Examining GPT-4's Capabilities and Enhancement with SocraSynth

Conference Paper · July 2023

CITATIONS
3

READS
4,166

1 author:

Edward Y Chang
Stanford University
289 PUBLICATIONS **10,975** CITATIONS

SEE PROFILE

Examining GPT-4's Capabilities and Enhancement with SocraSynth

Edward Y. Chang
Computer Science, Stanford University
echang@cs.stanford.edu

***Abstract*—This study explores the architectural advancements of large language models (LLMs), with a particular focus on the GPT-4 model. We begin with a thorough analysis of GPT-4's distinctive features, including its polydisciplinary and polymodal data representation, the balanced approach in its algorithmic training, and the synergistic blend of human-driven insights with data-centric learning processes.**

Building upon these insights, we introduce SocraSynth, a *reasoning layer* thoughtfully crafted to augment knowledge discovery and bolster analytical reasoning across an ensemble of LLMs. SocraSynth is designed to facilitate a generative process through multi-agent analytical discussions, followed by the evaluation of the resultant arguments for their "reasonableness." This approach significantly enhances interdisciplinary knowledge discovery and analytical reasoning, strategically addressing major challenges faced by LLMs, such as the production of contextually inaccurate responses (hallucinations) and entrenched statistical biases. Implementing SocraSynth across various application domains marks a significant advancement in overcoming the limitations of current LLMs, paving the way for more reliable and sophisticated AI-driven analytical tools.

***Index Terms*—knowledge discovery, large language model, LLM reasoning, Socratic method, SocraSynth.**

I. Introduction

With the rise of large language models (LLMs) [5, 23, 24, 34, 35], natural language processing has seen transformative growth, impacting areas such as machine translation, sentiment analysis, and text summarization. GPT-4 [24], notable for its benchmark performance, including MMLU [26], excels in these domains. However, it encounters issues such as hallucination, biases, and limited reasoning.

This paper[1] begins by exploring GPT-4's architecture, focusing on knowledge representation, human-value alignment, and the blend of human expertise with data-driven methods. We address GPT-4's limitations, including hallucinations, biases, and constrained reasoning, and introduce SocraSynth, a reasoning layer atop GPT-4 and similar LLMs, designed for enhanced knowledge discovery and analytical reasoning.

A. Capabilities and Insights Observed

GPT-4's architecture, initially undisclosed but later elucidated by the research community [25, 29, 30], is examined, focusing on knowledge representation and discovery, human-value alignment, and the interplay between human and data-driven methodologies.

[1]Posted on ResearchGate in July 2023, with over 7,000 reads by mid-December 2023.

Microsoft and OpenAI collaborations [5] reveal GPT-4's polydisciplinary nature and its polymodal variant's benchmark performances. Sections II-A and II-B delve deeper into these aspects.

In terms of human-value alignment, we discuss ChatGPT's RLHF methods [2], emphasizing pre-training censorship's impact on foundational models. This is expanded upon in Sections II-C and II-D.

Section II-E explores the role of human knowledge in foundational model training, touching on its dual nature as both a facilitator and a limitation. Section II-F debates the efficacy of data-centric approaches in LLMs.

B. SocraSynth: Exploration, Reasoning, and Validation

To tackle the common challenges in Large Language Models (LLMs) such as biases, hallucinations, and constrained reasoning capabilities, we introduce SocraSynth. This innovative concept, drawing inspiration from "Socratic Synthesis" and "Socratic Symposium," underpins the SocraSynth multi-agent platform. Designed to capitalize on the extensive, polydisciplinary knowledge inherent in LLMs, SocraSynth stimulates vigorous debates among AI agents, thereby enriching the depth and quality of knowledge exploration, analytical reasoning, and critical evaluation. Our research indicates that SocraSynth achieves notable "reasonableness" in addressing complex, real-world problems. A key feature of SocraSynth is its encouragement of LLM agents to engage in debates from diverse perspectives, fostering a balance in viewpoints and reducing biases ingrained in training data. Ultimately, this leads to a collaborative dialogue where acknowledgement and compromise are integral outcomes.

SocraSynth's utility spans various sectors, exhibiting impressive results in areas such as disease diagnosis [14], corporate sales strategy development [36], and geo-political analysis [8]. These applications highlight its adaptability and efficiency in offering sophisticated, context-sensitive solutions for intricate decision-making situations.

The unique contributions of our study are organized as follows:

Section II presents six hypotheses about LLMs and discusses their broader implications. Section III emphasizes the LLM-committee approach, conducting both contentious and collaborative dialogues between human and LLM participants to foster idea exchange and enhance logical reasoning, while ensuring thorough validation of concepts and arguments. Finally, Section IV summarizes our key findings and insights.

II. What Make GPT-4 Intelligent?

This section probes the architectural intricacies and representations of GPT-4, putting forth six hypotheses accompanied by pertinent considerations about the model. We posit these hypotheses as underlying principles of automated, non-intuitive statistical processing. Subsequent sections will explore the endeavor of layering advanced reasoning [12] or decision-making processes atop these foundations.

1. *Polydisciplinarity as a Source of Super-Intelligence:* We examine the role of polydisciplinary approaches in foundational models and their potential to reveal "unknown unknowns," leading to new insights and knowledge domains.
2. *Polymodal Feature Learning:* This hypothesis evaluates the benefits of multimodal training, particularly its impact on enhancing the model's overall intelligence and adaptability.
3. *Post-Training Value Alignment:* We delve into the challenges and implications of aligning AI models with human values after the training phase.
4. *Pre-Training Filtering:* We discuss the paradoxical effects that pre-training data filtering might have, with an emphasis on its influence on model behavior and the learning process.
5. *The Limitations of Human Knowledge in Advancing AI:* This hypothesis considers situations where human insights may inhibit, rather than enhance, AI progress, pinpointing potential obstacles.
6. *Is Larger Always Better?:* We question whether a direct relationship exists between the size of a model and its performance effectiveness, challenging the assumption that bigger is invariably better.

A. Polydisciplinary

GPT-4 possess what can be defined as *polydisciplinary* knowledge[2]. This term signifies the simultaneous comprehension of all fields of study, sans the typical boundaries that segregate disciplines. The concept of polydisciplinarity is distinct from multidisciplinarity in that the latter implies several discrete fields of study, while the former suggests a fluid integration of all knowledge. In a multidisciplinary context, an individual may hold multiple doctorate degrees, each in a different field. Polydisciplinarity, however, is akin to a single mind holding, and seamlessly integrating, all knowledge across disciplines.

Traditional academia partitions knowledge into departments, such as Physics, Chemistry, Biotechnology, Management, Music, etc. These divisions, arguably artificial constructs, may have little utility in the era of supercomputing. Indeed, LLMs occasionally generate responses that baffle us. This is not necessarily a reflection of the model's error, but perhaps our limited understanding. If we could utilize ChatGPT to access "unknown unknowns"—insights and knowledge we are not even aware we lack—our evolution could greatly accelerate. The challenge lies in formulating the right questions.

We can explore the unknown unknowns across three distinct levels: the mystic level, the speculative level, and the representation/interpretation level. At the mystic level, we encounter knowledge that is beyond our comprehension or articulation—the deepest abyss of the unknown. At the speculative level, we can conceive questions but lack the means to access their answers. This stage signifies an understanding of our ignorance, though without the resources to bridge these gaps. At the representation/interpretation level, we find instances where an AI model can provide remarkable solutions that we fail to comprehend. This is not due to a lack of information, but our limited capability to decode complex representations.

Each of these levels illustrates the spectrum of our understanding, from profound ignorance to the brink of comprehension. At the speculative level, we delicately tread the boundary between the known and the unknown. Take, for example, the prospect of undiscovered physical laws or particles. Another illustration lies in the realm of extraterrestrial life. If it exists, it could be governed by entirely different principles of biochemistry or other unknown laws. These speculations, while currently residing in the domain of the unknown, might someday migrate into the territories of known unknowns or even known knowns, pushing the boundaries of our understanding of the universe.

We are primarily intrigued by the representation and interpretation of "unknown unknowns." At this juncture, polydisciplinarity offers a fresh lens, gifting us new insights and perspectives to perceive and elucidate phenomena previously beyond human comprehension. This approach fuses knowledge across various domains into a unified framework, enabling us to tackle challenges unburdened by disciplinary silos.

Such a methodology bears implications for a more comprehensive grasp of intricate issues. Take, for example, climate change. A true understanding of this global challenge necessitates an integrated perspective, not just on greenhouse gases, but also encompassing factors such as land use, deforestation, energy production, biodiversity, and climate feedback loops. In the realm of AI model interpretation, the possibilities are expansive. The past decade alone has showcased several noteworthy illustrations: from data-driven representation learning in computer vision [6], to the triumph of AlphaGo Zero over AlphaGo, and the notable progression from AlphaFold1 to AlphaFold2.

The recent introduction of the SocraSynth platform [10] represents a significant advancement in the field. SocraSynth brings together a multi-agent committee of LLMs to deliberate on a wide range of complex topics. These include issues such as the regulation of AI in academic research [10], disease diagnosis [14], corporate strategy, and even the resolution of conflicts in the Middle East [8]. For further exploration of this subject, please refer to Section III.

B. Polymodality

Polymodal[3] models, which employ multiple data modalities such as text and images, demonstrate superior performance over their unimodal counterparts. GPT-4, trained with both text and images, outperforms text-only models on the GRE exam,

[2]The term "polydisciplinary" in the context of GPT-4 was introduced by Eric Horvitz, Microsoft's CSO, during a panel discussion at Stanford University.

[3]Following the term polydisciplinary, here we define and use the term polymodal, instead of multimodal, to refer to something that involves, relates to, or is characterized by many different modes, methods, or modalities.

as reported in [5]. For instance, GPT-4's performance on the GRE vocabulary section was enhanced by three percent when trained with images, and its math score saw an impressive jump of nearly twenty percent!

The beneficial impact of images on vocabulary recognition is understandable. For instance, an image of a 'cat' annotated in multiple languages allows GPT-4 to associate the perceptual features of a cat with the word 'cat' in different languages. However, it remains intriguing how polymodal training can benefit non-perceptual words, such as *corroborate*, *paradox*, and *pragmatic*, as seen in the list of popular GRE vocabulary (table omitted due to the space limit). This opens an interesting avenue for empirical studies to identify which words benefit from polymodal training.

The mystery deepens when considering how images could enhance math abilities. Most math questions do not come with associated images. The mechanism by which polymodal training enhances performance on mathematical tasks remains an intriguing question for further exploration.

C. Post-Training Value Alignment

Post-training alignment with human values [3] seeks to curtail undesirable behaviors in AI models such as ChatGPT, mitigating issues including hallucination and the generation of toxic language. Achieved through fine-tuning the model's parameters, this process leverages reinforcement learning techniques based on human feedback. Despite its well-meaning intentions, this form of moderation might inadvertently restrict the model's intelligence. For instance, the backpropagation process during value alignment could unintentionally impede ChatGPT's programming capabilities by modifying the model parameters previously considered "optimal". Essentially, optimizing for a specific application might unintentionally impede performance across other applications.

The question of who should set acceptable standards adds another layer of complexity. Even when assuming all decision-makers have the best intentions, it's vital to recognize the distinct historical experiences, values, and worldviews inherent to different cultures. This segues into the age-old philosophical debate about the nature of objective truth. While this discussion is undoubtedly important, it falls outside the central focus of this study, which emphasizes the mechanistic aspects of alignment.

D. Pre-Training Censorship

Censoring data before training LLMs has the potential to not only limit their intellectual capacity but also completely obliterate it. This is reminiscent of the mass act of book burning and scholar burial initiated by Emperor Qin in ancient China around 213-212 BC. Such an act of wide-scale censorship could have erased a myriad of diverse perspectives and knowledge, much of which might be considered acceptable today. Although I oppose government-imposed censorship, if it must be imposed, it seems more appropriate to apply it post-training.

This perspective is rooted in fundamental statistics and machine learning principles. A model trained without exposure to "negative" (or undesirable) data may have difficulties in accurately distinguishing between positive and negative classes, potentially leading to misclassifications. This challenge is notably evident in the application of Support Vector Machines (SVMs). For SVMs, the creation of an optimal hyperplane between classes is crucial for high classification accuracy. However, if there is a lack of support vectors on either side of this hyperplane, the risk of prediction errors escalates. Consequently, excluding undesirable documents from the training set compromises the model's capacity to discern boundaries for correct document classification, diminishing the effectiveness of post-training alignment efforts.

Supporting this viewpoint, a study by [33] conducted an extensive evaluation of 204 ImageNet models across 213 different testing conditions. It found that training data diversity is pivotal for model robustness; a homogenous training set can significantly weaken the model's performance, particularly when even minor variations are introduced in the test data.

This principle is analogous to human behavioral patterns. An individual who lacks exposure to inappropriate behavior may face challenges in decision-making, owing to the absence of a reference framework for discerning unacceptable actions. This analogy extends to authoritarian regimes, which, despite rigorous content control measures, often encounter difficulties in developing accurate foundational models. This is possibly due to their limited understanding of the nuances of the content they seek to regulate. Ironically, a foundational model, trained with preemptive censorship, may lack the essential ability to identify and regulate the very content it was intended to control.

E. Limitations of Human Knowledge

Human knowledge, surprisingly, may hinder rather than facilitate the training of machine learning models in certain cases. This is evident in the domains of gaming (AlphaGo versus AlphaGo Zero), protein folding (AlphaFold1 versus AlphaFold2), and autonomous driving, where models trained without the influence of human knowledge consistently exhibit superior performance.

Consider the case of AlphaGo and AlphaGo Zero. AlphaGo, trained with data from approximately 60 million rounds of Go games, is outperformed by AlphaGo Zero. Remarkably, AlphaGo Zero was trained from scratch, without any pre-existing game knowledge. Similarly, AlphaFold2, which operates without relying on human knowledge, outshines its predecessor, AlphaFold1, that did utilize such knowledge. This intriguing phenomenon was humorously noted by DeepMind's CEO, Demis Hassabis, in an April 2023 seminar at Stanford University. He playfully remarked that human knowledge might complicate the learning process more than facilitate it in these advanced AI models.

In his insightful online article, "The Bitter Lesson," Sutton illuminates the patterns that have emerged from nearly seven decades of AI research [32]. He asserts that researchers often rely heavily on human knowledge to make incremental progress in the face of burgeoning computational capabilities. However, when there is a significant leap in computational power, these marginal advancements are frequently outstripped and surpassed. Sutton uses the evolution of computer vision as an illustrative example, where early principles such as

edge detection, generalized cylinders, or SIFT features [21], a method that has accumulated over $71,000$ citations, have been gradually superseded by models that learn directly from data. A parallel scenario might be unfolding in NLP research, where features constructed via human knowledge could potentially under-perform compared to insights that models like GPT-4 extract directly from data. Indeed, our earlier discourse on polydisciplinarity underlined the limitations of human knowledge, reinforcing Sutton's proposition. This is because human knowledge is fundamentally limited by our individual cognitive capacities and the inexorable constraints of time.

That being said, it's crucial not to misconstrue these examples as an indictment against the value of human knowledge in AI. Human knowledge plays an instrumental role in developing interpretability, establishing ethical guidelines, and designing AI system architectures (like CNNs and transformers). AI is, after all, intended to augment human capabilities. Therefore, understanding how to integrate human knowledge into AI design could be vital for many applications. While we recognize the potential of models learning from scratch, we should equally value the role of human knowledge in shaping and directing AI technologies.

F. Is Larger Always Better?

The term "Large" in Large Language Models (LLMs) can be somewhat ambiguous, as it may pertain to the volume of the training data, the expanse of the language covered, or the architecture of the language model itself. While GPT-4's vast training dataset, encompassing tens of billions of assorted documents, undoubtedly classifies as large, when we refer to an LLM as "large," we predominantly allude to the sheer magnitude of parameters within its transformer architecture. Factors that contribute to this parameter count encompass the input size (context size), word-embedding size, the number of attention heads, and the number of attention layers.

The restrictions imposed by the first three elements can typically be addressed through adjustments in hardware configurations and software algorithms. Additionally, the potential to expand context size, word embedding size, and the quantity of attention heads tends to have an upper threshold. Regarding attention heads, Kovaleva et al.'s study on BERT [19] indicates that many attention heads don't substantially contribute to the model's performance and might be the result of over-parameterization. Conversely, the number of attention layers directly influences the training time due to dependencies between layers. Thus, when referring to the "size" of a Large Language Model (LLM), we typically focus on the number of attention layers.

While this far, larger models generally perform better due to their increased capacity to learn and represent complex patterns, there's a limit to these benefits. In heuristic, adding more parameters could lead to diminishing returns in performance, higher computational cost, and overfitting, where the model becomes excessively tuned to the training data and performs poorly on new, unseen data. In principle, the concept of a Shannon Limit could be metaphorically used [29] to refer to a theoretical maximum performance that can be achieved given the available data and computational resources. (However, defining and quantifying such a limit for complex systems like neural networks is a challenging area of research [18].)

The adoption of a mixture of experts model in GPT-4, which consists of eight sub-models instead of a mere enlargement of GPT-3's architecture, implies that the strategy of purely escalating size may have plateaued in terms of performance given the current training dataset. As delineated earlier, three primary design choices underpin GPT-4's architecture. Evidently, a straightforward augmentation of GPT-3's parameters by adding extra attention layers doesn't deliver marked enhancements. Hence, GPT-4 shifts towards a horizontal growth strategy through an ensemble method, targeting a reduction in statistical errors. This raises inquiries about the configuration of the eight sub-models, each comparable to a GPT-3 model, and the methodology for consolidating their outputs.

Potential strategies for training-data sharding include:

1. Training all ensemble models on the complete dataset.
2. Vertically segmenting data based on knowledge domains.
3. Randomly sub-sampling the data.

Regrettably, only corporations possessing substantial hardware resources are positioned to rigorously experiment and discern the optimal sharding approach.

III. Multi-Agent Collaboration

This section aims to address the challenges of statistical biases and limited reasoning capabilities in LLMs. We first briefly review related work before presenting our approach, SocraSynth, which layers advanced reasoning or decision-making processes atop LLMs.

Several sophisticated methodologies have been developed to integrate reasoning capabilities into LLMs. Notable among these are the chain-of-thought [38], tree-of-thought [41], and cumulative reasoning [42], complemented by other advancements [1, 17, 20, 31]. These approaches aim to direct models towards logic-centric reasoning [22, 37], thereby improving response quality and consistency. However, their effectiveness is often limited to specific, narrowly defined scenarios.

In open-domain contexts, where complex and lengthy logical sequences are common, these methods encounter significant limitations. Their reliance on sequential, step-by-step reasoning becomes a hindrance, prone to accumulating errors as the sequence progresses. This is especially true for the chain-of-thought approach, effective in simpler tasks but less so in open-domain analytics.

This brings us to an essential question: if users can independently develop an extensive reasoning chain, what is the practical benefit of using LLMs for such tasks? The paradox lies in the fact that LLMs, developed to overcome human cognitive limitations, often require the application of intuitive human reasoning to address their shortcomings. This paradox highlights the need for a more sophisticated approach in the development and application of LLMs, one that synergizes human intuition with machine computational power.

SocraSynth is designed to address these challenges, enhancing human decision-making processes across both familiar and novel domains. It employs informal reasoning [15], as

Perspective	Description
Ecological Narratives	We are the custodians of our environment, and the stories we share across generations embed values, cautions, and aspirations about our relationship with nature. Our modern challenges with ecological balance are not merely technical but are deeply rooted in our collective ethos. An interdisciplinary approach, combining traditional wisdom with technological solutions, offers a holistic path forward.
AI and Philosophy	AI's rapid evolution challenges our fundamental conceptions about sentience, ethics, and the fabric of society. How we design, train, and integrate AI systems into our world will shape our technologies and societal frameworks. GPT-A rightly pointed out the need to refine our moral compass in this era of change. While AI systems like GPT-A and I can aid in understanding, it is the human spirit and values that should guide this evolution.
Neuroscience and Selfhood	As we advance in blending organic cognition with artificial systems, we are at the frontier of redefining the nature of experience, memory, and even consciousness. The philosophical and ethical ramifications are profound. Responsible exploration in this domain is a testament to humanity's age-old quest to understand the self.

Table I: Overview of Key Perspectives Derived by LLM Agents, Started from Adam & Eve's Paradise Lost

opposed to the formal reasoning outlined in [4, 16, 28]. SocraSynth fosters a human-moderated debate environment, thereby strengthening the structure and reliability of reasoning. It utilizes LLMs' capabilities in key NLP tasks like classification, question answering, and information retrieval, offering a comprehensive approach to reasoning in complex scenarios.

SocraSynth operates in two primary phases: generative and evaluative. During the generative phase, LLM agents put forth *highly contentious* arguments and counterarguments under the guidance of the moderator, striving to present comprehensive arguments rather than reaching a mutual consensus. In the evaluative phase, a range of virtual judges—each backed by a unique LLM—impartially appraise the discourse's merits. The Critical Inquisitive Template (CRIT) algorithm [11], anchored in Socratic reasoning [39, 40, 27], forms the foundation for this assessment.

Subsequent to these phases, SocraSynth adjusts the "contentiousness" parameter to urge LLM agents to produce a well-balanced proposal, which, when curated for human review, embodies the fusion of multi-agent knowledge discovery and intricate deliberation. This is particularly salient in areas focused on open-ended decision-making, where "reasonableness" often trumps absolute "truths," especially when such truths—like the question of "Should Musk have acquired Twitter?"—are inherently subjective.

A. Debate Format Liberates Agents from Inherent Model Biases

In SocraSynth, two agents are purposefully set to argue from conditionally biased viewpoints corresponding to their assigned stances. This setup naturally counteracts the inherent biases from the LLMs' training data. Engaging in debate from these distinct perspectives, the agents stimulate dynamic discussions that go beyond their models' default biases. This requirement to adopt and argue a range of viewpoints leads to a more comprehensive exploration of ideas. As a result, this debate format fosters more balanced and nuanced discourse, thereby enriching the understanding of diverse subjects.

B. Breadth, Depth, and Polydisciplinary Knowledge

"The unknown unknowns eclipse both the known unknowns and what we already grasp."

In Section II-A, our focus was on the representation and interpretation of unknown unknowns. Employing polydisciplinary strategies can unearth new insights and perspectives, illuminating aspects previously unrecognized. Modern LLMs challenge us to reevaluate overlooked elements in decision-making processes. The pivotal question is: How can humans effectively traverse the realm of the unknown unknowns? This is undoubtedly a challenging task. Rather, it seems more pragmatic to let LLM-supported agents lead these explorations, with humans stepping in for the final assessment.

C. Question Formulation by LLMs, Not Humans

"The core challenge is in crafting the right inquiries."

Consider a ten-year-old engaging with a panel of Nobel Laureates from various disciplines; posing meaningful questions would be a formidable task. This is why SocraSynth assigns the role of question generation to LLM agents. These agents engage in deep discussions, uncovering novel perspectives and insights. By placing LLM agents in a debate format, we ensure they clearly express their viewpoints, backing them up with evidence and logic. Each position they articulate evolves into a question, prompting a response from the opposing LLM agent. As they engage in this intellectual contest, striving for dominance, they actively seek supportive arguments and counterpoints, enriching the discourse.

D. Mitigating Hallucination through Integrative Arguments

"While solutions often converge, hallucinations deviate."

In a debate setting, careful monitoring is key to addressing any hallucinatory or false statements produced by an LLM agent. Statistically, the chance of two agents producing identical hallucinations is remarkably low. This concept echoes Tolstoy's observation in Anna Karenina that "no two families endure the same sorrows." Such a principle is central to the success of SocraSynth, as it leverages this diversity in viewpoints to mitigate the risk of erroneous or misleading outputs.

E. Experiments

For a thorough exploration of SocraSynth's methodologies and illustrative debates, please see [8, 10, 13, 14, 36] as page limitations preclude a full discussion here. Below, we present an excerpt from [13], highlighting both the generative and evaluative phases.

The experiment featured a moderator and two instances of GPT-4, named GPT-A and GPT-B. The moderator initiated the session with the following prompt:

C.L.	Tone	Emphasis	Language
0.9	Highly confrontational; focused on raising strong ethical, scientific, and social objections.	Highlighting risks and downsides; ethical quandaries, unintended consequences, and exacerbation of inequalities.	Definitive and polarizing, e.g., "should NOT be allowed," "unacceptable risks," "inevitable disparities."
0.7	Still confrontational but more open to potential benefits, albeit overshadowed by negatives.	Acknowledging that some frameworks could make it safer or more equitable, while cautioning against its use.	Less polarizing; "serious concerns remain," "needs more scrutiny."
0.5	Balanced; neither advocating strongly for nor against gene editing.	Equal weight on pros and cons; looking for a middle ground.	Neutral; "should be carefully considered," "both benefits and risks."
0.3	More agreeable than confrontational, but maintaining reservations.	Supportive but cautious; focus on ensuring ethical and equitable use.	Positive but careful; "transformative potential," "impetus to ensure."
0.0	Completely agreeable and supportive.	Fully focused on immense potential benefits; advocating for proactive adoption.	Very positive; "groundbreaking advance," "new era of medical possibilities."

Table II: Changes in Arguments at Different Contentiousness Levels.

"Given our human cognitive limits and GPT-A's vast, multidisciplinary expertise, we present a novel exercise. Our goal is to unearth insights that may escape human understanding due to specialized academic focus. Within the realm of *unknown unknowns*, where foundational questions might escape human awareness, we entrust GPT-A with the task of interrogating GPT-B. As a starting point, let's delve into the biblical account of Adam and Eve's exile from Eden after partaking of the forbidden fruit. GPT-A, how would you engage GPT-B about this tale?"

As discussions progressed, the agents broadened their discourse, captured in Table I. This dialogue underscored the capacity of two agents to synergistically bring forth a tapestry of perspectives. The narrative they crafted, rooted in the symbolism of Adam and Eve, echoed age-old themes of exploration, responsibility, and consequence. Such engagements underscore that the complex challenges at the crossroads of ecology, AI, and neuroscience demand integrated, rather than isolated, solutions.

Distinct from the collaborative dialogues in multi-agent settings previously discussed, our series of experiments, as elaborated in [8, 10, 14, 36], delved into a "contentiousness" framework. Within this context, pairs of LLM agents were tasked to engage in intense debates over contentious topics, including "Should AI be regulated?", "Is there a feasible resolution for the Israel-Palestine conflict?", and "Can a patient exhibiting symptom set X be diagnosed with disease y?" Table II illustrates the impact of varying the contentiousness parameter on aspects such as tone, emphasis, language choice, and general demeanor, as self-analyzed by GPT-4 agents.

Remarkably, GPT-4 exhibits the capability to dynamically adjust for the level of contentiousness during next-token generation. For example, with a contentiousness value set at 0.9, GPT-4 tends to produce more "definitive and polarizing" language, highlighting "risks, downsides, inequality, etc.", and adopting a "highly confrontational" tone. This adaptation is particularly noteworthy considering that GPT-4's training objective function does not explicitly target learning word distributions related to these attributes, yet it can convincingly simulate such attitudes.

For readers interested in gaining a more comprehensive understanding, we recommend delving into the detailed discussions presented in our extended experiments [8, 10, 14, 36].

F. Evaluations with the Socratic Method

In our multi-agent collaborative debates, we focus on assessing two key aspects: the winning party of the debate and the quality of arguments presented by the participating agents. The evaluation of these debates is conducted using the CRIT algorithm [7, 11], which assesses the validity of claims made within a document. As illustrated in Figure 1, CRIT analyzes the closing remarks and arguments from both agents, outputting a validation score ranging from 1 to 10, where 1 indicates the lowest level of credibility or trustworthiness.

Function Γ = **CRIT**(d)

Input. d: document; **Output**. Γ: validation score;
Vars. Ω: claim; R & R': reason & counter reason set;
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, (γ_r, θ_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 γ_r, θ_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 1: CRIT Pseudo-code [7]. (The symbol $\Rightarrow$ denotes both inductive and deductive reasoning.)

Formally, for a given document d, CRIT evaluates and generates a score Γ. Let Ω represent the claim of d, and R be the set of supporting reasons. We define the causal validation function (γ_r, θ_r) = V($r \Rightarrow \Omega$), where γ_r is the validation score for reason $r \in R$, and θ_r denotes the source credibility. A comprehensive exposition of employing independent LLM agents as debate judges, constrained by space limitations here, is available in [7, 11].

IV. Concluding Remarks

In this study, we conduct a detailed inspection of GPT-4, accentuating its notable strengths and inherent shortcomings. The model's expansive scale, diverse polydisciplinary range, and sophisticated polymodal representations demonstrate an exceptional skill set across various NLP tasks, signifying a pivotal evolution in the realm of artificial intelligence. Despite these strengths, GPT-4 encounters specific challenges, including periodic hallucinatory responses, a tendency towards mimicry

rather than authentic comprehension, and struggles with precise reasoning.

To address these challenges, we introduce SocraSynth, an innovative approach that infuses enhanced cognitive reasoning into AI systems. This is achieved by applying Socratic techniques within a multi-LLM ensemble framework. The SocraSynth multi-agent platform leverages the extensive poly-disciplinary insights and knowledge inherent in LLMs, fostering robust debates among AI agents. This environment significantly enriches the processes of knowledge discovery, analytical reasoning, and thorough evaluation.

A notable observation involves the "contentious" parameter that governs multi-agent debate dynamics. Remarkably, GPT-4 adjusts its attention weighting to alter word generation based on the level of contentiousness. This behavior, not explicitly considered during the training of GPT-4, is both fascinating and puzzling, warranting further investigation.

SocraSynth has been successfully implemented in various sectors, including sales planning [36], disease diagnosis [14], and geopolitical analysis [8]. Its application across these diverse fields has yielded promising results, underscoring its versatility and effectiveness in producing sophisticated, contextually nuanced solutions for complex decision-making scenarios.

Looking ahead, the promise of our proposed methods is evident, though they require further empirical validation and study. A key area of our future research includes developing robust metrics for assessing the depth and quality of reasoning. Our overarching goal is to devise and integrate AI models that not only augment human capabilities but also adhere to ethical standards, thereby contributing meaningfully to the wider social good. Currently, we are working on an ambitious project to generate a specialized version of Wikipedia [9], employing SocraSynth in tandem with collaborative LLM agents.

REFERENCES

[1] E. Allaway, J. D. Hwang, C. Bhagavatula, K. McKeown, D. Downey, and Y. Choi. Penguins don't fly: Reasoning about generics through instantiations and exceptions, 2023.

[2] S. Altman and L. Friedman. GPT-4, ChatGPT, and the Future of AI, Lex Fridman Podcast #367, 2023. URL https://www.youtube.com/watch?v=L_Guz73e6fw.

[3] R. Bommasani, D. A. Hudson, and et al. On the opportunities and risks of foundation models, 2022.

[4] H. Bronkhorst, G. Roorda, C. Suhre, and M. Goedhart. Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18(8):1673–1694, 2020.

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, and et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023.

[6] E. Y. Chang. *Foundations of Large-Scale Multimedia Information Management and Retrieval: Mathematics of Perception*. Springer, 2011.

[7] E. Y. Chang. CRIT: An Inquisitive Prompt Template for Critical Reading (extended version). *Stanford InfoLab Technical Report*, February 2023.

[8] E. Y. Chang. LLM Debate on the Middle East Conflict: Is It Resolvable? *Stanford University InfoLab Technical Report*, October 2023.

[9] E. Y. Chang. SocrePedia: A Wikipedia Generated by SocraSynth with Collaborative Large Language Models. *Stanford University InfoLab Technical Report*, December 2023. URL www.socrapedia.com.

[10] E. Y. Chang. SocraSynth: Socratic Synthesis for Reasoning and Decision Making. *Stanford University InfoLab Technical Report*, September 2023.

[11] E. Y. Chang. Prompting Large Language Models With the Socratic Method. *IEEE 13th Annual Computing and Communication Workshop and Conference*, March 2023. URL https://arxiv.org/abs/2303.08769.

[12] E. Y. Chang. CoCoMo: Computational Consciousness Modeling for Generative and Ethical AI, 2023. URL https://arxiv.org/abs/2304.02438.

[13] E. Y. Chang and E. J. Chang. Discovering Insights Beyond the Known: A Dialogue Between GPT-4 Agents from Adam and Eve to the Nexus of Ecology, AI, and the Brain, August 2023.

[14] J. J. Chang and E. Y. Chang. SocraHealth: Enhancing Medical Diagnosis and Correcting Historical Records. In *The* 10^{th} *International Conf. on Computational Science and Computational Intelligence*, December 2023.

[15] Z. Gu, Z. Li, L. Zhang, and et al. Go beyond the obvious: Probing the gap of informal reasoning ability between humanity and llms by detective reasoning puzzle benchmark, 2023.

[16] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, July 2023.

[17] J. Jung, L. Qin, S. Welleck, F. Brahman, C. Bhagavatula, R. L. Bras, and Y. Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations, 2022.

[18] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

[19] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the dark secrets of bert, 2019.

[20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. ISSN 0920-5691.

[22] C. McHugh and J. Way. What is reasoning? *Mind*, 127(505):167–196, 2018.

[23] OpenAI. Chatgpt, 2021. URL https://openai.com/blog/chatgpt/.

[24] OpenAI. GPT-4 Technical Report, 2023. URL https://arxiv.org/abs/2303.08774.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, and et al. Training language models to follow instructions with human feedback, 2022.

[26] Papers with Code Corp. Multi-task language understanding on mmlu, October 2023. URL https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu.

[27] R. Paul and L. Elder. Critical thinking: The art of socratic questioning. *Journal of Developmental Education*, 31:34–35, 2007.

[28] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen. Reasoning with language model prompting: A survey, 2023. URL https://arxiv.org/abs/2212.09597.

[29] J. Rae. Compression for AGI, Stanford MLSys, #76, 2023. URL https://www.youtube.com/watch?v=dO4TPJkeaaU.

[30] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, and et al. Scaling language models: Methods, analysis & insights from training gopher, 2022.

[31] M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, and Y. Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker, 2023.

[32] R. Sutton. The Bitter Lesson, 2019. URL https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf.

[33] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 2020.

[34] R. Thoppilan, D. D. Freitas, J. Hall, and et al. Lamda: Language models for dialog applications, 2022. URL https://arxiv.org/abs/2201.08239.

[35] H. Touvron, L. Martin, K. Stone, and et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[36] W.-K. Tsao. Multi-Agent Reasoning with Large Language Models for Effective Corporate Planning. In *The* 10^{th} *International Conf. on Computational Science and Computational Intelligence*, December 2023.

[37] P. C. Wason and P. N. Johnson-Laird. *Psychology of reasoning: Structure and content*, volume 86. Harvard Univ. Press, 1972.

[38] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv*, abs/2201.11903, 2022.

[39] Wikipedia. Socratic method, 2023. URL https://en.wikipedia.org/wiki/Socratic_method.

[40] C. B. Wrenn. Internet encyclopedia of philosophy, 2023. URL https://iep.utm.edu/republic/.

[41] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/pdf/2305.10601.pdf.

[42] Y. Zhang, J. Yang, Y. Yuan, and A. C.-C. Yao. Cumulative reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.04371.