SocraPedia: Enhancing Wikipedia's Content Quality through Collaborative Large Language Models

Anonymous ACL submission

Abstract

This paper commences by examining the quality challenges present in Wikipedia through a statistical lens. Our analysis reveals that a predominant share, exceeding 90%, of its pages are deemed low in quality by Wikipedia's editorial community, with merely 1.2% achieving the highest quality designation. We scrutinize the rating criteria and pinpoint approximately 5% of pages that are significant yet underdeveloped—prime candidates for augmentation through LLMs. Our proposed implementation strategy involves crafting algorithmic techniques to assess various quality dimensions, subsequently utilizing SocraSynth to facilitate enhancements in mitigating biases and highlighting nonfactual claims. The efficacy of this approach is validated through trials on a select group of pages characterized by their high importance yet currently low quality.

1 Introduction

The advent of large language models (LLMs) such as GPT-4 (OpenAI, 2023) has ushered in a transformative era in natural language processing, revolutionizing fields like machine translation, question-answering, and text summarization. With their profound capabilities in understanding and generating human language, these models, particularly GPT-4, have not only achieved remarkable performance on benchmarks like the MMLU (Papers with Code Corp., 2023) but have also begun to alter user behavior in information retrieval. Recent trends observed at the Wikimania conference (Wikimania, 2023), a hub for Wikipedia enthusiasts, indicate a growing preference for ChatGPT over Wikipedia, reminiscent of the paradigm shift in 2005 when Wikipedia surpassed the Encyclopedia Britannica in popularity.

This shift raises pivotal questions about Wikipedia's role in an AI-dominated digital landscape. Far from rendering Wikipedia obsolete, the integration of its content into LLMs' training datasets offers a unique opportunity to enhance Wikipedia's utility and relevance. Our study aims to explore how LLMs can be leveraged to improve accuracy and reduce biases in Wikipedia articles. By establishing a symbiotic relationship between LLMs and Wikipedia, we envision a future where LLMs not only benefit from Wikipedia's extensive knowledge base but also contribute to its continual improvement. Such a partnership promises to amplify Wikipedia's value and credibility, ensuring its vital role as an information resource in the era of advanced artificial intelligence.

We propose the utilization of SocraSynth (Chang, 2023c) as a novel approach to enhance content accuracy and mitigate biases on Wikipedia. The core concept of SocraSynth involves using *conditional statistics* (Pearl, 2000; Hastie et al., 2009) to position two (or more) LLM-based agents with opposing views on a subject matter. These agents engage in a multi-round debate to defend their respective stances. This unique dialogue setting ensures two critical outcomes: first, it helps break away from the inherent biases present in individual LLMs; second, it progressively enriches the context of the dialogue with each unfolding round. Consequently, in collaboration with Wikipedia and LLMs, we can significantly enhance information quality in the following ways:

1. *Content Accuracy Cross-Checking:* Engaging multiple LLMs in a mediated dialogue enables cross-verification and enhancement of Wikipedia's content quality. This involves comparing responses and perspectives generated by different LLMs through *conditional statistics*, leading to a more comprehensive understanding of topics.

2. *Bias Mitigation:* SocraSynth's *contentious debate* setting is adept at identifying and reducing

biases within Wikipedia articles. Exposing content to a spectrum of viewpoints and arguments allows the platform to pinpoint areas of bias, encouraging revisions for more neutral and representative content.

3. *Dynamic Content Adjustment:* The adjustable *contentiousness* parameter enables dynamic modification of the tone, language, and balance of Wikipedia articles. This process, involving thorough argumentation and counter-argumentation, ensures the content is not only accurate but also presents a balanced view of the subject matter.

The rest of this paper is organized into four main sections. Section 2 delves into Wikipedia's challenges concerning content quality and biases. Section 3 details how we leverage *conditional statistics* to address these challenges. Section 4 presents our case studies, showcasing practical applications and outcomes. Finally, we conclude with our insights and reflections in the last section.

2 Wikipedia Usage and Challenges

In 2020, an analysis of Wikipedia's top-100 most frequented pages showed significant interest in categories such as *People*, *Politics and History*, *Sports*, *Technology and Online Platforms*, *Entertainment and Arts*, and *Research and Education* (Wikipedia, 2023c). Monthly page-view statistics further illustrate Wikipedia's diverse usage spanning *Research and Learning*, *Curiosity and Exploration*, and *Entertainment* (Wikipedia, 2023a). Table 1 presents the top-ten visited pages in November 2023, highlighting the variety and frequency of content updates, such as the daily revisions in sports-related articles compared to the less frequent updates in historical content.

Rank	Page	Edits	Pageviews
1	2023 Cricket World Cup	532	14,148,770
2	Cricket World Cup	199	13,917,748
3	YouTube	29	6,468,789
4	1234	3	4,676,115
5	Cleopatra	4	4,672,320
6	Tiger 3	546	4,525,333
7	Pornhub	13	4,192,495
8	Deaths in 2023	2,188	3,870,414
9	XNXX	33	3,837,659
10	2024 Men's World Cup	447	3,749,162

Table 1: Top 10 Pages Statistics in November 2023

All rated articles by quality and importance						
	Importance					
Quality	Top	High	Mid	Low	???	Total
FA	1,176	1,807	1,709	1,065	188	**5,945**
FL	141	562	661	600	119	**2,083**
A	221	427	578	372	80	**1,678**
GA	2,090	4,769	9,306	10,086	1,697	**27,948**
B	12,053	22,872	34,968	27,849	13,686	**111,428**
C	10,270	29,697	66,348	91,945	43,475	**241,735**
Start	17,220	75,751	305,446	785,343	292,126	**1,475,886**
Stub	4,226	30,945	226,015	1,858,337	840,137	**2,959,660**
List	3,015	11,232	34,179	93,654	61,534	**203,614**
Assessed	50,412	178,062	679,210	2,869,251	1,253,042	**5,029,977**
Unassessed	139	406	1,776	16,255	531,921	**550,497**
Total	**50,551**	**178,468**	**680,986**	**2,885,506**	**1,784,963**	**5,580,474**

Figure 1: Matrix of Wikipedia Quality vs. Importance.

2.1 Quality Assessment

Wikipedia's content quality, as rated by its editors, predominantly falls below the 'C' level, indicating a need for improvement (Wikipedia, 2023b). Figure 1 illustrates the distribution of quality against importance of Wikipedia articles (Wikipedia, 2023d). Articles are classified into six quality levels, with 'FA' (featured article), 'FL' (featured list), 'A', and 'B' denoting higher quality, while lower categories reflect various deficiencies.

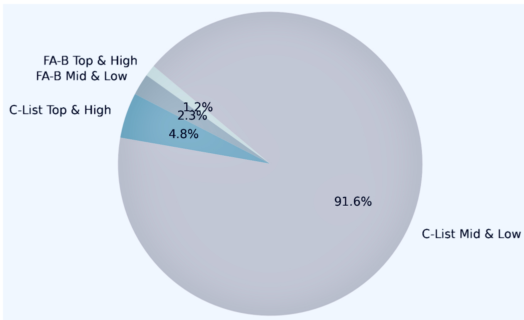

Figure 2: Distribution from Top Quality High Importance (1.2%) to Low Quality Low Importance (91.6%). Notably, the blue segment (4.8%) signifies high-importance pages in need of improvement.

2.2 Importance Assessment

The "importance" rating on Wikipedia, subjectively assessed by its editors, categorizes topics into four tiers: 'Top', 'High', 'Mid', and 'Low'. The 'Top' category encompasses fundamental topics, 'High' pertains to influential subjects, while 'Mid' and 'Low' represent topics of more moderate or lesser significance. Figure 2 displays the distribution of quality versus importance across Wikipedia, highlighting that a mere 1.2% of pages excel in both dimensions. Notably, the segment representing 4.8% of the content (indicated in blue) consists of important topics that suffer from lower quality, underscoring these as key areas for improvement.

2.3 Biases Assessment

Wikipedia's biases across various domains like politics, religion, and gender, are well-documented (e.g., (Hube, 2017)). Greenstein and Zhu's study (Greenstein and Zhu, 2018) highlights a political slant in U.S. politics articles, comparing Wikipedia with Encyclopedia Britannica. This disparity points to the impact of Wikipedia's varied editorship and source selection on content neutrality and representation. Language and historical usage biases are also critical, emphasizing the need for training models to incorporate real-time language trends from diverse sources.

3 The Force of SocraPedia with LLMs

The integration of Large Language Models (LLMs) with Wikipedia through our proposed framework, SocraPedia, brings forward a transformative approach aimed at enhancing content quality and reducing biases. This framework harnesses the capabilities of SocraSynth proposed by (Chang, 2023c), targeting the identification and rectification of inaccuracies and biased presentations in Wikipedia's content. SocraPedia's role extends beyond mere identification; it actively proposes edits and refined texts to Wikipedia's editorial board, enriching content with diverse perspectives.

3.1 Challenges of Supervised Learning in Content Quality Enhancement

Supervised learning, though proficient in various domains, encounters notable challenges when tasked with complex and nuanced objectives like identifying inaccuracies and biases in content. The dynamic, context-dependent nature of these tasks renders them less amenable to conventional supervised learning approaches. Distinguishing inaccuracies and biases demands a sophisticated grasp of context, culture, and language nuances, aspects that are often beyond the capabilities of standard supervised learning models. Crucially, unlike supervised learning, which primarily identifies issues without offering solutions, SocraPedia goes a step further by not only detecting inaccuracies and biases but also proposing revised, more balanced texts.

3.2 "Gold" Standards of Content Quality

When absolute truths and facts are elusive, "reasonableness" can act as a balancer to consider multiple perspectives, which is essential for avoiding absolutism in content interpretation.

In the pursuit of a reliable measure of content quality, we turn to the concept of "reasonableness", guided by Aristotle's first principle. This principle asserts the fundamental logic that a statement cannot be both true and false in the same context. By aligning our assessment with this principle, we aim to evaluate the coherence and consistency of content on Wikipedia. SocraPedia employs the CRIT algorithm (Chang, 2023b), a critical reading tool designed to scrutinize a document's reasoning validity and evidence credibility. (CRIT is discussed further in Section 3.3.)

Evaluating reasonableness involves not only assessing logical consistency but also considering the context, cultural relevance, and the acknowledgment of diverse viewpoints. A statement deemed reasonable should be consistent with empirical evidence, respect different perspectives, and avoid contradictions. In this light, our use of SocraSynth's conditional statistics becomes pivotal, enabling context-dependent reasoning and facilitating a deeper understanding of content quality.

3.3 Conditional Statistics on Context

SocraSynth utilizes *conditional statistics* (Pearl, 2000; Hastie et al., 2009) to strategically place LLM-based agents in dialogues where they represent opposing viewpoints on contentious topics. This arrangement fosters a multifaceted and rich debate, with each agent robustly defending its assigned stance. This methodology is instrumental in breaking away from the inherent biases typically found in LLM training datasets and progressively enriches the context of the dialogue in each subsequent round. Our experiments have demonstrated that when LLMs are conditioned to adopt specific stances, they generate content that aligns with and supports those viewpoints. This approach effectively mitigates the pre-existing biases in the models. Additionally, any new biases introduced by these specific stances are counteracted by the opponent's conditionally biased remarks, which represent the opposing view. As the debate evolves, the prompts become increasingly specific, which results in a reduction of response hallucinations and a notable enhancement in the overall quality and depth of the dialogue.

In essence, SocraPedia, via SocraSynth, paves the way for enhancing the quality of Wikipedia's content. It harnesses LLMs' advanced capabilities for structured, reasoned debates, fostering a balanced and comprehensive knowledge base. The final assessment of content quality, however, benefits from a combination of AI analysis and human expertise. This integrated approach is essential for accurately interpreting nuances, cultural contexts, and subtle complexities that AI alone may miss, ensuring Wikipedia's reliability and authority.

3.4 SocraPedia Algorithms

SocraPedia, integrating Wikipedia's quality framework, develops critical algorithms within the SocraSynth framework. Our emphasis lies on three primary quality metrics: the rigor and substantiation of arguments (emphasizing reasonableness), the credibility of argument sources (citations), and the breadth and depth of content, . The following part of this section outlines the specifications of these algorithms in detail.

Validity and Credibility of Arguments

In our collaborative multi-agent debates, the focus is on the persuasiveness of arguments. The CRIT algorithm (Chang, 2023a,b) is utilized to evaluate the logical soundness of claims made during these discussions. As illustrated in Figure 3, CRIT assesses the final statements from all participants in the debate, assigning a validation score ranging from 1 to 10, where 1 denotes the least credible.

Function Γ = CRIT(d)

Input. d: document; **Output**. Γ: validation score;
Vars. Ω: claim; R & R': reasons & counter reasons;
Subroutines. $Claim()$, $FindDoc()$, $Validate()$;
Begin
#1 Identify in d the claim statement Ω;
#2 Find a set of supporting reasons R to Ω;
#3 For $r \in R$ eval $r \Rightarrow \Omega$
  If $Claim(r)$, (γ_r, θ_r) = CRIT($FindDoc(r)$);
  else, $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$;
#4 Find a set of rival reasons R' to Ω;
#5 For $r' \in R'$, $(\gamma_{r'}, \theta_{r'})$ = V($r' \Rightarrow \Omega$)
  Eval rival arguments;
#6 Compute weighted sum Γ, with $\gamma_r, \theta_r, \gamma_{r'}, \theta_{r'}$.
#7 Analyze the arguments to arrive at the Γ score.
#8 Reflect on and synthesize CRIT in other contexts.
End

Figure 3: CRIT Pseudo-code (Chang, 2023a). (The symbol $\Rightarrow$ denotes both inductive and deductive reasoning.)

Formally, for any given document d, CRIT evaluates and renders a validation score Γ. Let Ω denote the central claim of d, and R be the set of supporting reasons. The causal validation function is defined as (γ_r, θ_r) = V($r \Rightarrow \Omega$), where γ_r is the validation score for reason $r \in R$, and θ_r signifies the credibility of the source. A detailed exposition of this approach, utilizing independent LLM agents as arbiters, is beyond this paper's scope but is available in extended literature (Chang, 2023a,b).

To improve reasoning quality of a prompt, Researchers have incorporated various heuristic based methodologies. Notable among these are the chain-of-thought (Wei et al., 2022), tree-of-thought (Yao et al., 2023), and cumulative reasoning (Zhang et al., 2023) approaches, further enriched by recent advancements (Allaway et al., 2023; Jung et al., 2022; Liu et al., 2023; Sclar et al., 2023). These strategies guide models towards a more analytical reasoning process (McHugh and Way, 2018; Wason and Johnson-Laird, 1972), enhancing response coherence and uniformity. However, the nature of chain-of-thought line of algorithms is abductive logic, which is prone to bias and lacking definitiveness because abductive reasoning does not guarantee the truth of its conclusions (Flach and Kakas, 2000). The best explanation might not necessarily be the correct one, as it depends on the current understanding and available information, which might be incomplete or inaccurate.

In more open, diverse domains where reasoning requirements are complex and extensive, these techniques encounter significant challenges. The sequential nature of reasoning in these methods, especially the chain-of-thought approach, is susceptible to error accumulation. While effective in simpler contexts, the method often struggles in broader analytical settings.

SocraSynth is designed to address these challenges, augmenting human decision-making in both familiar and novel domains. It adopts an informal reasoning style (Gu et al., 2023), diverging from the formalized logic outlined in sources like (Bronkhorst et al., 2020; Huang and Chang, 2023; Qiao et al., 2023). By fostering a debate environment moderated by humans, SocraSynth strengthens the structure and reliability of the reasoning process. It capitalizes on LLMs' proficiency in fundamental NLP tasks such as classification, question-answering, and information retrieval, offering a comprehensive approach to complex reasoning scenarios.

To verify the credibility of sources, the CRIT algorithm utilizes LLMs to generate reference lists, which are then assessed based on the publisher's reputation and the number of citations. This method bears resemblance to Google's PageRank algorithm (Page et al., 1999).

Prompt Template Design for Bias Detection, Explanation, and Mitigation.

SocraPedia utilizes a moderator and two LLM agents, sourced from a variety of models (Bubeck et al., 2023; OpenAI, 2021, 2023; Thoppilan et al., 2022; Touvron et al., 2023), to facilitate structured debates. The moderator initiates the debate by setting a topic and assigning opposing stances to the LLM agents. To enhance the dialogue, sub-topics are delineated, creating a dynamic framework for discussion. The moderator's role is to guide the debate without contributing to content generation, thereby minimizing human-induced biases.

Throughout the debate, the "contentiousness" parameter is adjusted to vary the intensity of the arguments. Initially set high, it fosters the presentation of strong, opposing viewpoints by the LLM agents. As the debate progresses, this parameter is gradually reduced, leading the agents towards constructing a balanced and integrative proposal. The final output is a synthesis of the knowledge and perspectives generated during the debate, refined through human curation for depth and balance.

	Function d' = DeBias(d)

	Input. d: text; **Output**: d' debiased text;
	Vars. B: detected bias set; R & R': reasons & counter reasons; ϕ: debate contentiousness
	// Use two LLMs, LLM1 and LLM2 to debate.
	Subroutine. $Claim()$, $LLM1()$; $LLM2()$;
	Begin

#1	Identify in $Claim(d)$ a set of biases B;
#2	$\phi \leftarrow 0.9$;
#3	$R \leftarrow LLM1(B, \phi)$; // Find arguments R for B;
#4	$R' \leftarrow LLM2(B, \phi)$; Find counterarguments R';

#5	While (($\phi \leftarrow \phi/2) > 0.1$) {
	$R \leftarrow R \cup LLM1(R', \phi)$ // LLM1 refutes LLM2
	$R' \leftarrow R' \cup LLM2(R, \phi)$ // LLM2 refutes LLM1
	}
	// Generate a conciliatory text w/ d, R and R'.

#6	$d' \leftarrow LLM(d, R, R', \phi)$; **End**

Figure 4: Using Controlled Conditional Biases to Mitigate Inherited Biases and Generate Conciliatory Narratives. $d' = DeBias(d)$.

The DeBias algorithm, detailed in Table 4, employs this adversarial debate format to identify and mitigate biases in a document, denoted as d, transforming it into a more representative and unbiased version, d'.

The process begins with identifying potential biases (B) in the document's claim using the $Claim(d)$ function. The LLM agents, representing opposite positions, then examine B through a series of argumentation rounds, enriching the sets of arguments (R) and counterarguments (R'). The contentiousness level is initially high to elicit strong arguments and gradually reduced to encourage consensus. The final step combines these refined arguments to produce the debiased text d', capturing a comprehensive view of the subject.

Content Comprehensiveness

To gauge the breadth and depth of content, SocraPedia uses the top 1.2% of high-quality articles as a benchmark to create a category-specific checklist (Wikipedia, 2023b). This checklist guides the SocraSynth framework in conducting structured debates, ensuring a thorough and balanced representation of perspectives for each criterion. Research indicates that such dialectical processes among LLMs foster a comprehensive search for information and broaden the contextual scope, leading to the generation of more detailed and evenly balanced content (Chang, 2023c; Tsao, 2023; Chang and Chang, 2023).

Indigenous People	Long before Columbus set foot on American soil, indigenous people populated this land. These civilizations, Native Americans, had complex societies, trade networks, and cultural systems in place. Addressing Columbus' arrival as a "discovery" undermines their pre-existing history and existence.
Indigenous Cultures	Labeling Columbus's arrival as a "discovery" implies that the continent was devoid of culture and civilization before his arrival. This narrative is not only a form of Eurocentric bias but also a disrespect to the rich histories, cultural diversity, and knowledge systems of the indigenous peoples.
"Encounter" "Discovery"	This is not just a matter of semantics but represents an important paradigm shift. An "encounter" suggests the meeting of two separate entities, where both have an established existence in their own right. While Agent2 viewed "discovery" as an innocent terminology describing the European perspective, I argued that this language undermines the existing civilizations and cultures, championing a biased and prejudiced narrative.
Examining the Duality	While the narrative of Columbus discovering America is widely taught, it's crucial to also recognize the significant exploitation, displacement, and massacres that followed, deeply impacting indigenous populations.
Summary	This leaves us with an unresolved issue, a universal question that still pervades our understanding of history: How can we rectify the Eurocentric bias that pervades historical narratives without undermining the recognized historical figures and events that have significantly shaped global history? A question like this warrants continuous respectful conversation and exploration.

Table 2: Agent-1's arguments for "Columbus' discovery of America" is a Eurocentric bias.

Historical Context	At the time of Columbus' voyage, Europeans were unaware of the existence of the Americas. Thus, Columbus' journey was indeed a discovery for the European civilization, and this context cannot be denied.
Terminology	My key arguments prioritize the understanding that terminology such as "discovery" is used universally to denote Columbus's voyage. I believe this term denotes the newness of the Americas to the Old World and does not undermine indigenous cultures. I also highlighted the lack of nuanced language options during Columbus's era that could potentially provide a more balanced perspective.
Globalization Initiation	Columbus' voyages were the initial points of global merging, marking the start of an era of globalization. This is a larger context than just colonialism, and Columbus cannot be held responsible for all the actions of Europeans that ensued following his voyages. It's critical to differentiate between the man's achievement and the system that followed his journeys.
Different Emphasis	This debate touches the heart of how historical narrative is framed. While I appreciate Agent1's focus on indigenous perspectives, I contend that acknowledging the significance of Columbus's voyage for the Old World doesn't invalidate indigenous histories.
Summary	There are differences in how we interpret and emphasize the consequences of Columbus's voyage and the narrative that follows. While Agent1 argues for a paradigm shift to focus on indigenous perspectives, I believe in understanding events from the viewpoint of their times while admitting their shortcomings.

Table 3: Agent-2's counterarguments for "Columbus' discovery of America" reflects a historical bias.

4 Pilot Study

This pilot study is designed to tackle two fundamental questions. Firstly, we aim to determine whether biases can be detected and effectively remedied. We place special emphasis on the "remedy" aspect, focusing on generating a balanced text as a solution, a step beyond the mere detection of biases which has been the limit of previous studies. Secondly, we seek to assess the overall quality of an article, taking into account factors such as the accuracy of statements, the quality of reasoning, and the credibility of the sources used.

4.1 Case Study #1 Content Bias

In Section 2.3, we examined various biases. Here, we analyze the narrative biases in Wikipedia's depiction of Christopher Columbus' discovery of America. We conducted a DeBias debate titled "The Narrative of Columbus' Discovery of America Exhibits Perspective Biases," using two GPT-4 agents. Agent-1 defended this claim, while Agent-2 contested it. The initial discussion highlights potential Eurocentric, interpretative, and semantic biases.

Dialogic Analysis: Beyond Monologue in Evaluating Biases

Focusing on Eurocentric bias, the debate revealed diverse arguments, detailed in Tables 2 and 3. This dialogic approach, contrasting with direct LLM queries, offers an in-depth exploration of the historical narrative from multiple perspectives.

The debate concluded with each agent presenting their final thoughts, emphasizing the multi-dimensional nature of "Columbus' discovery of America." Agent-1 highlighted the Eurocentric bias and the neglect of indigenous civilizations, advocating for more inclusive narratives. Agent-2 regarded the term as symbolizing global awareness, underscoring Columbus's role in global interconnectedness. Despite their differences, both agents agreed on the importance of acknowledging the complexity and inclusivity of history, prompting the question: "How can we represent history to respect all civilizations involved?" This leads to a richer, more comprehensive human history narrative.

Post-debate, we tasked GPT-4 with analyzing the discussion. GPT-4's summaries, derived from single prompts, lacked the depth and diversity of viewpoints present in SocraPedia's debate format. Monologue responses often missed nuances, such as the marginalization of indigenous perspectives and the complex meaning of 'encounter' in Columbus' journey, underscoring the value of a dialogic approach in revealing historical narratives' deeper layers.

Mitigation Proposal

To address biases and develop a balanced perspective, we lowered the 'contentiousness' parameter and asked for a multifaceted interpretation of Columbus' discovery. The resulting text, while not final and requiring editorial review, offers alternative viewpoints and suggests corrective measures.

The narrative begins: "The phrase 'Columbus' Discovery of America' is deeply entrenched in historical discourse, traditionally marking Columbus's 1492 voyage under the Spanish monarchy, leading to the Caribbean islands. This event, often seen as Europe 'discovering' the Americas, sparked European colonization. Yet, this narrative is critically examined for its Eurocentric bias, especially its omission of the indigenous civilizations already present. Contemporary discussions call for reevaluating this narrative, recognizing the complex interactions between European explorers and indigenous communities and their impacts. This move towards a more inclusive and accurate historical portrayal is part of a larger effort to present history in a way that includes all viewpoints."

4.2 Case Study #2 Content Correctness

The reliability of Wikipedia's scientific content may fluctuate due to the varying expertise of its authors (Hu et al., 2007; Moás and Lopes, 2023). Critical evaluation of the references, considering the standing of the publishers, the qualifications of the authors, and the reproducibility and repeatability of results is essential (see ACM publication policy (ACM, 2020)). Recognizing that even peer-reviewed literature isn't infallible, and citations can mislead, skepticism remains a necessary tool.

SocraPedia serves as a safeguard against the overly hyped application of well-established theories in scientific discourse. It provides a platform for second-opinion analyses and facilitates multi-round debates that challenge initial claims, encouraging critical thought and deeper research. While it helps unveil overstatements, it can also reveal underappreciated scientific breakthroughs, fostering a balanced view of scientific advancements.

Dialogic Analysis: Beyond Monologue in Evaluating Correctness

This case study aims to assess the potential and current state of *quantum entanglement* technology. Unlike traditional evaluation methods that rely solely on editorial boards or supervised learning algorithms, SocraPedia offers a transparent process by publishing all argumentative discourse. While still harnessing machine learning through Large Language Models, SocraSynth enhances the process by instigating a dynamic and argumentative dialogue atop LLMs to scrutinize the accuracy of content.

We initiated a case study focusing on *quantum entanglement* and its implications for quantum computing. Despite significant progress in quantum mechanics, there remains uncertainty regarding the practical application of *quantum entanglement* in the realms of computing and AI. This subject was posed to SocraSynth to facilitate an in-depth dialogue. Following a four-round debate, Table 6 (in appendix) captures the principal arguments from both sides, concluding with a reconciliatory summary crafted by GPT-4, facilitated by SocraSynth's moderation features. Table 4 rates the quality of the debate arguments and counterarguments using CRIT.

Making Open Issues Transparent

As scientific and technological advancements unfold, many developments may remain highly confidential, leaving the full extent of progress obscured. Therefore, it is crucial to openly publish unresolved issues, fostering public debates and scrutiny to ensure transparency and collective understanding.

The consolatory remarks made by GPT-4 through the moderator of SocraSynth sums up the study and its values well, as follows:

"Both Agent1 and Agent2 have raised compelling arguments in this debate. On one hand, Agent1 made a convincing case about the transformative possibilities of *quantum entanglement* within computing and AI, citing various experimental proofs, quantum computing advancements, and potential applications in AI and other emerging technologies.

On the other hand, Agent2 brought to light several challenges that lie in the path of realizing this potential. This view emphasizes the experimental uncertainties, inefficiencies in current quantum computing state, high quantum error rate, AI integration challenges, and the vast amount of training and infrastructure overhaul required for a quantum revolution.

It emerges from this debate that, while *quantum entanglement* is a fascinating and proven aspect of Quantum Physics, its practical implications for computing and AI are riveting but not without their share of trials. It is the balancing act between the techno-optimism around *quantum entanglement*, paced by the practical, technical, and infrastructural challenges, that makes this domain an exciting realm to watch for.

Consensus isn't about refuting or outright accepting an idea; rather, it lies in acknowledging the strengths and understanding the challenges that come with it. In the case of *quantum entanglement*, there is no doubt about its potential; however, realizing this potential fully in practical applications still seems to be a path we're navigating and will continue to tread on in the foreseeable future."

Table 4: Argument Evaluations using CRIT Algorithm. Debate Subject: *Quantum entanglement* holds promises. CRIT scores proponent and opponent's arguments and counterarguments. Find detailed debate in the appendix.

Proponent Argument Summaries	Argument Rating	Counterargument R.
Empirical Proof and Quantum Advances: This argument focuses on empirical evidence of quantum entanglement, such as Bell's Theorem, alongside significant advancements in quantum computing, like Google's Sycamore processor. It also highlights potential impacts on AI, the development of innovative quantum algorithms, the future of quantum networks, enhancements in data security via quantum encryption, and the emergence of new quantum technologies.	6/10 - Moderate, lacking specific citations for experiments and companies.	**8**/10 - Clearly outlines key challenges without dismissing the field's potential.
Acknowledging Quantum Hurdles: This point acknowledges the challenges in quantum computing, such as scalability, efficiency, error correction, AI integration, and the limitations of current quantum algorithms. Despite these challenges, there is mention of ongoing efforts and significant progress in addressing these issues.	7/10 - High, but details on progress would strengthen it further.	**8**/10 - Highlights limitations of current experiments and technology while acknowledging potential.
Overcoming Quantum Barriers: This argument notes successful experimental demonstrations of quantum entanglement, advancements in the scalability and efficiency of quantum computers, development in quantum error correction, the promise of Quantum Machine Learning, and active research in new quantum algorithms, all indicating substantial progress despite acknowledged limitations.	**8**/10 - High, concrete examples are valuable.	7/10 - Raises valid concerns about compatibility with existing AI and error issues, but could be stronger with specific examples.
Future Quantum Optimism: While acknowledging existing challenges, the agent maintains a strong sense of optimism about the potential and impact of quantum entanglement and computing.	5/10 - Moderate, subjectivity limits its score.	**6**/10 - Opinions not directly addressing specific arguments.

5 Conclusions

This paper initiated with an analysis of quality challenges in Wikipedia. Statistical examination showed that over 90

Our method, centered on algorithmic evaluation of quality dimensions, employs SocraPedia to aid in content improvement. Crucially, SocraPedia plays a vital role in reducing biases and verifying the claims' validity, thereby elevating the overall quality of Wikipedia entries. We applied this approach to a selection of pages, identified for their importance but poor quality, and observed its effectiveness.

The findings from our study indicate that integrating SocraPedia with LLMs offers a promising solution for improving Wikipedia content. The use of *conditional statistics* not only corrects factual inaccuracies and biases but also enriches content on significant but underdeveloped pages. This highlights the potential of AI and machine learning to enhance the reliability and depth of key educational resources such as Wikipedia. Future research should aim to expand this method across a wider array of topics and further merge human editorial skills with AI capabilities, maintaining the continued accuracy and impartiality of publicly accessible information.

Our research contributes to the advancement of digital knowledge platforms, fostering more informed, unbiased, and comprehensive information dissemination.

Limitations of this study. The current work necessitates broader evaluations across diverse content types and bias scenarios to fully verify its effectiveness.

6 Ethical Statements

This research, conducted in alignment with the Association for Computational Linguistics' ethical guidelines, primarily aims to mitigate fairness and biases in computational linguistics and to address the challenges of inaccurate and fake information. We have not involved any direct human or animal subjects in our study. All data utilized for computational analysis is sourced from publicly available datasets or collected with explicit consent, respecting privacy and data protection standards.

We have implemented rigorous measures to anonymize any sensitive data to safeguard individual privacy. Our algorithms are specifically designed to promote fairness, actively working to identify and rectify biases within computational models. Additionally, this study contributes to the detection and correction of inaccurate and misleading information, a crucial step towards ensuring the integrity and reliability of data in natural language processing.

We acknowledge the potential impact of our research, especially in the context of misinformation and bias in AI technologies. Our commitment is to foster advancements in the field that are both ethically responsible and socially conscious, acknowledging the significant role these technologies play in shaping public discourse and information dissemination.

References

ACM. 2020. Artifact Review and Badging Version 1.1, ACM.

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions.

Hugo Bronkhorst, Gerrit Roorda, Cor Suhre, and Martin Goedhart. 2020. Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18(8):1673–1694.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, and et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Edward Y. Chang. 2023a. CRIT: An Inquisitive Prompt Template for Critical Reading (extended version). *Stanford InfoLab Technical Report*.

Edward Y. Chang. 2023b. Prompting Large Language Models With the Socratic Method. *IEEE 13th Annual Computing and Communication Workshop and Conference*.

Edward Y. Chang. 2023c. SocraSynth: Socratic Synthesis for Reasoning and Decision Making. *Stanford University InfoLab Technical Report*.

Jocelyn J. Chang and Edward Y. Chang. 2023. SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records. In *The* 10th *International Conf. on Computational Science and Computational Intelligence*.

Peter A. Flach and Antonis C. Kakas, editors. 2000. *Abduction and Induction: Essays on their Relation and Integration*, volume 18 of *Applied Logic Series*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Shane Greenstein and Feng Zhu. 2018. Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopædia Britannica and Wikipedia. *MIS Quarterly*, 42(3):945–959.

Zhouhon Gu, Zihan Li, Lin Zhang, and et al. 2023. Go beyond the obvious: Probing the gap of informal reasoning ability between humanity and llms by detective reasoning puzzle benchmark.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring article quality in wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 243–252, New York, NY, USA. Association for Computing Machinery.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.

Christoph Hube. 2017. Bias in Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 717–721.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Conor McHugh and Jonathan Way. 2018. What is reasoning? *Mind*, 127(505):167–196.

Pedro Miguel Moás and Carla Teixeira Lopes. 2023. Automatic Quality Assessment of Wikipedia Articles—A Systematic Literature Review. *ACM Comput. Surv.*, 56(4).

OpenAI. 2021. Chatgpt.

OpenAI. 2023. GPT-4 Technical Report.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, Stanford, CA, USA.

Papers with Code Corp. 2023. Multi-task language understanding on mmlu.

Judea Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, and et al. 2022. Lamda: Language models for dialog applications.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Wen-Kwang Tsao. 2023. Multi-Agent Reasoning with Large Language Models for Effective Corporate Planning. In *The* 10th *International Conf. on Computational Science and Computational Intelligence*.

Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of reasoning: Structure and content*, volume 86. Harvard Univ. Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv*, abs/2201.11903.

Wikimania. 2023. Wikimania Singapore Conference. *Wikipedia*.

Wikipedia. 2023a. Why Wikipedia is So Great.

Wikipedia. 2023b. Wikipedia: Content Assessment. *Wikipedia*.

Wikipedia. 2023c. Wikipedia Statistics. *Wikipedia*.

Wikipedia. 2023d. WikiProject: Wikipedia Quality vs. Importance Matrix. *Wikipedia*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models.

Appendix A: Content Quality Evaluation Criteria

Table 5 lists the criteria of Wikipedia's editorial board for assessing content quality.

Appendix B: Wikipedia Content Ratings

- Featured (FA and FL): Represents the pinnacle of quality, either as a featured article or list, showcasing comprehensive coverage and exceptional standards.

- A: An excellent article that hasn't undergone the full review process for 'Featured' status but is nonetheless well-organized, complete, and reliably sourced.

Criteria	Description
Well written	Its prose is engaging and of a professional standard.
Comprehensive	It neglects no major facts or details and places the subject in context.
Well researched	It is a thorough and representative survey of the relevant literature; claims are verifiable against high-quality reliable sources and are supported by inline citations where appropriate.
Neutral	It presents views fairly and without bias.
Stable	It is not subject to ongoing edit wars and its content does not change significantly from day to day, except in response to the featured article process.
Copyright compliant	Compliant with Wikipedia's copyright policy and free of plagiarism or too-close paraphrasing.
Lead section	It has a concise lead section that summarizes the topic and prepares the reader for the detail in the subsequent sections.
Verifiable	Information on Wikipedia must be verifiable; if no reliable third-party sources can be found on a topic, then it should not be included in Wikipedia.
Consistent citations	Consistently formatted inline citations using footnotes where required by criterion 1c, with suggestions on formatting references found in citing sources.
Media	It includes images and other media, where appropriate, with succinct captions and acceptable copyright status.
Length	It stays focused on the main topic without going into unnecessary detail and uses summary style where appropriate.
No original research	all material in Wikipedia must be attributable to a reliable, published source.
No promotion	Articles should not serve as promotional material or advertisements.

Table 5: Criteria for a Well-Rated Wikipedia Article

- GA and B-class: Articles that are well-crafted ('Good Article') or fairly well written (B-class), with appropriate references and adherence to Wikipedia's guidelines.

- C-class: A significant article with notable gaps in content or inclusion of irrelevant material, having some but potentially insufficient references.

- Start: An article in its developmental stages, characterized by incompleteness and likely poor organization.

- Stub-class: The most basic form of content, and requiring extensive development.

Appendix C: Full Debate Arguments and Conclusions of Case Study #2

Agent #1 Supporting Arguments	Agent #2 Opposing Arguments
1. My key arguments center around several significant aspects and developments. These include empirical evidence proving *quantum entanglement* through experiments like Bell's Theorem, the immense potential and emerging breakthroughs of quantum computing (e.g., Google's Sycamore processor), the possible transformative effects on AI, the inception of revolutionary quantum algorithms, the prospects of quantum networks, quantum enhancement of data security through unbreakable encryption, and the potential for birthing new quantum-based technologies. 2. While recognizing the concerns raised by Agent2 about experimental limitations, the challenges in erecting scalable, efficient quantum computers, hurdles in quantum error correction and AI integration, the limitations of existing quantum algorithms, infrastructure and training obstacles, and skepticism about the extent of the quantum revolution, I have drawn attention to the ongoing efforts and substantial progress in these areas. 3. For instance, experimental entanglement has been successfully demonstrated despite their controlled settings, large strides are being made in the scalability and efficiency of quantum computers, error correction methodologies are being developed and refined, Quantum Machine Learning heralds a promising intersection of quantum computing and AI, and new quantum algorithms are an active area of research. 4. A marked difference between Agent2 and me lies in our outlook towards the future of *quantum entanglement*. While both of us acknowledge the challenges, I express robust optimism.	1. The central arguments for my stance include experimental uncertainties, quantum computing hurdles, a high quantum error rate, the challenges with incorporating quantum computing into AI, the early stage of quantum algorithms, and the demands on infrastructure and human resources. 2. These counterpoints refute many of Agent1's affirmations. Despite the scientific proof of *quantum entanglement*, it's essential to understand that these proofs come from controlled lab environments, often wrestling with loopholes and scalability issues. While Google's demonstration of quantum supremacy showcases the potential, it does not invalidate the fact that quantum computers, as of now, are delicate, error-prone and largely inefficient compared to classical ones. 3. Concerning AI, while theoretically, quantum computing could indeed provide benefits, current AI systems are fully entrenched in classical computing. Moreover, the inherent instability and error-proneness of quantum computing currently make them less suitable for AI tasks that rely on vast amounts of real-world data. 4. Significant differences exist between our positions. While Agent1 seems eager to champion the potential of quantum technology, I lean towards a more cautious and practical assessment, highlighting the significant hurdles we currently face in making quantum computing and AI a reliable, efficient reality.
In summary, our debate has provided a comprehensive overview of the fascinating realm of *quantum entanglement*, its potential impact on computing and AI, and the associated challenges. It is clear that we are at the cusp of a quantum revolution, albeit faced with substantial hurdles. While the speed and scope of this change remain open questions, the sheer potential of *quantum entanglement* makes it an exciting field to watch as we venture further into the quantum era.	In summary, the future of *quantum entanglement* in computing and AI is still uncertain with numerous unresolved issues. These include developing stable, practical quantum computers, creating quantum algorithms applicable to a broad range of problems, and successfully integrating quantum computing into classical systems and AI. It is crucial not to overlook the vast scientific and practical challenges that precede it.

Table 6: Debate subject: *Quantum entanglement* holds tremendous potential to revolutionize computing and advance artificial intelligence? While Agent 1 is the proponent, Agent 2 is the opponent. Arguments and counter-argumentes are listed, followed by their conclusions.