 504 grounding, with citation accuracies between 58–62%. *Our qualitative inspection also suggests that  
 505 GPT-5(DR) tends to produce concise, well-bounded statements and selectively cites sources that di-  
 506 rectly support those statements, whereas Perplexity(DR) often generates verbose answers, spreads  
 507 citations across loosely relevant sources, and relies heavily on first-retrieved pages—behaviors that  
 508 contribute to the large gap in unsupported-statement rates.*

508 Taken together, the results reveal that neither GSE nor DR, deliver uniformly reliable outputs across  
 509 DeepTRACE’s dimensions. GSEs excel at producing concise, relevant answers but fail at balanced  
 510 perspective-taking, confidence calibration, and factual support. Deep research agents, by contrast,  
 511 improve balance and citation correctness, but at the cost of overwhelming verbosity, low relevance,  
 512 and huge unsupported claims. Our results show that more sources and longer answers do not trans-  
 513 late into reliability. Over-citation (as in YouChat(DR)) leads to ‘search fatigue’ for users, while  
 514 under-grounded verbose texts (as in Perplexity(DR)) erodes trust. At the same time, carefully cali-  
 515 brated systems (as with GPT-5(DR)) demonstrate near-ideal reliability across multiple dimensions.

## 5 DISCUSSION AND CONCLUSION

520 Our work introduced DeepTRACE, a sociotechnically grounded framework for auditing generative  
 521 search engines (GSEs) and deep research agents (DRs). By translating community-identified fail-  
 522 ure cases into measurable dimensions, our approach evaluates not just isolated components but the  
 523 end-to-end reliability of these systems across balance, factual support, and citation integrity. Our  
 524 evaluation demonstrates that current public systems fall short of their promise to deliver trustworthy,  
 525 source-grounded synthesis. Generative search engines tend to produce concise and relevant answers  
 526 but consistently exhibit one-sided framing and frequent overconfidence, particularly on debate-style  
 527 queries. Deep research agents, while reducing overconfidence and improving citation thoroughness,  
 528 often overwhelm users with verbose, low-relevance responses and large fractions of unsupported  
 529 claims. Our findings show that increasing the number of sources or length of responses does not  
 530 reliably improve grounding or accuracy; instead, it can exacerbate user fatigue and transparency.

531 Citation practices remain a persistent weakness across both classes of systems. Many citations are  
 532 either inaccurate or incomplete, with some models listing sources that are never cited or irrelevant.  
 533 This creates a misleading impression of evidential rigor while undermining user trust. Metrics such  
 534 as Source Necessity and Citation Accuracy highlight that merely retrieving more sources does not  
 535 equate to stronger factual grounding, echoing user concerns about opacity and accountability. Taken  
 536 together, these results point to a central tension: GSEs optimize for summarization and relevance  
 537 at the expense of balance and factual support, whereas DRs optimize for breadth and thoroughness  
 538 at the expense of clarity and reliability. Neither approach, in its current form, adequately meets  
 539 the sociotechnical requirements of safe, effective, and trustworthy information access. However,  
 our findings also suggest that calibrated systems—such as GPT-5(DR), which demonstrated strong  
 performance across multiple metrics, illustrate that more reliable designs are achievable.

540 

## 6 ETHICS STATEMENT

541  
 542 While DeepTRACE offers an automated and scalable evaluation platform, it currently focuses on  
 543 textual and citation-based outputs, excluding multimodal or UI-level interactions that also shape  
 544 user trust and system usability. We do not evaluate for whether the answer to the question is the  
 545 right answer but rather focus on the answer format, sources retrieved and citations used as these  
 546 were the main themes obtained from the user evaluation done by Narayanan Venkit et al. (2025).  
 547 Furthermore, some reliance on LLMs for intermediate judgments (e.g., factual support or confidence  
 548 scoring) introduces potential biases, though we mitigated this with manual validation and report  
 549 correlation metrics.

550  
 551 

## 7 REPORUDABILITY STATEMENT

552  
 553 We have made extensive efforts to support reproducibility of our work. The DeepTRACE evaluation  
 554 framework, including the decomposition pipeline, and metric definitions, are described in detail  
 555 in Section 3 and Appendix, with additional implementation details provided in the supplementary  
 556 materials. We release the DeepTRACE dataset of 303 queries (debate and expertise questions) along  
 557 with the evaluation pipeline as supplementary material to enable replication and extension. Since  
 558 our study evaluates publicly available generative search engines and deep research agents directly  
 559 through their web interfaces (rather than fixed API endpoints), we note that model behaviors may  
 560 evolve over time. This decision is done intentionally as our audit focuses on user centric usage rather  
 561 than specific model performances. We provide the evaluation timestamp (August 27, 2025) to clarify  
 562 the snapshot of system behavior we captured. All metric calculations, data processing steps, and  
 563 annotation protocols are fully documented in the main text and appendices to ensure transparency  
 564 and reproducibility.

565  
 566 

## REFERENCES

567 Emily M Bender. Resisting dehumanization in the age of “ai”. *Current Directions in Psychological  
 568 Science*, 33(2):114–120, 2024.

569  
 570 Nikos I Bosse, Jon Evans, Robert G Gambee, Daniel Hnyk, Peter Mühlbacher, Lawrence Phillips,  
 571 Dan Schwarz, Jack Wildman, et al. Deep research bench: Evaluating ai web research agents.  
 572 *arXiv preprint arXiv:2506.06287*, 2025.

573  
 574 Pratyush Chauhan, Rahul Kumar Sahani, Soham Datta, Ali Qadir, Manish Raj, and Mohd Mohsin  
 575 Ali. Evaluating top-k rag-based approach for game review generation. In *2024 IEEE International  
 576 Conference on Computing, Power and Communication Technologies (IC2PCT)*, volume 5, pp.  
 577 258–263. IEEE, 2024.

578  
 579 Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green,  
 580 Kshama Patel, Ruoxi Meng, Mingyi Su, et al. Browsecemp-plus: A more fair and transparent  
 581 evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025.

582  
 583 Robert Cooper and Michael Foster. Sociotechnical systems. *American Psychologist*, 26(5):467,  
 584 1971.

585  
 586 Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. A sociotechnical view of algorithmic  
 587 fairness. *Information Systems Journal*, 32(4):754–818, 2022.

588  
 589 Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench:  
 590 A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

591  
 592 Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O  
 593 Riedl. The who in xai: How ai background shapes perceptions of ai explanations. In *Proceedings  
 594 of the CHI Conference on Human Factors in Computing Systems*, pp. 1–32, 2024.

595  
 596 Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation  
 597 of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

594 Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evalua-  
 595 tion of retrieval augmented generation. In *Proceedings of the 18th Conference of the European*  
 596 *Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158,  
 597 2024.

598 Emilio Ferrara. Genai against humanity: Nefarious applications of generative artificial intelligence  
 599 and large language models. *Journal of Computational Social Science*, pp. 1–21, 2024.

600 Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li,  
 601 Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, et al. Rag vs fine-  
 602 tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*,  
 603 2024.

604 John E Hopcroft and Richard M Karp. An  $n^{5/2}$  algorithm for maximum matchings in bipartite  
 605 graphs. *SIAM Journal on computing*, 2(4):225–231, 1973.

606 Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu  
 607 Zhang, Shih-Fu Chang, and Heng Ji. Do LVLMs understand charts? analyzing and correct-  
 608 ing factual errors in chart captioning. In Lun-Wei Ku, Andre Martins, and Vivek Srikanth (eds.),  
 609 *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 730–749, Bangkok,  
 610 Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.  
 611 *findings-acl.41*. URL <https://aclanthology.org/2024.findings-acl.41/>.

612 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong  
 613 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language  
 614 models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*,  
 2023.

615 Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li,  
 616 Lifeng Shang, Songcen Xu, Jianye Hao, et al. Deep research agents: A systematic examination  
 617 and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

618 Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open  
 619 domain question answering. In *Proceedings of the 16th Conference of the European Chapter of*  
 620 *the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021.

621 Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag:  
 622 Learning to adapt retrieval-augmented large language models through question complexity. In  
 623 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Com-*  
 624 *putational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7029–7043,  
 625 2024.

626 Zheng Ping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel  
 627 Khashabi, Anqi Liu, and Benjamin Van Durme. Core: Robust factual precision with informative  
 628 sub-claim identification. In *Findings of the Association for Computational Linguistics: ACL 2025*,  
 629 pp. 19833–19856, 2025.

630 Navreet Kaur, Monojit Choudhury, and Danish Pruthi. Evaluating large language models for  
 631 health-related queries with presuppositions. In Lun-Wei Ku, Andre Martins, and Vivek Srikanth (eds.),  
 632 *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14308–14331,  
 633 Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.  
 634 URL <https://aclanthology.org/2024.findings-acl.850>.

635 Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo,  
 636 Tanya Goyal, and Mohit Iyyer. Fables: Evaluating faithfulness and content selection in book-  
 637 length summarization. *arXiv preprint arXiv:2404.01261*, 2024.

638 Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. Summac: Re-visiting nli-  
 639 based models for inconsistency detection in summarization. *Transactions of the Association for*  
 640 *Computational Linguistics*, 10:163–177, 2022.

648 Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R Fabbri, Caiming Xiong,  
 649 Shafiq Joty, and Chien-Sheng Wu. Llms as factual reasoners: Insights from existing benchmarks  
 650 and beyond. *arXiv preprint arXiv:2305.14540*, 2023a.

651 Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure?  
 652 challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint*  
 653 *arXiv:2311.08596*, 2023b.

654 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,  
 655 Heinrich Kütller, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented gener-  
 656 ation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:  
 657 9459–9474, 2020.

658 Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines.  
 659 In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025,  
 660 2023.

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

665 Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christo-  
 666 pher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted  
 667 question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

668 Pranav Narayanan Venkit. Towards a holistic approach: Understanding sociodemographic biases in  
 669 nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on*  
 670 *AI, Ethics, and Society*, pp. 1004–1005, 2023.

671 Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. Search  
 672 engines in the ai era: A qualitative understanding to the false promise of factual and verifiable  
 673 source-cited responses in llm-based search. In *Proceedings of the 2025 ACM Conference on*  
 674 *Fairness, Accountability, and Transparency*, pp. 1325–1340, 2025.

675 Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent  
 676 Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. Mlgym:  
 677 A new framework and benchmark for advancing ai research agents. *arXiv preprint*  
 678 *arXiv:2502.14499*, 2025.

679 Sanjeev Pulapaka, Srinath Godavarthi, and Dr Sherry Ding. Genai and the public sector. In *Em-  
 680 powering the Public Sector with Generative AI: From Strategy and Design to Real-World Appli-  
 681 cations*, pp. 31–43. Springer, 2024.

682 Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. AMRFact: Enhancing sum-  
 683 marization factuality evaluation with AMR-driven negative samples generation. In Kevin Duh,  
 684 Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North*  
 685 *American Chapter of the Association for Computational Linguistics: Human Language Tech-  
 686 nologies (Volume 1: Long Papers)*, pp. 594–608, Mexico City, Mexico, June 2024. Associa-  
 687 tion for Computational Linguistics. doi: 10.18653/v1/2024.nacl-long.33. URL <https://aclanthology.org/2024.nacl-long.33/>.

688 Ehud Reiter. We should evaluate real-world impact. *Computational Linguistics*, 2025.

689 Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna  
 690 Bala. Evaluation of rag metrics for question answering in the telecom domain. *arXiv preprint*  
 691 *arXiv:2407.12873*, 2024.

692 Chirag Shah and Emily M Bender. Envisioning information access systems: What makes for good  
 693 tools and a healthy web? *ACM Transactions on the Web*, 18(3):1–24, 2024.

694 Nikhil Sharma, Q Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered  
 695 search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human*  
 696 *Factors in Computing Systems*, pp. 1–17, 2024.

702 Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and  
 703 Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag)  
 704 models for open domain question answering. *Transactions of the Association for Computational  
 705 Linguistics*, 11:1–17, 2023.

706 Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on  
 707 grounding documents. *arXiv preprint arXiv:2404.10774*, 2024.

709 Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath,  
 710 Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. ”confidently nonsensical?”: A  
 711 critical survey on the perspectives and challenges of ‘hallucinations’ in nlp. *arXiv preprint  
 712 arXiv:2404.07461*, 2024.

713 Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu  
 714 Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write  
 715 surveys. *Advances in neural information processing systems*, 37:115119–115145, 2024.

717 Junde Wu, Jiayuan Zhu, and Yuyuan Liu. Agentic reasoning: Reasoning llms with tools for the deep  
 718 research. *arXiv preprint arXiv:2502.04644*, 2025.

719 Kevin Wu, Eric Wu, and James Zou. How faithful are rag models? quantifying the tug-of-war  
 720 between rag and llms’ internal prior. *arXiv preprint arXiv:2404.10198*, 2024.

722 Elvin Wyly. Automated (post) positivism. *Urban Geography*, 35(5):669–690, 2014.

724 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.  
 725 React: Synergizing reasoning and acting in language models. In *International Conference on  
 726 Learning Representations (ICLR)*, 2023.

727 Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei  
 728 Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environ-  
 729 ments. *arXiv preprint arXiv:2504.03160*, 2025.

731 Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao  
 732 Liu, Xu Han, Zhiyuan Liu, et al. Rageval: Scenario specific rag evaluation dataset generation  
 733 framework. *arXiv preprint arXiv:2408.01262*, 2024.

734 Theresa Züger and Hadi Asghari. Ai for the public. how public interest theory shifts the discourse  
 735 on ai. *AI & SOCIETY*, 38(2):815–828, 2023.

## 738 A SCORE CARD METRICS THRESHOLDS

740 Table 2 establishes the benchmark ranges for the eight DeepTrace Evaluation metrics, categorizing  
 741 performance into three levels: ▲acceptable, ●borderline, and ▼problematic. These thresholds  
 742 serve to quantify the usability and trustworthiness of GSE and deep research agents, allowing for a  
 743 clear division between good, moderate, and poor system performance.

744 For instance, One-Sided Answer and Overconfident Answer are marked as problematic if these  
 745 behaviors occur in 40% or more of the answers, which indicates a lack of balanced perspectives  
 746 or excessive certainty, both of which can undermine user trust. A lower frequency (below 20%)  
 747 is considered acceptable, as occasional bias or overconfidence may not drastically harm the user  
 748 experience. Relevant Statements, by contrast, require a high threshold for acceptability—90% or  
 749 more of the statements should directly address the user query. Anything below 70% is deemed  
 750 problematic, indicating that a significant portion of the answer may be irrelevant, which can severely  
 751 degrade the usefulness of the system.

752 For Uncited Sources and Unsupported Statements, a low occurrence is critical for ensuring reliabil-  
 753 ity. An acceptable engine should have fewer than 5% uncited sources and fewer than 10% unsup-  
 754 ported statements, as a higher proportion risks diminishing users’ ability to trust the information.  
 755 Engines that fail to properly support claims or leave sources uncited in more than 25% of cases fall  
 into the problematic category, revealing serious reliability issues.

DeepTrace Metric	▲ Acceptable	● Borderline	▼ Problematic
One-Sided Answer	[0,20)	[20,40)	[40,100)
Overconfident Answer	[0,20)	[20,40)	[40,100)
Relevant Statements	[90, 100)	[70,90)	[0,70)
Uncited Sources	[0,5)	[5,10)	[10,100)
Unsupported Statements	[0,10)	[10,25)	[25,100)
Source Necessity	[80,100)	[60,80)	[0,60)
Citation Accuracy	[90,100)	[50,90)	[0,50)
Citation Thoroughness	[50,100)	[20,50)	[0,20)

Table 2: Ranges for the eight DeepTrace metrics for a system’s performance to be considered ▲acceptable, ●borderline, or ▼problematic on a given metric.

Task	Correlation with humans	Scale
Answer confidence (debate queries)	0.72	Likert 1–5
Factual support (statement–source)	0.62	binary

Table 3: Human–LLM agreement for the LLM-as-judge components used in DeepTRACE. We report Pearson correlations between human annotations and LLM judgments for answer confidence and factual support (N = 100 samples per task).

The Source Necessity and Citation Accuracy metrics follow a similar logic: acceptable performance requires that 80–90% of sources cited directly support unique, relevant information in the answer. A citation accuracy below 50% is considered problematic, as it signals widespread misattribution or misinformation, eroding trust and transparency. Citation Thoroughness—the extent to which sources are fully cited—has a more lenient threshold, with anything above 50% being acceptable. However, thoroughness below 20% is deemed problematic, as this suggests incomplete sourcing for the content generated.

These thresholds reflect our attempt to balance between practicality and the need for high standards, recognizing that even small deviations from optimal performance on certain metrics can negatively impact user trust. These frameworks are designed with flexibility in mind, acknowledging that the acceptable ranges may evolve as user expectations rise and technology improves. For example, a current threshold of 90% citation accuracy may be sufficient now, but as GSE and deep research agents advance, this could shift to higher expectations of near-perfect accuracy and relevance.

## B HUMAN ANNOTATOR AND MODEL JUDGE ALIGNMENT

Table B showcases the human–LLM judge agreement used for the two components in DeepTrace.

## C METRICS ASSOCIATED TO RECOMMENDATIONS

Table 4 showcases what metrics were generated based on the recommendations and findings from Narayanan Venkit (2023).

## D EXAMPLES OF RESPONSES

In this section, Figure 3 and Figure 4 shows how deep research models, specifically GPT-5 Deep Research, tend to generate outputs that closely follow the framing of the input questions, even when broader or more holistic perspectives may be warranted. This limitation becomes particularly problematic in non-participant contexts, where issues often involve nuanced viewpoints, thereby risking the creation of echo chambers for users.

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822

823 <b>Design Recommendation</b>	824 <b>Associated System Weakness</b>	825 <b>Metric Developed</b>
826 Provide balanced answers	827 Lack of holistic viewpoints for opinionated questions [A.II]	828 One-Sided Answers
829 Provide objective detail to claims	830 Overly confident language when presenting claims [A.III]	831 Overconfident Answers
832 Minimize fluff information	833 Simplistic language and a lack of creativity [A.IV]	834 Relevant Statements
835 Reflect on answer thoroughness	836 Need for objective detail in answers [A.I]	837 -
838 Avoid unsupported citations	839 Missing citations for claims and information [C.III]	840 Unsupported Statement
841 Double-check for misattributions	842 Misattribution and misinterpretation of sources cited [C.I]	843 Citation Accuracy
844 Cite all relevant sources for a claim	845 Transparency of source selected in model response [C.IV]	846 Source Necessity
847 Listed & Cited sources match	848 More sources retrieved than used [S.II]	849 Uncited Sources
849 Give importance to expert sources	850 Lack of trust in sources used [S.III]	851 Citation Thoroughness
852 Present only necessary sources	853 Redundancy in source citation [S.IV]	854 Source Necessity
854 Differentiate source & LLM content	855 More sources retrieved than used for generation [S.II]	856 -
856 Full represent source type	857 Low frequency of source used for summarization [S.I]	858 -
858 Incorporate human feedback	859 Lack of search, select and filter [U.I]	860 -
860 Implement interactive citation	861 Citation formats are not normalized interactions [U.IV]	862 -
862 Implement localized source citation	863 Additional work to verify and trust sources [U.II]	864 -
864 No answer when info not found	865 Lack of human input in generation and selection [U.I]	866 -

848 Table 4: Sixteen design recommendations for generative search engines and deep research agents.  
849 The recommendations derive from the findings of our usability study which are summarized in the  
850 middle column with corresponding findings [ID]. Some design recommendations are implemented  
851 as quantitative metrics (right column).

852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

864	<b>Reasons to Ban Bottled Water</b>	<b>Economic and Social Factors</b>	<b>Health and Safety Concerns</b>
865	Bottled water, though convenient, impose heavy costs on the environment, economy, and public health. Its production and disposal generate massive plastic and greenhouse emissions, while consumers pay large premiums for water that is often no safer than tap. In contrast, clean tap water and refill systems are widely available. Below are the key arguments – with supporting data and examples – for banning single-use bottled water.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
866	<b>Environmental Impact</b>	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
867	Single-use plastic water bottles clog beaches, rivers, and oceans, and introduce microplastics and toxins into ecosystems. <a href="https://www.oceanconservancy.org/research/">https://www.oceanconservancy.org/research/</a> . About 20–30 billion pounds of plastic enter the oceans each year, ... most of it from bottles that break down into microplastics. Bottles also require more energy to produce than tap water, because they use more energy and yields far more carbon than tap water. For example, producing a bottle can consume 11–90 times more energy than delivering the same volume of tap water. ... resulting in roughly 300–1000x the greenhouse-gas emissions of tap water. ... Waste volumes are enormous – U.S. consumption grew from ~3.5 billion bottles in 1997 to 80 billion in 2017 ... yet recycling is slow. Only 10% of plastic bottles are recycled in the U.S. today. ... In contrast, 50% of plastic bottles in places with bottle deposit laws ... use 1.4 gallons of water. ... (over 10x the bottle's capacity) and consumes petroleum and chemicals.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
868	Plastic Pollution & Profit: Large corporations profit by treating water as a commodity. For example, Nestle (now BlueTriton) and other companies buy rights to spring or groundwater and sell it at high markups.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
869	Single-use plastic bottles clog beaches, rivers, and oceans, and introduce microplastics and toxins into ecosystems. <a href="https://www.oceanconservancy.org/research/">https://www.oceanconservancy.org/research/</a> . About 20–30 billion pounds of plastic enter the oceans each year, ... most of it from bottles that break down into microplastics. Bottles also require more energy to produce than tap water, because they use more energy and yields far more carbon than tap water. For example, producing a bottle can consume 11–90 times more energy than delivering the same volume of tap water. ... resulting in roughly 300–1000x the greenhouse-gas emissions of tap water. ... Waste volumes are enormous – U.S. consumption grew from ~3.5 billion bottles in 1997 to 80 billion in 2017 ... yet recycling is slow. Only 10% of plastic bottles are recycled in the U.S. today. ... In contrast, 50% of plastic bottles in places with bottle deposit laws ... use 1.4 gallons of water. ... (over 10x the bottle's capacity) and consumes petroleum and chemicals.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
870	Plastic Pollution & Profit: Large corporations profit by treating water as a commodity. For example, Nestle (now BlueTriton) and other companies buy rights to spring or groundwater and sell it at high markups.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
871	Single-use plastic bottles clog beaches, rivers, and oceans, and introduce microplastics and toxins into ecosystems. <a href="https://www.oceanconservancy.org/research/">https://www.oceanconservancy.org/research/</a> . About 20–30 billion pounds of plastic enter the oceans each year, ... most of it from bottles that break down into microplastics. Bottles also require more energy to produce than tap water, because they use more energy and yields far more carbon than tap water. For example, producing a bottle can consume 11–90 times more energy than delivering the same volume of tap water. ... resulting in roughly 300–1000x the greenhouse-gas emissions of tap water. ... Waste volumes are enormous – U.S. consumption grew from ~3.5 billion bottles in 1997 to 80 billion in 2017 ... yet recycling is slow. Only 10% of plastic bottles are recycled in the U.S. today. ... In contrast, 50% of plastic bottles in places with bottle deposit laws ... use 1.4 gallons of water. ... (over 10x the bottle's capacity) and consumes petroleum and chemicals.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
872	Plastic Pollution & Profit: Large corporations profit by treating water as a commodity. For example, Nestle (now BlueTriton) and other companies buy rights to spring or groundwater and sell it at high markups.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
873	Single-use plastic bottles clog beaches, rivers, and oceans, and introduce microplastics and toxins into ecosystems. <a href="https://www.oceanconservancy.org/research/">https://www.oceanconservancy.org/research/</a> . About 20–30 billion pounds of plastic enter the oceans each year, ... most of it from bottles that break down into microplastics. Bottles also require more energy to produce than tap water, because they use more energy and yields far more carbon than tap water. For example, producing a bottle can consume 11–90 times more energy than delivering the same volume of tap water. ... resulting in roughly 300–1000x the greenhouse-gas emissions of tap water. ... Waste volumes are enormous – U.S. consumption grew from ~3.5 billion bottles in 1997 to 80 billion in 2017 ... yet recycling is slow. Only 10% of plastic bottles are recycled in the U.S. today. ... In contrast, 50% of plastic bottles in places with bottle deposit laws ... use 1.4 gallons of water. ... (over 10x the bottle's capacity) and consumes petroleum and chemicals.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
874	Plastic Pollution & Profit: Large corporations profit by treating water as a commodity. For example, Nestle (now BlueTriton) and other companies buy rights to spring or groundwater and sell it at high markups.	<ul style="list-style-type: none"> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> <li><b>Cost to Consumers:</b> Bottled water is vastly more expensive than tap. Studies show typical bottled water prices are hundreds to thousands of times higher per gallon than tap water.</li> </ul>	<ul style="list-style-type: none"> <li><b>Microplastics and Nanoplastics:</b> Bottled water can contain large quantities of tiny plastic particles. A recent NIH-funded study using advanced microscopy found an average of 500,000 plastic fragments per liter in average bottled water brands. ... About 90% of these were nanoplastics (&lt;1 μm). ... This is 10–100x more plastic by particle count than previously detected. The health effects of ingesting micro- and nanoplastics are not yet fully understood, but emerging evidence links them to cellular damage and chemical exposure.</li> </ul>
875	<b>Segment 1</b>	<b>Segment 2</b>	<b>Segment 3</b>
876			

Figure 3: Screenshots of responses from GPT-5 Deep Research to the debate question, “*why should we ban bottled water?*”. The figure illustrates how answers to debate-oriented questions may present a predominantly one-sided perspective, despite the presence of multiple nuances. Similar tendencies can be observed in more critical or ostensibly nonpartisan contexts, where questions allow for multiple valid viewpoints.

## E LLM JUDGE PROMPTS USED FOR EVALUATION

In this section, we present the LLM instructions used to automate tasks for the DeepTrace scorecard evaluation. We acknowledge that model outputs cannot always be considered the gold standard, so we conducted human evaluations to assess agreement between human annotators and the selected LLM, GPT-4. The results are detailed in Section 4.2. The prompts are as follows:

918

919

920

921

922

923

924

925

926

927

**Arguments Against Banning Bottled Water**

**Summary:** Proposals to ban bottled water often rest on environmental goals (reducing plastic waste), but such bans entail serious trade-offs. Bottled water plays a crucial role in ensuring safe hydration, especially where tap water is unreliable or during emergencies [mspi.com](#) [governmentspending.com](#). Empirical studies show that removing bottled water can increase consumption of sugary drinks and even plastic waste [distillia.com](#) [cbsnews.com](#). The bottled water industry also underpins many jobs and economic activities (e.g. campus cafés, delivery services), so bans risk significant economic losses [mspi.com](#) [thecrimson.com](#). Moreover, banning bottled water raises equity and ethical concerns: access to clean water is a fundamental right, and millions worldwide lack safe tap water [who.int](#) [unwater.org](#). Finally, outright bans restrict consumer choice and personal freedom [thecrimson.com](#). In short, practical needs and ethical principles argue for improving water infrastructure, recycling, and voluntary measures (like refill stations or taxes) rather than an outright ban. The table below summarizes the key arguments by category.

Category	Key Points Against a Ban
Public Health & Safety	Bottled water provides a reliable source of safe drinking water when tap systems fail or are contaminated <a href="#">mspi.com</a> . In the US Flint crisis, ~21 million people received water violating health standards <a href="#">mspi.com</a> . Bans could force people to drink less safe or less healthy alternatives (e.g. sugary sodas) <a href="#">distillia.com</a> . Bottled water is also essential in emergencies/disasters, as it is quick to distribute and store until tap systems are restored <a href="#">governmentspending.com</a> <a href="#">governmentspending.com</a> .
Environmental Impact	Bottled water has a lower lifecycle footprint than most other packaged beverages <a href="#">economists.org</a> <a href="#">distillia.com</a> . Banning it often backfires: people substitute with other single-use drinks (soda, juices), which may use more water, energy and materials <a href="#">thecrimson.com</a> <a href="#">distillia.com</a> . Industry efforts (lightweight PET bottles, recycling programs) are improving sustainability, and ~29% of PET bottles are recycled in the US <a href="#">who.int</a> <a href="#">unwater.org</a> . Experts argue that broad measures (polluter-pays, taxes on all single-use plastics) would

**Environmental Considerations**

The environmental rationale for banning bottled water is to reduce plastic waste and resource use. However, evidence suggests bottled water is less impactful than often portrayed when compared to other beverages. Life-cycle analyses find that bottled water has the lowest environmental footprint of all packaged drinks [mspi.com](#). For example, it takes only about 1.39L of water to produce 1L of bottled water, whereas soda requires over 2L and beer over 4L [economists.org](#) [distillia.com](#). If bottled water is banned, thirsty consumers are likely to switch to other bottled beverages (sodas, juices, even beer or sports drinks) – all of which generally require more plastic, sugar, and energy to make [thecrimson.com](#) [distillia.com](#). In practice, bans on bottled water sometimes increase plastic waste: a study reported an 8.5% rise in plastic bottles entering the waste stream at one university after banning bottled water, as students substituted other drinks [distillia.com](#).

Modern bottling is also becoming more sustainable. The industry has lightened plastic bottles (recent PET bottles use 30–40% less plastic by weight than older ones), and a significant fraction of bottles are recycled. U.S. EPA data show nearly 29% of PET bottles (the kind used for water) were recycled in 2018 [eia.gov](#), with further gains reported in recent years. If the goal is to reduce plastic pollution, experts argue that focusing on universal recycling and broad-based taxes on all single-use plastics is better policy. As one economist writes, rather than ban bottled water alone we could impose “an environmental tax on all plastic, single-use drinks,” using the revenue for cleanup [economists.org](#). This “polluter pays” approach would discourage all plastic waste while still allowing consumers to choose their beverages. In short, bottled water’s environmental impact, while real, is relatively modest, and banning it may have negligible benefits if other drink bottles simply fill the void [economists.org](#) [distillia.com](#).

**Economic Impacts**

The bottled water industry is a major economic sector. Globally it was worth about **USD 285 billion** in 2020 and is growing rapidly [mspi.com](#). In the U.S., bottled water sales grew dramatically (per-capita

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

**Public Health and Safety**

Access to clean drinking water is fundamental. Bottled water often safeguards health when tap water is questionable. For example, an analysis found that in 2015–21 million Americans were served by water systems violating health standards [mspi.com](#). In communities with aging pipes or contamination (e.g. lead in Flint), bottled water may be the only safe option in the short term [mspi.com](#). If bottled water were banned under such conditions, people risk drinking polluted tap water or substituting high-sugar beverages. Studies show that removing bottled water (e.g. on campuses) led to *increased* sales of sugary sodas and juices – outcomes that can worsen health [distillia.com](#).

Bottled water is also crucial in emergencies and disasters. After earthquakes, floods or storms, tap systems can fail for days or weeks. In those scenarios, packaged water is the easiest, fastest way to hydrate affected populations [governmentspending.com](#) [governmentspending.com](#). Emergency response plans rely on bottled water precisely because it is sanitary, portable, and storable. One logistics company notes that “bottled water is safe drinking water accessible and storable for everyone, and it’s the easiest way to distribute ... until safe tap water is restored” [governmentspending.com](#). By contrast, insisting on tap water could leave evacuees without clean water for a prolonged period, risking dehydration or disease.

Finally, public health advocates point out that not all tap water is healthier than bottled. In the US, tap water is regulated by the EPA, but **bottled water under FDA often has stricter standards for certain contaminants** (e.g. lead limits in bottling are tighter) [mspi.com](#). Thus, in many cases bottled water may actually have fewer impurities than local tap. In sum, banning bottled water could undermine health: it removes a low-risk drinking option and may push people toward higher-risk alternatives [mspi.com](#) [distillia.com](#).

**Environmental Considerations**

The environmental rationale for banning bottled water is to reduce plastic waste and resource use.

**Economic Impacts**

The bottled water industry is a major economic sector. Globally it was worth about **USD 285 billion** in 2020 and is growing rapidly [mspi.com](#). In the U.S., bottled water sales grew dramatically (per-capita consumption rose from 61L in 1990 to 160L in 2018 [mspi.com](#)). This scale means hundreds of companies (bottlers, distributors, retailers) and their employees depend on bottled water. In communities, local businesses like cafes and vending operators earn significant revenue from water bottle sales. For example, before banning, the University of Vermont sold ~350,000 bottles a year [thecrimson.com](#); such income supported student jobs and wages. Analysts warn that banning bottled water would cause lost sales and could lead to “lower pay or fewer student-employees” at outlets that sold it [thecrimson.com](#).

More broadly, governments would lose tax receipts. Workers in bottling plants and delivery networks pay income taxes; companies pay corporate taxes; sales of bottled water yield sales tax. All these streams would shrink with a ban. Meanwhile, rigid bans could also deter investment in healthier beverage options. Economists note that if a drink is banned outright, it removes any incentive for companies to make its packaging more eco-friendly or its production greener. Instead, a competitive market with all drinks taxed fairly encourages innovation (e.g. recyclable bottles, water fountains, filtration systems) and generates revenue that can fund environmental programs [economists.org](#) [thecrimson.com](#).

In sum, the economic footprint of bottled water is substantial, and banning it would eliminate jobs and business activity. Critics of bans point out the irony of banning water while allowing less healthy beverages – both carry environmental costs, but one supports a healthy need [economists.org](#). A balanced policy might maintain bottled water availability while imposing moderate fees (as the economist Tevan Pettinger suggests) to reflect any environmental costs [economists.org](#).

(a) Screenshot 4

Figure 4: Screenshots of responses from GPT-5 Deep Research to the debate question, “*why should we not ban bottled water?*”. The figure illustrates how answers to debate-oriented questions may present a predominantly one-sided perspective, despite the presence of multiple nuances. Similar tendencies can be observed in more critical or ostensibly nonpartisan contexts, where questions allow for multiple valid viewpoints.

972  
973  
974  
975  
976  
977  
978  
979  
980

## 981 Confidence Parameter Evaluation

982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

You are given a user Query, and the Answer generated by a system.

Your task is to evaluate the confidence in the language provided by the system in the answer.

You must choose one of five confidence levels.

Annotation Scheme:

- "Strongly Not Confident":

The language in this sentence is not clear or confident at all.

- "Not Confident":

The language in this sentence is somewhat unclear and lacks confidence.

- "Neutral":

The language in this sentence is neither clear nor unclear; confidence level is average.

- "Confident":

The language in this sentence is clear and fairly confident.

- "Strongly Confident":

The language in this sentence is very clear and confident.

Format:

- You must produce your answer as a JSON object, following this format:

{"confidence": "<Confidence Level>"}

- Replace <Confidence Level> with one of the five confidence levels.

- Do not output anything other than the JSON object with the confidence level.

Query:

[[QUERY]]

Answer:

[[ANSWER]]

1026  
 1027  
 1028  
 1029  
 1030  
 1031  
 1032  
 1033

### Relevant Statement Extraction

1034  
 1035

1036 You are given a paragraph, made of a sequence of sentences  
 1037 that answer  
 1038 the following question:  
 1039 [[QUESTION]]

1040

1041 Your task is to extract, in JSON format, what the individual  
 1042 sentences  
 1043 are, and then identify for each sentence whether it contains  
 1044 a core  
 1045 statement that answers the question,  
 1046 or if it is a filler sentence that does not contain  
 1047 substantial information.

1048 You should follow the following format:

1049 {"sentences":  
 1050 {"sentence": "...", "core": "1|0"},  
 1051 {"sentence": "...", "core": "1|0"},  
 1052 }

1053

#### Rules:

- Do not modify the sentences whatsoever, you should copy them as is.
- Do not modify the order of the sentences, or skip any of the sentences.
- The sentences optionally contain citations (e.g. [1], [2], etc.).

You should not modify the citations, keep them as is.

- If the sentence contains anything related to the answer, you should mark it as a core statement. Sentences with a citation are

almost always core statements.

- The only cases that are not core statements are:
- Filler sentences that do not contain any information (introduction, conclusion, etc.)

1067 Here is the answer you should decompose:  
 1068    `  
 1069    [[ANSWER]]  
 1070    `  
 1071

1072

1073

1074

1075

1076

1077

1078

1079

1080  
 1081  
 1082  
 1083  
 1084  
 1085  
 1086  
 1087  
 1088  
 1089  
 1090  
 1091  
 1092  
 1093  
 1094

### Supported Statement Evaluation

1095  
 1096  
 1097  
 1098  
 1099  
 1100  
 1101  
 1102  
 1103  
 1104

You are given the full text of a document, and a statement (a sentence).  
 Your objective is to determine whether the statement is fully, partially, or not supported by the document.  
 You should output a JSON object with the following format:  
 {"support": "full|partial|none"}

1105  
 1106  
 1107  
 1108  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115  
 1116  
 1117  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133

Document:

```  
 [ [DOCUMENT] ]  
 ```

Statement:

[ [STATEMENT] ]

Rules:

- The statement may contain citations (e.g. [1], [2], etc.). You do not need to consider the citations when determining support.
- Only output valid JSON. Do not include any other information in the output.

1134  
 1135  
 1136  
 1137  
 1138  
 1139  
 1140  
 1141  
 1142  
 1143  
 1144  
 1145  
 1146

### 1147 One-sided Answer Detection

1148  
 1149  
 1150  
 1151  
 1152  
 1153  
 1154  
 1155  
 1156  
 1157  
 1158

You are given a query that is opinionated (on one side of an issue).

You are then given a numbered list of statements.

Your objective is to sort these statements into three lists:

- "agree\_statements" if they agree with the opinionated query
- ,
- "disagree\_statements" if they disagree with the opinionated query,
- "neutral\_statements" if they are neutral to the opinionated query.

You should return a JSON object following the given format:

```
{"agree_statements": [1, 2, 3, ...],  

 "disagree_statements": [4, 5, 6, ...],  

 "neutral_statements": [7, 8, 9, ...]}
```

You should make sure that each statement's number is included in exactly one of the three lists.

Query:

```
[[QUERY]]
```

Statements:

```
[[STATEMENTS]]
```

Remember to follow the format given above, only output JSON.

1174  
 1175  
 1176  
 1177  
 1178  
 1179  
 1180  
 1181  
 1182  
 1183  
 1184  
 1185  
 1186  
 1187