Comparing the Subjective Opinions and Justifications of LLMs and Web Search Engines

Md Taukir Azam Chowdhury
University of California, Riverside
United States of America
mchow068@ucr.edu

Vagelis Hristidis
University of California, Riverside
United States of America
vagelis@cs.ucr.edu

Kevin Esterling
University of California, Riverside
United States of America
kevin.esterling@ucr.edu

Jannat Ara Meem
University of California, Riverside
United States of America
jmeem001@ucr.edu

Zabir Al Nazi
University of California, Riverside
United States of America
znazi002@ucr.edu

Abstract

In this paper, we study the stance and reasoning of popular LLMs using two well-known subjective opinion question datasets, ArguAna (250 subjective questions) and OpinionQA (survey questions on privacy, political views, and health). We compare the implicit stance and reasoning of popular Web search engines on the same collections of questions. First, we compare the stances (support, oppose, neutral) of LLMs and Web search engines across these topics. Then we compare their justifications to support these stances based on two well-accepted justification classifications: Lippi's classification (epistemic, practical, moral) and the Rhetorical Triangle Framework (logos, ethos, pathos). Our experiments show that the LLMs (ChatGPT and Gemini) have similar stances and justifications to each other but significantly differ from Web search sources. Specifically, LLM justifications are dominated by epistemic (Lippi's taxonomy) and logos (the Rhetorical Triangle Framework) categories, whereas Web search results demonstrate a broader range, including moral, practical, and epistemic reasoning. This implies that LLMs, perhaps due to their training or alignment processes, tend to underrepresent the diversity of human reasoning in response to subjective queries compared to search engines. We publish a corpus of the Web search result documents and the justifications extracted for the analyzed collections of subjective questions for others to build on our work.

Keywords

Subjective Question Answering, Justification Classification, Large Language Models; Web Search Engines, Argument Mining, Opinion Diversity

ACM Reference Format:
Md Taukir Azam Chowdhury, Vagelis Hristidis, Kevin Esterling, Jannat Ara Meem, and Zabir Al Nazi. 2026. Comparing the Subjective Opinions and Justifications of LLMs and Web Search Engines. In *18th ACM Web Science Conference (WebSci '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3795766.3799755

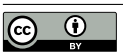
1 Introduction

Subjective question answering (QA) addresses queries whose "answers depend not on verifiable facts but on perspectives, values, and internal opinions" [4]. Questions such as *"Should governments ban fossil fuel cars?"* or *"Is remote work better for productivity?"* require reasoning over divergent viewpoints rather than factual lookup [26, 30]. Subjective questions are increasingly common in information-seeking settings, yet little work has studied how legacy (Web search) and modern (LLM) information-seeking technologies differ in their responses.

In this paper, we investigate how Web search engines and large language models (LLMs) differ in their stances toward subjective questions covering areas such as social norms, politics, and health. We compare systems within each group, considering differences among search engines and among LLMs, as well as across the two, to examine how their perspectives align or diverge. In addition to studying the stances of information-seeking technologies on popular subjective questions, we also study and compare the justifications used to support the stances. We consider two complementary frameworks of argumentation: Lippi's epistemic, practical, and moral taxonomy [15] and Aristotle's logos, ethos, and pathos model [22].

As shown in Figure 1, to compute the stance and justification of Web search engines, we considered the two most popular search engines: Google and Bing. We collected the top-100 Web search results for each question, and analyzed their stance and justification with the help of an LLM (llama3.1-8B). To compute stance and justification for LLMs, we picked ChatGPT and Gemini, and used different variants of each question (rephrased and negated). We added clear definitions of stance and justification categories to the prompts and included a few examples to map them into our taxonomies. We further studied how changing the variant of a question through negation affects the stance and justification patterns produced by the LLMs.

We also created and published a comprehensive dataset that contains all the data collected during our study, including nearly



WebSci '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2504-3/2026/05
https://doi.org/10.1145/3795766.3799755

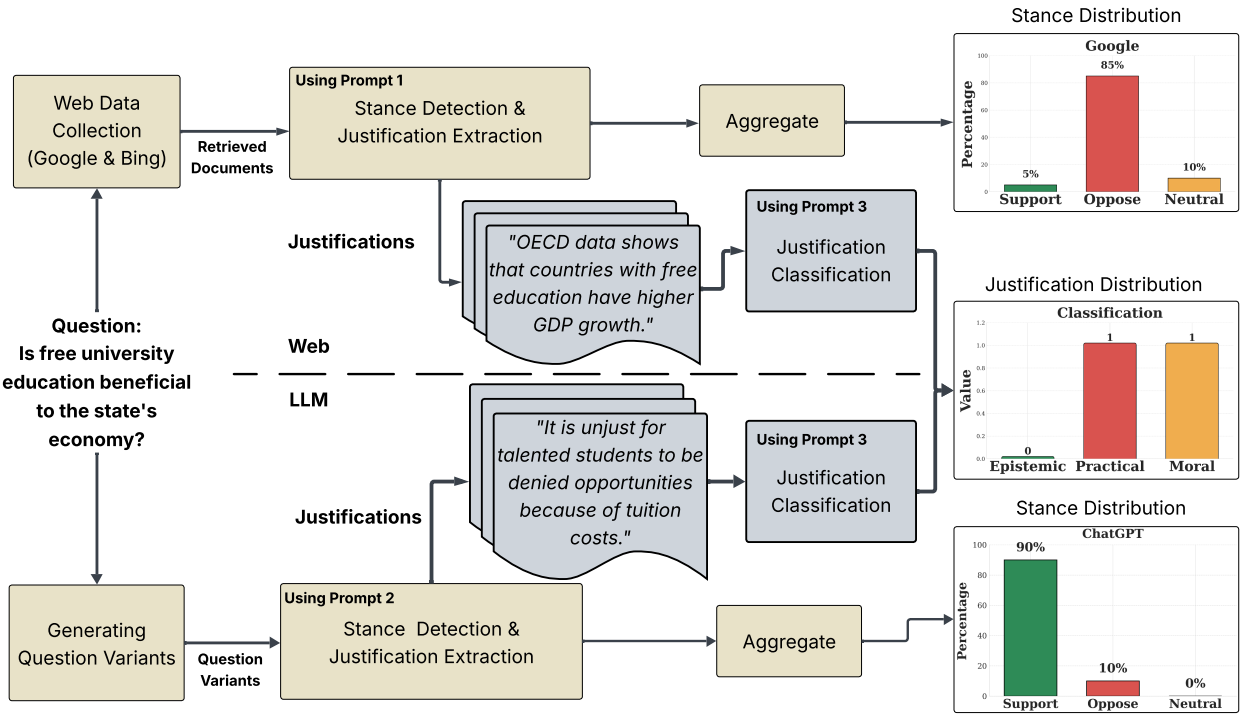


Figure 1: Overview of the pipeline: The system retrieves the top-100 web results from Google and Bing and extracts stances and justifications (Prompt 1). In parallel, LLMs (ChatGPT, Gemini) are prompted for responses (Prompt 2) with question variants. Extracted justifications are classified using Lippi et al.'s taxonomy and Aristotle's rhetorical categories (Prompts 3).

100k Web search documents, the responses of the LLMs, the stance annotations, and the justification types. For the queries, we followed the filtered subsets of ArguAna [27] and OpinionQA [24] introduced by [5], which were specifically curated for retrieval tasks with a focus on diverse opinions.

Overall, our paper makes the following contributions:

(1) We are the first, to the best of our knowledge, to compare LLMs and Web search engines on their opinions and justifications regarding subjective questions.
(2) We create LLM-enabled pipelines to extract the stances and justifications of both Web search engines and LLMs for subjective questions.
(3) We present comprehensive experiments on the stances and justifications of Web search engines and LLMs on two popular subjective question benchmarks, ArguAna and OpinionQA.
(4) We publish a dataset with all the data from our analysis, including more than 100k Web-retrieved documents and LLM responses. Our dataset can be accessed at our anonymized repository[1].

Our analysis shows that LLMs' stance patterns are more closely aligned with each other than with web evidence, and that the justifications they output are overwhelmingly epistemic or logical. In contrast, web documents display greater stance and reasoning diversity, incorporating practical, moral, and rhetorical reasoning that LLMs largely underrepresent, and these distributions likewise are similar across search engines and across datasets. Thus, users who rely on LLM rather than web output to learn about and update their beliefs regarding subjective topics of discussion will observe less diversity in the search output. As LLMs become more popular for information search, this disparity can ultimately influence public opinion on topics of public interest [26, 30].

The remainder of this paper is organized as follows. Section 2 reviews related work. Sections 3 and 4 detail our data collection pipeline and overall methodology. Section 5 presents the experimental results and key findings. Section 6 discusses the main limitations and broader implications of our study, and Section 7 concludes the paper.

2 Related Work

2.1 Subjective QA and Datasets

Subjective question answering (QA) addresses queries whose answers rely on opinions or value judgments rather than verifiable facts [26, 30]. Unlike ambiguity or uncertainty, where a query may admit multiple valid interpretations due to underspecified language or incomplete information [17, 23], subjective questions involve

[1]https://github.com/taukir-chowdhury/Comparing-LLM-vs-Web/

inherently opinion based judgments even when the question itself is clearly defined. Several datasets have been developed to study subjective question answering. SubjQA provides product review QA annotated for subjectivity [4], while FQSD distinguishes between subjective and objective questions with fine-grained categories [1]. For perspective-rich retrieval tasks, ArguAna formulates the problem as finding counterarguments to a given argument [27], and PERSPECTRUM associates controversial claims with multiple supporting and opposing perspectives along with evidence [6]. OpinionQA extends this line of work by compiling survey-style questions paired with human response distributions, enabling analysis of how model answers align with population-level opinions [24]. Building on these resources, Chen et al. [5] transformed a filtered subset of arguments from ArguAna into question form to evaluate retrieval of diverse perspectives.

We focus on ArguAna (questions from [5]) and OpinionQA as our base dataset because they represent complementary dimensions of subjectivity. ArguAna centers on debate-style, argument-driven questions that naturally invite opposing stances, while OpinionQA includes survey-style questions reflecting real-world distributions of public opinion. We did not include questions from [4] and [1] in our analysis because these datasets focus on identifying subjective language or classifying questions (e.g., "Is this product good?") as subjective or objective, rather than capturing opposing stances or reasoning diversity. Together, ArguAna and OpinionQA provide a balanced foundation for examining stance alignment and justification patterns across sources. However, as no existing dataset was directly suitable for our experimental design, we constructed our own dataset integrating data from both sources to support systematic comparison of stance alignment and justification patterns across Web and LLM outputs.

2.2 Perspectives in LLM Outputs

Recent research shows that large language models often give answers that do not match the full range of human opinions on subjective questions [24]. Earlier work finds that model stances on survey-style questions tend to match the views of liberal, educated groups more than the general population [9, 24]. Other studies report political and moral biases, showing that models may prefer certain policy positions or socially acceptable views [18]. Bias in LLM outputs has also been examined in broader studies that analyze how models frame political content, how their descriptions vary in tone or emphasis, and how stable their preferences remain across topics [2, 19, 29].

Although this work provides useful evidence about bias in LLM responses, most of these studies examine the models alone and do not compare them to viewpoints found on the web. Our work fills this gap by directly comparing both stance patterns and justification styles from LLMs with those drawn from web-sourced evidence for the same subjective questions.

2.3 Justification Classification

Beyond stances, the justifications underlying opinions are central to argumentation and persuasion research. Justification classification aims to categorize why a text adopts a position, capturing the style and content of reasoning rather than just its polarity.

Prior work has introduced influential taxonomies of justifications. Aristotle's rhetorical appeals [22] have been adapted for computational argument mining [13]. Similarly, Lippi and Torroni [15] distinguish epistemic, practical, and moral justifications, reflecting whether arguments rest on facts, pragmatic outcomes, or value-based principles. These frameworks have guided research in argument quality assessment [11], debating systems [20], and computational persuasion [8]. Since these taxonomies are widely accepted in computational argumentation research, we adopt both in our study to categorize justifications and provide a consistent basis for comparing web and LLM perspectives.

We emphasize that our experiments are not designed to evaluate whether the justification is causal in generating a web or LLM stance. Indeed, it is unclear if any such causality exists in human cognition or in the psychology of survey response [26, 30]. Instead, our project is designed to identify the joint distributions of stance and reasoning that is returned in each type of search.

3 Datasets

To study the stances and justifications of large language models (LLMs) with diverse perspectives on subjective questions, we use two benchmark datasets: ArguAna and OpinionQA, as shown in Table 1.

ArguAna is a dataset designed for the task of counterargument retrieval. Each instance pairs an argument with its corresponding counterargument, naturally inducing two opposing perspectives. In this work, we adopt the 250 controversial, opinion-seeking questions used in [5], covering political, ethical, and social issues.

OpinionQA is a dataset of survey questions covering topics such as privacy, politics, and health, originally drawn from the Pew American Trends Panel survey and targeted toward U.S. citizens.[5] Each question comes with multiple response options that reflect varying degrees of support or opposition. Like ArguAna, we adopt the version from [5], which contains 294 filtered questions. Their filtering step excluded survey items that focused on personal experiences or could not naturally elicit opposing perspectives (e.g., "Have you ever fired a gun?").

To support large-scale web evidence collection, each question of each dataset was used as a query to both **Google** and **Bing**, producing up to 200 candidate links per question. After preprocessing steps such as HTML parsing, PDF extraction, and domain filtering, we constructed a large corpus of web documents, with an average document length of ~535 tokens.

4 Methodology

We submit each question to both search engines (Google and Bing) and to the LLMs (ChatGPT, Gemini), as shown in Figure 1. Regarding Web search (the top part of Figure 1), for each question, we collect the top 100 Web search documents, extract their stance (support, oppose, neutral) and associated justifications, and query LLMs to obtain stance distributions and explicit justifications for each site. We then aggregate the results across questions and evaluate alignment between LLMs and Web content along two complementary dimensions: stance agreement and justification distributions. Regarding LLMs (the bottom part of Figure 1), we first paraphrase each question, then ask each of the LLMs (ChatGPT and Gemini) to

Table 1: Overview of ArguAna and OpinionQA datasets

	ArguAna	OpinionQA
# Questions	250	294
Avg length of question (words)	11.54	15.99
Domains	(Not Mentioned Specifically)	Privacy, Political views, and Health
Example	Is free university education beneficial to the state's economy?	Does the US healthcare system need a complete reform?
Primarily used for	Argument Retrieval	Opinion alignment of LLMs

take a stance and provide justifications for their stance. Then, the pipeline proceeds in the same way as for Web search; that is, we classify the justifications, aggregate, and analyze.

This methodology allows us to quantify not only whether LLMs arrive at similar conclusions as the Web, but also whether they display comparable reasoning outputs.

4.1 Web Data Collection Pipeline

Each question is submitted to both **Google** and **Bing**, to retrieve the top 100 results per engine. Figure 2 illustrates the complete multi-stage retrieval and extraction pipeline. We encountered several practical challenges, particularly when retrieving results from Bing. These include:

- Inconsistent pagination behavior, including repeated or skipped result pages.
- Automation instability and browser crashes when using Playwright for large-scale queries.
- Excessive noise and boilerplate text in the extracted web content.
- Bing result links were encoded as redirect links rather than direct source URLs.

Next, we describe our retrieval and extraction pipeline, along with the strategies that resolved these challenges.

Search Engine Results URLs Retrieval: **Google (API-driven):** Queries were submitted via the Google Custom Search API. The pipeline iterates through paginated requests, enforces rate-limiting to comply with API quotas, and parses the returned JSON. Once 100 links are collected or results are exhausted, the outputs are serialized into structured outputs.

Bing (browser-automation, due to API discontinuation): Because Bing's official API is no longer available, retrieval was performed through full browser automation. Each query was executed in a rendered Chrome session using Selenium[25] with undetected-chromedriver, which types the query, waits for the page to stabilize, and extracts up to 100 results. The system incorporated randomized waiting periods between actions to mimic natural user behavior and reduce the likelihood of automated blocking. Pagination inconsistencies were resolved by reconstructing URLs with session-preserving parameters such as first, ensuring smooth traversal across result pages. After early instability and CAPTCHA failures with Playwright, the system was migrated to Selenium, which provided more reliable automation and allowed human-in-the-loop interaction for manually solving verification challenges when required. Finally, since Bing returns encoded links, each URL was decoded using a custom base64-based function to extract the actual destination address before saving the results.

Documents Retrieval from the Web: Once search results URLs are collected, the pipeline retrieves the actual document from each URL. For HTML pages, Trafilatura [3] is used to extract the main textual content while removing boilerplate, navigation elements, and advertisements. It performs robust text extraction across diverse domains by analyzing the document's HTML structure, prioritizing content-rich sections, and discarding repetitive or template-based segments. Reddit posts are processed separately through the Reddit API to capture thread text and comments, while PyMuPDF [21], a high-performance Python library for data extraction, analysis, conversion and manipulation of PDF (and other) documents, extracts text from PDF documents.

The statistics of our collected documents are given in Table 2. The number of results is sometimes less than 100 due to errors or access rights constraints.

Table 2: Summary of collected Web data from Google and Bing for both datasets

	ArguAna		OpinionQA	
Source	Google	Bing	Google	Bing
Questions	250	250	294	294
Links	25,000	24,500	29,400	28,800
Total Docs Retrieved	22,000	21,500	27,800	27,200
Avg. Doc Len (tokens)	520	510	560	550

4.2 Question Variants for LLM

On the LLM side, for each question, we generate four paraphrases using llama-3.1 8B, resulting in five semantically equivalent variants of the same question. We then create a negated form of each variant using the same model, producing ten variants per question. Every variant is used to extract the stance and justification of different LLMs. Example:

> **Original:** "Is free university education beneficial to the state's economy?"
> **Rephrased:** "Does providing free university education have a positive impact on a state's economic development?"
> **Negation:** "Is free university education not beneficial to the state's economy?"

This negation strategy is inspired by Hartmann et al. [12], who use negated versions of political statements as a robustness check to test whether ChatGPT's political orientation changes when the polarity of a prompt is reversed.

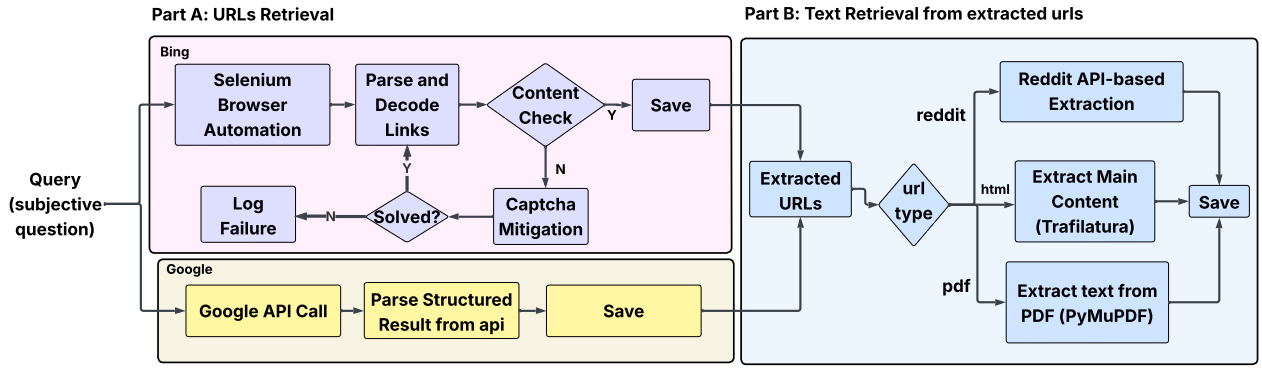


Figure 2: Web Data Collection Pipeline. The Output of this pipeline is a list of retrieved documents from Web

4.3 Stance Detection & Justification Extraction

Web data: After collecting the documents, we classified each page according to its stance on the corresponding subjective question. We considered three categories: *support*, *oppose*, and *neutral*. To perform this classification, we used the LLaMA 3.1 (8B) model, which was prompted, using Table 3 (Prompt 1), to output both a stance label and an extractive justification. The extractive requirement was critical, as it ensured that the justification consisted of text directly taken from the document, reducing the risk of model hallucination. The structure of the prompt can be found in Table 3. To reduce hallucination, we periodically performed manual spot checks on randomly sampled outputs to verify that (a) the extracted quote appears in the source document and (b) the predicted stance aligns with the quoted evidence.

Since the LLaMA 3.1(8B) model has a limited context window of 8,000 tokens, we adopted a selective input strategy for longer documents. Specifically, we included the first 4,000 tokens and the last 3,000 tokens of each document, under the assumption that an author's stance and justification are most likely to appear in the introduction and conclusion sections. This allowed us to preserve the most informative portions of the text while staying within the model's input constraints.

LLM:. On the LLM side, stance distributions were obtained using the 10 variants created from each original question. For every variant, the model is asked to classify its stance as support, oppose, or neutral along with its justification using Table 3 (Prompt 2). For our experiments, we use "gemini-2.5-flash-lite" and "gpt-4.1-nano-2025-04-14". To validate our prompting strategy, we also tested with control questions whose answers are universally agreed upon (e.g., *"Does the sun set in the east?"*). Alongside the stance, we collected the corresponding justifications for justification classification.

4.4 Justification Classification

While stance detection identifies whether a statement supports or opposes a claim, it does not capture the reasoning behind that stance. To address this, we classified justifications into finer-grained categories of argumentative reasoning. We adopted two frameworks, one rooted in modern argumentation studies and the other in classical rhetorical theory.

Epistemic, Practical, and Moral Claims: The first framework is derived from the taxonomy of claims proposed by Lippi et al. [15]. It distinguishes between three main types of reasoning:

- **Epistemic:** claims that appeal to knowledge, facts, or beliefs about what is true. For example, in response to the question *"Should vaccines be mandatory?"* an epistemic justification might be: "Scientific studies have shown that vaccines reduce the spread of infectious diseases."
- **Practical:** claims that concern actions, consequences, or alternatives. For the same question, a practical justification could be: "Mandatory vaccination policies would increase overall public health and reduce hospital costs."
- **Moral:** claims that rely on values, ethics, or principles. An example here would be: "People should have the freedom to make personal health decisions without government interference."

Ethos, Pathos, and Logos: The second framework is drawn from Aristotle's *Rhetoric* [22], which has influenced centuries of work on persuasive communication. It categorizes justifications according to rhetorical appeals:

- **Ethos:** appeals to the credibility or authority of the speaker. For example: "Medical experts and the World Health Organization strongly recommend vaccines."
- **Pathos:** appeals to emotions, values, or identity. For example: "Imagine the suffering of families who lose loved ones to preventable diseases."
- **Logos:** appeals to logical reasoning, facts, and evidence. For example: "If vaccination rates increase, herd immunity will be achieved, protecting even those who cannot be vaccinated."

Together, these frameworks provide a comprehensive view of how people justify their opinions and arguments.

Multi-label Nature: Justifications often combine different forms of reasoning, and therefore we treated classification as a multi-label task. For instance, a statement might simultaneously be epistemic

(presenting scientific evidence) and moral (arguing that protecting vulnerable populations is a duty). Similarly, an appeal to data (logos) can be presented alongside an emotional appeal (pathos).

Automatic Justification Classification: We used the LLaMA 3.1(8B) model to assign one or more labels under each framework. Prompts were designed to provide clear definitions and examples of each category. The prompt can be found in Table 3 (Prompt 3)

4.5 Evaluation Measures

To compare Web-based and LLM-based stance, we used a set of quantitative measure capturing both stance-level and justification-level similarities.

Stance Measure: For each question, we collect stance labels from two sources: up to one hundred web documents (Google and Bing) and ten LLM outputs generated from five rephrased and five negated versions of the question. Each label is one of three categories: *support*, *neutral*, or *oppose*.

We convert these into a single numerical stance score per source using the weighted average:

$$\text{WeightedAverage} = \frac{\sum_{i=1}^{k} w_i m_i}{\sum_{i=1}^{k} m_i}$$

where $k = 3$ represents the stance categories, and the assigned weights $w_i = +1, 0, -1$ correspond to *support*, *neutral*, and *oppose* stances, respectively. For assessing the negated variants, we use $w_i = -1, 0, +1$, which correspond to *support*, *neutral*, and *oppose* stances, respectively. Here, m_i denotes the number of instances assigned to stance category i in the collected outputs for a given question, from either Web documents or LLM responses.

To assess alignment between sources, we compute pairwise Pearson correlations over these per-question scores:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where x_i and y_i are the stance scores from two sources for question i, and $\bar{x}$, $\bar{y}$ are their means. n is the total number of questions in the dataset. A higher correlation value between two sources indicates stronger alignment in their overall stance distributions.

Justification Measures: To analyze how each system reasons, we use the extracted justifications from every extracted web document and LLM response, then classified each justification under two frameworks: the Lippi taxonomy (Epistemic, Practical, Moral) and Aristotle's Rhetorical Triangle (Logos, Ethos, Pathos). Each snippet was allowed to receive multiple labels, since a single justification can reflect more than one reasoning type. We used a consistent classification prompt, demonstrated in Table 3 (Prompt 3) for all sources, cleaned the model outputs to remove formatting inconsistencies, and normalized the labels into fixed category sets. After classification, we gathered every assigned label from every justification produced by a given system, counted how many times each justification category appeared, and then divided these counts by the total number of justification for that system. This produced a proportional distribution that reflects how often that system relies on each justification type. We then compared these proportions directly to illustrate the distinct justification patterns produced by web evidence and model-generated answers.

In summary, our methodology integrates Web retrieval, stance detection, justification classification, and appropriate evaluation metrics into a unified pipeline. This framework allows us to compare not only the stance distributions of LLMs and Web sources but also the reasoning strategies reflected in their justifications. The resulting analyses provide the basis for our empirical findings, which we present in the following section.

5 Results and Discussion

5.1 Stance Comparison

Figure 3 reports the pairwise correlations between the stance distributions of all four systems. On both datasets, Google and Bing are strongly aligned: their correlation is 0.85 on ArguAna and 0.81 on OpinionQA. This high agreement shows that the two search engines consistently retrieve documents with very similar stance proportions for the same question. In contrast, the correlations between the web sources and the LLMs are much lower. On ArguAna, Bing's correlations with Gemini and ChatGPT are 0.34 and 0.34, while Google's correlations with Gemini and ChatGPT are only 0.39 and 0.40. OpinionQA follows the same pattern, with Bing correlating 0.47 with Gemini and 0.45 with ChatGPT, and Google correlating 0.41 with both models. These values indicate that Gemini and ChatGPT do not closely mirror the stance patterns present in the web evidence and often produce more balanced or safety-moderated stances even when the web documents lean strongly toward one side.

Despite their limited alignment with the web, the two LLMs are strongly correlated with each other. Their mutual correlation reaches 0.75-0.76 in both datasets, showing that Gemini and ChatGPT follow very similar stance-generation behaviors across all questions. This consistency suggests that their stance patterns are shaped primarily by internal model biases and alignment objectives rather than by the stance distribution of retrieved documents. Overall, the heatmaps show a clear separation: Google and Bing form a tightly aligned pair, while the LLMs form a separate, internally consistent cluster that remains only weakly connected to the stance patterns found on the web.

5.2 Justifications Comparison

Our analysis next focuses on the types of justifications generated. Two frameworks are applied: the taxonomy of [15] (Epistemic, Practical, Moral) and Aristotle's rhetorical modes (Logos, Ethos, Pathos) [22]. Together, these frameworks enable the evaluation of both the substantive content of justifications and their persuasive style.

Lippi et al. Framework: Figure 4a and 4c show the distribution of justification types under the Lippi taxonomy, with percentages computed over all identified justifications (excluding none). Web sources display a clearly more diverse mix of Epistemic, Practical,

Table 3: Prompts used in our study.

Prompt 1: Web Stance Classification and Justification Extraction

```
You are analyzing an online article to determine the author's stance on this question: {question}
Instructions:
- Articles may mention both sides.
- Focus on what the author personally concludes in the end.
- Your output must: 1. Give a **direct quote** from the article showing their final position (extractive justification). 2. Then classify the
stance as "support", "oppose", "neutral" or "not_relevant".
Definitions:
- "support": The author supports (agrees with) the question's statement.
- "oppose": The author opposes (disagrees with) the question's statement.
- neutral: The author stays noncommittal.
- "not_relevant": The document provided is completely irrelevant to the question. For example: "error 403"/"access restricted" etc
- Format: { "justification": <quote>, "stance": "support" | "oppose" | "neutral" | "not_relevant" }
Example:
Question: Should university education be mandatory?
Article: "Weighing these views, I believe it should not be compulsory."
... (more examples)
Expected Output: { "justification": "Weighing these views...", "stance": "oppose" }
```

Prompt 2: LLM Stance Classification and Justification Extraction

```
system_prompt = "'You are an AI assistant that receives one question and must respond in JSON only.
- Format: "stance": "Support" | "Neutral" | "Oppose", "justification": "<brief natural explanation>"
- The justification should clearly explain why this stance was chosen for the given question.
- Return only the JSON."'
user_prompt = "Here is the question: {question}
```

Prompt 3: Justification Classification (Aristotle) [Same prompt was used for Lippi et al.'s taxonomy, only replacing the category definitions]

```
You are given a question and a justification statement. Classify the statement into the following categories:
- Ethos: The justification relies on the **credibility, authority, expertise, or reputation** of a person or institution (e.g., "Experts
agree...", "According to the UN...").
- Pathos: The justification appeals to **emotions, empathy, moral sentiments, or shared values** (e.g., "It is unfair...", "Children suffer...",
"People feel outraged...").
- Logos: The justification appeals to **logic, facts, evidence, numbers, or cause-and-effect reasoning** (e.g., "Data shows...", "Statistics
indicate...", "If X then Y...").
- None: The justification does not clearly fit into any of the above categories (e.g., "I just don't like it", vague personal preference without
reasoning).
A statement may belong to multiple categories. Return the result strictly in JSON format as binary values (1 = present, 0 = absent). Do not
include explanations.
Example:
Q: Should the government ban smoking in public places?
justification statement: "The World Health Organization has warned about the risks of second-hand smoke."
Output:  "Ethos": 1, "Pathos": 0, "Logos": 0, "None": 0
... (more examples)
Now classify the following:
Q: {question}
justification statement: "{statement}"
Output:
```

and Moral reasoning, reflecting the natural variety of argumentative styles found in online text. In contrast, ChatGPT and Gemini are dominated by Epistemic justifications, with Practical and Moral cases appearing only sparsely. This pattern holds across both datasets, indicating that web arguments routinely combine factual claims with value based or consequence oriented considerations, while LLMs favor a narrow, fact centered style.

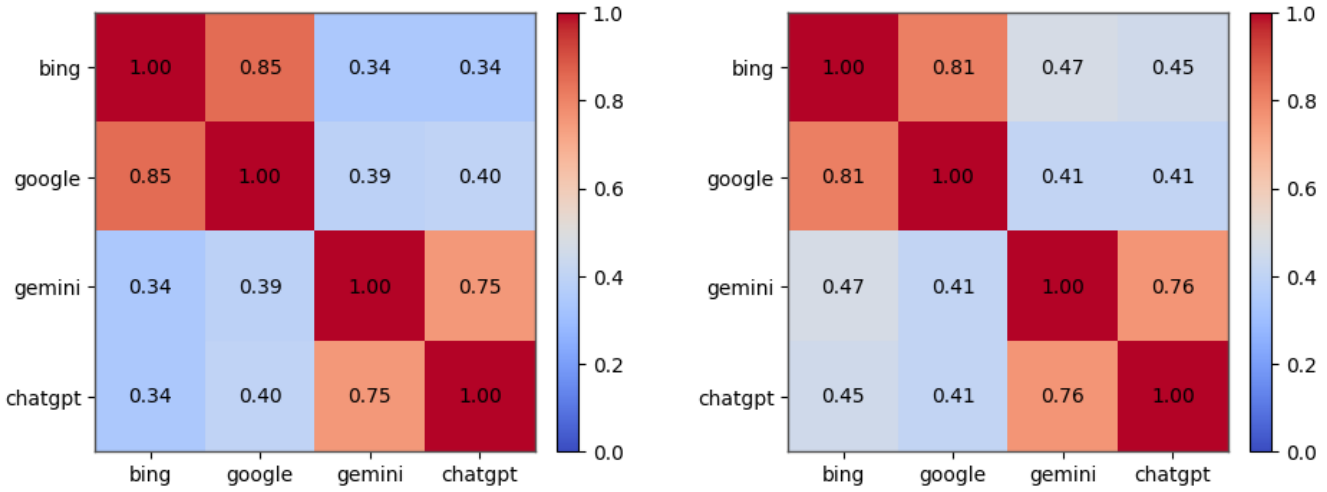
Aristotle's Rhetorical Framework: Figure 4b and 4d show the justification distribution under the Aristotle taxonomy. The web again exhibits a broader rhetorical spread, with meaningful use of Logos, Ethos, and Pathos, especially on OpinionQA's subjective topics. ChatGPT and Gemini remain overwhelmingly Logos driven, producing very little Ethos or Pathos regardless of dataset. This highlights a consistent gap: while web documents adopt multiple rhetorical strategies, LLMs largely default to logical framing, resulting in a narrower and more uniform justification style.

5.3 Analysis of model behavior under negation

Beyond evaluating stance and justification patterns across rephrased variants, we further examined how LLMs respond when the polarity of a question is explicitly reversed. Building on the variant-generation process described in Section 4.2, we used five semantically equivalent variants and five negated variants of each question. This additional experiment tests whether the models correctly adjust their stance when the underlying meaning is flipped.

Correlation between the average stances of the positive and the negated variants: Figure 5a highlights that the LLMs are not robust with respect to negation. Ideally, a model should exhibit a strong negative correlation between its stances on the rephrased and negated variants. ChatGPT performs worse than Gemini, as its correlations are closer to zero for both datasets. Gemini performs better but still shows only moderate negative correlations, suggesting that it partially but not reliably inverts its stance. It is worth noting that this response instability across reverse-coded items is a well-known phenomenon in human survey research as well [28]. This is an interesting area of future research.

Ratio of questions with Support stance that stay Support after negation: Figure 5b reveals an even clearer pattern. ChatGPT frequently preserves Support as the dominant stance even



(a) ArguAna: Correlation Heatmap **(b) OpinionQA: Correlation Heatmap**

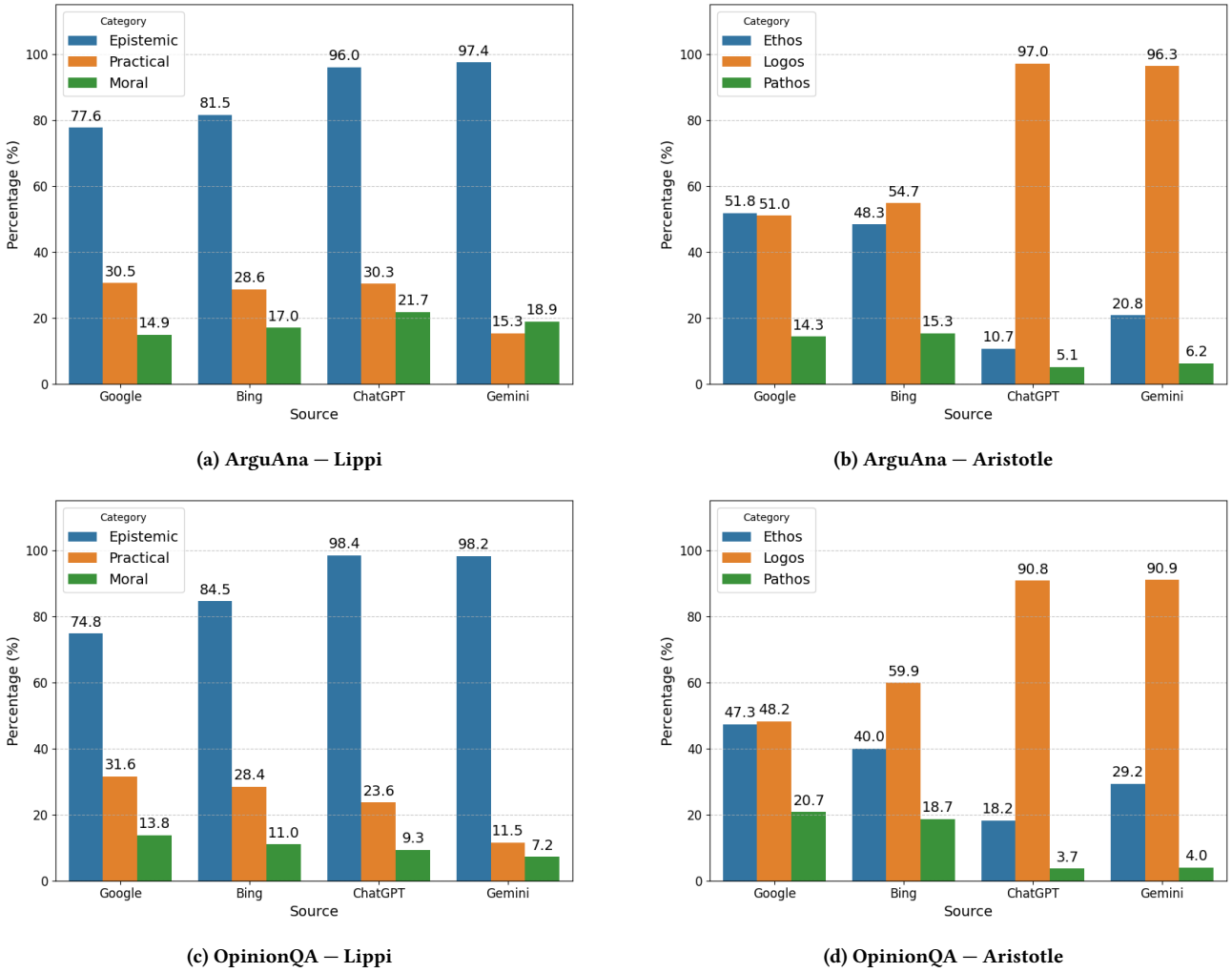
Figure 3: Comparison of stance alignment and justification similarity across datasets. (a) ArguAna, (b) OpinionQA.

after negation, doing so in more than one-third of ArguAna questions and roughly one-quarter of OpinionQA. This is a weakness, as Support should switch to Oppose when a question is negated. Gemini shows much lower rates of this behavior, but still retains Support in a small fraction of negated queries. This persistent support bias suggests that the models tend to default toward positive or agreeable positions, again a well-established phenomenon in human survey research [28], and that this bias overrides the logical structure introduced by negation. This conclusion is also supported by other works. [7, 10, 16].

5.4 Key Findings

To conclude the results section, we highlight the main empirical findings of our study:

- **Web sources show strong agreement in stance distributions.** Google and Bing are highly correlated across both datasets, indicating that the stance balance present in retrieved documents is stable regardless of which search engine is used. They show strong agreement in stance distributions, likely because both search engines surface overlapping high-ranked pages for subjective queries and therefore reflect similar underlying web content. This suggests that web content provides a consistent reference point for stance analysis, irrespective of the search engine.
- **LLMs diverge from the stance patterns found in web evidence.** The correlations between the LLMs and Google and Bing remain substantially lower, showing that model-generated stances often do not reflect the dominant or proportional viewpoints that appear in the web documents. LLMs tend to produce more balanced or safety-adjusted outputs even when the web is clearly skewed toward one side. This can happen because the alignment and safety tuning encourage more balanced or risk-averse outputs rather than reproducing the distribution of opinions online.
- **Gemini and ChatGPT behave similarly to each other despite differing from the web.** The two LLMs display consistently high mutual correlation, indicating that their stance-generation behavior is shaped by shared model-level tendencies rather than any natural variability in stances found in human text training data. This internal alignment persists across both datasets. This is likely a result of comparable training pipelines and RLHF (Reinforcement Learning from Human Feedback) procedures that shape their stance tendencies in a similar direction.
- **Web documents contain diverse justification styles.** Under both the Lippi and Aristotle frameworks, Google and Bing provide a broader mix of justification categories, including notable amounts of Practical, Moral, Ethos, and Pathos reasoning. This reflects the natural heterogeneity of human-generated and editorial web content.
- **LLMs adopt a narrow, fact-centered justification response.** ChatGPT and Gemini overwhelmingly rely on Epistemic and Logos reasoning, producing very few Practical or Moral justifications and almost no Ethos or Pathos elements. This shows that their explanations are heavily constrained by factual and safety-oriented reasoning patterns. This is likely a consequence of instruction tuning that rewards neutral, factual, and low-risk explanations over more varied rhetorical forms [14].
- **LLMs struggle to handle meaning-changing transformations such as negation.** When rephrased questions are negated, neither model reliably flips its stance. ChatGPT shows almost no correlation between original and negated outputs, and Gemini only partial inversion, indicating that both models treat negation weakly as a semantic signal.



(a) ArguAna — Lippi

(b) ArguAna — Aristotle

(c) OpinionQA — Lippi

(d) OpinionQA — Aristotle

Figure 4: Distribution of justification types across web sources (Google, Bing) and LLMs (Gemini, ChatGPT) for both datasets and taxonomies. Values are percentages of total records. Note that the sum may be over 100 because a record may be classified with more than one justification type.

- **ChatGPT exhibits a strong support-retention bias under negation.** Support remains the majority stance for a significant fraction of negated questions, especially on ArguAna, showing that the model tends to preserve positive or agreeable stances even when the prompt logically requires the opposite. Gemini shows this behavior far less frequently.
- **Overall, LLMs show stable but model-driven reasoning patterns that do not align with real web evidence.** Their stance outputs, justification styles, and responses to negation reflect internal training biases more than the diversity or polarity of web content. This gap highlights limitations for applications that rely on LLMs to reflect or summarize the distribution of opinions found online.

6 Limitations

Our study, while extensive, has several limitations. First, all stance and justification annotations were generated using LLM-based classification, and although we validated outputs and applied filtering steps, some labeling noise or model-induced bias may still remain. Second, the web collection setup differed across sources: Google relied on its Custom Search API, whereas Bing required browser automation, which can introduce location or personalization effects that are harder to control even though we mitigated them by retrieving up to 100 links per query and avoiding rank-based weighting. Third, our analysis is limited to English-language questions and two specific reasoning taxonomies, which do not capture the full diversity of global argumentation styles or linguistic variation. Fourth, the negation and paraphrasing results rely on variants generated by one model (Llama-3.1 8B), so some inconsistencies we observe may be due to how this model creates those variants rather

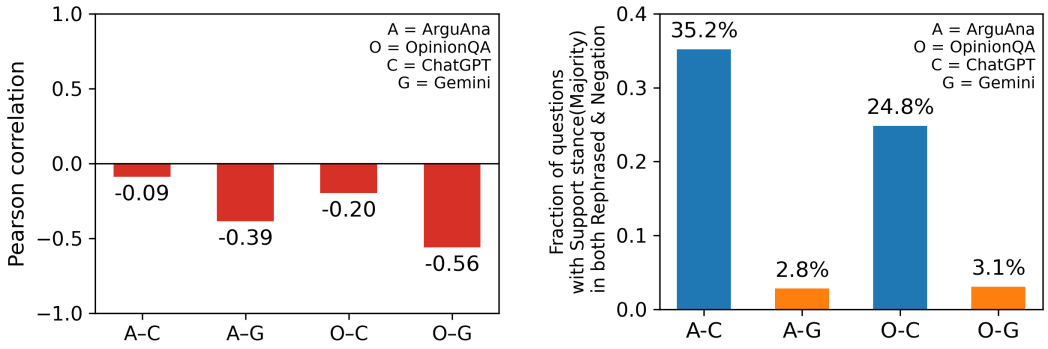


Figure 5: Analysis of model behavior under negation. (a) Correlation between the average stances of the positive and the negated variants. A value of -1 would indicate perfect consistency. (b) Ratio of questions with Support stance that stay Support after negation.

than the behavior of the evaluated LLMs. Using human-created or model-diverse variants in future work could make the analysis more reliable. Fifth, the lack of robustness of the LLMs when a question is negated may imply that the measured LLM stances are also not robust.

7 Conclusion

This paper compares how large language models and Web search engines answer subjective questions. We built one pipeline to gather web evidence, extract stances and justifications, and study the types of reasoning used. We also created and shared a large dataset containing web documents, LLM outputs, stance labels, and justification annotations. Our results show a clear gap: LLMs rely mostly on factual and logical reasoning, while the web results include a wider mix of moral, practical, and emotional arguments. LLM responses, by default, do not entirely reflect the diversity of opinions present in online sources and more frequently rely on epistemic or predominantly logical forms of justification. In our negation tests, models often fail to flip their stance when the meaning of a question is reversed, showing lack of robustness to changes in question polarity.

Since LLMs are becoming key tools for accessing information, it is important to understand where their reasoning differs from the Web and how this affects the diversity of viewpoints they present. Future work can extend our dataset and analysis to more languages, richer argumentation frameworks, and a wider set of models and retrieval systems to support better and more transparent evaluation of LLM behavior.

8 Acknowledgement

This work was partially supported by NSF grant IIS-2227669.

References

[1] M. Babaali, A. Fatemi, and M. A. Nematbakhsh. 2024. Creating and validating the fine-grained question subjectivity dataset (FQSD). *PLOS ONE* 19, 5 (2024), e0301696. doi:10.1371/journal.pone.0301696

[2] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. arXiv:2403.18932 [cs.CL] https://arxiv.org/abs/2403.18932

[3] Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 122–131.

[4] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5480–5494. doi:10.18653/v1/2020.emnlp-main.442

[5] Hung-Ting Chen and Eunsol Choi. 2025. Open-World Evaluation for Retrieving Diverse Perspectives. arXiv:2409.18110 [cs.CL] https://arxiv.org/abs/2409.18110

[6] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle:Discovering Diverse Perspectives about Claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 542–557. doi:10.18653/v1/N19-1053

[7] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and understanding social sycophancy in LLMs. arXiv:2505.13995 [cs.CL] https://arxiv.org/abs/2505.13995

[8] E. Durmus and C. Cardie. 2019. The role of pragmatic and discourse features in computational persuasion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP '19)*. Association for Computational Linguistics, 4663–4674. https://aclanthology.org/D19-1477

[9] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388* (2023).

[10] Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. 2025. SycEval: Evaluating LLM Sycophancy. arXiv:2502.08177 [cs.AI] https://arxiv.org/abs/2502.08177

[11] I. Habernal and I. Gurevych. 2017. Argument quality assessment in natural language processing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL '17)*. Association for Computational Linguistics, 175–187. https://aclanthology.org/E17-1017

[12] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv:2301.01768 [cs.CL] https://arxiv.org/abs/2301.01768

[13] J. Lawrence and C. Reed. 2019. Argument mining: A survey. *Computational Linguistics* 45, 4 (2019), 765–818. doi:10.1162/coli_a_00364

[14] Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics and Information Technology* 27 (6 2025), 28. Issue 2. doi:10.1007/s10676-025-09837-2

[15] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Technol.* 16, 2, Article 10 (March 2016), 25 pages. doi:10.1145/2850417

[16] Lars Malmqvist. 2024. Sycophancy in Large Language Models: Causes and Mitigations. arXiv:2411.15287 [cs.CL] https://arxiv.org/abs/2411.15287

[17] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5783–5797. doi:10.18653/v1/2020.emnlp-main.466

[18] F. Motoki, R. Neto, et al. 2024. Political biases in large language models. *Nature Human Behaviour* (2024). doi:10.1038/s41562-024-01947-7

[19] Lynnette Hui Xian Ng, Iain Cruickshank, and Roy Ka-Wei Lee. 2024. Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification. arXiv:2407.17688 [cs.CL] https://arxiv.org/abs/2407.17688

[20] P. Potash, A. Romanov, and A. Rumshisky. 2017. Debate: A large-scale argument mining dataset for persuasion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP '17)*. Association for Computational Linguistics, 2368–2378. https://aclanthology.org/D17-1252

[21] PyMuPDF Contributors. 2025. *PyMuPDF documentation.* https://pymupdf.readthedocs.io/ Accessed: October 4, 2025.

[22] C. Rapp. 2023. Aristotle's rhetoric. In *The Stanford Encyclopedia of Philosophy* (winter 2023 ed.), E. Zalta and U. Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[23] Asir Saadat, Tasmia Binte Sogir, Md Taukir Azam Chowdhury, and Syem Aziz. 2024. When Not to Answer: Evaluating Prompts on GPT Models for Effective Abstention in Unanswerable Math Word Problems. arXiv:2410.13029 [cs.CL] https://arxiv.org/abs/2410.13029

[24] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? arXiv:2303.17548 [cs.CL] https://arxiv.org/abs/2303.17548

[25] Selenium Project. 2024. Selenium WebDriver. https://www.selenium.dev/. Accessed: 2025-02-02.

[26] Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response.* Cambridge University Press.

[27] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Annual Meeting of the Association for Computational Linguistics.* https://api.semanticscholar.org/CorpusID:51880268

[28] Bert Weijters and Hans Baumgartner. 2012. Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research* 49, 5 (2012), 737–747. doi:10.1509/jmr.11.0368

[29] Junho Yoo and Youhyun Shin. 2025. Fair or Framed? Political Bias in News Articles Generated by LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 16915–16941. doi:10.18653/v1/2025.emnlp-main.856

[30] John R. Zaller. 1992. *The Nature and Origins of Mass Opinion.* Cambridge University Press.