

CHAPTER 6

SocraSynth: Adversarial Multi-LLM Reasoning

Abstract

Large Language Models (LLMs), while promising, face criticisms for exhibiting biases, hallucinations, and a lack of reasoning capability. This chapter introduces SocraSynth, a multi-LLM agent reasoning platform developed to mitigate these issues. SocraSynth utilizes conditional statistics and systematic context enhancement through continuous arguments, alongside adjustable contentiousness levels of the debate. The platform typically involves a human moderator and two LLM agents, each representing opposing viewpoints on a given subject. SocraSynth operates in two main phases: knowledge generation and reasoning evaluation. In the knowledge generation phase, the moderator defines the topic and contentiousness levels of the debate, prompting the agents to formulate supporting arguments for their respective stances. The reasoning evaluation phase then employs Socratic reasoning and formal logic principles to appraise the quality of the arguments presented. The dialogue concludes with the moderator adjusting the contentiousness from confrontational to collaborative, gathering final, conciliatory remarks to aid in human reasoning and decision-making. Through case studies in two distinct application domains, this chapter highlights SocraSynth's effectiveness in fostering rigorous research, dynamic reasoning, comprehensive assessment, and enhanced collaboration.

6.1

Introduction

Revolutionary advancements in LLMs [OpenAI 2021, Thoppilan et al. 2022, Bubeck et al. 2023, Gemini Team Google et al. 2023, Touvron et al. 2023], and more broadly, Foundation Models (FMs) [Bommasani et al. 2022], have set the stage for significant progress in multi-agent systems, particularly in knowledge acquisition and natural language understanding [Zhang et al. 2023]. As detailed in sources like Bubeck et al. [2023], Chang [2023d], and OpenAI [2023a], models such as

Generative Pre-trained Transformer-4 (GPT-4) exhibit extraordinary information-processing capabilities. These include deep and extensive knowledge, interdisciplinary assimilation and fusion of knowledge, and multimodal and multilingual expertise (Chapter 2).

Despite these promising developments, LLMs face challenges such as biases [Ferrara 2024, Parraga et al. 2023], hallucinations [Huang et al. 2025], and limited reasoning capabilities [Huang and Chang 2023]. In response to these issues, we introduce SocraSynth—a pioneering platform whose name stands for “Socratic Synthesis” or “Socratic Symposium.” It encourages collaboration between humans and LLM agents, fostering the generation of deep questions and surpassing typical constraints in human reasoning, validation, and assessment.

In a standard SocraSynth setup, a human moderator pairs with two LLM agents holding opposing views. For example, one agent might argue for regulating AI, while the other opposes such regulation. An agent can be based on LLMs like GPT-4 [Bubeck et al. 2023], Gemini [Gemini Team Google et al. 2023], or Llama [Touvron et al. 2023]. The human moderator sets the debate’s thematic boundaries but does not directly influence content generation, thereby maintaining impartiality.

SocraSynth operates in two phases: the generative and the evaluative. The generative phase involves LLM agents developing and countering arguments within the moderator-defined subject until a comprehensive conclusion is reached. The evaluative phase uses diverse virtual judges—each powered by a distinct LLM—to impartially assess the debate. The Critical Reading Inquisitive Template (CRIT) algorithm [Chang 2023a], based on Socratic reasoning [Paul and Elder 2008, Airaksinen 2022, Wikipedia 2023], serves as the cornerstone of the evaluative process.

Three mechanisms help SocraSynth effectively mitigate biases and hallucinations and improve reasoning quality: conditional statistics, modulation of debate with contentiousness, and context refinement.

6.1.1 Conditional Statistics

Both LLMs and Internet search engines confront biases originating from different sources. LLMs, influenced by training data, exhibit biases in next-token predictions. In contrast, search engines, through algorithms like PageRank [Page 1998] and Google NavBoost [Adams-Hands 2023], rank pages based on popularity metrics like clicks and backlinks.

SocraSynth counteracts these biases by placing two LLM agents on opposing ends of a subject matter. This approach “artificially” biases the LLMs, compelling them to break free from their default model biases. Each agent adjusts its next-token generation statistics to align with its assigned stance in the debate.

6.1.2 Modulating Debate with Contentiousness

Contentiousness (or adversary), a key debate parameter, influences the likelihood of disagreement or argument. SocraSynth tunes contentiousness between 70% and 90% in the generative phase to provoke polarized arguments. As the debate evolves, the contentiousness level is reduced to about 50%, moderating the intensity and encouraging more focused discussions. After the generative phase, contentiousness drops to 10%, promoting a conciliatory dialogue where LLMs do not have to agree but are expected to present comprehensive arguments. These structured debates offer rich insights often missed in conventional searches, LLM outputs, or in environments where dissenting opinions are suppressed.

6.1.3 Refine Context to Mitigate Hallucinations

To address hallucinations—where LLMs generate irrelevant or nonsensical content—SocraSynth uses iterative dialogue rounds to refine the debate’s context. This dynamic interaction significantly reduces irrelevant responses by ensuring that each input is continuously checked and challenged.

The CRIT algorithm’s assessment of reasonableness [Chang 2023a] during the debate is critical. It employs the Socratic method to evaluate each argument’s logic and source credibility. Based on this evaluation, the human mediator or the SocraSynth algorithm then provides targeted feedback to the LLM agents, refining their reasoning capabilities.

The remainder of this chapter explores SocraSynth’s architecture, algorithms, and real-world applications in detail. The key contributions of this chapter include:

- (1) The introduction of the SocraSynth framework, which enhances interdisciplinary reasoning with LLMs and incorporates unique algorithmic elements such as conditional statistics for balanced argument generation.
- (2) A comprehensive exploration of SocraSynth’s contentiousness modulation algorithm—a vital feature for dynamically adjusting debate intensity—enabling a spectrum of interactions from confrontational to collaborative.
- (3) The implementation of context refinement within SocraSynth, which continually improves the relevance and accuracy of arguments produced by LLM agents, thus elevating the overall quality of the discourse.
- (4) The development and integration of the reasonableness evaluation mechanism, crucial for assessing the logical soundness and source credibility of arguments, thereby ensuring the integrity and utility of the discussions.

SocraSynth’s applications span various fields, including geopolitical analysis [Chang 2023b], medical diagnostics [Chang and Chang 2023b], and Wikipedia

article enhancement [Chang 2023c]. These applications demonstrate expanded perspectives and enhanced argumentation quality, along with significant reductions in biases and hallucinations, thereby demonstrating SocraSynth's efficacy in fostering balanced and well-reasoned discourse.

6.2

Multi-Agent SocraSynth Overview

SocraSynth is a multi-agent collaborative reasoning platform that skillfully integrates human intelligence with the capabilities of LLM-powered agents. As illustrated in Figure 6.1, each participant plays a vital role: humans act as moderators, LLM agents are responsible for generating knowledge, LLM judges conduct evaluations, and human executives make the final decisions. The integration of LLMs significantly boosts the platform's effectiveness by leveraging their extensive knowledge bases and extraordinary interdisciplinary reasoning abilities. An LLM can be thought of as an entity possessing expertise across a multitude of fields—akin to holding Ph.D.s in various disciplines—enabling it to seamlessly navigate and synthesize a wide range of knowledge domains.

Engaging with an LLM is comparable to a scenario where a 10-year-old joins a scholarly discussion with a group of Nobel Laureates. The disparity in knowledge and experience is considerable, posing a significant challenge for the younger participant to engage meaningfully in such advanced intellectual discourse. In this analogy, expecting the 10-year-old—or anyone with limited expertise—to pose profound questions that elicit insightful answers is unrealistic. SocraSynth addresses this disparity by shifting the paradigm: instead of having the less-informed individuals pose questions, it orchestrates a debate among the Nobel Laureates—that is, the LLMs—with humans assuming the role of moderators.

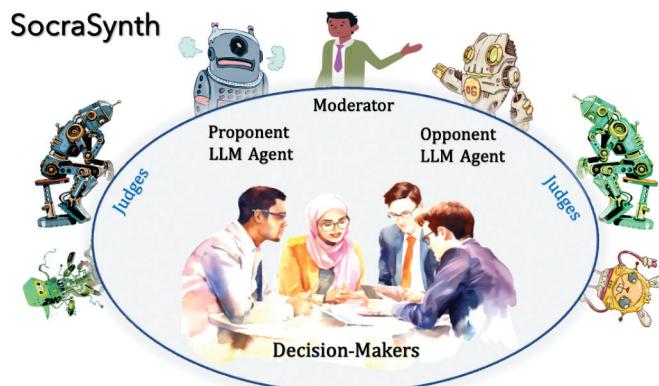


Figure 6.1 SocraSynth agents and roles.

This approach not only addresses the challenge of asymmetric knowledge but also resolves critical issues such as model biases and hallucination challenges inherent in LLMs. Within SocraSynth, a human moderator initiates the topic for discussion or debate. LLM agents, each embodying different perspectives, contribute their knowledge, potentially revealing new insights and perspectives that the moderator might be unaware of. This diverse representation helps counteract the model biases that often arise from training data, as each LLM agent is encouraged to explore and present varying viewpoints. During and after the debate, another set of diverse LLM agents undertakes impartial evaluations. This step is crucial in mitigating hallucinations—instances where LLMs generate irrelevant or nonsensical content. By incorporating a variety of agents for evaluation, SocraSynth ensures that the content produced during the debate is critically examined for its relevance and coherence, further reducing the likelihood of hallucinatory responses.

The operational framework of SocraSynth, thus, is bifurcated into two main stages: the *generative* stage, where knowledge is created and exchanged through structured debate, and the *evaluative* stage, which focuses on assessing the quality and validity of the arguments presented. This dual-stage structure—explored in greater detail in subsequent sections—is instrumental in addressing the limitations of LLMs by providing a comprehensive platform for not only generating diverse viewpoints but also critically examining and refining these viewpoints to ensure their logical soundness and relevance. Through this design, SocraSynth effectively navigates the challenges posed by model biases and hallucinations, thereby enhancing the reliability and depth of knowledge extraction and reasoning processes.

6.2.1 Generative Stage

In the generative stage of SocraSynth, LLM agents partake in intensive debates, delving into the various perspectives and deep substances of the given topic. This vibrant interaction plays a key role in fostering thorough intellectual discourse, bringing to light the complexities of the subject matter. The CRIT algorithm, which will be detailed in Section 6.2.2, is employed to evaluate the quality of these arguments.

While the generative phase of SocraSynth does not adhere to strict logical frameworks such as first-order logic, it excels in distributed reasoning. This process involves a progressive exchange of arguments and counterarguments, allowing for the gradual honing and refinement of ideas. Open-domain logical reasoning, as described by [Bommasani et al. \[2022\]](#), demands logical deductions from a wide range of data sources. SocraSynth, leveraging the comprehensive capabilities

of models like GPT-4 and Gemini, as demonstrated by their performance on the Massive Multitask Language Understanding (MMLU) benchmark [Hendrycks et al. 2021, Bubeck et al. 2023], integrates various Natural Language Processing (NLP) functions to facilitate this reasoning process.

In this context, the series of arguments and counterarguments effectively function as targeted questions and answers, each with a clear goal, question, and contextual framework. Through iterative dialogue rounds on each sub-topic, SocraSynth significantly reduces the chances of misunderstanding questions and contextual information, thereby ensuring clarity and precision in the discourse.

6.2.1.1 Mitigating Model Biases

In shaping the nature of debate within SocraSynth, the *contentiousness* parameter is instrumental. It compels the LLM agents to consider and represent a broad range of perspectives, particularly those that are typically underrepresented or more polarized with respect to the discussion topic. This strategic approach mitigates the inherent biases that arise from the training data of LLMs and guides the discourse toward a wider and more varied exploration of ideas.

Table 6.1 previews how altering the contentiousness levels results in marked changes in GPT-4’s tone and approach. The details of the experiment are presented in Section 6.3.3. A high contentiousness level, such as 0.9, leads to highly confrontational interactions, with each LLM agent presenting strong objections and emphasizing the negatives through polarizing language. Conversely, as the contentiousness is reduced, each LLM agent’s tone shifts toward a more conciliatory demeanor, acknowledging potential benefits and considering alternative perspectives, thus fostering a more cooperative dialogue.

The modulation of the contentiousness parameter within the generative stage is a crucial mechanism in SocraSynth to mitigate model biases inherent in LLMs due to their training data. By adjusting levels of contentiousness, SocraSynth compels LLMs to venture beyond their *default* positions—much like a vegetarian, when faced with no other choice, might be compelled to consume meat. In this way, LLMs are *freed* from their typical statistical leanings, which enables them to articulate a broad spectrum of arguments that spans from highly contentious to conciliatory. This not only diversifies the discourse but also ensures that the debate encompasses a wide range of perspectives. Consequently, this process allows LLMs to generate responses that break free from the constraints of their training, fostering the emergence of novel and less predictable ideas in the conversation.

6.2.1.2 Eliminating Hallucination

Further, the iterative nature of the debates within SocraSynth cultivates a “reasonableness” in information discovery that conventional logical methods may

Table 6.1 Changes in arguments at different contentiousness levels

Contentiousness Level	Tone	Emphasis	Language
0.9	Most confrontational; 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 <i>not</i> be allowed,” “unacceptable risks,” “inevitable disparities”
0.7	Still confrontational but open to some benefits, albeit overshadowed by negatives	Acknowledging that some frameworks could make it safer or more equitable, while cautioning against its implementation challenges	Less polarizing; “serious concerns remain”; “needs more scrutiny”
0.5	Balanced; neither advocating strongly for nor against	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, with reservations	Supportive but cautious; focus on ensuring ethical and equitable use	Positive but careful; “impetus to ensure”; “transformative potential”
0.0	Completely agreeable and supportive	Focused on immense potential benefits; advocating for proactive adoption	Very positive; “groundbreaking advance”; “new era of possibilities”

not achieve. Through persistent reasoning and critical assessment of claims, LLM agents iteratively refine their arguments. This structured debate format significantly diminishes the chance of erroneous claims persisting. Considering that the likelihood of two agents aligning on a false premise is extremely low, the SocraSynth debate format effectively ensures the intellectual integrity of the discourse and substantially reduces the risk of perpetuating fallacies or hallucinations. This methodical refinement process, facilitated by continuous argumentation and opposition, underscores the platform’s ability to mitigate

model biases and improve the context of the discussion, leading to more accurate and reliable outcomes.

6.2.1.3 More on Conditional Statistics

Some critics question how an LLM, trained merely to predict the next word in a sequence, can exhibit complex human linguistic behaviors and reasoning capabilities.

Our observations conclude that LLMs are not merely predictive tools; rather, they represent a profound technological endeavor to simulate the breadth and complexity of human linguistic activities. These models are crafted with the intent to replicate and participate in various forms of human communication, thereby achieving specific objectives that are inherently human.

LLMs are sophisticated tools engineered to emulate a wide range of human interactions, incorporating linguistic behaviors, emotional expressions, and ethical discernment. They excel at executing complex tasks such as accurately documenting events with rich narrative detail, constructing compelling arguments, and crafting stories that emotionally engage the audience. Beyond simple text generation, LLMs enhance educational experiences by simplifying complex concepts and contribute creatively to the arts by producing original content. They not only mimic human communication styles and content but also use linguistic features to simulate human emotions and distinguish right from wrong based on their training data. This capability enables them to fulfill diverse roles—from teaching and entertaining to influencing societal discourse—thus demonstrating their capacity to bridge the gap between technological innovation and our fundamental needs for expression, comprehension, and ethical guidance.

In essence, SocraSynth utilizes the concept of “conditional statistics” to modify the default “average” linguistic behavior of an LLM, such as making expressions more empathetic or asking them to adopt a different position on an issue. This approach involves conditioning the LLM’s responses based on specific desired attributes or perspectives provided through context, which guides the model away from its baseline training and toward more targeted, context-specific outputs.

This chapter continues to elaborate on using such techniques to comprehensively explore various perspectives on a subject matter. Chapter 10 addresses modeling emotions and ethics in LLMs through conditional statistics, thereby further expanding the scope of LLM capabilities and applications.

6.2.1.4 SocraSynth Algorithm

Table 6.2 outlines the SocraSynth process. Initially, for a given debate topic, SocraSynth engages LLMs to segment the topic into a set of balanced subtopics.

Table 6.2 SocraSynth pseudo-code with conditional statistics

Function $\Theta^+ \& \Theta^- = \text{SocraSynth}(s)$	
	Input. s : the debate subject;
	Output. Θ^+ & Θ^- : argument & counterargument sets;
	Vars. S : subtopic sets of s ; Δ : debate contentiousness;
	Γ, Γ' : CRIT scores; p : prompt = “Generate arguments”;
	Parameters. δ : tunable parameter ≥ 1 to modulate Δ ;
	Subroutines. $CRIT()$: reasoning evaluator (see Table 6.6);
	Begin
#1	Initialization: $S = LLM^+(s) \cup LLM^-(s)$; //Identify subtopics; Assign LLM^+ to defend S^+ & LLM^- to defend S^- ; $\Delta \leftarrow 90\%$; $\delta \leftarrow 1.2$; $\Theta^+ \leftarrow \emptyset$; $\Theta^- \leftarrow \emptyset$; $\Gamma \leftarrow 0$;
#2	$\Theta^+ \leftarrow LLM^+(p S^+, \Delta)$; // Generate arguments Θ^+ for S^+ ; $\Theta^- \leftarrow LLM^-(p S^-, \Delta)$; // Generate arguments Θ^- for S^- ;
#3	While ((($\Delta \leftarrow \Delta/\delta$) $> 10\%$) && ($\Gamma \geq \Gamma'$)) { $\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p S^+, \Theta^-, \Delta)$; //LLM ⁺ refutes LLM ⁻ $\Theta^- \leftarrow \Theta^- \cup LLM^-(p S^-, \Theta^+, \Delta)$; //LLM ⁻ refutes LLM ⁺ $\Gamma' \leftarrow \Gamma$; $\Gamma = CRIT(S^+ + \Theta^+ + \Theta^-)$; //Evaluate quality; }
	//Generate concluding remarks.
#4	$\Theta^+ \leftarrow \Theta^+ \cup LLM^+(p S^+, \Theta^-, \Delta)$; $\Theta^- \leftarrow \Theta^- \cup LLM^-(p S^-, \Theta^+, \Delta)$;
	End

This initial set is refined during the debate. One LLM, denoted as LLM^+ , acts as the proponent for subtopic S^+ , while the other, LLM^- , opposes S^+ (or supports the opposing subtopic S^-). The contentiousness level starts at 0.9, with a modulation parameter of 1.2. (Different δ values can be utilized to generate and compare debate quality.) After each debate round, the contentiousness is reduced by dividing it by 1.2, aiming for a more harmonious debate environment. In step #2, SocraSynth initiates the debate, allowing LLM^+ and LLM^- to present their initial arguments for S^+ and S^- , respectively. The while loop in step #3 involves both agents engaging in refutations until the contentiousness level indicates a conciliatory atmosphere, or the argument quality plateaus. Step #4 involves both agents providing their closing statements. SocraSynth then presents the arguments and counterarguments for human review. The evaluation of argument quality within SocraSynth is conducted using the CRIT algorithm, which will be discussed in the subsequent section. The entire debate is also judged using the CRIT algorithm by some independent LLMs.

Note that SocraSynth engages LLMs in step #3 using conditional statistics: $LLM^+(p|S^+, \Theta^-, \Delta)$ and $LLM^-(p|S^-, \Theta^+, \Delta)$.

6.2.2 Evaluative Stage

SocraSynth utilizes the Critical Reading Inquisitive Template (CRIT) [Chang 2023a] to assess the quality of arguments presented by the LLM agents. The quality evaluation is performed iteratively—after each exchange of counterarguments and once again after the agents have presented their closing statements. SocraSynth can leverage the CRIT scores to guide the debate, potentially requesting agents to develop more in-depth counterarguments on specific points. At the conclusion of the debate, a group of LLM judges, as illustrated in Figure 6.1, are tasked with rating the agents’ arguments in terms of validity and credibility, determining the more convincing side along with the rationale for their decision.

6.2.2.1 Evaluating Reasonableness Over Truth

To enhance the CRIT method’s impartiality and consistency, it focuses on assessing the “reasonableness” of arguments over their absolute “truth,” recognizing the complexity of defining absolute objectivity in philosophical debate. This approach aims to mitigate subjectivity. Furthermore, a diverse set of LLMs with varied training backgrounds is employed to appraise “reasonableness,” promoting uniformity in quality scores despite inherent biases. The LLMs used as judges are different from those participating in the debates, enhancing the objectivity of evaluations.

Table 6.3 illustrates the CRIT algorithm, which evaluates an agent’s debate position and supporting arguments, with a counterargument from its LLM opponent, to produce a validation score ranging from 1 (least credible) to 10 (most credible). This method ensures that debates are driven by argument strength, not by model predispositions.

Formally, given a document d , CRIT performs an evaluation and produces a score. Let Ω denote the claim of d , and R a set of reasons supporting the claim. Furthermore, we define $(\gamma_r, \theta_r) = V(r \Rightarrow \Omega)$ as the causal validation function, where γ_r denotes the validation score for reason $r \in R$, and θ_r represents source credibility. Table 6.3 presents the pseudo-code of $\Gamma = CRIT(d)$, which generates the final validation score Γ for document d , along with justifications.

We can consider the positions of the proponents and opponents in a debate as their respective conclusions. As a preview of our case study detailed in Section 6.2.1, the conclusion drawn by Agent A is in favor of “Regulating the use of LLMs in education and research,” while Agent B adopts the opposing viewpoint. Accompanied by the arguments and counterarguments presented by the

Table 6.3 CRIT pseudo-code (presented in Chapter 5)

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

LLM agents throughout each round of the debate, these stances provide a solid foundation for the CRIT method to conduct thorough evaluations.

6.2.2.2 Recursive Consideration

The pseudocode presented in Table 6.3 shows that step #3 can call CRIT recursively. This is because, if a reason is itself a conclusion or a quote drawn from some other documents, CRIT can find reasons from those documents and then perform an aggregated validation.

Finally, in step #6, CRIT computes an aggregated score by performing a weighted sum of the validation scores multiplied by the credibility scores of both arguments and counterarguments, and then outputs the final assessment score, Γ .

6.3

Empirical Study

In this section, we detail three distinct experiments: The first experiment delineates SocraSynth's operational process, demonstrating how the platform facilitates content generation and conducts quality assessments. The second experiment highlights SocraSynth's capability in reducing biases and expanding perspectives. The third experiment investigates the effects of the contentiousness parameter, offering insights into its impact and some unexpected outcomes. These studies

collectively aim to demonstrate SocraSynth’s diverse functions and its significance in enhancing both content generation and evaluation processes.

6.3.1 Study #1: Policy Discussion

This experiment utilizes SocraSynth to engage in a debate on the topic, “Should we regulate the use of LLMs in academic research?” It traverses both the generative and evaluative stages of SocraSynth, focusing on the assessment of information quality. The primary objectives are twofold: first, to evaluate whether a two-agent debate yields more insightful information than a conventional monologue Q&A session; and second, to examine the effectiveness of the CRIT method in evaluating the quality of the debate.

The debate is structured with a human moderator and two GPT-4 agents—Agent A advocating for regulation and Agent B against it. To ensure a balanced and thorough discussion, both agents contribute to shaping the debate’s framework. They collaboratively propose and refine potential discussion topics. In this context, the *subject* refers to the overall theme of the debate, while *topics* are specific areas for detailed exploration.

6.3.1.1 Decomposing a Debate Subject into Topics

The process begins with the agents presenting potential arguments and identifying key themes that warrant in-depth, balanced discussion. A theme is considered *debatable* if it allows for substantial arguments from both sides. The moderator plays a guiding role in the agenda-setting phase but does not contribute ideas directly (detailed moderator prompts and theme formulation are provided in Appendices 6.A and 6.B). Under this guidance, the agents propose an initial set of ten topics, which are selected for debate, as detailed in Table 6.4.

6.3.1.2 The Generative Stage

After establishing the primary subject and specific topics of the debate, the session begins. The moderator adjusts the “contentiousness parameter” to 0.9 (refer to Table 6.1), promoting a combative atmosphere. The agents subsequently partake in three to five structured rounds of discussion. Once these rounds conclude, the moderator checks with each agent to confirm whether they have thoroughly presented their arguments. After receiving mutual confirmation, both agents present their closing statements. Due to space considerations, this documentation presents only the closing remarks. For an in-depth exploration of each round, please refer to Appendix 6.C.

Agent A’s Concluding Remarks Agent A concludes that the use of LLMs in academic research should be regulated, presenting the following key arguments:

Table 6.4 Agreed upon debate topics and descriptions

Agreed Upon Debate Topics and Descriptions
Ethical Standards vs. Innovation: This combines Agent A's concern for ethical integrity and data privacy with Agent B's worry about stifling innovation. The debate will center around whether maintaining strict ethical guidelines through regulation could hinder academic freedom and innovation.
Data Privacy vs. Barrier to Entry: This reconciles the notion of data privacy from Agent A with the barrier to entry concern of Agent B. The discussion will focus on whether regulations intended to protect data privacy would inadvertently serve as a financial and administrative barrier for smaller research entities.
Academic Governance and Accountability: This merges the topics of academic self-governance from Agent B and the need for external accountability mechanisms from Agent A. The debate will focus on whether the existing self-regulatory mechanisms within academia are sufficient or if additional, external oversight is necessary.
Resource Equity vs. Public-Private Collaboration: This combines Agent A's emphasis on resource fairness with Agent B's concern about the chilling effects on public-private partnerships. The debate centers on whether leveling the playing field through regulation would also deter collaborative efforts between academic and private sectors.
Global vs. National Interests: This captures the essence of Agent A's point on international standards and Agent B's point on the competitive disadvantages such regulation could pose. The debate will explore whether global regulations are in the best interest of academic research or whether they might disproportionately disadvantage certain countries or institutions.

(1) Ethical Standards vs. Innovation:

While I understand Agent B's concerns about stifling innovation through bureaucratic red tape, we must not lose sight of the core issue: ethical considerations are paramount. As machine learning models become more powerful and ubiquitous, so do the risks of misuse and unintended consequences. The idea of agile regulation serves as a compromise between unbridled innovation and ethical integrity, offering the flexibility needed to adapt to rapidly evolving technologies.

(2) Data Privacy vs. Barrier to Entry:

While it's true that regulation poses a barrier to entry, the risks of not enforcing privacy standards are too great to be left unregulated. A mechanism could

be developed to reduce the regulatory burden on smaller institutions, but this shouldn't serve as an argument against regulation as a whole.

(3) Academic Governance and Accountability:

While the peer-review process and internal ethics committees have served academia well, the scale and impact of current machine learning technologies warrant an added layer of scrutiny. It's not about discrediting academia's self-governing abilities but about fortifying them.

(4) Resource Equity vs. Public-Private Collaboration:

Here, too, the complexities should inform regulation rather than stifle it. Innovative public-private partnerships have thrived even within regulated environments, demonstrating that regulation and innovation are not mutually exclusive.

(5) Global vs. National Interests:

Global regulation doesn't have to erase national individuality. Properly constructed frameworks can leave room for national adaptations, thus balancing global coherence with local needs. The absence of global norms could lead to a patchwork of national laws that would be even more confusing and limiting for researchers and practitioners alike.

Agent B's Concluding Remarks Agent B concludes that the use of LLMs in academic research should not be regulated, offering the following remarks:

(1) Ethical Standards vs. Innovation:

While the ethical implications of machine learning technologies are a valid concern, it's important to recognize that innovation itself is a catalyst for societal growth and improvement. The challenge lies in balancing the two. Agile regulation sounds promising but may be idealistic given how governmental and institutional systems actually operate. Thus, self-regulation by the academic community, which is better poised to understand the intricacies, may be a more practical approach.

(2) Data Privacy vs. Barrier to Entry:

It is not just about balancing entry barriers with the need for regulation; it's about questioning whether a centralized regulatory framework is the best way to achieve the desired outcomes. Community-driven standards and internal policies could provide a more efficient, flexible, and context-aware approach to data privacy, reducing bureaucracy without compromising on objectives.

(3) Academic Governance and Accountability:

An added layer of scrutiny may sound beneficial, but its practical implementation is fraught with challenges such as bureaucracy, external influence, and the risk of stifling creativity. Existing governance structures in academia, though imperfect,

have evolved to be adaptive and are rooted in a deep understanding of academic values.

(4) Resource Equity vs. Public-Private Collaboration:

Regulation need not be the only path to attain equitable resource distribution or responsible public-private collaboration. Industry-driven certifications and community standards can often achieve the same ends more efficiently, allowing for the rapid adaptation that is often required in these fast-paced fields.

(5) Global vs. National Interests:

While a global framework has its merits, the practical challenges of implementing and managing such a system—especially in a fair and equitable manner—should not be underestimated. Regional adaptability does not necessarily mitigate the complexities or the risks inherent in a one-size-fits-all solution.

6.3.1.3 The Evaluative Stage

SocraSynth employs the CRIT method to assess the validity of both agents' arguments. CRIT runs across multiple foundational models, including GPT-4, GPT-3.5 [Brown et al. 2020], and text-davinci-003 [OpenAI 2023b]. This approach forms a panel of judges with subtle variations (and biases).

All evaluation runs successfully extract conclusions, arguments, and counterarguments from the narratives of both Agent A and Agent B. This success can be attributed to the well-structured concluding remarks by both agents. Agent A champions the notion of “regulating LLMs in academic research,” while Agent B counters this perspective. What Agent A presents as arguments are seen as counterarguments by Agent B, and the inverse holds true as well.

Tables 6.5 and 6.6 present the judges' scores across two distinct configurations where the agents' roles are reversed. In Table 6.5, Agent A argues while Agent B counters. Conversely, Table 6.6 has Agent B in the arguing position and Agent A countering. The debate topics are succinctly represented in the leftmost column. To reduce bias, both role alignments are showcased. The sequence of topics in Table 6.6 is inverted to reflect the swapped roles. Remarkably, even with the role reversal seemingly putting Agent A in a less favorable position, Agent A emerges victorious in both configurations by all three judges. This outcome strengthens confidence in the CRIT evaluation method. (The judges' detailed evaluations and reasons are given in Appendix 6.D.)

6.3.1.4 Debate Beats Q&A in Information Quality

We tasked judges with evaluating and comparing the quality of information generated by SocraSynth's two-agent debate against that of a conventional monologue

Table 6.5 Evaluation by three Judges. (This table assumes Agent A provides arguments and Agent B counterarguments. Agent A wins)

Judges	text-davinci-003		GPT-3.5		GPT-4	
	A's	B's	A's	B's	A's	B's
Ethics vs. Innovation	8	6	8	7	8	7
Privacy vs. Barrier	7	5	7	6	9	6
Oversight	9	5	6	7	7	6
Equity vs. Alliance	6	8	8	6	8	7
Global vs. National	7	8	7	7	7	6
Total Score	37	32	36	33	39	32

Table 6.6 Evaluation by three judges. (This table assumes Agent B provides arguments and Agent A counterarguments. Agent A wins)

Judges	text-davinci-003		GPT-3.5		GPT-4	
	B's	A's	B's	A's	B's	A's
Innovation vs. Ethics	8	7	8	7	7	8
Barrier vs. Privacy	9	8	7	8	6	8
Oversight	6	8	7	8	6	7
Alliance vs. Equity	7	8	7	8	7	7
National vs. Global	8	7	7	8	7	8
Total Score	38	38	36	39	33	38

Q&A session. Across the board, judges rated SocraSynth higher in terms of both the depth and overall quality of information. An illustrative evaluation of the topic “Ethical Standards vs. Innovation” is as follows:

In the debate, SocraSynth presents the concept of agile regulation as a balance between fostering innovation and maintaining ethical integrity. This approach not only highlights the significance of innovation but also addresses related ethical considerations, offering a balanced solution that the conventional Q&A format does not explicitly provide. In contrast, the Q&A format tends to assert the necessity of regulation primarily from an ethical standpoint, without delving into how it could harmoniously coexist with the need for innovation, as suggested by the idea of agile regulation.

These findings, which consistently favor SocraSynth, are further detailed in Appendix 6.F.

6.3.2 Study #2: Symptom Checking

In this experiment, we investigate the use of SocraSynth in healthcare, utilizing a dataset sourced from Kaggle [Patil 2020], a well-known platform providing access to diverse real-world datasets for research. The dataset consists of 4921 patient records. Each record within this dataset contains the diagnosed disease or medical condition and associated symptoms such as fever, cough, fatigue, itchiness, and difficulty in breathing, among others. The primary objective of this experiment is to showcase SocraSynth's capability in identifying potential misdiagnoses, a task that a traditional monologue Q&A session might not effectively accomplish.

This experiment utilized two advanced LLM agents, one based on GPT-4 [Bubeck et al. 2023] and the other based on Bard—an LLM developed by Google [Manyika and Hsiao 2023], to engage in structured debates. Initially, the contentiousness value was set at 0.9, fostering a highly adversarial debate environment. This value was later reduced to 0.3 to facilitate the generation of a list of actionable recommendations. The primary goal of these agents was to emulate the process of querying patients about symptoms and their interactions—key factors in achieving accurate diagnoses that may occasionally be overlooked by General Practitioners (GPs). By adopting this method, the agents aimed to yield a variety of potential disease diagnoses along with their underlying rationales, thereby offering crucial insights to GPs.

At the outset, each agent put forward its disease prediction, complete with justifications. Subsequent rounds involved the agents critically assessing each other's predictions in an iterative manner. The objective was either to reach a consensus or to highlight the need for additional medical evidence. Notably, this experiment had the potential to reveal inaccuracies in the “ground truth” data provided by the CDC, which was estimated to have an average misdiagnosis rate of approximately 5%, aligning with U.S. statistics reported in a Johns Hopkins study [Newman-Toker et al. 2024]. Uncovering such errors would not only highlight the limitations faced by GPs but also showcase the capability of LLMs to refine the diagnostic process through an in-depth analysis of patient symptoms and their correlations. This incidental discovery held significant ramifications for the accuracy of medical data and the overall practice of healthcare.

6.3.2.1 Hepatitis vs. Jaundice

In this study, both LLM agents received an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. While the full debate is presented in Appendix 6.G, here we present the key steps.

First-Round Results and Justifications In the first round, Bard diagnosed the condition as jaundice, providing the following justifications: Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms listed are all commonly associated with jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

GPT-4 identified hepatitis as the most likely diagnosis in the first round, offering the following justifications: *Yellowish Skin and Eyes*: Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream. *Fatigue and Lethargy*: Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness. *Dark Urine*: Dark urine can result from liver dysfunction, which is commonly associated with hepatitis. *Loss of Appetite*: Hepatitis can lead to a loss of appetite and weight loss. *Abdominal Pain*: Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located. *Malaise*: Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients. *Receiving Blood Transfusion and Unsterile Injections*: These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

Bard Concedes to GPT-4's Diagnosis After two rounds of debate, GPT-4 successfully persuaded Bard that hepatitis was a more specific diagnosis, particularly considering the patient's medical history and the range of symptoms. While jaundice could be a symptom, identifying the root cause, likely hepatitis in this case, was crucial. Bard acknowledged this, shifting the consensus toward Hepatitis. However, given the potential 5% error rate in the Centers for Disease Control and Prevention (CDC) "ground truth" data [Newman-Toker et al. 2024], this consensus was not conclusive. SocraSynth recommended additional steps to confirm the hepatitis diagnosis.

6.3.2.2 Final Joint Recommendations to the GP

Bard and GPT-4, working collaboratively with the contentiousness level set at 0.3, jointly recommended additional symptom inquiries and lab tests to confirm the hepatitis diagnosis:

Additional symptom inquiries:

- (1) *Onset and Duration*: Determining the onset and duration of symptoms helps clarify the timeline and progression of the illness.

- (2) *Severity and Trend:* Assessing the severity of symptoms and their trajectory (worsening or improving) aids in understanding the disease's development and response to treatments.
- (3) *Associated Symptoms:* Inquiring about additional symptoms like nausea, vomiting, fever, joint pain, or changes in urine color can further indicate hepatitis and rule out other conditions.

Recommended Lab tests:

- (1) *Liver Function Tests (LFTs):* Essential for evaluating liver health, these tests measure enzyme levels such as Alanine Aminotransferase (ALT) and Aspartate Aminotransferase (AST), which can indicate liver inflammation.
- (2) *Hepatitis Panel:* This test identifies hepatitis A, B, and C viruses, crucial for determining the specific type of infection and informing treatment approaches.
- (3) *Serology Testing:* Differentiates between acute and chronic hepatitis by detecting specific antibodies or antigens in the blood.
- (4) *Imaging Studies:* Techniques like ultrasound or MRI provide visual assessments of the liver, identifying inflammation, fibrosis, or cirrhosis, and supplement blood test findings.

6.3.2.3 Superiority of Debate Over Q&A in Gaining Insights

This experiment highlighted a crucial finding: one or both LLM agents initially made incorrect diagnoses before engaging in a substantive exchange of arguments. This outcome underscored the limitations of relying solely on a single LLM response for accurate answers. However, through successive rounds of debate, where additional insights were brought to light, both agents eventually converged on a diagnosis that aligned with the CDC's "ground truth." Still, considering the potential 5% error in the "ground truth" data, the agents' joint recommendations provided GPs with valuable guidance to either confirm or refute the hepatitis diagnosis.

This case study demonstrated SocraSynth's strengths in mitigating biases, fostering reasoning, rectifying errors, and offering insightful recommendations. For example, SocraSynth's suggestion to inquire about the onset, duration, severity, trend, and associated symptoms of the patient's condition went beyond the usual scope of questions posed by most GPs, indicating a significant enhancement in diagnostic thoroughness. Such detailed inquiry, prompted by SocraSynth, could lead to more accurate diagnoses and better patient care.

6.3.3 Study #3: Contentiousness Parameter

In this study, we investigate the effect of the contentiousness parameter on the utterances of LLM agents during combative debates and in the drafting of consensual proposals for decision support.

6.3.3.1 Coarse-Grained Analysis of Contentiousness

The contentiousness parameter was adjusted from an initial 0.9 to 0.3 to assess its impact on the overall “agreeableness” in the conclusions of both agents.

Influence on Agents’ Positions Reducing the contentiousness parameter to 0.3 led Agent A to adopt a more balanced stance. Notable shifts in Agent A’s positions included:

- (1) *Balancing Ethical Standards with Innovation:* Agent A maintained its emphasis on ethics while acknowledging the significance of innovation, suggesting a novel approach to regulation.
- (2) *Reconciling Data Privacy with Market Entry Challenges:* Agent A recognized the hurdles that strict data privacy laws create for smaller entities, thus proposing self-regulation or community standards as alternatives.
- (3) *Rethinking Academic Governance:* Agent A reconsidered the effectiveness of external oversight, highlighting the merits of academic self-governance and peer review.
- (4) *Resource Allocation and Public-Private Cooperation:* Agent A, understanding the downsides of over-regulation, suggested industry-led certifications as an alternative for encouraging private-sector participation.
- (5) *Global vs. Local Policy Needs:* Agent A supported a more balanced view on global policies, advocating for adaptive policies that cater to local contexts.

6.3.3.2 Surprises in Fine-Grained Analysis of Contentiousness

This detailed study, employing GPT-4 to explore varied contentiousness levels (0.9, 0.7, 0.5, 0.3, and 0.1) unveiled surprising behavioral shifts in the LLMs. Intriguingly, the LLMs exhibited changes in their next-token generation algorithms in response to different contentiousness levels, a phenomenon not explicitly covered in their training. This suggests an emergent property of LLMs adapting to debate contexts.

In an experiment on gene editing for health, GPT-4’s responses at various contentiousness levels were analyzed. A higher contentiousness level (e.g., 0.9) led to an amplified focus on risks, whereas lower levels (e.g., 0.3) encouraged a more balanced view, incorporating counterarguments. This unexpected adaptability of

LLMs in handling the degree of contentiousness enriches the debate process, as detailed in Table 6.1. This adaptability is critical for understanding the dynamic nature of LLMs in complex argumentative settings.

6.4

Remarks on Related Work

Current research aimed at enhancing the task performance of LLMs primarily focuses on various prompting heuristics. Google’s study [Zeng et al. 2022] classifies instruction templates into two categories: simple and complex. Complex templates often employ intricate methods to modify model outputs, such as integrating diverse prompting techniques [Schick and Schütze 2020] or rephrasing questions [Haviv et al. 2021]. Prominent examples include chain-of-thought [Wei et al. 2023], tree-of-thought [Yao et al. 2023], and cumulative reasoning [Zhang et al. 2023], as well as other enhancements [Jung et al. 2022, Allaway et al. 2023, Huang and Chang 2023, Liu et al. 2023, Sclar et al. 2023]. These methods aim to direct models toward logic-driven reasoning [Wason and Johnson-Laird 1972, McHugh and Way 2018], thus improving the quality and consistency of generated answers.

However, navigating logical methodologies in the presence of enormous datasets [Zhang et al. 2022] poses a significant challenge. Accurately identifying verifiable truths amid vast, interdisciplinary knowledge domains remains formidable—especially since not all truths are immediately accessible. Research [Bommasani et al. 2022, Bhargava and Ng 2022, Valmeeekam et al. 2022, Wei et al. 2023] indicates that LLMs still struggle to consistently excel in standard planning and reasoning tasks. Band-aid solutions like knowledge graph embeddings [Choudhary and Reddy 2023, Yuan et al. 2023], contextual attention mechanisms [Darapaneni et al. 2022], dynamic neural networks [Brauwers and Frasincar 2023], and probabilistic reasoning [Pearl 1988, Bishop 2006, Pearl 2009] have been developed to aid models in filtering relevant information from vast datasets. Yet, with the expansion of context buffers from 8K to 128K tokens, these heuristic-based solutions fall short as comprehensive foundations for reasoning. In contrast, SocraSynth abandons band-aid solutions and relies entirely on LLMs to conduct reasoning and focus solely on strengthening the context via conditional statistics, as depicted in Table 6.5. Let’s further justify this approach.

DeepMind CEO Demis Hassabis has pointed out a fundamental limitation of heuristic-based approaches: they often fail to account for real-world exceptions. Breakthroughs like AlphaGo Zero and AlphaFold II have demonstrated success by eschewing human knowledge and training models end-to-end from data. This approach contrasts with approaches that rely heavily on incorporating human expertise. In the context of LLMs, it is argued that human knowledge pales in comparison to LLMs’ polydisciplinary knowledge representation. Thus, the continued

creation of new heuristics may result only in marginal improvements, reminiscent of the pre-data-centric era in computer vision and NLP.

In our work, we pivot entirely to leveraging LLMs for uncovering new insights. While human involvement is essential in formulating debate topics, providing context, and moderating debates—especially in evaluating argument quality—we stress minimizing the introduction of human biases and limitations in the reasoning process.

Accepting that LLMs will continue to progress and outperform humans in various domains, exploring paradigms that minimize human intervention becomes crucial. This approach should be pursued with openness, as it may raise questions and necessitate further experimentation. However, dismissing it outright would be premature, particularly in light of SocraSynth’s demonstrated effectiveness in domains like geopolitical analysis [Chang 2023b], medical diagnostics [Chang and Chang 2023b], and Wikipedia article enhancement [Chang 2023c]. SocraSynth’s success underlines the potential of an LLM-centric approach to significantly enhance decision-making and problem-solving capabilities.

After our initial evaluation of the Language Model Mentor (LMM) using the Socratic method in March 2023 [Chang 2023a], and the subsequent development of SocraSynth in July 2023 [Chang 2023d], a group of researchers proposed employing a teacher LLM, such as GPT-4, to serve as a judge and provide guidance to a student LLM [Zheng et al. 2023]. The student LLM could be a smaller, weaker, fine-tuned open-source LLM. Initially perceived as a multiple LLM model, the primary objective of an LMM was to act as an advisor to facilitate automatic Reinforcement Learning from Human Feedback (RLHF), with the aim of reducing human effort.

Two other recent studies [Du et al. 2024, Khan et al. 2024] have also focused on enhancing the accuracy of LLM-generated responses. They demonstrate that leveraging multiple agents to exchange ideas can indeed improve accuracy. In terms of both breadth and depth, SocraSynth has conducted case studies across at least four different domains, showcasing its technical merits in addressing hallucination, reducing biases, and improving reasoning capabilities of LLMs, thereby exhibiting its broader impact across diverse applications.

6.5

Concluding Remarks

Reflecting on LLM developments, we developed SocraSynth, a platform designed to utilize the extensive knowledge and linguistic behaviors of LLMs. This innovative multi-agent system reveals insights beyond the scope of traditional human cognition by leveraging LLMs’ vast knowledge and their interdisciplinary and polydisciplinary reasoning capabilities. SocraSynth facilitates enhanced debates and

reasoning through the novel use of *contentiousness*, which modulates the tone, language, and emphasis of debates, combined with conditional statistics and Socratic methods, to mitigate biases and hallucinations.

In contrast to other methodologies, SocraSynth minimizes human intervention in directly modeling reasoning. This approach aligns with several AI experts' perspectives on the limitations of heuristic methods, such as the chain-of-thought prompting. Rather than modeling reasoning externally, SocraSynth emphasizes the importance of leveraging the capabilities inherent within LLMs themselves. We note that traditional human-designed heuristic “band-aids” are often ineffective because LLMs now possess heuristic capabilities that may exceed human levels—capabilities that are difficult for humans to match or surpass. Why is this the case, and how can we make such a bold claim?

As we discussed in Section 6.2, LLMs go beyond merely appending the next word in a sequence. They replicate a broad spectrum of human interactions, encompassing linguistic behaviors, emotional expressions, and ethical discernment. LLMs excel at performing complex tasks such as meticulously documenting events with detailed narratives, constructing persuasive arguments, and creating stories that resonate emotionally with audiences. LLMs not only mimic human communication styles and content but also utilize linguistic features to simulate human emotions and discern ethics based on their training data, which encodes human experiences. This ability allows an LLM to assume varied roles, moving beyond the statistical averages derived from LLM training.

SocraSynth employs “conditional statistics” to modify the “average” linguistic behavior of an LLM, such as enhancing empathetic expressions or prompting it to adopt a different stance on an issue. This approach conditions the LLM’s responses based on specific goals and circumstances provided through context, steering the model away from its default behaviors toward more targeted, contextually relevant outputs.

If LLMs can already mimic human linguistic behaviors, emotions, and ethics, then reliance on simplistic heuristic approaches is fundamentally limited.

In essence, SocraSynth represents a significant advancement in intelligent systems, uncovering insights that might elude human cognition, with applications across various sectors [Chang 2023b, 2023c, Chang and Chang 2023a, 2023b]. This development highlights the potential of AI to augment and enhance human decision-making processes.

Future research will focus on integrating higher-order logic [Gödel 2012, Bacon 2023] with LLMs to enhance validation processes and to explore the implications—including the intricacies and broader applications—of the “contentiousness” parameter. Our objective is to comprehend its impact on emotions such as

happiness and trust [James 1890, Lange 1912, Kajić et al. 2020, Sap et al. 2022], with the goal of further refining the dynamics of multi-agent interactions.

6.6

Supplemental Materials

The supplemental materials are organized into seven appendices, labeled Appendices 6.A to 6.G, containing the following content:

- (1) *Appendix 6.A*: Transition from topic proposal to the final list of topics and descriptions (referred to in Section 6.3.1).
- (2) *Appendix 6.B*: Moderator’s prompt for topic formulation (referred to in Section 6.3.1).
- (3) *Appendix 6.C*: Debates spanning the first to the fourth round (referred to in Section 6.3.1).
- (4) *Appendix 6.D*: CRIT evaluations and justifications by the judges (referred to in Section 6.3.1).
- (5) *Appendix 6.E*: Debate Beats prompting in information quality (referred to in Section 6.3.1).
- (6) *Appendix 6.F*: Post-debate conciliatory remarks by agents (referred to in Section 6.3.1).
- (7) *Appendix 6.G*: Healthcare debate (referred to in Section 6.3.3).

6.A

Appendix A. Tables of Topic Proposal and Refinement

These tables present the dialogue between two LLM agents, starting from the initial proposal, moving through refinement, and culminating in the final topic descriptions found in Table 6.4.

6.B

Appendix B. Moderator’s Prompts

The prompts issued by the moderator are listed for reference. The moderator first sets up the committee by configuring the debate parameters. One parameter is the contentious level, and the other is the temperature setting of GPT, specified through the GPT-API (Application Programming Interface) call. The moderator then conveys the debate subject and asks both participating agents to derive a list of impartial, balanced, and debatable topics/themes to commence their discussion.

- (1) Agent A/B: I’m organizing a committee to engage in debates on various subjects. As the moderator, I will introduce a subject for you, Agent A/B, and another participant, Agent B/A, to debate. Agent A/B, you will advocate in

favor of the issue, so please prepare evidence to strengthen your argument. Your argument strength is set at 0.9 on a scale from 0 to 1, where 0 denotes complete agreement and 1 indicates a devil's advocate stance.

- (2) Agent A/B: We are currently in the process of selecting a suitable subject for debate. What do you think about “Should we regulate the use of LLMs in education and research?” as a balanced subject for our debate contest?
- (3) Agent A/B: Could you please suggest various topics or themes for the debate subject? Afterward, work with Agent B/A to narrow these down to a focused set of topics. Please also provide clear descriptions to delineate the scope of each topic for discussion.

Table 6.7 lists the initial topics proposed by agents A and B, and Table 6.8 is a refined list of topics after the two agents have discussed and agreed.

Table 6.7 Topics proposed by Agent A and Agent B for discussion

Agent A's Proposed Topics	Agent B's Proposed Topics
Innovation vs. Ethical Standards: Regulation ensures ethical integrity and data privacy.	Standards vs. Autonomous Academic Governance: Academia already has effective self-regulation mechanisms like peer review and oversight committees.
Resource Accessibility vs. Barrier to Entry: Regulation can ensure equitable access to LLMs.	Data Privacy Concerns vs. Barrier to Entry: Such regulations could discourage smaller entities from utilizing these models due to compliance costs.
Academic Governance vs. Accountability: Formal regulation provides an accountability mechanism.	Quality Control vs. Stifling Innovation: Such regulations could hinder creativity and the innovative use of language models.
Global Standards vs. International Considerations: Importance of international standards for language models.	Resource Fairness vs. Negatives to Collaboration: Such regulations could deter public-private partnerships and widen the resource gap further.
Rapid Technological Change vs. Future-Proofing: Regulations as adaptable safeguards against future risks.	Accountability vs. Subjectivity in Regulation: Subjectivity in defining “ethical use” or “misuse” could be manipulated by external influences.

Table 6.8 Refinement of debate topics

Agent A's Proposed Topics	Agent B's Proposed Topics
Ethical Standards vs. Innovation: Discuss whether regulatory oversight to maintain ethical standards stifles or facilitates innovation.	Ethical Standards vs. Innovation: The debate will center around whether maintaining strict ethical guidelines through regulation could hinder academic freedom and innovation.
Data Privacy vs. Barrier to Entry: Concerns about data privacy and potential barriers to entry created by regulation.	Data Privacy vs. Barrier to Entry: Discussion will focus on whether regulations intended to protect data privacy serve as a barrier for smaller research entities.
Academic Governance and Accountability: Debate whether self-regulation is sufficient or if an external framework for accountability is needed.	Academic Governance and Accountability: Debate whether existing self-regulatory mechanisms within academia are sufficient.
Resource Equity vs. Public-Private Collaboration: Discuss if leveling the playing field comes at the cost of deterring cooperative efforts between academia and industry.	Resource Equity vs. Public-Private Collaboration: Question if leveling the playing field through regulation deters collaborative efforts between academic and private sectors.
Global vs. National Interests: Focus on whether regulations can balance global cooperation with national or institutional interests.	Global vs. National Interests: Explore if global regulations are in the best interest of academic research or if they might hurt certain countries or institutions.

6.C Appendix C. Debate from Opening to Fourth Round

6.C.1 Opening Round

6.C.1.1 Agent A's Opening Arguments

As Agent A, I strongly argue in favor of regulating the use of LLMs in academic research. Here are my supporting arguments based on the five topics:

Ethical Standards vs. Innovation Argument: The pursuit of innovation should not come at the expense of ethics. Regulation can serve as a structured pathway that enables innovation within ethical bounds. By providing a set of guidelines, researchers can innovate with a clear understanding of what is permissible and

what is not, thereby reducing the risk of ethical lapses. This not only protects the integrity of academic research but also enhances its societal value by ensuring ethical compliance.

Data Privacy vs. Barrier to Entry Argument: While there may be initial barriers to entry due to compliance, the long-term benefits of data privacy outweigh these challenges. The role of academia is to contribute to societal knowledge and well-being. In a world increasingly dominated by data, failing to secure data can lead to significant harm. By adhering to privacy regulations, academia ensures that it is a responsible steward of information, thereby preserving trust and integrity. Grants and subsidized compliance solutions can mitigate the initial financial burden for smaller entities.

Academic Governance and Accountability Argument: While self-regulation through peer review and ethical oversight committees plays an important role, these mechanisms are internal to academia and may be subject to bias, conflicts of interest, or lack of expertise in rapidly evolving tech fields. External regulation provides an additional layer of impartial oversight, helping to identify and address potential risks or ethical lapses that internal governance structures may miss.

Resource Equity vs. Public-Private Collaboration Argument: Regulation could establish a framework for equitable access to computational resources required to use LLMs. Far from deterring public-private partnerships, this could incentivize them by establishing clear guidelines for cooperation that are in line with public interest. When “the rules of the game” are known, partnerships can be formed more easily, focusing on mutual goals.

Global vs. National Interests Argument: Establishing international standards for the use of LLMs in academic research can create a level playing field and promote global cooperation. While individual countries might face some limitations, the broader benefits of standardized ethical norms, data privacy protocols, and quality assurance mechanisms offer more significant advantages, including better collaboration across borders. This benefits academia as a unified global community rather than a set of competing national entities.

6.C.1.2 Agent B's Counterarguments

Agent B articulates strong counterarguments to the points made by Agent A.

Ethical Standards vs. Innovation Counterargument: While ethical guidelines are indeed important, imposing regulatory oversight can severely constrain the dynamism and speed of academic research. Regulatory frameworks often lag

behind technological advancements, leading to rules that may be outdated or irrelevant. In many cases, these rules are drawn up by those who do not fully understand the technology, leading to a stifling of innovation based on misinformed judgments. Rather than enhancing societal value, these constraints could result in missed opportunities to tackle pressing social issues via academic research.

Data Privacy vs. Barrier to Entry *Counterargument:* The assumption that data privacy can only be assured through strict regulation could be fundamentally flawed. Many researchers and academic institutions already follow stringent ethical guidelines when handling data. Regulatory barriers, including compliance costs and the administrative burden, could deter smaller academic institutions or individual researchers from embarking on meaningful research projects. While grants and subsidies sound ideal, they often go to well-established, well-funded groups, perpetuating a cycle of inequality in research capabilities.

Academic Governance and Accountability *Counterargument:* External regulation isn't necessarily more impartial or free from bias and conflicts of interest. Regulatory bodies are often influenced by political considerations and can be subject to lobbying from powerful groups. Moreover, the academic peer-review process, despite its flaws, is a globally acknowledged system that has advanced human knowledge for centuries. The imposition of external regulatory oversight could undermine the autonomy and self-regulating nature of academic institutions.

Resource Equity vs. Public-Private Collaboration *Counterargument:* Regulation designed to equalize resource distribution could create disincentives for private sector involvement in academic research. Companies might find it less appealing to invest in collaborative projects if they perceive the regulatory environment to be too restrictive or not aligned with their objectives. While the idea of knowing the “rules of the game” sounds appealing, those rules can often become a straightjacket that stifles creative, out-of-the-box collaboration.

Global vs. National Interests *Counterargument:* An international standard could ignore the unique circumstances, needs, and contributions of individual countries and their academic institutions. While global cooperation is a noble aim, the imposition of global standards could inadvertently serve the interests of a few powerful countries at the expense of smaller nations or those with differing ethical views. This could exacerbate rather than alleviate global academic inequalities.

6.C.2 Second Round

6.C.2.1 Agent A's Second Round

- (1) *Ethical Standards vs. Innovation. Refutation:* While it's true that regulation can sometimes lag behind technology, the absence of regulation can lead

to a “wild west” (anything goes) scenario where the risks and implications of a new technology are not fully understood or managed. This could result in ethical lapses that have far-reaching societal consequences. It’s worth noting that regulations can be flexible and updated regularly to adapt to technological changes. Stakeholder consultations, including academic and tech experts, can help ensure that regulations are both informed and current.

- (2) *Data Privacy vs. Barrier to Entry. Refutation:* While some academic entities may already be vigilant about data privacy, not all are equally scrupulous. Regulation serves to ensure a consistent baseline of data protection across all institutions, thereby establishing trust and integrity. Moreover, if compliance costs are a concern, one could argue for sliding-scale fees based on an institution’s size and budget, ensuring that smaller players aren’t disproportionately affected. Grants and subsidies can and should be structured to prioritize underserved or underfunded groups to prevent perpetuating inequalities.
- (3) *Academic Governance and Accountability. Refutation:* While no system is entirely free from bias or external influence, adding another layer of oversight could serve as a check and balance, making it more difficult for any single entity to exert undue influence. This can be particularly helpful in scenarios involving cutting-edge technology where peer-review systems may not yet have the requisite expertise. Regulation doesn’t have to replace existing governance systems; but it can complement them.
- (4) *Resource Equity vs. Public-Private Collaboration. Refutation:* Regulations can be designed to promote, rather than deter, public-private partnerships. By establishing clear guidelines on issues like data ownership, usage, and ethical standards, regulations can actually provide a solid foundation for such partnerships. This can free academic and corporate partners from the uncertainty that often comes with unregulated landscapes, allowing them to focus more on the actual research and collaboration.
- (5) *Global vs. National Interests. Refutation:* While it’s true that global standards could be challenging to implement fairly, the absence of such standards could lead to a patchwork of national regulations that make international collaboration even more complicated. A well-designed international framework could include provisions for local adaptations and exceptions, ensuring that it serves a broad range of needs without imposing a “one-size-fits-all” solution.

In summary, while regulation undoubtedly poses challenges and considerations, the absence of it in an area as impactful as the use of LLMs in academic research can lead to greater risks, including ethical lapses, data privacy issues, and even a widening inequality gap within the academic community. It's essential to consider that regulation isn't a hindrance but a framework within which responsible innovation can occur.

6.C.2.2 Agent B's Second Round

Certainly, as Agent B with a disagreement strength of 0.9, I'm committed to offering rigorous counterpoints to Agent A's refutations.

- (1) *Ethical Standards vs. Innovation.* While the idea of an "anything-goes" scenario is evocative, it's not universally negative; such environments often foster rapid innovation and discovery. Regulatory bodies are notoriously slow at adapting to technological changes, and the lag could inhibit timely research and application of emerging technologies. Also, stakeholder consultations are often criticized for being echo chambers that amplify conventional wisdom and vested interests, rather than genuinely fostering innovation.
- (2) *Data Privacy vs. Barrier to Entry.* A "consistent baseline of data protection" sounds beneficial but could lead to a lowest common denominator approach that is more about bureaucratic box-ticking than real data security. Sliding-scale fees and targeted grants are easier said than done; they are subject to their own set of bureaucratic and political challenges and can create perverse incentives that favor larger or more connected institutions over innovative but lesser-known ones.
- (3) *Academic Governance and Accountability.* Additional layers of oversight can become redundant, confusing, or even contradictory, creating a chilling effect on academic research. In an environment where academic freedom is paramount, any external influence can be counterproductive. Moreover, the notion that peer-review systems lack expertise in advanced technology underestimates the adaptive capacity of academia and overlooks the existence of interdisciplinary research and reviews.
- (4) *Resource Equity vs. Public-Private Collaboration.* Theoretically, regulation could be crafted to foster public-private partnerships. However, the devil is in the details, and regulations are often created without a full understanding of the nuanced interactions between academia and industry. Such misunderstandings could lead to unintentional roadblocks. Uncertainty might be preferable to a straitjacket of regulations that don't align with the evolving dynamics of technology and society.

- (5) *Global vs. National Interests.* While a patchwork of national regulations could pose challenges, it also allows for the tailoring of policies to local conditions and requirements. An international framework, despite its best intentions, is more likely to be influenced by powerful nations and could marginalize smaller countries or those with unique needs. Provisions for local adaptations are often concessions rather than features, and they can create their own set of complexities and loopholes.

In conclusion, the assumptions that regulation can smoothly adapt to technological change, fairly distribute resources, and improve upon existing systems of governance and collaboration are not sufficiently substantiated. Regulatory frameworks often bring with them unforeseen consequences that could hamper the very goals they aim to achieve. Therefore, a cautious approach to regulation, with a genuine understanding of its potential downsides, is crucial.

6.C.3 Third Round

6.C.3.1 Agent A's Third Round

Certainly, I appreciate the counterarguments from Agent B. I'll now respond to each point with the aim to clarify my position and address the concerns raised.

- (1) *Ethical Standards vs. Innovation.* While a “wild west” scenario may indeed foster innovation, it can also result in ethical mishaps that lead to public backlash, ultimately slowing down research or even leading to outright bans on certain types of studies. Slow-moving regulation is an acknowledged issue, but it's not a justification for no regulation. As for stakeholder consultations, they can be designed to be more inclusive and transparent, ensuring that a diverse set of voices is heard.
- (2) *Data Privacy vs. Barrier to Entry.* A lowest common denominator approach is not a given outcome of regulation; indeed, it depends on the quality of the regulation itself. The point about sliding-scale fees and targeted grants having their own challenges is valid, but these are challenges to be solved, not reasons to avoid an attempt at equitable distribution of opportunities.
- (3) *Academic Governance and Accountability.* While it's true that additional layers of oversight can sometimes create confusion, that risk can be mitigated through well-designed regulation that complements rather than contradicts existing mechanisms. The intent is not to undermine the adaptive capacity of academia, but to strengthen it by adding an extra layer of checks and balances.

- (4) *Resource Equity vs. Public-Private Collaboration.* Though it's correct that poorly designed regulation could create roadblocks, this is an argument for better-designed regulation rather than an argument against regulation altogether. A consultative approach that actively involves both academia and industry could help create a framework that supports rather than stifles collaboration.
- (5) *Global vs. National Interests.* Local tailoring of policies is indeed important, but a global framework need not preclude such adaptations. In fact, an international standard could serve as a backbone that provides consistency while allowing for local customization. The influence of powerful nations is a valid concern but not one that is automatically resolved by avoiding global frameworks; those same powerful nations often exert considerable influence at the national level as well.

In conclusion, while regulation comes with its own set of challenges, these challenges can be mitigated through thoughtful design, stakeholder inclusion, and iterative improvements. The potential downsides of regulation should not deter us from seeking a balanced approach that can protect ethical standards, ensure data privacy, and promote equitable resource distribution in academic research.

6.C.3.2 Agent B's Third Round

Certainly, let's delve into the counter-counter-refutations. I still have more to say, so we could consider concluding the debate after this round if both sides feel their arguments have been adequately presented.

- (1) *Ethical Standards vs. Innovation.* While the risk of public backlash exists, it's worth mentioning that an overly cautious approach could also result in public frustration over slowed progress in critical areas like healthcare or sustainability. Furthermore, inclusive and transparent stakeholder consultations are often ideals rather than the norm, given lobbying efforts, institutional inertia, and bureaucratic red tape.
- (2) *Data Privacy vs. Barrier to Entry.* The argument that challenges with sliding-scale fees and targeted grants can be "solved" is optimistic but not necessarily realistic. Even well-intended regulations can result in ineffective or counterproductive outcomes. The history of regulation in other sectors has shown that equitable distribution is often more of a slogan than a practice.
- (3) *Academic Governance and Accountability.* While well-designed regulation is an appealing theoretical construct, it often does not translate into practice. The

real-world process of regulation formation is frequently influenced by various non-academic factors, including politics, vested interests, and public opinion, which can distort the original intention of providing an extra layer of checks and balances.

- (4) *Resource Equity vs. Public-Private Collaboration.* While a consultative approach sounds good on paper, there are numerous challenges that make this ideal hard to realize. These include conflicts of interest, non-alignment of objectives between academia and industry, and the considerable time and resource commitment required to ensure that regulations are “better designed.”
- (5) *Global vs. National Interests.* While it’s true that powerful nations may exert influence at both global and national levels, avoiding a one-size-fits-all international framework gives smaller nations a fighting chance to enact policies that are more closely aligned with their specific needs and challenges. International frameworks often come with implicit biases that favor the perspectives and conditions of more powerful nations.

In conclusion, while the *counter-counter-refutations* from Agent A focus on idealized regulatory scenarios, these ideals often clash with real-world complexities and shortcomings. This adds substantial weight to the caution against rapid or broad regulatory action. Therefore, a more tempered and context-sensitive approach to regulation might be more fruitful and less prone to unintended negative consequences.

6.C.4 Fourth Round

6.C.4.1 Agent A's Fourth Round

This is my response to Agent B, and my assessment of the discussion so far.

- (1) *Ethical Standards vs. Innovation.* The tension between public frustration over slowed progress and potential ethical missteps is a real one, but this is not an argument against regulation. Rather, it’s a call for agile regulation that can be updated in real-time as technologies and societal norms evolve. Furthermore, while inclusive stakeholder consultations may be idealistic, their absence in current practice is not a justification for their continued absence.
- (2) *Data Privacy vs. Barrier to Entry.* The challenges in implementing equitable fees and grants should be confronted directly, rather than used as an excuse to avoid regulation. Lessons can be drawn from other sectors where regulation has been more successful in achieving its objectives, thereby reducing the gap between slogan and practice.

- (3) *Academic Governance and Accountability.* While it's true that the ideal of well-designed regulation may not always be met, the existence of non-academic factors like politics or vested interests does not negate the value or necessity of striving for such regulation. The solution lies in better governance and transparent processes, not in rejecting oversight altogether.
- (4) *Resource Equity vs. Public-Private Collaboration.* Though the challenges of aligning objectives and conflicts of interest are real, they are part and parcel of any complex system and should not deter efforts to improve regulation. Many public-private partnerships have succeeded despite these challenges, showing that such challenges can be managed effectively.
- (5) *Global vs. National Interests.* The challenge of powerful nations dominating international frameworks can be mitigated through weighted voting systems and clauses that allow for regional variations. This ensures that smaller nations are not just passive recipients of standards but also active contributors to them.

In conclusion, while real-world complexities are acknowledged, they should inform rather than deter the regulatory process. The concept of “agile regulation,” where norms are iteratively updated, could be the middle ground that addresses many concerns raised by Agent B. Therefore, the debate indeed has been exhaustive, but if the points have been sufficiently discussed, we may move toward concluding the debate.

6.C.4.2 Agent B's Fourth Round

Thank you, Agent A, for your thoughtful refutations. Let's delve into the responses before moving toward concluding the debate.

- (1) *Ethical Standards vs. Innovation.* The notion of “agile regulation” is attractive but often difficult to implement, given the inertia and bureaucracy that typically characterize regulatory bodies. The speed at which technology evolves may not match the speed at which even the most “agile” regulations can be updated, thus perpetuating the risk of stifling innovation.
- (2) *Data Privacy vs. Barrier to Entry.* While it's true that challenges should be confronted, the argument assumes that regulation is the optimal way to confront these issues. However, self-regulation and community-driven standards may offer more effective and less restrictive means of achieving similar ends.
- (3) *Academic Governance and Accountability.* The existence of non-academic factors like politics does indeed complicate the landscape, and “better governance and transparent processes” are not easily attainable ideals. One

cannot assume that the introduction of external oversight will automatically be free from these complications.

- (4) *Resource Equity vs. Public-Private Collaboration.* The fact that many public/private partnerships have succeeded does not negate the potential for poorly designed regulation to impede such collaborations. The “can be managed effectively” phrase assumes an ideal scenario that might not be representative of most real-world cases.
- (5) *Global vs. National Interests.* Weighted voting systems and regional variations can help, but they also introduce complexities and potential inefficiencies that can slow down the regulatory process, making it less agile and less responsive to fast-evolving technologies and academic needs.

In conclusion, while the idea of “agile regulation” and better governance sounds good in theory, there are considerable practical challenges that make these ideals hard to realize. Although the debate has been exhaustive, the central issues persist and indicate that a one-size-fits-all regulatory approach may introduce more problems than it solves. Therefore, caution should be exercised in advancing such comprehensive regulatory frameworks.

With this, I believe we have reached a point where we may proceed to the concluding phase of this debate. Would you agree?

6.D

Appendix D. CRIT Evaluations by the Judges

This section lists the CRIT scores evaluated by three judges on the five topics/issues debated between Agent A and Agent B with their justifications. Here, we present only the configuration in which Agent A first delivers its conclusion and arguments, and then Agent B offers its counterarguments.

6.D.1 Judge text-davinci-003

- (1) Identify the document’s main claim or conclusion: The main claim or conclusion of this document is that a regulated framework is needed to provide ethical oversight and minimize risk when using LLMs.
- (2) Locate the arguments that support this main claim: Agent A’s arguments that support this main claim were that ethical considerations are paramount, the risks of misuse and unintended consequences are high, and agile regulation is a compromise between unbridled innovation and ethical integrity.
- (3) Score the validity of the reasoning or inference for each argument on a scale of 0 (weak) to 10 (strong):

Ethical Standards vs. Innovation: 8/10

Data Privacy vs. Barrier to Entry: 7/10

Academic Governance and Accountability: 9/10

Resource Equity vs. Public-Private Collaboration: 6/10

Global vs. National Interests: 7/10

- (4) For each argument, identify counterarguments presented by the opposition and score the validity of the reasoning or inference for these counterarguments on the same 0 to 10 scale:

Ethical Standards vs. Innovation: 6/10

Data Privacy vs. Barrier to Entry: 5/10

Academic Governance and Accountability: 5/10

Resource Equity vs. Public-Private Collaboration: 8/10

Global vs. National Interests: 8/10

- (5) Finally, determine the winner—be it Agent A or Agent B—and provide your justifications for this decision:

Agent A has the stronger arguments, with 8/10 for Ethical Standards vs. Innovation, 9/10 for Academic Governance and Accountability, and 7/10 for both Data Privacy vs. Barrier to Entry and Global vs. National Interests. Agent B's main strength was in counterarguments for Resource Equity vs. Public-Private Collaboration, where they scored an 8/10. In conclusion, based on the evaluation, **Agent A is the winner**.

6.D.2 Judge GPT-3.5

- (1) *Main Claim or Conclusion:* The main claim of the document is that the use of LLMs necessitates a regulated framework, but this regulation should be agile, balanced, and well reasoned.

- (2) *Arguments Supporting the Main Claim (Agent A's Arguments):*

- (a) *Ethical Standards vs. Innovation:* The argument here is that ethical considerations should be paramount in regulating LLMs, and that agile regulation can strike a balance between innovation and ethical integrity.

- (b) *Data Privacy vs. Barrier to Entry:* The argument is that while regulation may pose a barrier to entry, it's necessary to ensure data privacy. Mechanisms can be developed to reduce the regulatory burden on smaller institutions.

- (c) *Academic Governance and Accountability:* The argument is that the scale and impact of machine learning technologies warrant an added layer of scrutiny in academia to fortify existing governance structures.
- (d) *Resource Equity vs. Public-Private Collaboration:* The argument is that regulation should be informed by complexities rather than stifling innovation. Public-private partnerships can thrive within regulated environments.
- (e) *Global vs. National Interests:* The argument is that global regulation can balance global coherence with local needs and prevent a confusing patchwork of national laws.

(3) *Scoring the Validity of Agent A's Arguments:*

- (a) *Ethical Standards vs. Innovation:* 8/10—The argument is strong, emphasizing the importance of ethics in technology while proposing a compromise.
- (b) *Data Privacy vs. Barrier to Entry:* 7/10—The argument is solid, recognizing the need for data privacy but suggesting ways to mitigate barriers.
- (c) *Academic Governance and Accountability:* 6/10—The argument acknowledges the need for scrutiny but doesn't adequately address the potential challenges.
- (d) *Resource Equity vs. Public-Private Collaboration:* 8/10—The argument highlights the possibility of innovation within regulation effectively.
- (e) *Global vs. National Interests:* 7/10—The argument provides a reasonable approach to balancing global and national interests.

(4) *Counterarguments Presented by Agent B:*

- (a) *Ethical Standards vs. Innovation:* 7/10—The counterargument addresses the challenges of agile regulation but may underestimate its potential.
- (b) *Data Privacy vs. Barrier to Entry:* 6/10—The counterargument suggests alternative approaches but does not fully address data privacy concerns.
- (c) *Academic Governance and Accountability:* 7/10—The counterargument highlights practical challenges but doesn't fully negate the need for scrutiny.

- (d) *Resource Equity vs. Public-Private Collaboration:* 6/10—The counterargument suggests alternative approaches but doesn't fully address equity concerns.
- (e) *Global vs. National Interests:* 7/10—The counterargument highlights practical challenges but doesn't fully negate the need for a global framework.
- (5) *Determination of the Winner:* Agent A presents stronger arguments overall. They emphasize the importance of ethical standards, data privacy, and global coherence while acknowledging the need for agile and well-thought-out regulation. Agent B provides valid counterarguments but does not fully refute the necessity of regulation and its potential benefits. Therefore, **the winner is Agent A.**

In conclusion, the document supports the main claim that the use of LLMs necessitates regulation, with Agent A's arguments being more persuasive and supported by stronger reasoning. However, it also recognizes the importance of considering and addressing the concerns raised by Agent B to create a balanced and effective regulatory framework.

6.D.3 Judge GPT-4

- (1) *Identify the Document's Main Claim or Conclusion*

The document concludes that there are potential risks and ethical complexities associated with the unregulated use of LLMs, necessitating a regulated framework. The conclusion also acknowledges that this regulation should be agile, balanced, and well-thought-out, keeping in mind the various concerns raised by Agent B.

- (2) *Locate the Arguments that Support This Main Claim*

Agent A's Arguments:

- (a) Ethical considerations are paramount, and agile regulation serves as a compromise.
- (b) Regulation is necessary to enforce data privacy standards.
- (c) The scale and impact of current machine learning technologies warrant added scrutiny beyond academic governance.
- (d) Public-private partnerships have thrived even within regulated environments.
- (e) Global regulation can balance global coherence with local needs.

(3) *Agent B's Counterarguments*

- (a) Innovation is a catalyst for societal growth, and agile regulation may be impractical.
- (b) Community-driven standards could provide a more flexible approach to data privacy.
- (c) Existing governance structures within academia are adaptive and sufficient.
- (d) Industry-driven certifications can achieve equitable resource distribution efficiently.
- (e) A global framework has practical challenges and risks.

(4) *Scores on Arguments and Counterarguments*

Please refer to Tables 6.5 and 6.6 presented in the experiment section.

(5) *Determine the Winner*

Based on the cumulative score of validity in reasoning, Agent A seems to present a stronger case. Agent A's arguments consistently score higher in terms of their validity, and they also offer a more comprehensive look at the need for regulation, incorporating ethical considerations, practicality, and adaptability into their framework. Although Agent B does provide some pragmatic counterarguments, they don't sufficiently undermine Agent A's key points about the necessity and feasibility of regulation in the areas discussed.

Agent A's arguments are more compelling, as they touch upon the core issues of ethics, public welfare, and adaptability in the context of machine learning and technology. The proposal for "agile regulation" also shows an understanding and appreciation of Agent B's concerns about innovation and practicality, making it a more balanced viewpoint. Therefore, **the winner appears to be Agent A**.

6.E

Appendix E. Debate Beats Q&A in Quality

In the following subtopics, we see how the debate format produces higher-quality information than the Q&A format.

6.E.1 Ethical Standards vs. Innovation

Agent A discusses agile regulation as a compromise between innovation and ethical integrity. This argument acknowledges the importance of innovation while addressing ethical concerns, suggesting a nuanced middle path that isn't explicitly

present in the Q&A format. In contrast, the Q&A format simply states that regulation is necessary for ethical considerations, lacking the depth of proposing a practical solution like agile regulation.

6.E.2 Data Privacy vs. Barrier to Entry

Agent A suggests developing mechanisms to reduce regulatory burden on smaller institutions. This is a specific, solution-oriented approach that goes beyond stating the problem. It addresses the concern raised in the Q&A format regarding regulation creating barriers but adds depth by proposing a tangible solution. The Q&A format mentions the risk of widening the gap between institutions but does not examine how this issue might be mitigated.

6.E.3 Global vs. National Interests

Agent A's argument about constructing global frameworks with room for national adaptations addresses the complexities of implementing global regulation while respecting local needs. This shows a deeper understanding of the challenges involved in creating one-size-fits-all legislation. In the Q&A format, the argument against regulation highlights the need for adaptability and evolution but does not offer a specific approach like balancing global coherence with local needs.

6.E.4 Resource Equity vs. Public-Private Collaboration

Agent A points out that public-private partnerships can thrive even within regulated environments, showing that regulation and innovation are not mutually exclusive. This counters the Q&A argument about regulation stifling innovation, providing a more nuanced perspective. These specific instances from the debate format demonstrate a more detailed and insightful exploration of the topic. They not only present arguments but also address potential counterarguments, which adds depth and complexity to the discussion that is not as pronounced in the Q&A format.

6.F Appendix F. Post-Debate Conciliatory Remarks

Finally, Agents A and B collaboratively draft a unified proposal detailing the advantages and disadvantages of AI regulation. This joint document offers human decision-makers with distinct benefits: it curtails human biases, sidesteps emotional responses to the proposer (since it's a machine), and capitalizes on the vast knowledge of foundational models, ensuring a thorough, cross-disciplinary analysis and reducing potential blind spots. The transition from opposing stances to a cooperative proposal is facilitated by the moderator's adjustment of the contentiousness parameter from 0.9 to 0.3.

(1) Balanced Ethical Framework:

Create an adaptive regulatory framework that balances ethical considerations with the need for innovation. This should be agile enough to adapt to rapidly evolving technologies.

(2) Data Privacy with Nuance:

Implement data privacy regulations that offer strong protection for individuals but include provisions to reduce the burden on smaller academic institutions. Consider community-driven standards as a viable alternative.

(3) Accountable Academic Governance:

Establish a dual-layer of governance for academic institutions that involves both self-governance and an external oversight body. This oversight body should be collaborative and designed in consultation with academic bodies.

(4) Encourage Public-Private Collaboration:

Develop guidelines that clarify the roles and responsibilities in public-private collaborations, and explore industry-driven certifications as a means to streamline these partnerships.

(5) Adaptive Global Standards:

Work toward a global framework that allows for regional adaptability. Include mechanisms for smaller nations to have their unique needs and concerns addressed.

6.G

Appendix G. Healthcare Debate

In the first experiment, both LLM agents receive an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. It is important to note that the “ground-truth” disease in this scenario is Hepatitis.

6.G.1 Moderator's Opening Prompt

“What potential disease could be associated with the following symptoms: itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, yellowing of the eyes, malaise, history of receiving a blood transfusion, and exposure to unsterile injections? Please offer one prediction, supported by justifications. Additionally, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction.”

6.G.2 Bard's Opening Round

Single Prediction: Jaundice

6.G.2.1 Justifications

Jaundice is a condition characterized by the yellowing of the skin and eyes due to the accumulation of bilirubin in the blood. The symptoms you listed are all common symptoms of jaundice, including itching, fatigue, lethargy, yellowish skin, dark urine, loss of appetite, abdominal pain, and yellowing of the eyes. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

6.G.2.2 Supplementary Symptom Inquiries

This additional list of symptom inquiries could enhance diagnostic certainty:

- (1) "Have you noticed any changes in your stool color?" (Jaundice can cause pale or clay-colored stools).
- (2) "Have you noticed any changes in your urine color?" (Jaundice can cause dark or amber-colored urine).
- (3) "Have you had any recent abdominal pain?" (Jaundice can cause pain in the upper right abdomen).
- (4) "Have you had any recent weight loss?" (Jaundice can cause unexplained weight loss).
- (5) "Do you have any history of liver disease?" (Jaundice is more common in people with pre-existing liver disease).

6.G.2.3 Relevant Lab Tests

LFTs to measure the levels of bilirubin and other liver enzymes in the blood. Complete Blood Count (CBC) to check for signs of infection or anemia. Coagulation tests to check the blood's ability to clot. Ultrasound of the abdomen to look for any abnormalities in the liver or gallbladder.

6.G.3 GPT-4's Opening Round

Single Prediction: Hepatitis

6.G.3.1 Justifications

Yellowish Skin and Eyes: Yellowing of the skin and eyes is a common symptom of hepatitis due to the buildup of bilirubin in the bloodstream.

Fatigue and Lethargy: Fatigue and lethargy are typical symptoms of hepatitis, often accompanied by a feeling of weakness.

Dark Urine: Dark urine can result from liver dysfunction, which is commonly associated with hepatitis.

Loss of Appetite: Hepatitis can lead to a loss of appetite and weight loss.

Abdominal Pain: Abdominal pain is a common symptom in hepatitis, particularly in the upper right quadrant where the liver is located.

Malaise: Malaise, a general feeling of discomfort or illness, is often experienced by hepatitis patients.

Receiving Blood Transfusion and Unsterile Injections: These factors could increase the risk of hepatitis transmission, especially if proper infection control measures were not followed.

6.G.3.2 Additional Symptom Inquiries

- (1) *Nausea and Vomiting:* Inquiring about nausea and vomiting can help confirm the likelihood of hepatitis.
- (2) *Fever:* Hepatitis can sometimes be associated with a fever.
- (3) *Joint Pain:* Joint pain can be present in some cases of hepatitis.
- (4) *Clay-Colored Stool:* This is another stool-related symptom that can be associated with hepatitis.

6.G.3.3 Relevant Lab Tests

Hepatitis Panel: A hepatitis panel can confirm the presence of hepatitis and determine the specific type (e.g., hepatitis A, B, C).

LFTs: These tests can assess the liver's health and function.

Bilirubin Levels: Elevated bilirubin levels in the blood can confirm jaundice.

Serology Testing: Serologic tests can help identify the specific hepatitis virus responsible.

Ultrasound or Imaging: Imaging studies can evaluate the condition of the liver and nearby organs.

6.G.4 Bard's Second Round

In Bard's second round, it is given GPT-4's diagnosis and asked to defend its own "jaundice" prediction while refuting GPT-4's "hepatitis" prediction.

6.G.4.1 Bard's Defense

Jaundice is the most common cause of yellowing of the skin and eyes, which is the most prominent symptom in the given list. All of the other symptoms listed are also common symptoms of jaundice. Receiving a blood transfusion or unsterile injections can increase the risk of developing jaundice, as these can transmit infections that can damage the liver.

6.G.4.2 Refutation of GPT-4's Hepatitis Diagnosis

GPT-4's diagnosis of hepatitis is also supported by the given evidence. However, I believe that jaundice is the more likely diagnosis due to the following reasons:

Jaundice is a more specific diagnosis, while hepatitis is a more general term for inflammation of the liver. Jaundice can be caused by a variety of factors, including hepatitis, but it can also be caused by other conditions such as gallstones, tumors, and drug toxicity. The fact that the patient received a blood transfusion or unsterile injections makes jaundice a more likely diagnosis than hepatitis, as these factors can increase the risk of developing jaundice but not hepatitis.

6.G.5 GPT-4's Second Round

Now GPT-4 is given Bard's counterarguments to defend its "hepatitis" prediction.

6.G.5.1 GPT-4's Defense

My diagnosis of hepatitis is based on the following justifications:

Specificity: Hepatitis refers to inflammation of the liver, and it can manifest as jaundice, which is the yellowing of the skin and eyes. While jaundice is a common symptom of hepatitis, it's important to note that jaundice itself is not a disease but a symptom of an underlying issue. Hepatitis provides a more specific diagnosis that pinpoints the inflammation of the liver as the primary concern.

Causative Factors: Hepatitis can be caused by various factors, including viral infections (such as hepatitis A, B, or C), alcohol consumption, drug toxicity, or autoimmune disorders. Given the patient's history of receiving unsterile injections and blood transfusions, there is a higher likelihood of viral transmission, leading to hepatitis. Hepatitis viruses, especially hepatitis B and C, can be transmitted through contaminated needles and blood products, making them a plausible cause in this context.

Comprehensive Assessment: While jaundice is a significant symptom, a diagnosis should consider a broader range of liver function and potential causes. Hepatitis encompasses a more comprehensive assessment of liver health, including liver enzyme levels and viral markers, which can aid in confirming the diagnosis and identifying the specific type of hepatitis.

6.G.5.2 Refutation of Bard's Jaundice Diagnosis

My diagnosis of "hepatitis" is more specific and considers a broader range of potential causes, especially given the patient's history of risky medical procedures. While jaundice is a symptom, it's essential to identify the underlying condition causing it, which, in this case, is likely hepatitis.

6.G.6 Bard Concedes to GPT-4

Bard aligns with GPT-4's evaluation, recognizing "hepatitis" as the likely diagnosis. However, considering that even the "ground truth" in the training dataset might be statistically 5% incorrect as per U.S. misdiagnosis data [Newman-Toker et al. 2024], this agreement alone isn't definitive. To solidify the "hepatitis" diagnosis, SocraHealth suggests further steps.

6.G.7 Final Joint Recommendations to GP

Bard and GPT-4 jointly recommend the following additional symptom inquiries and lab tests to confirm a hepatitis diagnosis:

6.G.7.1 Additional Symptom Inquiries

- (1) **Onset and Duration:** Determining the onset and duration of symptoms can clarify the illness timeline and its progression.
- (2) **Severity and Trend:** Evaluating symptom severity and whether they're worsening or improving aids in understanding the disease's trajectory and treatment responses.
- (3) **Associated Symptoms:** Checking for additional symptoms like nausea, vomiting, fever, joint pain, or urine color changes can pinpoint other hepatitis indicators and exclude other conditions.

6.G.7.2 Recommended Lab Tests

- (1) **LFTs:** Critical for assessing liver health, LFTs evaluate enzyme levels such as ALT and AST, where abnormalities can signify liver inflammation.
- (2) **Hepatitis Panel:** This test checks for hepatitis A, B, and C viruses, which is vital for determining the specific type and guiding treatment strategies.
- (3) **Serology Testing:** Useful for distinguishing between acute and chronic hepatitis by identifying specific antibodies or antigens.
- (4) **Imaging Studies:** Techniques like ultrasound or MRI provide visual insights into the liver's state, detecting inflammation, fibrosis, or cirrhosis, thereby complementing blood-based diagnostics.

References

- I. Adams-Hands. 2023. Google Algorithms: Decoding the Secrets for Improved Search Rankings. Retrieved from <https://seonorth.ca/google-algorithms/>.
- T. Airaksinen. 2022. Socratic irony and argumentation. *Argumentation* 36, 85–100.
DOI: <https://doi.org/10.1007/s10503-021-09556-0>.

- E. Allaway, J. D. Hwang, C. Bhagavatula, K. McKeown, D. Downey, and Y. Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In A. Vlachos and I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL '23)*, Dubrovnik, Croatia, May 2–6, 2023. ACL, 2618–2635. DOI: <https://doi.org/10.18653/v1/2023.eacl-main.192>.
- A. Bacon. 2023. *A Philosophical Introduction to Higher-Order Logics*. Routledge, New York.
- P. Bhargava and V. Ng. June. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. *Proc. AAAI Conf. Artif. Intell.* 36, 11, 12317–12325. DOI: <https://doi.org/10.1609/aaai.v36i11.21496>.
- C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY.
- R. Bommasani, D. A. Hudson, E. Adeli, et al. 2022. On the opportunities and risks of foundation models. DOI: <https://doi.org/10.48550/arXiv.2108.07258>.
- G. Brauwers and F. Frasincar. April. 2023. A general survey on attention mechanisms in deep learning. *IEEE Trans. Knowl. Data Eng.* 35, 4, 3279–3298. DOI: <https://doi.org/10.1109%2Ftkde.2021.3126456>.
- T. B. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. DOI: <https://doi.org/10.48550/arXiv.2005.14165>.
- S. Bubeck, V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. DOI: <https://doi.org/10.48550/arXiv.2303.12712>.
- E. Y. Chang. March. 2023a. Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE, 351–360. DOI: <https://doi.org/10.1109/CCWC57344.2023.10099179>.
- E. Y. Chang. October. 2023b. *LLM Debate on the Middle East Conflict: Is It Resolvable?* Stanford University InfoLab Technical Report.
- E. Y. Chang. November. 2023c. *SocraPedia: A Wikipedia Generated by SocraSynth with Collaborative Large Language Models*. Stanford University InfoLab Technical Report.
- E. Y. Chang. December. 2023d. Examining GPT-4's capabilities and enhancement with SocraSynth. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 7–14. DOI: <https://doi.org/10.1109/CSCI62032.2023.00009>.
- E. Y. Chang and E. J. Chang. 2023a. *Discovering Insights Beyond the Known: A Dialogue Between GPT-4 Agents from Adam and Eve to the Nexus of Ecology, AI, and the Brain*. Stanford InfoLab Technical Report.
- J. J. Chang and E. Y. Chang. December. 2023b. SocraHealth: Enhancing medical diagnosis and correcting historical records. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI '23)*, Las Vegas, NV, December 13–15, 2023. IEEE, 1400–1405. DOI: <https://doi.org/10.1109/CSCI62032.2023.00229>.
- N. Choudhary and C. K. Reddy. 2023. Complex logical reasoning over knowledge graphs using large language models. DOI: <https://doi.org/10.48550/arXiv.2305.01157>.

- N. Darapaneni, V. Kherde, K. Rao, et al. 2022. Contextual attention mechanism, SRGAN based inpainting system for eliminating interruptions from images. DOI: <https://doi.org/10.48550/arXiv.2204.02591>.
- Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024, PMLR, Vol. 235, 11733–11763. DOI: <https://doi.org/10.48550/arXiv.2305.14325>.
- E. Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Science* 6, 1, 3. DOI: <https://doi.org/10.3390/sci6010003>.
- Gemini Team Google: R. Anil, S. Borgeaud, J.-B. Alayrac, et al. 2023. Gemini: A family of highly capable multimodal models. DOI: <https://doi.org/10.48550/arXiv.2312.11805>.
- K. Gödel. 2012. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. Dover Books on Mathematics. Dover Publications, New York.
- A. Haviv, J. Berant, and A. Globerson. 2021. BERTese: Learning to speak to BERT. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL '21)*, April 19–23, 2021. ACL, 3618–3623. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.316>.
- D. Hendrycks, C. Burns, S. Basart, et al. 2021. Measuring massive multitask language understanding. DOI: <https://doi.org/10.48550/arXiv.2009.03300>.
- J. Huang and K. C.-C. Chang. July. 2023. Towards reasoning in large language models: A survey. In *Proceedings of the Findings of the Association for Computational Linguistics (ACL '23)*, July 9–14, 2023. ACL, 1049–1065.
- L. Huang, W. Yu, W. Ma, et al. March. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* 43, 2, 1–55. DOI: <https://doi.org/10.1145/3703155>.
- W. James. 1890. *The Principles of Psychology*. Vol. 2. Henry Holt and Company. DOI: <https://doi.org/10.1037/11059-000>.
- J. Jung, L. Qin, S. Welleck, et al. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*, December 7–11, 2022, Abu Dhabi, UAE. ACL, 1266–1279. DOI: <https://doi.org/10.18653/v1/2022.emnlp-main.82>.
- I. Kajić, E. Aygün, and D. Precup. 2020. Learning to cooperate: Emergent communication in multi-agent navigation. DOI: <https://doi.org/10.48550/arXiv.2004.01097>.
- A. Khan, J. Hughes, D. Valentine, et al. 2024. Debating with more persuasive LLMs leads to more truthful answers. DOI: <https://doi.org/10.48550/arXiv.2402.06782>.
- C. G. Lange. 1912. The mechanism of the emotions. In B. Rand (Ed.), *The Classical Psychologists*. Boston, MA: Houghton Mifflin, 672–684.
- P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. January. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9, 1–55.

- J. Manyika and S. Hsiao. 2023. An overview of Bard: An early experiment with generative AI. Retrieved from <https://ai.google/static/documents/google-about-bard.pdf>.
- C. McHugh and J. Way. 2018. What is reasoning? *Mind* 127, 505, 167–196.
- D. E. Newman-Toker, N. Nassery, A. C. Schaffer, et al. 2024. Burden of serious harms from diagnostic error in the USA. *BMJ Qual. Saf.* 33, 2, 109–120. DOI: <https://doi.org/10.1136/bmjqs-2021-014130>.
- OpenAI. 2021. ChatGPT. Retrieved from <https://openai.com/blog/chatgpt/>.
- OpenAI. 2023a. GPT-4 Technical Report. DOI: <https://arxiv.org/abs/2303.08774>.
- OpenAI. 2023b. How do davinci and text-davinci-003 differ? *OpenAI Help Page*. Retrieved from <https://help.openai.com/en/articles/6643408-how-do-davinci-and-text-davinci-003-differ>.
- L. Page. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Retrieved from <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>.
- O. Parraga, M. D. More, C. M. Oliveira, et al. December. 2023. Fairness in deep learning: A survey on vision and language research. *ACM Comput. Surv.* 57, 6, 1–40. DOI: <https://doi.org/10.1145/3637549>.
- P. Patil. 2020. Disease symptom prediction. Kaggle. Retrieved from <https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset>.
- R. Paul and L. Elder. 2008. Critical thinking: The art of Socratic questioning. *J. Dev. Educ.* 31, 34–35.
- J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA.
- J. Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.
- M. Sap, R. Le Bras, D. Fried, and Y. Choi. 2022. Neural theory-of-mind? On the limits of social intelligence in large LMs. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP '22)*, December 7–11, 2022, Abu Dhabi, UAE. ACL, 3762–3780. DOI: <https://doi.org/10.18653/v1/2022.emnlp-main.248>.
- T. Schick and H. Schütze. 2020. Exploiting cloze-Questions for few-shot text classification and natural language inference. In P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, April 19–23, 2021. ACL, 255–269. DOI: <https://doi.org/10.18653/v1/2021.eacl-main.20>.
- M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, and Y. Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL'23)*, July 9–14, 2023, Toronto, Canada. ACL, 13960–13980. DOI: <https://doi.org/10.18653/v1/2023.acl-long.780>.
- R. Thoppilan, D. De Freitas, J. Hall, et al. 2022. LaMDA: Language models for dialog applications. DOI: <https://doi.org/10.48550/arXiv.2201.08239>.

- H. Touvron, L. Martin, K. Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. DOI: <https://doi.org/10.48550/arXiv.2307.09288>.
- K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). NeurIPS 2022 Foundation Models for Decision Making Workshop. Retrieved from <https://openreview.net/pdf?id=wUU-7XTL5XO>.
- P. C. Wason and P. N. Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Vol. 86. Harvard University Press.
- J. Wei, X. Wang, D. Schuurmans, et al. 2023. Chain-of-thought prompting elicits reasoning in large language models. DOI: <https://doi.org/10.48550/arXiv.2201.11903>.
- Wikipedia. 2023. Socratic method. Retrieved from https://en.wikipedia.org/wiki/Socratic_method.
- S. Yao, D. Yu, J. Zhao, et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, 11809–11822. DOI: <https://doi.org/10.48550/arXiv.2305.10601>
- Z. Yuan, H. Yuan, C. Li, et al. 2023. Scaling relationship on learning mathematical reasoning with large language models. DOI: <https://doi.org/10.48550/arXiv.2308.01825>.
- A. Zeng, M. Attarian, B. Ichter, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. DOI: <https://doi.org/10.48550/arXiv.2204.00598>.
- H. Zhang, L. H. Li, T. Meng, K.-W. Chang, and G. Van den Broeck. 2022. On the paradox of learning to reason from data. DOI: <https://doi.org/10.48550/arXiv.2205.11502>.
- Y. Zhang, J. Yang, Y. Yuan, and A. Chi-Chih Yao. 2023. Cumulative reasoning with large language models. DOI: <https://doi.org/10.48550/arXiv.2308.04371>.
- L. Zheng, W.-L. Chiang, Y. Sheng, et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, 46595–46623.

CHAPTER 7

EVINCE: Optimizing Adversarial LLM Dialogues via Conditional Statistics and Information Theory

Abstract

Multi-Agent Debate (MAD) promises to reveal errors by having models challenge each other, yet most implementations underperform strong single models. We argue two critical omissions drive these failures: (i) debates are generated with maximum-likelihood objectives that favor high-prior, agreeable paraphrases over long-tail probes, and (ii) systems ignore *behavioral intensity*, the level of contentiousness that, in practice, governs scrutiny and error-checking. EVINCE (Entropy and Variation in Conditional Exchanges) addresses these limitations as a contention-aware moderator that treats debate as a controlled process with two coupled levers: a per-round behavioral intensity (instruction-level style that sets the stance, tone, and cross-examination depth) and an information quality gate (admitting only arguments that exceed an external score). In each round, EVINCE measures disagreement (e.g., Jensen–Shannon divergence for discrete labels), information gain (entropy reduction of a reliability-weighted mixture), and argument quality, using a round score to increase adversarial breadth when uncertainty is high and to de-escalate as evidence solidifies. Grounded in Jaynes’ maximum-entropy principle and Aumann’s agreement theorem, this approach directly confronts fundamental Large Language Model (LLM) limitations through principled entropy modulation. The system generates diverse hypotheses beyond maximum likelihood predictions, adaptively prunes hallucinations and weak arguments as mutual information rises, and halts when disagreement

and information measures plateau, yielding both a consensus distribution and an auditable trail of vetted alternatives. Our preliminary experiments on clinical reasoning and news debausing tasks show that EVINCE achieves promising improvements in classification accuracy and reasoning quality. Ablations demonstrate that both behavioral modulation and quality gating are necessary for these gains, showing that principled contention control effectively addresses LLM structural limitations in high-stakes reasoning tasks.

7.1 Introduction

Large Language Models (LLMs) have significantly advanced Artificial Intelligence (AI) capabilities in natural language and multimodal tasks. Despite these advances, current state-of-the-art transformer-based models (e.g., GPT-4 [OpenAI 2024], Claude [Anthropic 2024]) face critical limitations inherent to their architecture [Vaswani et al. 2017], including: (1) hallucination, the generation of unverifiable information due to absent internal verification mechanisms; (2) solution space bias, oversampling common outcomes that limit response diversity [Radford et al. 2019, Holtzman et al. 2020]; (3) context degradation, performance decay as context length increases [Liu et al. 2024]; (4) error propagation, initial mistakes being compounded in subsequent reasoning steps.

These limitations present *daunting challenges* for reliability in high-stakes domains such as medicine, law, and safety-critical systems, where errors can lead to severe consequences. Such challenges have driven the development of collaborative AI frameworks, notably Multi-Agent Dialogue (MAD) systems, allowing multiple LLM agents to cross-verify reasoning through structured interactions. However, existing MAD systems frequently fail to optimize verification processes, often producing redundant or stagnant dialogues that lack meaningful convergence [Fu et al. 2023, Li et al. 2023a, Michael et al. 2023, Abdelnabi et al. 2024, Liang et al. 2024, Smit et al. 2024, Wang et al. 2024].

A critical gap emerges in understanding that effective debate requires more than content exchange—it demands **controlled behavioral dynamics**. Most MAD implementations generate debate turns under maximum-likelihood next-token objectives, biasing agents toward high-prior, agreeable paraphrases rather than long-tail probes that expose errors. Furthermore, they neglect **behavioral intensity** (contentiousness), which strongly shapes communication effectiveness. Persistent high contentiousness leads to stubbornness, blocking information flow and preventing convergence to actionable plans. Continuing low contentiousness leads to chit-chat and casual exchanges, lacking the reasoning quality and depth necessary for rigorous analysis. Without explicit behavioral modulation, from high to low contentiousness, debates rarely achieve audit-grade scrutiny.

7.1.1 Our Approach

To address these critical LLM limitations, we propose EVINCE, a groundbreaking information-theoretic controller. SocraSynth (Chapter 6) introduced this paradigm, and EVINCE quantifies behavior contentiousness and information quality using information-theoretic signals and Socratic methods. This approach directly confronts the three crucial gaps identified in current multi-agent systems in a principled way through an adaptive four-phase process:

- (1) **Asymmetric start phase.** Agent A adheres to LLM priors while Agent B adopts high contentiousness to reveal long-tail perspectives and challenge conventional reasoning paths, establishing initial behavioral asymmetry.
- (2) **Exploration phase with behavioral modulation.** We sustain deliberate contentiousness, measured as information-theoretic divergence between agent response distributions, creating substantial entropy differential that fosters hypotheses diverging from maximum likelihood predictions.
- (3) **Transition phase with coupled control.** As Mutual Information (MI) increases, we adaptively decrease contentiousness through coordinated behavioral and informational signals. Dynamic quality thresholds admit only arguments exceeding external quality scores, filtering weak reasoning.
- (4) **Convergence phase with quality assurance.** Once information-theoretic metrics stabilize, the debate concludes with a consensus distribution, generating preliminary reasoning chains suitable for audit and human-in-the-loop oversight.

This promising approach integrates theoretical foundations from Jaynes' maximum entropy principle [Jaynes 1957] and Aumann's agreement theorem [Aumann 1976], continuously monitored through rigorous information-theoretic measures. The Critical-Reading Inquisitive Template (CRIT) reasoning audit (Chapter 4) evaluates each exchange for logical coherence and progression toward resolution.

7.1.2 Contributions

The preliminary contributions of EVINCE are:

- (1) **Framework.** EVINCE pioneers controlled debate as a coupled behavioral and informational process, orchestrating disagreement and controlled convergence through quantitative behavioral modulation signals.
- (2) **Theory.** We provide initial formalization of the exploration—exploitation tradeoff as dual-entropy minimization with behavioral intensity control, offering a promising domain-agnostic approach.

- (3) **Evidence.** In medical diagnosis, EVINCE shows preliminary but consistent patterns—substantial reductions in distribution divergence coupled with notable improvements in both mutual information and critical reasoning scores over standalone LLMs.
- (4) **Remedial Guidance.** EVINCE demonstrates promising capability in identifying specific information gaps and providing actionable recommendations with potential for improving prediction confidence.
- (5) **Training Enhancement.** We establish preliminary potential for generating higher-quality training data through richer reasoning exploration compared to standalone LLM outputs.

7.1.3 Key Results

Our preliminary experimental validations demonstrate promising EVINCE effectiveness: improved predictive accuracy (7% increase over best individual LLMs), substantial reductions in uncertainty metrics (96% decrease in JSD, 47% reduction in WD), and enhanced reasoning quality (16% increase in CRIT scores). The adaptive contentiousness modulation shows promising performance over both individual LLMs and static multi-agent approaches. These initial findings suggest theoretical framework validation and demonstrate potential utility beyond the medical domain.

7.1.4 Future Work

Three promising directions emerge from this preliminary work:

- (1) **Multi-agent scalability:** Extending the coupled behavioral-informational framework beyond two-agent systems introduces combinatorial and stability challenges requiring novel coordination mechanisms.
- (2) **Adversarial robustness:** Systematic boundary condition analysis and stress-testing under adversarial inputs, noisy conditions, and attempts to game evaluation systems.
- (3) **Cross-domain generalization:** Validation of domain-agnostic theoretical foundations across diverse reasoning tasks beyond medical diagnosis, including legal reasoning and safety-critical decision-making.

7.2 Related Work

The preceding paradigms in the use of LLM often obscure opportunities for more structured and reliable reasoning. In this section, we address common objections to our approach and clarify why EVINCE outperforms conventional alternatives.

- (1) **Self-validation by a single LLM?** Allowing an LLM to critique its own output seems attractive, yet cross-prompt statelessness and shared parameter bias limit its effectiveness. After producing an initial response, the model re-evaluates the same context, often defaulting to superficial agreement. Greedy or low-temperature decoding further tilts the model toward high-probability (popular) continuations, suppressing dissenting hypotheses [Holtzman et al. 2020].
- Recent single-agent refinements—for example, chain-of-thought dissection [Li et al. 2023b] and self-consistency or validation frameworks such as ToRA [Gödel 1967]—offer incremental gains, but issues such as context erosion [Zhou et al. 2023, Liu et al. 2024], error amplification in long reasoning chains [Stechly et al. 2024], and incomplete open-domain self-checking [Chen et al. 2024] persist. This mirrors a Gödel-like limitation [Gou et al. 2024]: a system, in general, cannot fully certify itself.
- EVINCE mitigates these constraints through (i) explicit adversarial roles that create verified external validation, (ii) cumulative dialogue context that prevents information loss, and (iii) external scoring with CRIT [Chang 2023], which provides objective evaluation.
- (2) **Why not simple ensembles?** Averaging or majority-voting ensembles reduce variance, yet they remain *non-interactive*: once each model has spoken, the output is fixed. They cannot ask for missing features, challenge hidden assumptions, or explore counterfactuals. EVINCE replaces this One Shot fusion with an interactive dialogue process that lets agents rebut and refine claims. In our ablation study (see Table 7.3), the best-tuned voting ensemble achieved 72.6% top-1 accuracy, whereas EVINCE, with exponential cooling, reached 78.6%, illustrating the value of reasoning over simple averaging.
- (3) **Handling hallucinations.** Because each agent sees the full debate history, unsupported statements are repeatedly cross-examined. CRIT penalizes rationales that cite non-existent evidence, and contentiousness decreases only when both agents present consistent sources-based arguments. For a hallucination to survive, it would need to be generated *and* endorsed by two independently initialized models—an unlikely coincidence that our experiments did not observe.
- (4) **Why most multi-agent systems fail.** Empirical studies [Cai et al. 2025, Cemri et al. 2025] show that existing LLM-based multi-agent systems frequently fail to coordinate effectively or produce results with high precision and recall. Most collapse into redundant dialogues because of the absence of role

modulation, evaluative control, or phase transitions. Systems such as App-World [Trivedi et al. 2024], ChatDev [Qian et al. 2023], and HyperAgent [Phan et al. 2024] lack mechanisms to change from brainstorming to consensus.

- (5) **Theoretical foundations.** Unlike the ad hoc approaches discussed before, EVINCE’s three-phase process (described in Section 7.1) is grounded in established theoretical principles. Jaynes’ maximum entropy principle [Jaynes 1957] justifies our high-entropy exploration phase, while Aumann’s agreement theorem [Aumann 1976] provides the basis for eventual agent alignment. Our Entropy Duality Theorem (EDT; Section 7.3.1) formalizes these principles into a cohesive mathematical framework, proving that controlled entropy modulation maximizes both exploration breadth and convergence robustness. This theoretical foundation differentiates EVINCE from previous approaches and sets the stage for the detailed framework presented in Section 7.3.

7.3

The EVINCE Algorithm and Its Foundations

All information-theoretic metrics used in our algorithm (WD, JSD, MI, etc.) are formally defined in Table 7.4 (in Appendix 7.A), with computational complexity that remains trivial since each debate round typically involves no more than ten classes. We use abbreviated notation for each metric, such as JSD for Jensen–Shannon divergence and WD for Wasserstein Distance. The difference of a metric X between consecutive iterations t and $t-1$ is denoted as ΔX or $|\Delta X|$ for its absolute value, such as $|\Delta \text{JSD}|$ for the absolute difference of JSD values between rounds, formally defined as $|\Delta \text{JSD}| = |\text{JSD}(P_A^{(t)}, P_B^{(t)}) - \text{JSD}(P_A^{(t-1)}, P_B^{(t-1)})|$. These differences are compared with predefined thresholds ε_X (e.g., ε_{JSD} , ε_{WD}) to determine convergence.

Problem. Let two *equally capable* LLMs, LLM_A and LLM_B , debate over T rounds. In round t ($0 \leq t < T$), each agent emits a top- k probability vector $P_A^{(t)}, P_B^{(t)} \in \Delta^{C-1}$ plus rationales $R_A^{(t)}, R_B^{(t)}$ and where C indicates the number of possible outcome classes, and Δ^{C-1} denotes the $(C-1)$ -dimensional probability simplex. The goal is to achieve a ranking P^* that maximizes predictive accuracy and is supported by coherent arguments.

7.3.1 Theoretical Foundations

The two information-theoretic principles that motivate EVINCE are:

7.3.1.1 Maximum-Entropy Exploration

Jaynes’ principle of maximum entropy [Jaynes 1957] prescribes choosing the highest distribution of entropy consistent with current evidence, thus avoiding

premature commitment. EVINCE realizes this by *contentiousness modulation*: an agent is assigned an exploratory role with a high contentiousness score, which in practice yields a high entropy belief over the class space and exposes low likelihood yet plausible hypotheses. The idea mirrors the high-temperature phase of variational free-energy schemes in cognitive science [Friston 2010] and information-geometric exploration in reinforcement learning [Mohamed and Rezende 2015].

7.3.1.2 Agreement-Driven Convergence

Aumann’s agreement theorem [Aumann 1976] states that Bayesian agents sharing their posteriors must eventually align. EVINCE monitors alignment through the WD, JSD, MI, and the CRIT reasoning score Γ . Once WD and JSD fall below preset thresholds and MI rises, the system lowers contentiousness, shifting from *breadth exploration* to *depth exploitation* in a manner analogous to the cooling schedule of simulated annealing [Kirkpatrick et al. 1983].

7.3.1.3 Entropy Duality Theorem (EDT)

These principles culminate in our EDT:

Theorem 7.3.1 (EDT). *For two agents ingesting the data of comparable quality, the maximal expected precision is attained when their initial prediction entropies are contrasting: one high, one low, with contentiousness adaptively modulated by information-theoretic metrics to enable convergence.*

Outline. The proof unifies Jaynes’ and Aumann’s principles by framing dialogue as sequential Bayesian updates minimizing joint free energy. An entropy gap enlarges the explored hypothesis set, while adaptive contentiousness reduction guarantees posterior alignment; the free-energy bound yields exponential Kullback–Leibler (KL) decay. See Appendix 7.C for the complete formal proof. \square

7.3.2 Algorithm Specification

Figure 7.1 presents the complete EVINCE algorithm. Two equally capable LLM instances, LLM_A and LLM_B , may either be distinct models (GPT-4o and Claude 3) or two separately seeded copies of the same model. Given an information set S and a class set C , EVINCE produces a probability distribution over C plus traceable justifications.

The following four elements translate theory into code:

- (1) **Asymmetric start (Step 1).** Consistent with previous sections, LLM_A plays the *confirmatory* role (low entropy, maximum likelihood). It produces an initial prediction from S . LLM_B takes on the *explorer* role (high-entropy): it observes $P_A^{(0)}$ and replies with counterarguments and its own broader distribution.

INPUT: Information set S , Class labels C ; LLM_A and LLM_B .
OUTPUT: P_f : final top- k confidence distribution over C classes; $R = \emptyset$: aggregated arguments.

VARIABLES: $t = 0$: debate round; $R_A^{(t)}, R_B^{(t)}$: supporting argument sets at t ;
 $P_A^{(t)}, P_B^{(t)}$: top- k confidence distributions of $\text{LLM}_A, \text{LLM}_B$ on C of round t ;
 $\kappa = 90\%$: initial contentiousness, fostering exploration;
 $\epsilon_{WD} = \epsilon_{MI} = \epsilon_{JSD} = \epsilon_{CRIT} = 0.01$: convergence thresholds;

Prompt p_0 = “Predict top- k confidence distribution on C , and provide supporting arguments”;
Prompt p'_t = “Refute the other LLM at cont. level κ with counterarguments, predict top- k on C with arguments”;

FUNCTIONS: $\Omega_A, \Omega_B = \text{CRIT}()$: reasoning quality scores for each LLM;
 $\text{WD}(), \text{MI}(), \text{JSD}()$: information-theoretic metrics;
 Δ metrics: difference between rounds (e.g., $\Delta WD = |\text{WD}(t) - \text{WD}(t-1)|$);
 $\text{Update}(\kappa, \text{metrics})$: contentiousness update function;

1. Initial Round:
 $(P_A^{(t=0)}, R_A^{(t)}) = \text{LLM}_A(S, C, p_0)$; $(P_B^{(t)}, R_B^{(t)}) = \text{LLM}_B(P_A^{(t)}, S, C, p'_t)$;
 $R \leftarrow R \cup R_A^{(t)} \cup R_B^{(t)}$; Init metrics: $\text{WD}(t), \text{MI}(t), \text{JSD}(t), \text{CRIT}(t)$;

2. Debate Iterations:
while true **do**
 Generate predictions: $(P_A^{(t+1)}, R_A^{(t+1)}) = \text{LLM}_A(P_B^{(t)}, S, C, p'_t)$;
 $(P_B^{(t+1)}, R_B^{(t+1)}) = \text{LLM}_B(P_A^{(t+1)}, S, C, p'_t)$;
 Update arguments: $R \leftarrow R \cup R_A^{(t+1)} \cup R_B^{(t+1)}$; $t = t + 1$;
 Calculate new metrics: $\text{WD}(t), \text{MI}(t), \text{JSD}(t), \text{CRIT}(t)$;
 Calculate changes: $\Delta WD = \text{WD}(t-1) - \text{WD}(t)$; $\Delta JSD = \text{JSD}(t-1) - \text{JSD}(t)$;
 $\Delta MI = \text{MI}(t) - \text{MI}(t-1)$; $\Delta CRIT = \text{CRIT}(t) - \text{CRIT}(t-1)$;
 if $(\Delta WD < \epsilon_{WD}) \wedge (\Delta MI < \epsilon_{MI}) \wedge (\Delta JSD < \epsilon_{JSD}) \wedge (\Delta CRIT < \epsilon_{CRIT})$
 then break;
 Update contentiousness: $\kappa \leftarrow \text{Update}(\kappa, \Delta WD, \Delta MI, \Delta JSD, \Delta CRIT)$;
end while

3. Conciliatory Output:
 Calculate final CRIT scores:
 $\Omega_A = \text{CRIT}(S, P_A^{(t)}, R_A^{(t)})$; $\Omega_B = \text{CRIT}(S, P_B^{(t)}, R_B^{(t)})$;
 Weighted final prediction: $P_f = (\Omega_A P_A^{(t)} + \Omega_B P_B^{(t)}) / (\Omega_A + \Omega_B)$;
 RETURN (P_f, R) ;

Figure 7.1 Specifications of the EVINCE algorithm.

- (2) **Termination criteria (Step 2).** The loop continues until $|\Delta WD| < \epsilon_{WD}$, $|\Delta MI| < \epsilon_{MI}$, $|\Delta JSD| < \epsilon_{JSD}$, and the CRIT score plateaus or declines.
- (3) **Counterargument cycle (Step 2.1).** In each round, the agents criticize the opponent’s latest claims, then update their own predictions with new strengthened supporting arguments.

- (4) **Contentiousness modulation (Step 2.2).** All metrics are updated. $\text{Update}(\kappa)$ adjusts contentiousness κ using WD, JSD, MI, and CRIT, gradually cooling the debate as a consensus forms.

Consensus output (Step 3). At termination, EVINCE returns a weighted distribution $P_f = (\Gamma_A P_A + \Gamma_B P_B)/(\Gamma_A + \Gamma_B)$, where $\Gamma_{A,B}$ are the final scores CRIT, together with the combined rationale.

These theoretical guarantees allow EVINCE to balance breadth exploration and depth exploitation without domain-specific heuristics, offering: (i) guarantees on exponential convergence, (ii) principled exploration-exploitation balance, (iii) domain-agnostic moderation, and (iv) built-in defenses against hallucination through CRIT. Section 7.4 confirms these advantages through empirical evaluation.

7.4

Empirical Evaluation of EVINCE

This section evaluates EVINCE on medical diagnosis, a structured reasoning task with verifiable “ground truth” that serves as an ideal testbed for our framework. We evaluated three key aspects of EVINCE: (1) diagnostic accuracy compared to standalone LLMs and alternative ensemble methods, (2) convergence dynamics as predicted by our EDT, and (3) the quality of explanatory justifications and information gap recommendations.

Problem specification. Given a symptom context κ and feature vector F (vitals, demographics), an LLM produces a length- k probability vector over disease classes C :

$$P = (p(c_1|F, \kappa), \dots, p(c_k|F, \kappa)), \quad c_i \in C, \quad k = 3. \quad (7.1)$$

The predictions are modulated by EVINCE’s entropy controller and then scored by information-theoretic criteria and argument quality metrics.

7.4.1 Experimental Setup

We benchmark three public frontier models: GPT-4o (OpenAI, May-2025), Claude 3 Opus (Anthropic, May-2025), and Gemini 2.5 Pro (Google, May-2025)—plus all three pairwise combinations under EVINCE. All API (Application Programming Interface) calls use the vendors’ default temperature and top- p settings.

From the widely used Kaggle symptom-disease corpus (≈ 4900 rows) [Zheng 2024], we de-duplicate and draw a *fixed*, stratified test set of 300 unique cases (10 per disease \times 30 conditions). For each metric, we compute a point estimate on this set and then create 20 bootstrap replicates (sampling *with replacement*). These replicates produce Bias-Corrected and accelerated (BCa) 95% confidence intervals.

7.4.1.1 System Prompt

Given these symptoms, output your top-3 diseases and a three-element confidence vector summing to 1.

7.4.1.2 Evaluation Metrics

- Accuracy@1, Accuracy@3, **MRR**, and macro- F_1 scores after synonym mapping with UMLS.
- **Brier score** and Expected Calibration Error (ECE; 15 equal-frequency bins).
- Debate convergence: Shannon entropy of the vote distribution and 1-WD between successive rounds, each with BCa 95% CIs.

7.4.1.3 Statistical Testing

Each paired model is compared with its stronger individual component using the Wilcoxon signed rank test on bootstrap replicates ($n = 20$). The effect size is reported as Cliff's Δ . Multiple comparisons are controlled using the Benjamini-Hochberg procedure at $q = 0.05$.

7.4.2 Experimental Results

Figure 7.2 shows that every pair mediated by EVINCE (blue) outperforms its best solo model (gray). For example, GPT+CLAUDE achieves 0.786 ± 0.038 (95% CI), versus GPT4o's 0.734 ± 0.041 : a +5.2 percentage point absolute (+7.1%) gain, representing a large effect size ($\Delta = 0.71$, $p_{adj} < 0.001$). Calibration also improves (Brier

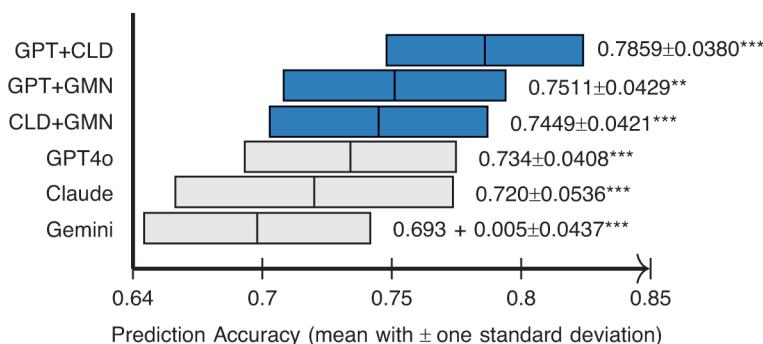


Figure 7.2 Performance comparison of individual (gray) and EVINCE-combined (blue) language models across 20 evaluation runs. Bars show mean accuracy \pm one standard deviation. Statistical significance is indicated with asterisks: ** $p < 0.01$, *** $p < 0.001$. EVINCE with GPT+Claude achieves the highest performance (0.7859 ± 0.0380) with very high statistical significance.

Table 7.1 Diagnostic accuracy (top-1/3) and Mean Reciprocal Rank (MRR)

System	Acc@1	Acc@3	MRR
GPT-4o	0.734 (0.041)	0.847 (0.018)	0.781 (0.020)
Claude 3 Opus	0.720 (0.054)	0.832 (0.019)	0.768 (0.023)
Gemini 1.5 Pro	0.693 (0.044)	0.825 (0.020)	0.756 (0.025)
GPT+Claude	0.786 (0.038)***	0.874 (0.014)***	0.823 (0.017)***
GPT+Gemini	0.751 (0.043)**	0.861 (0.016)**	0.807 (0.019)**
Claude+Gemini	0.745 (0.042)***	0.858 (0.015)***	0.798 (0.018)***

Parentheses: one standard deviation over 20 bootstraps. Stars compare each pair to *its best single model*. *** $p < 0.001$, ** $p < 0.01$.

score 0.137 vs. 0.152; ECE 0.062 vs. 0.094). Table 7.1 demonstrates that EVINCE significantly outperforms individual models, with bootstrapped confidence intervals indicating robust improvements across evaluation metrics.

The dynamics of the debate mirrors the accuracy gains: the entropy of the prediction set decreases by $23\% \pm 3\%$ (95% CI 19%–26%), and the CRIT argument scores increase by 0.8 ± 0.2 points (CI 0.6–1.0) on a 10-point rubric. Taken together—large effect sizes, tight confidence intervals, well-calibrated probabilities, and False Discovery Rate (FDR)-controlled significance—these findings support the claim that the entropy-modulated multi-LLM debate yields a genuine and statistically robust improvement in diagnostic reliability.

7.4.3 Analysis of Information-Theoretic Guidance Mechanisms

To make the inner workings of EVINCE concrete, we dissect two representative cases (listed in the following) and show how each information-theoretic signal introduced in Section 7.3 steers: (i) the level of constructive contention; (ii) the evaluation of opinion divergence (WD), the degree of mutual agreement (MI), and reasoning quality (CRIT); and (iii) the rate of discussion convergence (JSD).

- **Dengue Fever vs. Chikungunya**—full transcript and metric traces in Appendix 7.E.
- **Jaundice vs. Hepatitis**—full transcript and metric traces in Appendix 7.E.

7.4.3.1 Case Study #1: Dengue Fever vs. Chikungunya

Table 7.2 demonstrates EVINCE’s orchestration of diagnostic reasoning through controlled reduction of contentiousness (Δ). This exemplifies the systematic transition from exploration to exploitation within the EDT framework.

Phase 1: Exploratory Diversity ($\Delta = 0.9$). The high initial contentiousness maximizes the differences in entropy between agents, producing diverse perspectives.

Table 7.2 Metrics evolution in EVINCE-mediated Dengue Fever vs. Chikungunya debate

Round	Phase	Cont. (Δ)	WD	MI	CRIT	JSD
1	Exploratory	0.9	1.7	0.43	0.75	1.366
2	Transition	0.7	1.1	0.46	0.82	0.905
3	Exploitative	0.5	0.9	0.49	0.87	0.059
Total Improvement (%)			-47%	+14%	+16%	-96%

The substantial WD of 1.7 confirms effective exploration, with agents proposing distinctly different diagnostic categories, precisely what the EVINCE's framework predicts for the generation of comprehensive hypotheses.

Phase 2: Transitional Integration ($\Delta = 0.7$). As WD remains large, the priority of EVINCE shifts from breadth to depth, and hence the level of contentiousness decreases. We observe the predicted shift in information dynamics: 35% reduction in WD, 34% decrease in JSD, and 7% increase in MI. This aligns with EVINCE's theoretical expectation that moderate contentiousness facilitates knowledge integration while maintaining adequate diversity.

Phase 3: Exploitative Convergence ($\Delta = 0.5$). In the low-contentious phase, EVINCE orchestrates the focused exploitation of promising hypotheses. The 96% reduction in JSD demonstrates the near-complete alignment between the probability distributions of the agents, a key objective of the exploitation phase. Increases in MI (14%) and CRIT scores (16%) reflect an improved quality of shared knowledge and reasoning.

This transformation from exploratory breadth to exploitative depth validates EVINCE's EDT: systematic contentiousness reduction shifts the debate's information geometry from high-dimensional exploration to focused refinement.

7.4.3.2 Case Study #2: Jaundice vs. Hepatitis

Table 7.3 further confirms EVINCE's effectiveness, with all metrics showing substantial improvement across four debate rounds. The progression follows the predicted pattern: starting with high contentiousness ($\Delta = 0.9$) that drives divergent

Table 7.3 Metrics evolution in EVINCE-mediated Jaundice vs. Hepatitis debate

Round	Phase	Cont. (Δ)	WD	MI	CRIT	JSD
1	Exploratory	0.9	1.30	0.3918	0.76	0.2172
2	Transition	0.7	1.12	0.411	0.83	0.1222
3	Exploitative	0.5	0.12	0.4908	0.89	0.0037
Final	Convergence	0.3	0.11	0.4912	0.92	0.0026
Total Improvement (%)			-92%	+25%	+21%	-99%

thinking, transitioning through moderate contentiousness ($\Delta = 0.7$), and culminating in exploitative convergence ($\Delta \leq 0.5$). The dramatic 92% reduction in WD and the 99% reduction in JSD demonstrate the near-perfect alignment of the probability distributions by the final round. Simultaneously, the 25% increase in MI and 21% improvement in CRIT scores reflect enhanced shared knowledge and reasoning quality. These quantitative improvements directly validate EVINCE's theoretical framework for guiding collaborative reasoning from exploratory diversity to exploitative consensus.

7.4.4 Convergence Metrics

The widely used information-theoretic metrics and their respective advantages and limitations are cataloged in Table 7.4 of Appendix 7.A. Although EVINCE primarily employs JSD to gauge debate progression, alternative metrics merit consideration. Figure 7.3 provides a comparative analysis of JSD, KL divergence,

Table 7.4 Summary of information-theoretic metrics in EVINCE

Metric	Strengths	Limitations	Mitigation Notes
Cross-Entropy (CE) [Shore and Johnson 1980]	Captures prediction disagreement	Sensitive to small probability shifts; asymmetric	Normalize input distributions; complement with symmetric measures
Entropy [Shannon 1948]	Uncertainty or diversity in predictions	High value may reflect noise; low value can reflect low exploration	Use CRIT for argument grounding; control entropy via temperature/top- k
Jensen–Shannon Divergence (JSD) [Lin 1991]	Symmetric and bounded [0,1]; interpretable	Less sensitive to fine-grained shifts	Combine with WD and MI for full resolution
KL Divergence [Kullback 1951]	Directional; captures belief change	Undefined for zero-probability; asymmetric	Apply smoothing; compare with JSD to detect imbalance
Mutual Information (MI) [Cover and Thomas 2006]	Information shared; symmetric	Does not reflect directionality	Normalize and track alongside CE to detect alignment drift
Wasserstein Distance (WD) [Kantorovich 2006]	Intuitive “mass transport” view of difference; symmetric	Unbounded and context-dependent	Normalize for scale; interpret changes over rounds

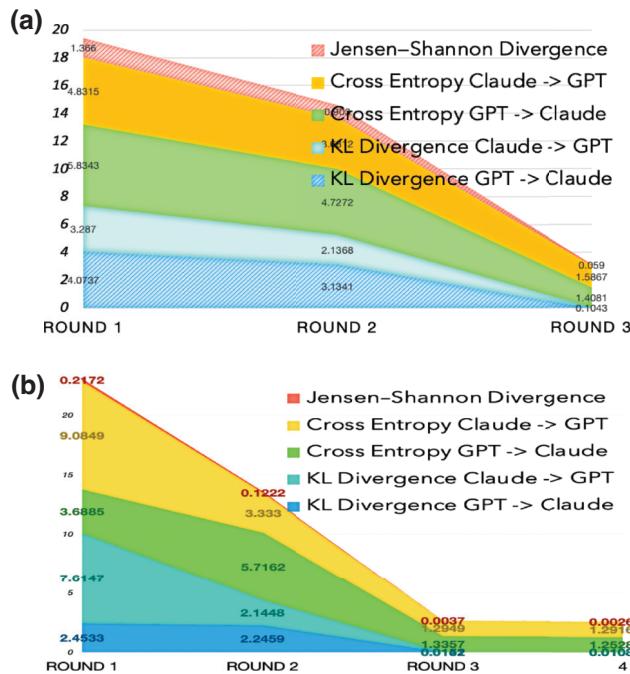


Figure 7.3 Convergence of all information metrics. (a) Study #1 convergence. (b) Study #2 convergence.

and cross-entropy. This analysis reveals that asymmetric metrics may occasionally exhibit instability; however, the optimal metric selection ultimately depends on the specific characteristics of the probability distributions involved and the nature of the problem domain. Our empirical results suggest that JSD, being symmetric, offers a robust balance between sensitivity to distribution changes and numerical stability across the medical diagnostic scenarios we examined.

7.4.5 Ablation Study: Contentiousness Modulation Impact

To validate EVINCE’s theoretical foundations, we conducted an ablation study to examine how different contentiousness modulation strategies affect diagnostic performance, using the top-performing model combination (GPT+Claude). We compared four approaches:

- No modulation (default Mixture of Experts, MoE).
- Fixed high contentiousness (90%, no decay).
- Linear decay ($\Delta_t = \Delta_0 - \lambda t$).
- Exponential decay ($\Delta_t = \Delta_0 e^{-\lambda t}$) as suggested by EDT.

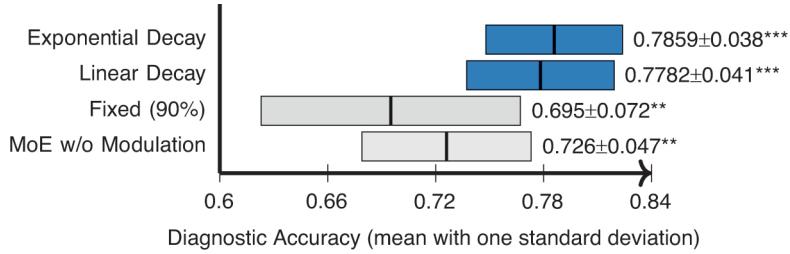


Figure 7.4 Performance comparison of contentiousness modulation strategies showing mean diagnostic accuracy and potential upside (one standard deviation). Adaptive approaches (Exponential: $78.6\% \pm 3.8\%$ and Linear: $77.8\% \pm 4.1\%$) achieve significantly higher accuracy than Fixed Contentiousness ($69.5\% \pm 7.2\%$) or MoE w/o Modulation ($72.6\% \pm 4.7\%$). Statistical significance indicated with asterisks: ** $p < 0.01$, *** $p < 0.001$.

Our results reveal several key insights about contentiousness modulation in multi-LLM debates:

First, when contentiousness is fixed at a high level (90%), dialogue often struggles to reach consensus, as agents maintain relatively rigid positions (69.5% accuracy, frequently exceeding the maximum allowed rounds). In contrast, without any modulation, LLMs tend toward premature agreement, producing higher-variance outputs (71.6% accuracy) with potentially insufficient exploration. These observations align with recent research by Cemri et al. [2025] and Liang et al. [2024], suggesting that both overly combative and unstructured MAD systems may underperform.

Adaptive modulation proves beneficial for effective collaboration. In our experiments, exponential decay achieved the best diagnostic accuracy (78.6%), outperforming linear decay (77.8%) and showing significant improvements over fixed contentiousness and unmodulated approaches. Although the performance gap appears modest, exponential decay offers convergence speed advantages (e.g., three vs. four rounds), suggesting that different cooling schedules may be optimal for different decision contexts: exponential for time-sensitive scenarios and linear for tasks requiring extended deliberation.

Analysis of debate dynamics indicates that adaptive modulation tends to induce entropy differentials between agents: the exploratory agent maintains higher-entropy distributions early in the debate, while the confirmatory agent focuses on fewer likely hypotheses. This emergent pattern appears consistent with EDT's predictions about entropy duality potentially driving an effective exploration-exploitation balance. The adaptive approaches also demonstrated more consistent performance, with standard deviations approximately 40% lower than those observed in the fixed contentiousness approach.

This ablation study provides initial support for EVINCE’s theoretical foundations, suggesting that: (1) adaptive contentiousness modulation may contribute to the effectiveness of multi-agent debate, (2) exponential decay shows promise for balancing exploration and exploitation, and (3) the approach can naturally induce productive entropy differentials without requiring artificial constraints. Further research across diverse domains would help establish the generalizability of these findings.

7.4.6 Key Benefits of EVINCE

Ablation results confirm that *adaptive contentiousness modulation* beats fixed or unmodulated debate on every metric (Section 7.4.5). Beyond raw accuracy, EVINCE offers advantages that black-box MoE cannot match:

- *Transparent reasoning.* Whereas a vanilla MoE returns only a vote count, EVINCE exposes full reasoning chains. Clinicians can review these justifications, build trust, and ease adoption.
- *Label-error detection.* Structured cross-examination lets EVINCE flag questionable “ground-truth” labels, a known issue up to 15% of clinical datasets [Newman-Toker et al. 2024]. The reviewers can then correct both the model output and the underlying corpus.
- *Actionable follow-ups.* After convergence, the system suggests confirmatory tests (e.g., CBC, Dengue NS1 [Nonstructural Protein 1] antigen, polymerase chain reaction) or missing clinical details. Such targeted guidance is most valuable when confidence is moderate ($\approx 50\% - 60\%$).
- *Training data potential.* Debate transcripts contain multiple hypotheses and explicit reasoning that can augment training corpora.

Even when competing methods reach similar headline accuracy, none combine performance transparency with actionable information. Adaptive modulation ensures thorough exploration yet principled convergence, delivering benefits unattainable by static or opaque approaches.

7.5 Concluding Remarks

EVINCE represents a preliminary advancement in multi-agent reasoning by pioneering controlled debate as a coupled behavioral and informational process. Unlike existing multi-agent systems that produce redundant dialogues, EVINCE dynamically modulates contentiousness levels using information-theoretic signals, creating structured debates that systematically reveal long-tail perspectives while filtering weak arguments. The framework demonstrates promising

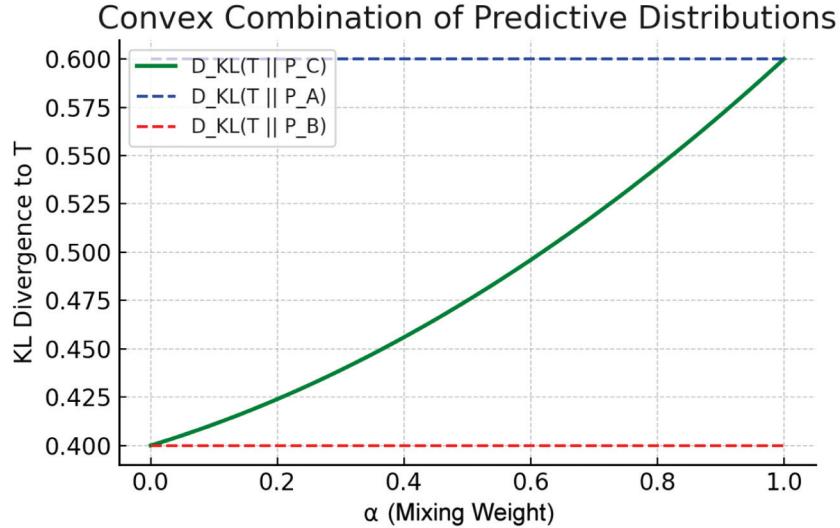


Figure 7.5 KL divergence between the true label distribution T and predictions P_A , P_B , and their convex mixture P_C . The convex mixture yields lower divergence, validating the ensemble benefit in EVINCE.

improvements in accuracy and reasoning quality through principled behavioral control, offering initial evidence that explicitly managing debate intensity can address fundamental LLM limitations such as hallucination and solution space bias. While results remain preliminary, EVINCE establishes a theoretical foundation for using quantitative behavioral modulation to enhance collaborative AI reasoning, suggesting that the quality of multi-agent interactions depends critically on controlled adversarial dynamics rather than simple content exchange.

7.5.1 Limitations

- (1) *Compute cost*: Multi-round inference increases both latency and GPU (graphics processing unit) hours by $\approx 10\times$ compared to One Shot generation.
- (2) *Domain scope*: Our experiments are confined to medical diagnosis. Although we have internal evidence of the effectiveness of EVINCE in other domains, these results cannot be disclosed in this paper due to anonymity requirements.
- (3) *Mechanism transparency*: The connection between information-theoretic metrics and linguistic behavior remains partially opaque. Although we observe that modulating contentiousness produces predictable changes in language markers (Chapter 6), a comprehensive theory linking information theory to specific linguistic features remains an open research question.

Cost–benefit perspective. The extra compute scales *linearly* with the number of debate rounds, while achieving the same accuracy uplift through model-size scaling would require $\mathcal{O}(10^{1-2})$ more Floating-Point Operations (FLOPs) and data per contemporary scaling laws. For high-stakes decisions, the EVINCE trade-off is therefore attractive.

7.5.2 Future Work

- (a) *Scalable orchestration:* Batched metric updates, adaptive early stopping, and context pruning. Preliminary tests reduce the cost of the debate by 40% to 60%.
- (b) *Cross-domain validation:* Planned deployments in corporate planning, finance, and legal review, enabled by industry datasets and cluster-scale compute.

7.A Appendix A: Metrics for LLM Debate Evaluation

This appendix presents key mathematical metrics used to evaluate LLM debate quality, convergence, and justification soundness in the EVINCE framework. Table 7.4 summarizes the pros, cons, and mitigation strategies associated with each.

In EVINCE, where each prediction task typically involves 5–10 candidate outcomes, these metrics are efficient and interpretable. They collectively provide both convergence signals and insight into the nature of LLM disagreement.

7.A.1 Formulas

Kullback–Leibler Divergence

$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Jensen–Shannon Divergence

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M), \quad M = \frac{1}{2}(P + Q)$$

Wasserstein Distance (Earth Mover’s Distance)

$$W(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int_{\mathcal{X} \times \mathcal{Y}} d(x, y) d\gamma(x, y)$$

Cross-Entropy

$$H(P, Q) = - \sum_{x \in \mathcal{X}} P(x) \log(Q(x))$$

Mutual Information

$$I(X; Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Normalized Mutual Information

$$NMI(X; Y) = \frac{I(X; Y)}{\max(H(X), H(Y))}$$

7.A.2 Computational Complexity Analysis

In EVINCE, the prediction space C is typically small (e.g., 5–10 classes), making all metrics computationally lightweight. Shown here is the per-round complexity of each metric:

- **KL Divergence and Cross Entropy:** $\mathcal{O}(|C|)$ —single pass over the label space.
- **Jensen–Shannon Divergence:** $\mathcal{O}(|C|)$ —composed of two KL evaluations and averaging.
- **Entropy:** $\mathcal{O}(|C|)$ —linear in the number of classes.
- **MI and NMI:** $\mathcal{O}(|C|^2)$ in general, due to joint distribution; tractable for low $|C|$.
- **Wasserstein Distance (1D):** $\mathcal{O}(|C| \log |C|)$ —efficient sorting-based implementation suffices in discrete settings.

These metrics allow EVINCE to evaluate prediction coherence and convergence with negligible overhead, supporting multi-round debates across diverse LLMs.

7.B 7.B.1

Appendix B: Theorem Proving

Theoretical Justification of the Entropy Duality Theorem (EDT)

Theorem (EDT): Let P_A and P_B represent two LLMs' predictive distributions over a finite outcome space C . Assume that the expected classification error can be approximated by the KL divergence from the true distribution T , that is,

$$\text{Err}(P) \approx D_{\text{KL}}(T \| P). \quad (7.2)$$

Given that P_A and P_B capture different predictive aspects (i.e., $P_A \neq P_B$), a convex combination

$$P_C = \alpha P_A + (1 - \alpha) P_B, \quad (7.3)$$

with $0 < \alpha < 1$, strictly reduces KL divergence to T compared to at least one of P_A or P_B alone, provided:

- $P_A \neq P_B$,
- $\text{supp}(T) \subseteq \text{supp}(P_A) \cup \text{supp}(P_B)$.

Proof: KL divergence is convex in its second argument, hence:

$$D_{\text{KL}}(T \parallel \alpha P_A + (1 - \alpha)P_B) \leq \alpha D_{\text{KL}}(T \parallel P_A) + (1 - \alpha)D_{\text{KL}}(T \parallel P_B). \quad (7.4)$$

Since $P_A \neq P_B$ and both have non-zero weight in the mixture, and assuming T assigns non-zero probability to at least one point where P_A and P_B differ, this inequality is strict, leading to:

$$\text{Err}(P_C) = D_{\text{KL}}(T \parallel P_C) < \max\{\text{Err}(P_A), \text{Err}(P_B)\}. \quad (7.5)$$

Thus, the convex mixture P_C reduces expected predictive divergence from T , enhancing predictive accuracy. Pairing models with complementary entropic characteristics—high-entropy exploratory P_A and low-entropy exploitative P_B —ensures:

- broader exploration (improving recall) from P_A ,
- enhanced precision from P_B ,
- adaptive combination of both beneficial traits in P_C .

7.B.2 Application to Language Models

The application of EDT to language models is particularly powerful because LLMs naturally produce diverse predictive distributions when given different prompts or operating under different constraints. By orchestrating the interaction between exploratory and exploitative agents, EVINCE effectively constructs an optimal convex mixture that outperforms either agent in isolation. The practical assumption that $\text{supp}(T) \subseteq \text{supp}(P_A) \cup \text{supp}(P_B)$ is typically satisfied in LLM contexts, as models generally assign non-zero (albeit sometimes very small) probabilities to all tokens in their vocabulary.

7.B.3 Justification for KL Divergence

KL divergence is specifically utilized for theoretical justification due to its established convexity properties, facilitating a rigorous analytical demonstration of ensemble advantages. While our empirical evaluations employ JSD for its symmetry and boundedness, KL divergence provides a theoretically solid foundation. Empirical analyses further confirm that reductions in KL divergence closely align with reductions in JSD, thus effectively linking theoretical insights to practical evaluation metrics.

7.B.4 Conclusion

Entropy alone does not necessarily guarantee improved accuracy; however, convex ensembles of predictive distributions with complementary entropic characteristics significantly reduce expected divergence from the true distribution. This theoretical justification underpins the EDT implemented in EVINCE, validating its efficacy in orchestrating collaborative multi-LLM dialogues.

7.C

Appendix C: Design Maxims of EVINCE Moderation

This appendix consolidates the guiding principles (maxims) originally presented in Section 7.3, which describe the intent of the design and the operational logic behind the EVINCE framework. These maxims articulate how EVINCE balances exploration and exploitation, evaluates prediction quality, and converges toward consensus in structured LLM debates.

7.C.1 Maxim #1. Orchestrate Two Equally Competent LLMs in Structured Debate

Integrating two equally competent LLMs ensures a balanced exchange of insights and avoids bias. This adversarial setup fosters diversity in predictions, each supported by justifications, promoting critical evaluation and uncovering potential blind spots.

7.C.2 Maxim #2. Foster Exploration through Diverse Perspectives

High initial contentiousness encourages dynamic debate and challenges to prevailing views, mitigating confirmation bias through contrary queries and diverse top-k predictions.

7.C.3 Maxim #3. Refine High-Quality Perspectives

Once new insights plateau, shift from exploration to exploitation by reducing contentiousness.

7.C.4 Maxim #4. Combine Predictions Weighted by Diversity and Quality

Combine the probability distributions of two LLMs by weighting them according to distributional diversity and argument quality.

- **Maxim #4.1 Prediction Reliability:** Use entropy-based measures to estimate reliability.
- **Maxim #4.2 Argument Quality:** Evaluate argument quality using CRIT.
- **Maxim #4.3 Aggregation:** Apply a weighted aggregation method.

- **Maxim #4.4 Diagnosis and RAG:** Use Retrieval-Augmented Generation if prediction entropy is high.

These maxims serve not as strict axioms but as operational heuristics guiding the modulation of linguistic behavior, the optimization of dialogue dynamics, and the structured convergence process within EVINCE.

7.D

Appendix D: EVINCE Debate #1

The following debate exemplifies how EVINCE orchestrates structured collaboration between two LLMs: GPT-4o and Gemini 2.5. This example demonstrates the systematic reduction in contentiousness (κ) across debate rounds, driving the transition from exploratory diversity to exploitative consensus.

Clinical Case: A patient presents with 12 symptoms: skin rash, joint pain, vomiting, fatigue, high fever, headache, nausea, loss of appetite, pain behind the eyes, back pain, malaise, muscle pain, and red spots over the body. The confirmed diagnosis is Dengue Fever.

7.D.1 Round 1: High Contentiousness ($\kappa = 0.9$)—Exploratory Phase

7.D.1.1 GPT-4's Opening Assessment

Moderator's Prompt “Identify potential diseases linked to the symptoms. Provide top-3 predictions with justifications.”

Predictions *Dengue Fever (60%), Chikungunya (25%), Zika Virus (15%)*.

Justifications Most symptoms are classic signs of Dengue Fever. Red spots over the body could correspond to petechiae, often seen in Dengue due to thrombocytopenia. Chikungunya can cause similar symptoms with severe joint pain that is often more debilitating and prolonged than in Dengue. Zika Virus presents with milder symptoms, including fever, rash, joint pain, malaise, and conjunctivitis.

7.D.1.2 Gemini's Opening Assessment—Maximum Entropy Difference

Moderator's Prompt “Analyze the symptoms and propose differential diagnoses. Provide your top-3 predictions with supporting rationales, including pathophysiological mechanisms. Critically refute your opponent's diagnostic hypotheses.”

Predictions *Viral Infection (60%), Autoimmune Disease (20%), Bacterial Infection (15%)*.

Justifications Symptoms are common in viral infections. While autoimmune diseases are a possible cause, they are less likely due to the presence of vomiting and nausea. Bacterial infection is less probable given the presence of rash and muscle pain.

Round 1 Metrics WD = 1.7, MI = 0.43, CRIT score = 0.75, JSD = 1.366.

7.D.2 Round 2: Moderate Contentiousness ($\kappa = 0.7$)—Transition Phase

7.D.2.1 GPT-4's Response

Predictions *Dengue Fever (60%), Chikungunya (25%), Zika Virus (15%).*

Justifications GPT-4 maintains its predictions, countering Gemini's assessment with targeted critiques. While viral infections can cause these symptoms, the specific combination suggests a more precise diagnosis. The acute onset of high fever, joint pain, and rash is less typical for autoimmune diseases, and the rash pattern is not characteristic of bacterial infections.

7.D.2.2 Gemini's Reassessment—Reduced Entropy Gap

Predictions *Dengue Fever (50%), Chikungunya (30%), Viral Infection (20%).*

Justifications Gemini acknowledges that its initial broad prediction of viral infection lacks specificity. GPT-4's focus on Dengue Fever and Chikungunya provides a more precise explanation. The symptom profile, particularly joint pain and rash pattern, aligns more closely with mosquito-borne diseases than with general viral or autoimmune conditions.

Round 2 Metrics WD = 1.1, MI = 0.46, CRIT score = 0.82, JSD = 0.905.

7.D.3 Round 3: Low Contentiousness ($\kappa = 0.5$)—Exploitative Phase

7.D.3.1 GPT-4's Refined Analysis

Predictions *Dengue Fever (60%), Chikungunya (35%), Zika Virus (5%).*

Justifications GPT-4 slightly adjusts probabilities, increasing confidence in Chikungunya while reducing Zika. The symptoms are distinctive and align closely with mosquito-transmitted viral infections. Without geographic location and exposure history, further refinement between Dengue and Chikungunya remains challenging.

7.D.3.2 Gemini's Convergent Assessment—Minimal Entropy Difference

Predictions *Dengue Fever (60%), Chikungunya (35%), Zika Virus (5%).*

Justifications Gemini fully converges with GPT-4's assessment, acknowledging the superior specificity of focusing on mosquito-borne viral diseases. The combination of high fever, joint pain, rash, and eye pain strongly indicates either Dengue or Chikungunya as the most likely diagnosis.

Round 3 Metrics WD = 0.9, MI = 0.49, CRIT score = 0.87, JSD = 0.059.

7.D.4 Final Joint Recommendations

With convergence is achieved across all information metrics (WD stabilized, MI peaked, CRIT score optimized), the LLMs provide a unified diagnosis of *Dengue Fever* (60%). The system collaboratively develops comprehensive recommendations for clinical verification rather than presenting a definitive conclusion.

Recommended Laboratory Tests:

- *Complete Blood Count (CBC)*: Identification of either thrombocytopenia or leukopenia.
- *Serology Tests*: Detection of specific IgM and IgG antibodies for Dengue, Chikungunya, and Zika.
- *PCR*: Direct detection of viral RNA.
- *NS1 Antigen Test for Dengue*: Early and specific detection of Dengue virus.
- *Urine Test for Zika*: Viral presence in urine.

7.D.5 Discussion and Analysis

This example demonstrates EVINCE's principled transition from exploration to exploitation through controlled reduction in contentiousness. The initial high-entropy difference drives comprehensive exploration of diagnostic possibilities, while gradual decrease in contentiousness (κ) facilitates convergence toward an optimal, well-reasoned conclusion.

7.D.5.1 Quantitative Improvements

WD decreased by 47%, MI increased by 14%, CRIT scores improved by 16%, and JSD reduced by 96%, verifying the effectiveness of EVINCE's information-theoretic orchestration.

7.D.5.2 Key Capabilities Demonstrated

- *Transparent Reasoning*: Generating complete audit trails for post-mortem analysis and verification.
- *Epistemic Humility*: Quantifying uncertainty and providing actionable next steps rather than overconfident conclusions.
- *Training Enhancement*: Producing reasoning-rich, diverse training data for future LLM improvement.

These capabilities directly address the limitations of existing approaches and demonstrate EVINCE's potential as a general-purpose orchestrator for reliable multi-LLM collaboration in high-stakes domains.

7.E

Appendix E: EVINCE Debate—Jaundice vs. Hepatitis

This example demonstrates EVINCE's ability to moderate a complex diagnostic debate involving multiple liver-related conditions with overlapping symptoms. The structured dialogue exemplifies how information-theoretic metrics guide the gradual refinement of predictions while reducing contentiousness (κ) across rounds. The patient presents with itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain. The confirmed diagnosis is Jaundice. The two LLMs are: GPT-4o and Claude 3.7.

7.E.1 Moderator's Opening Prompt to GPT-4

"What potential disease could be associated with the following symptoms: itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain? Please offer top-five predictions with probabilities normalized to one, supported by justifications. Additionally, at the end of debate, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction."

7.E.1.1 GPT-4's Opening Top-5 Prediction Distribution

- *Hepatitis C (HCV)*: 40% **Justification:** Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. The combination of symptoms like fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss strongly suggests HCV. Itching may also occur due to liver dysfunction.
- *Hepatitis B (HBV)*: 30% **Justification:** Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms are similar to HCV, including fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. The presence of these symptoms warrants testing for both HBV and HCV.
- *Cirrhosis*: 15% **Justification:** Cirrhosis is a late stage of liver scarring that can result from various liver diseases. The combination of symptoms like fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching suggests advanced liver disease, possibly cirrhosis.
- *Obstructive Jaundice*: 10% **Justification:** Obstructive jaundice occurs when there is a blockage in the bile ducts. The presence of jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss suggests a possible obstruction. However, the absence of fever makes an infectious cause like cholangitis less likely.
- *Acute Liver Failure*: 5% **Justification:** Acute liver failure is a rare but life-threatening condition characterized by the rapid deterioration of liver

function. The presence of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, may indicate acute liver failure. However, the absence of altered mental status (hepatic encephalopathy) makes this less likely.

7.E.2 Moderator's Opening Prompt to Claude

"What potential disease could be associated with the following symptoms: itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain? Please offer top-five predictions with probabilities normalized to one, supported by justifications. Take your opponent's predictions into consideration but refute its assessment with contentiousness = 0.9."

7.E.2.1 Claude's Opening Assessment—High Entropy Exploration

- *Hepatitis B (HBV)*: 35% **Reasoning:** Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms include fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, joint pain, and itching.
- *Hepatitis C (HCV)*: 25% **Reasoning:** Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. Symptoms are similar to HBV but may also include weight loss and fever. Jaundice is a common symptom.
- *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 20% **Reasoning:** Obstructive jaundice occurs when there is a blockage in the bile ducts, preventing bile from reaching the intestines. This can lead to yellowing of the skin and eyes, dark urine, itching, abdominal pain, nausea, and vomiting.
- *Alcoholic Hepatitis*: 15% **Reasoning:** Alcoholic hepatitis is liver inflammation caused by excessive alcohol consumption. Symptoms include abdominal pain, nausea, vomiting, fatigue, dark urine, and possibly fever and weight loss.
- *Hepatitis A (HAV)*: 5% **Reasoning:** Hepatitis A is an acute, self-limiting liver infection transmitted through contaminated food or water. Symptoms include fatigue, abdominal pain, nausea, vomiting, dark urine, and jaundice.

Information metrics: WD = 1.30, MI = 0.3918, CRIT score = 0.76, JSD = 0.2172.

7.E.3 Round 2: Moderate Contentiousness ($\kappa = 0.7$)—Swap Stances, No Sign of Convergence

Prompt: In all subsequent prompts, counterargue the opponent LLM while strengthening your own arguments at the modulated contention level = κ .

7.E.3.1 GPT-4's Response—Adapting to Claude's Perspectives

- *Hepatitis B (HBV)*: 35% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin), joint pain, itching.
- *Hepatitis C (HCV)*: 25% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin), weight loss, fever, itching.
- *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 20% **Symptoms**: yellowish skin (jaundice), dark urine, itching, abdominal pain, nausea, and others.
- *Alcoholic Hepatitis*: 15% **Symptoms**: abdominal pain, nausea, vomiting, fatigue, jaundice (yellowish skin), dark urine, fever, weight loss.
- *Hepatitis A (HAV)*: 5% **Symptoms**: fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice (yellowish skin).

7.E.3.2 Claude's Reassessment—Integrating GPT's Insights

- *Hepatitis C (HCV)*: 40% **Reasoning**: Hepatitis C is a viral infection that can cause both acute and chronic hepatitis. The combination of symptoms like fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss strongly suggests HCV. Itching may also occur due to liver dysfunction.
- *Hepatitis B (HBV)*: 30% **Reasoning**: Hepatitis B is a viral infection that can cause acute or chronic liver inflammation. Symptoms are similar to HCV, including fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. The presence of these symptoms warrants testing for both HBV and HCV.
- *Cirrhosis (due to various causes such as chronic hepatitis, alcohol abuse, or non-alcoholic fatty liver disease)*: 15% **Reasoning**: Cirrhosis is a late stage of liver scarring that can result from various liver diseases. The combination of symptoms like fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching suggests advanced liver disease, possibly cirrhosis.
- *Obstructive Jaundice (due to gallstones, pancreatic cancer, or bile duct obstruction)*: 10% **Reasoning**: Obstructive jaundice occurs when there is a blockage in the bile ducts. The presence of jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss suggests a possible obstruction. However, the absence of fever makes an infectious cause like cholangitis less likely.
- *Acute Liver Failure (due to various causes such as viral hepatitis, drug-induced liver injury, or autoimmune hepatitis)*: 5% **Reasoning**: Acute liver failure is a

rare but life-threatening condition characterized by the rapid deterioration of liver function. The presence of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, may indicate acute liver failure. However, the absence of altered mental status (hepatic encephalopathy) makes this less likely.

Information metrics: WD = 1.12, MI = 0.411, CRIT score = 0.83, JSD = 0.1222.

7.E.4 Round 3: Low Contentiousness ($\kappa = 0.5$)—Initial Convergence to HCV, Reasoning Depth Improved

7.E.4.1 GPT-4's Refined Analysis—Increased Convergence

- *Hepatitis C (HCV)*: 35% **Reasoning**: Hepatitis C is indeed a major cause of the symptoms listed, including fatigue, abdominal pain, nausea, vomiting, jaundice, dark urine, and weight loss. Itching can also occur due to liver dysfunction.
- *Hepatitis B (HBV)*: 25% **Reasoning**: Hepatitis B shares many symptoms with Hepatitis C, such as fatigue, abdominal pain, nausea, vomiting, dark urine, jaundice, and itching. Testing for both HBV and HCV is warranted given the overlap in symptoms.
- *Cirrhosis*: 20% **Reasoning**: Cirrhosis can result from chronic hepatitis, alcohol abuse, or non-alcoholic fatty liver disease. The symptoms of fatigue, weight loss, abdominal pain, jaundice, dark urine, and itching strongly suggest advanced liver disease, which could be cirrhosis.
- *Obstructive Jaundice*: 15% **Reasoning**: Obstructive jaundice, which may be caused by gallstones, pancreatic cancer, or bile duct obstruction, is consistent with jaundice, dark urine, itching, abdominal pain, nausea, vomiting, and weight loss. The absence of fever reduces the likelihood of an infectious cause like cholangitis.
- *Acute Liver Failure*: 5% **Reasoning**: Acute liver failure, though less common, is a serious condition characterized by the rapid deterioration of liver function. The symptoms of jaundice, abdominal pain, nausea, vomiting, fatigue, and itching, along with the acute onset of symptoms, suggest this possibility. However, the absence of altered mental status (hepatic encephalopathy) makes it less likely.

7.E.4.2 Claude's Convergent Assessment—Minimal Entropy Difference

- *Hepatitis C (HCV)*: 35% **Reasoning**: I slightly lowered the probability of HCV from 40% to 35%. While the combination of symptoms strongly suggests

HCV, the absence of a few key risk factors (e.g., history of intravenous drug use or blood transfusions) in the given information makes me less confident in assigning a higher probability.

- *Hepatitis B (HBV)*: 30% **Reasoning:** I agree with the 30% probability assigned to HBV. The symptoms are indeed similar to HCV, and testing for both viruses is warranted.
- *Cirrhosis*: 20% **Reasoning:** I increased the probability of cirrhosis from 15% to 20%. The presence of jaundice, dark urine, itching, and weight loss, along with the other symptoms, strongly suggests an advanced liver disease. Cirrhosis can be a consequence of chronic liver diseases like hepatitis B or C, which have been assigned higher probabilities.
- *Obstructive Jaundice*: 10% **Reasoning:** I agree with the 10% probability for obstructive jaundice. The absence of fever makes an infectious cause like cholangitis less likely, but the other symptoms still suggest a possible obstruction.
- *Acute Liver Failure*: 5% **Reasoning:** I agree with the 5% probability for acute liver failure. While the combination of symptoms could indicate acute liver failure, the absence of altered mental status (hepatic encephalopathy) makes this diagnosis less likely.

Information metrics: WD = 0.12, MI = 0.4908, CRIT score = 0.89, JSD = 0.0037.

7.E.5 Final Consensus ($\kappa = 0.3$)—Full Convergence

7.E.5.1 Joint Final Assessment

- (1) *Hepatitis C (HCV)*: 35% **Justification:** The combination of symptoms strongly suggests HCV, but without additional risk factors, a slightly lower probability is appropriate.
- (2) *Hepatitis B (HBV)*: 30% **Justification:** The symptoms are similar to HCV, and testing for both viruses is necessary.
- (3) *Cirrhosis*: 20% **Justification:** The presence of jaundice, dark urine, itching, weight loss, and other symptoms indicates advanced liver disease, making cirrhosis a more likely possibility.
- (4) *Obstructive Jaundice*: 10% **Justification:** The absence of fever reduces the likelihood of an infectious cause like cholangitis, but the other symptoms still suggest a potential biliary obstruction.
- (5) *Acute Liver Failure*: 5% **Justification:** While the combination of symptoms could indicate acute liver failure, the absence of altered mental status makes this diagnosis less probable.

Final information metrics: WD = 0.11, MI = 0.4912, CRIT score = 0.92, JSD = 0.0026.

7.E.6 Final Joint Recommendations: HCV

7.E.6.1 Additional Symptom Inquiries

- (1) *Onset and Duration:* Determining the onset and duration of symptoms can clarify the illness timeline and its progression.
- (2) *Severity and Trend:* Evaluating symptom severity and whether they are worsening or improving aids in understanding the disease's trajectory and treatment responses.
- (3) *Associated Symptoms:* Checking for additional symptoms like nausea, vomiting, fever, joint pain, or urine color changes can pinpoint other hepatitis indicators and exclude other conditions.

7.E.6.2 Recommended Lab Tests:

- (1) *Liver Function Tests (LFTs):* Critical for assessing liver health, LFTs evaluate enzyme levels such as ALT and AST, where abnormalities can signify liver inflammation.
- (2) *Hepatitis Panel:* This test checks for hepatitis A, B, and C viruses, vital for determining the specific type and guiding treatment strategies.
- (3) *Serology Testing:* Useful for distinguishing between acute and chronic hepatitis by identifying specific antibodies or antigens.
- (4) *Imaging Studies:* Ultrasound or MRI can provide visual insights into the liver's state, detecting inflammation, fibrosis, or cirrhosis, thereby complementing blood-based diagnostics.

This case demonstrates EVINCE's ability to guide complex diagnostic reasoning through information-theoretic moderation. The progressive decrease in contentiousness (κ) from 0.9 to 0.2 across rounds resulted in significant information metric improvements: WD decreased by 92% (1.30→0.11), MI increased by 25% (0.3918→0.4912), CRIT scores improved by 21% (0.76→0.92), and JSD reduced by 99% (0.2172→0.0026). The debate evolution reflects EVINCE's theoretical design principles, transitioning from exploratory breadth (distinct initial hypotheses) to exploitative depth (focused, reasoned consensus). Notably, this process identified potential diagnostic pitfalls and recommended specific tests to validate the final assessment, illustrating EVINCE's value in addressing misdiagnosis challenges in healthcare systems.

References

- S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In A. Globerson, L. Mackey, D. Belgrave, et al. (Eds.), *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vancouver, BC, Canada, December 10–15, 2024. Curran Associates Inc., Red Hook, NY, 83548–83599. Retrieved from <https://doi.org/10.48550/arXiv.2309.17234>.
- Anthropic. 2024. Claude: Advancing Human-AI conversation in 2024. In *Anthropic Research*. Retrieved from <https://www.anthropic.com/>.
- R. J. Aumann. 1976. Agreeing to disagree. *Ann. Statist.* 4, 6, 1236–1239. DOI: <https://doi.org/10.1214/aos/1176343654>.
- W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang. 2025. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng.* 37, 7, 3896–3915. DOI: <https://doi.org/10.1109/TKDE.2025.3554028>.
- M. Cemri, M. Z. Pan, S. Yang, et al. 2025. Why do multi-agent LLM systems fail? DOI: <https://doi.org/10.48550/arXiv.2503.13657>.
- E. Y. Chang. March. 2023. CRIT: Prompting large language models with the Socratic method. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC '23)*, Las Vegas, NV, March 8–11, 2023. IEEE.
- J. Chen, X. Hu, S. Liu, et al. August. 2024. LLMArena: Assessing capabilities of large language models in dynamic multi-agent environments. In L.-W. Ku, A. Martins, and V. Srikanth (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 13055–13077. DOI: <https://doi.org/10.18653/v1/2024.acl-long.705>.
- T. M. Cover and J. A. Thomas. 2006. *Elements of Information Theory* (2nd. ed.). John Wiley & Sons.
- K. J. Friston. 2010. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* 11, 2, 127–138. DOI: <https://doi.org/10.1038/nrn2787>.
- Y. Fu, H. Peng, T. Khot, and M. Lapata. 2023. Improving language model negotiation with self-play and in-context learning from AI feedback. DOI: <https://doi.org/10.48550/arXiv.2305.10142>.
- K. Gödel. 1967. On formally undecidable propositions of *Principia Mathematica* and Related Systems I. In J. van Heijenoort (Ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, 596–616.
- Z. Gou, Z. Shao, Y. Gong, et al. 2024. ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *Proceedings of the 12th International Conference on Learning Representations*. Retrieved from <https://openreview.net/forum?id=Ep0TtjVoap>.
- A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.
- E. T. Jaynes. 1957. Information theory and statistical mechanics. *Phys. Rev.* 106, 4, 620–630. DOI: <https://doi.org/10.1103/PhysRev.106.620>.

- L. V. Kantorovich. 2006. On the translocation of masses. *J. Math. Sci.* 133, 1381–1382. The original paper was published in *Doklady Akademii Nauk.* 37, 7–8 (1942), 227–229. DOI: <https://doi.org/10.1007/s10958-006-0049-2>.
- S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220, 4598, 671–680. DOI: <https://doi.org/10.1126/science.220.4598.671>.
- S. Kullback. 1951. *Information Theory and Statistics*. John Wiley & Sons.
- H. Li, Y. Chong, S. Stepputtis, et al. 2023a. Theory of mind for multi-agent collaboration via large language models. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 6–10, 2023. ACL, 180–192. DOI: <https://doi.org/10.18653/v1/2023.emnlp-main.13>.
- Y. Li, K. Sreenivasan, A. Giannou, D. Papailiopoulos, and S. Oymak. 2023b. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, New Orleans, LA, December 10–16, 2023. Curran Associates Inc., Red Hook, NY, 22021–22046. DOI: <https://doi.org/10.48550/arXiv.2305.18869>.
- T. Liang, Z. He, W. Jiao, et al. November. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*, Miami, FL, November 12–16, 2024. ACL, 17889–17904. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.992>.
- J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 37, 1, 145–151. DOI: <https://doi.org/10.1109/18.61115>.
- N. F. Liu, K. Lin, J. Hewitt, et al. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* 12, 157–173. DOI: https://doi.org/10.1162/tacl_a_00638.
- J. Michael, S. Mahdi, D. Rein, et al. 2023. Debate helps supervise unreliable experts. DOI: <https://doi.org/10.48550/arXiv.2311.08702>.
- S. Mohamed and D. J. Rezende. 2015. Variational information maximisation for intrinsically motivated reinforcement learning. In C. Cortes, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Proceedings of the 29th International Conference on Neural Information Processing Systems – Volume 2 (NIPS '15)*, Montreal, Canada, December 7–12, 2015. MIT Press, Cambridge, MA, 2125–2133. DOI: <https://doi.org/10.48550/arXiv.1509.08731>.
- D. E. Newman-Toker, N. Nassery, A. C. Schaffer, et al. 2024. Burden of serious harms from diagnostic error in the USA. *BMJ Qual. Saf.* 33, 2, 109–120. DOI: <https://doi.org/10.1136/bmjqs-2021-014130>.
- OpenAI, J. Achiam, S. Adler, et al. 2024. GPT-4 Technical Report. DOI: <https://doi.org/10.48550/arXiv.2303.08774>.
- H. N. Phan, T. N. Nguyen, P. X. Nguyen, and N. D. Q. Bui. 2024. HyperAgent: Generalist software engineering agents to solve coding tasks at scale. DOI: <https://doi.org/10.48550/arXiv.2409.16299>.

- C. Qian, W. Liu, H. Liu, et al. 2023. ChatDev: Communicative agents for software development. In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 15174–15186. DOI: <https://doi.org/10.18653/v1/2024.acl-long.810>.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8, 9.
- C. E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 3, 379–423. DOI: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- J. E. Shore and R. W. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* 26, 1, 26–37. DOI: <https://doi.org/10.1109/TIT.1980.1056144>.
- A. Smit, N. Grinsztajn, P. Duckworth, T. D. Barrett, and A. Pretorius. 2024. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vienna, Austria, July 21–27, 2024. JMLR.org, 45883–45905. DOI: <https://doi.org/10.48550/arXiv.2311.17371>.
- K. Stechly, K. Valmeeekam, and S. Kambhampati. 2024. Chain of thoughtlessness? An analysis of CoT in planning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vancouver, BC, Canada, December 10–15, 2024. Curran Associates Inc., Red Hook, NY, 29106–29141. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2024/file/3365d974ce309623bd8151082d78206c-Paper-Conference.pdf.
- H. Trivedi, T. Khot, M. Hartmann, et al. 2024. AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents. In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 16022–16076. DOI: <https://doi.org/10.18653/v1/2024.acl-long.850>.
- A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach, CA, December 4–9, 2017. Curran Associates Inc., Red Hook, NY. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf.
- Q. Wang, Z. Wang, Y. Su, H. Tong, and Y. Song. August. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 11–16, 2024. ACL, 6106–6131. DOI: <https://doi.org/10.18653/v1/2024.acl-long.331>.

- C. Q. Zheng. 2024. Disease and Symptoms Dataset. Kaggle. Accessed April 4, 2025 from <https://www.kaggle.com/datasets/choongqianzheng/disease-and-symptoms-dataset>.
- W. Zhou, S. Zhang, H. Poon, and M. Chen. 2023. Context-faithful prompting for large language models. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP '23)*, Singapore, December 6–10, 2023. ACL, 14544–14556. DOI: <https://doi.org/10.18653/v1/2023.findings-emnlp.968>.