# Entropy Variation and Information Competence: Enhancing Predictive Accuracy of Collaborative Language Models

Anonymous ACL submission

## Abstract

This paper introduces EVINCE (Entropy Variation and INformation CompetencE), a cutting-edge dialogue framework that orchestrates adversarial debates and collaborative insights among multiple large language models (LLMs). By integrating advanced principles from conditional statistics, information theory, and in-context learning, EVINCE masterfully balances the exploration of diverse perspectives with the exploitation of established priors. Central to our innovation is the validation of the dual entropy theory, which we developed to determine the optimal pairing of LLMs with one high and one low entropy for enhanced probabilistic prediction accuracy. We also employ several information-theoretic metrics, such as mutual information, cross-entropy, Wasserstein distance, and Jensen-Shannon divergence, to measure communication opportunities, dialogue progress, and convergence. This meticulous approach fosters an interpretable and productive multi-LLM dialogue, leading to more informed and reliable outcomes. We illustrate EVINCE's potential by applying it to healthcare, demonstrating its effectiveness in improving disease diagnosis, and discuss its broader implications for enhancing decision-making across various domains.

## 1 Introduction

Ensemble approaches in machine learning, where multiple predictors combine to address classification and regression tasks, have consistently demonstrated superior performance compared to individual models (Kuncheva and Whitaker, 2003; Dietterich, 2000; Krogh and Vedelsby, 1995). The diversity of errors across these models is a crucial factor in their effectiveness. Recent research has explored extending this ensemble concept to Large Language Models (LLMs) collaborating on classification, question answering, and other tasks (Michael et al., 2023; Chan et al., 2023; Liang et al., 2023; Du et al., 2023). While initial findings suggest accuracy improvements similar to traditional ensemble methods, multi-LLM collaboration holds the potential for much broader impact. As noted by (Chang, 2023a), this approach can unearth novel perspectives, mitigate biases, and even contribute to creative endeavors like writing a novel, thereby extending its capabilities far beyond accuracy gains.

Achieving optimal performance in multi-LLM ensembles requires more than simply maximizing error diversity. A critical balance must be struck between confident, well-supported predictions and the exploration of novel and diverse perspectives. To facilitate this balanced approach, we introduce EVINCE (Entropy Variation through INformation CompetencE), a framework designed to foster structured debates among multiple LLMs, thereby maximizing prediction accuracy while encouraging the exploration of alternative viewpoints to mitigate biases. EVINCE represents a new paradigm in collaborative LLM research, effectively navigating the trade-off between exploration and exploitation in joint predictions. EVINCE rests on three key theoretical pillars:

**Conditional Statistics:** Conditional Statistics: By placing LLMs in adversarial stances and demanding rigorous justification for their positions, EVINCE leverages in-context learning to elicit from the opposing LLMs diverse perspectives backed by robust reasoning and evidence. This method, rooted in the Bayesian framework of conditional statistics (Finn et al., 2017; Brown et al., 2020; Xie et al., 2021), effectively modifies the linguistic behaviors of LLMs, shifting them away from the default optimization for maximum likelihood next-token prediction.

**Dual Entropy:** Our theoretical proof (via Jensen's Inequality) (Section 3.3) and empirical studies (Section 4) reveal a key insight: optimal accuracy in a two-LLM ensemble is achieved when the agents begin with differing levels of entropy. Specifically, one LLM should initially exhibit high prediction entropy, signaling a willingness to explore diverse perspectives, while the other should maintain low entropy, emphasizing precision and stability. This dual entropy configuration maximizes the ensemble's ability to balance exploration and exploitation, as the high-entropy LLM introduces a wider range of possibilities, including those that may challenge or counteract potential biases in the low-entropy LLM's initial predictions. Meanwhile, the low-entropy LLM acts as a stabilizing force, grounding the exploration in a foundation of established knowledge. Through a process of communication and reasoning, evaluated by the Socratic method and metrics from information theory (which we will elaborate on in the subsequent discussion), the two agents converge towards a collaborative and accurate prediction, ideally mitigating biases that may have been present in either agent's initial viewpoints. This finding challenges the traditional notion that faster agreement among agents necessarily leads to better outcomes, highlighting the importance of initial diversity in avoiding tunnel vision and fostering robust decision-making.

**From Divergence to Conciliatory:** EVINCE begins by positioning two agents in a state of dual entropy, then fosters effective information exchange between LLMs to gradually reduce cross entropy and Wasserstein distance, and maximize mutual information in their prediction distributions. This enhances the depth and breadth of their predictions. The framework initiates debates with high contentiousness (Chang, 2023a), using mutual information to quantify the potential for productive communication. As the diversity of predictions, measured by the divergence metrics, decreases below a threshold, contentiousness is modulated, encouraging collaboration. This culminates in a joint prediction, accompanied by explainable arguments and counterarguments.

Diversity in predictive modeling can introduce noise, while an overly strong belief in existing perspectives may hinder the exploration of new ideas. To address these challenges, EVINCE employs several proxy metrics in conjunction with a "contentiousness" parameter to achieve a balance. By reasoning through and analyzing several case studies, we demonstrate how EVINCE enhances prediction accuracy, robustness, and stability. The framework facilitates a debate process where rigorous arguments and counterarguments are recorded, making the decision-making process transparent. Transparency allows humans to understand the recommendations clearly, provide feedback, and make final predictions that are well-informed, encompassing a comprehensive range of pros and cons.

The main contributions of this paper are:

1. **EVINCE Framework Design**: Different from using debate as a way to improve accuracy via redundancy, EVINCE's approach is vastly different and thus facilitates information discovery, bias mitigation, and decision-making that requires both breadth and depth of information.

2. **Theoretical Foundations**: We establish a theoretical basis for EVINCE, rooted in conditional Bayesian statistics, mutual information, and dual entropy. These principles are applied to measure, monitor, and modulate collaborative LLM interactions, contributing to a deeper understanding of how LLMs can effectively cooperate for improved decision-making. The dual entropy theory is novel and ground-breaking, illustrating how a productive decision-making process should start with room for diverse input and stable objectives, and then, through information exchange, converge to optimal decision/prediction.

3. **Empirical Validation**: We provide empirical validation of EVINCE's underlying maxims and theories, highlighting the framework's effectiveness in balancing exploration and exploitation to enhance prediction accuracy. We also introduce a set of maxims derived from our empirical findings, offering practical guidance for optimizing mutual information and minimize various divergence measures.

## 2 Related Work

The core objective of adversarial debate, as embodied in EVINCE, is to foster diverse opinions and challenge assumptions, ultimately leading to more comprehensive and informed decision-making. This contrasts with traditional ensemble learning methods, which prioritize error diversity for improved accuracy.

### 2.1 Ensemble and Multi-Agent Learning

Ensemble methods like Bagging (Breiman, 1996), Boosting (Freund and Schapire, 1997), and Mixtures of Experts (Jacobs et al., 1991) have focused on combining predictions from multiple models to improve overall accuracy. Early LLM debate frameworks also followed this trend (Michael et al., 2023; Chan et al., 2023; Liang et al., 2023; Du et al., 2023).

EVINCE distinguishes itself by prioritizing the generation of diverse predictions to explore a wider range of perspectives. Recent research on multi-LLM collaboration, building on in-context learning and Bayesian frameworks (Xie et al., 2021; Zhang et al., 2023), has shown promising results. However, the challenge remains in effectively moderating communication between LLMs. EVINCE addresses this by employing quantitative measures to calibrate and adjust individual LLM behaviors, contributing to the growing field of multi-agent LLM communication (Abdelnabi et al., 2024; Chan et al., 2023; Fu et al., 2023; Li et al., 2023; Liang et al., 2023; Michael et al., 2023; Smit et al., 2024).

### 2.2 Metrics for Managing Diversity, Contentiousness, Information Quality, and Convergence

EVINCE employs various metrics to manage the debate's dynamics and progress:

- **Fostering Diversity & Quality**: Shannon entropy and relative entropy measure diversity of perspectives (Cover and Thomas, 2006; Shannon, 1948), while the CRIT algorithm assesses argument quality (Chang, 2023b).

- **Balancing Exploration & Stability**: Correlation coefficients track opinion evolution and debate stability (Brown et al., 2005), Wasserstein Distance measures prediction distribution differences (Kantorovich, 1942; Rubner et al., 2000; Villani, 2008), and Mutual Information quantifies information overlap (Cover and Thomas, 2006).

- **Examining Information Overlap & Termination**: Jensen-Shannon Divergence assesses distribution similarity (Lin, 1991), Cross Entropy measures asymmetric differences (Shore and Johnson, 1980), and Kullback-Leibler Divergence reveals asymmetric differences between probability distributions (Kullback, 1951).

Section 3 details how EVINCE utilizes these metrics to balance exploration and exploitation, leading to optimal predictions. The dual entropy theorem provides further theoretical justification for the framework.

## 3 Maxims, Algorithm, and Theorem

**Problem Statement:** Organize a structured debate between two equally competent large language models (LLMs), LLM$_A$ and LLM$_B$, to conduct $t$ rounds. At each round $t$, each model produces a probability distribution, denoted as $P_A^{(t)}$ and $P_B^{(t)}$, over $C$ possible outcomes, accompanied by supporting arguments $R_A^{(t)}$ and $R_B^{(t)}$. The goal is to design an iterative debate process that leverages the structured exchange of arguments to enable the models to converge on an optimal prediction distribution $P^*$ across the $C$ classes.

### 3.1 Maxims with Theoretical Foundations

Progress towards the optimality goal is guided and measured by metrics introduced in Section 2. This section explains how they can be used in complementary ways to facilitate proper trade-offs between diversity and convergence, exploration and exploitation, and several other factors.

**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.

*How?* Choosing LLMs with comparable performance on a shared validation set, a balanced debate can be ensured. Suitable models include GPT-4, Claude, and Gemini. Conditioning different instances of the same LLM to support opposing stances on a subject matter can also be effective due to the theoretical justification of in-context learning with conditional Bayesian statistics (Xie et al., 2021).

**Maxim #2: Encourage the Accurate Rather Than the "Popular" Prediction:** Typically, LLMs, with their maximum likelihood next-token prediction objective, tend to favor the most popular predictions. By conditioning LLMs within specific contexts, we can prioritize specific stance over popularity, mitigating confirmation biases.

*How?* Using the proxy metrics in Table 1, EVINCE dynamically adjusts the "contentiousness" level in debates (see Appendix G for details). These metrics quantify agreement, diversity, and mutual information, promoting productive information exchange and enhancing prediction quality.

| Metric | Pros | Cons | Remedies |
|---|---|---|---|
| Cross Entropy (CE) (Shore and Johnson, 1980) | Measures how well the predictions of one model fit the actual distribution of another model's outputs (asymmetric). | Computationally intensive especially with large models and data sets; sensitive to the exact nature of probability distributions. | Optimize computation strategies; use approximations or sampling methods to manage large data sets or complex models. |
| Entropy Shannon (Shannon, 1948) | Indicates level of diversity; high suggests exploration of possibilities, and low for confidence on few choices | High entropy might indicate noise rather than useful diversity; low entropy might mask important variability. | Use critical reading methods (Appendix A) to assess argument quality; implement noise detection to differentiate between useful diversity and noise. |
| Jensen-Shannon Divergence (JS) (Lin, 1991) | Symmetric and bounded (0 to 1), providing an interpretable measure of distributional differences. | May be less sensitive to small differences between distributions. | Increase sensitivity settings or resolution of the metric; combine with other metrics to capture finer distinctions between distributions. |
| KL Divergence (Kullback, 1951) | Measures difference between two probabilistic distributions. | Asymmetric; not well-defined if a distribution has zero probabilities | Use smoothing techniques to avoid zero probabilities; consider symmetric alternatives like JS divergence |
| Mutual Info (MI) (Shore and Johnson, 1980) | Measures reduction of uncertainty; symmetric. | Does not indicate the directionality of information flow. | Supplement with directional information metrics; normalized with max entropy of A and B. |
| Wasserstein Distance (WD) (Kantorovich, 1942) | Direct measure of how similar or different the model outputs are; it depicts symmetric relationship. | Not bounded but can be normalized or bounded for consistent interpretation. | Define context-specific bounds for low, medium, and high divergence; consider normalizing it for non-directional comparisons. |

Table 1: Summary of metrics for assessing LLM debates (equations are presented in Appendix E)

**Maxim #3. Combine Predictions Weighted by Diversity and Quality:** Weighting the probability distributions from two LLMs based on diverse probabilistic insights and argument quality.

*How?* Following these three maxims:

- **Maxim #3.1 Prediction Reliability:** Estimate the reliability of predictions using entropy-based measures to quantify uncertainty and information content. Typically, lower entropy indicates higher confidence in a prediction, suggesting higher reliability.
- **Maxim #3.2 Argument Quality:** Evaluate the quality of supporting arguments using techniques inspired by the Socratic method. This includes identifying logical fallacies, assessing the relevance and credibility of evidence.
- **Maxim #3.3 Aggregation:** Employ a weighted aggregation method, such as a Bayesian model to combine weighted predictions accounting for both probabilistic insights and the quality of supporting arguments.

**Maxim #4. Evaluating the Convergence Rate of the Predictions Across the Rounds:** This aspect focuses on measuring how quickly and effectively the predictions from the LLMs converge over successive rounds, assessing the efficiency of the debate and aggregation mechanisms.

*How?* Convergence is assessed by measuring mutual information and using proxy metrics such as Wasserstein distance. When the mutual information is low or the similarity between predictions is high, the debate is considered to be converging.

### 3.2 Algorithm Specifications

With all proxy metrics and their pros, cons, and combined strengths comprehensively surveyed, and also examined by our two experiments documented in Sections 4.2 and 4.3, Algorithm 1 formally specifies the algorithm of EVINCE with the maxims.

### 3.3 Entropy Duality Theorem (EDT)

**Theorem EDT: Optimal Pairing of LLMs for Probabilistic Prediction Accuracy.** The optimal pairing of LLMs for diagnosis accuracy, in terms of stability, accuracy, and robustness, occurs when the LLMs are 1) equivalent in the quality of the information they process, and 2) exhibit contrasting entropy values in their prediction distributions—one high and one low.

[**Proof**]: In Appendix B.

## 4 Empirical Study

This empirical study investigates the application of EVINCE to disease diagnosis, leveraging large language models (LLMs) as diagnostic tools. We aim to validate the following three hypotheses:

1. *Contentiousness & Prediction Quality*: Initial LLM disagreement (measured by Wasserstein distance) increases with higher initial contentiousness but decreases as debate progresses. Individual LLM prediction uncertainty (Shannon entropy) will follow a similar pattern.

**Algorithm 1** Specifications of Algorithm EVINCE

1: **Input:** Information set $S$, Class labels $C$; Two equally competent LLMs: $\text{LLM}_A$ and $\text{LLM}_B$ (**Maxim #1**);
2: **Output:** $P_f$, final probability distribution over $C$;
3: **Variables:** $t$: debate round; $R = \emptyset$ aggregated arguments;
$P_A^{(t)}$, $P_B^{(t)}$: prediction distributions of $\text{LLM}_A$ and $\text{LLM}_B$ on $C$ of round $t$; $R_A^{(t)}$, $R_B^{(t)}$: supporting reason sets;
$\Delta = 90\%$: debate contentiousness, initialize to high to foster adversary between LLMs (**Maxim #2**);
$p$: prompt = "Predict top-$k$ probability distribution on $C$ with $S$ and $R$ at contentiousness $\Delta$";
4: **Functions:** CRIT($d$) (Chang, 2023b), Critical Reading Inquisitive Template for evaluating argument quality;
ARA (Guo et al., 2024), Algorithmic Robust Aggregation for optimal prediction aggregation (**Maxims #3**);
5: **Initial Predictions** $t = 0$**:**
LLMs generate their predictions in probability distributions with supporting reasons:

$$(P_A^{(t=0)}, R_A^{(t)}) = \text{LLM}_A(S, p), \quad (P_B^{(t=0)}, R_B^{(t)}) = \text{LLM}_B(S, p).$$

6: **Debate Iterations:**
6.1. **Update Predictions:**
Calculate the confidence-based weights using the inverse of entropy (**Maxim #3.1**):

$$\alpha = 1/(H(P_A^{(t)}) + 1), \quad \beta = 1/(H(P_B^{(t)}) + 1).$$

Use the blending mechanism to update predictions (**Maxim #3.3**):

$$P_A'^{(t)} = \alpha P_A^{(t)} + (1-\alpha) P_B^{(t)}, \quad P_B'^{(t)} = \beta P_B^{(t)} + (1-\beta) P_A^{(t)}.$$

6.2. **LLMs Generate New Predictions:** Both LLMs use accumulated $R = R \cup R_A^{(t)} \cup R_B^{(t)}$.

$$(P_A^{(t+1)}, R_A^{(t+1)}) = \text{LLM}_A((P_B'^{(t)}), R, p), \quad (P_B^{(t+1)}, R_A^{(t+1)}) = \text{LLM}_B((P_A'^{(t)}), R, p).$$

6.3. **Exit Condition Check with Wasserstein distance (Maxim #4):**

$$\textbf{If WD}(P_A^{(t+1)}, P_B^{(t+1)}) < \epsilon \quad \textbf{EXIT}; \quad t = t + 1, \quad \Delta = \Delta \times 80\%.$$

7: **Final Decision:** Weighted prediction by quality scores of the evaluator e.g., CRIT (Appendix A) (**Maxim #3.2**):

$$P_f = \Omega_A P_A^{(t+1)} + \Omega_B P_B^{(t+1)} / \Omega_A + \Omega_B.$$

2. *EDT Effectiveness & Confusion Matrices*: LLM pairs following the Entropy Duality Theorem (EDT) will have complementary error patterns, leading to higher combined prediction accuracy than non-EDT pairs.
3. *EVINCE & Historical Misdiagnoses*: EVINCE, applied to real-world data, will improve diagnostic accuracy and identify potential misdiagnoses or ambiguities within the ground truth.

**Problem Statement:** Given a set of symptoms, denoted as $S$, and a context $\kappa$, the goal is to predict a probability distribution of top-$k$ diseases over $C$ possible diseases. This is represented as $P = \text{LLM}(S, \kappa)$, where each LLM generates top-$k$ predictions on $C$ ($k \leq C$) based on the input symptoms $S$ and context $\kappa$.

$$P = (p(\text{top } 1 \text{ to } k \in D \mid S, \kappa)\,.$$

Context $\kappa$ is where dual entropy is adjusted through three knobs: temperature, the $k$ of top-$k$, and the contentious level $\Delta$. A distribution tends to have high entropy when all three knobs are set high, and vice versa.

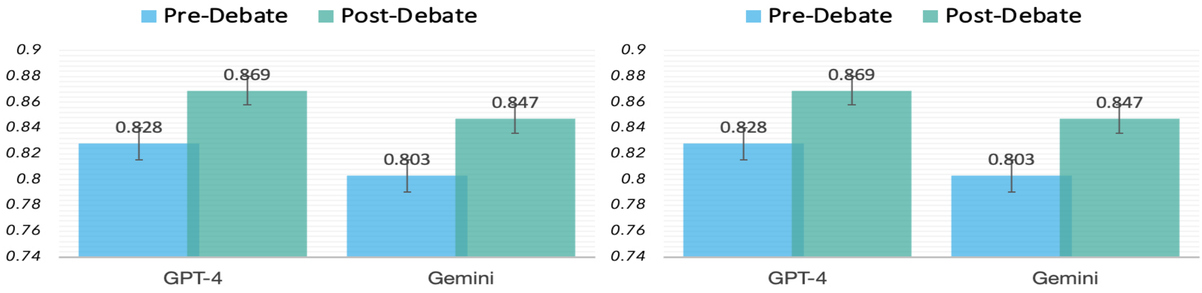
**Resources, Dataset & Data Preparation:** Our study utilizes a dataset obtained from Kaggle (Patil, 2020), which comprises 4,921 patient records. Each record includes the diagnosed disease along with up to 17 symptoms such as fever, cough, fatigue, itchiness, and difficulty breathing. We first remove duplicates from the dataset, resulting in 304 unique diagnostic instances spanning 40 diseases. (The refined dataset is uploaded as supplementary data.) Each instance acts as a test case where EVINCE utilizes the inherent knowledge of LLMs (GPT-4, Gemini, and Claude3) instead of training them through few-shot techniques on this specific dataset. Our computing resources are sponsored by Azure, with a monthly budget of US$500.

**Evaluation:** We evaluate the quality of predictions using the top-k Mean Reciprocal Rank (MRR). If one of the top-k predicted diseases matches the ground truth diagnosis, the score is the reciprocal of its rank (1 for the top prediction, 1/2 for the second, 1/3 for the third, etc.). If none of the top-k predictions are correct, the score is 0.

### 4.1 Study #1: Post vs. Pre-Debate Accuracy

For each of the 304 patient instances, we employ GPT-4, Gemini, and Claude3, to perform independent disease predictions and then use EVINCE to pair them to evaluate performance gain.

In our first experiment, we set $k = 5$ for both LLM agents. One agent had a high temperature




(a) GPT4 pairs Claude (b) GPT4 pairs Gemini

Figure 1: Pre-/post-debate accuracy on all patients on all diseases shows EVINCE helps

while the other had a low temperature. The contentiousness level was set very high ($\Delta = 0.9$ out of 1) to encourage significant cross entropy. Setting $k = 5$ ensures some minimal common ground, meaning the probability of shared information is sufficient to foster meaningful interaction. High contentiousness promotes counterarguments and information exchange.

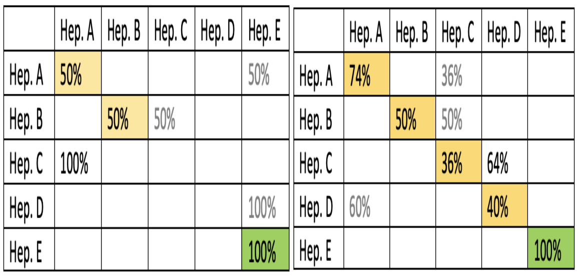
**Pre- and Post-Debate Evaluation** We conducted two sets of experiments. First, as a baseline, we constrained disease predictions to the 40 labels in the dataset, mimicking common supervised learning assumptions. While this yielded high accuracy (95-97%), it's unrealistic for real-world diagnosis where a general practitioner considers all possibilities. This constraint also highlights the flexibility of LLMs, which are not confined by training data labels and thus less prone to overfitting some erroneous labels (further discussed in the next two studies).

Next, we removed the label constraint to better simulate real-world conditions. In this unconstrained scenario, all 304 patient cases yielded stable results across GPT-4, Gemini-3, and Claude-3, with a standard deviation of just 1.5%. Prior to debate (light blue bars in Figure 1), GPT-4 led in accuracy (82.8%), followed by Gemini (80.3%) and Claude (79.5%).

Implementing EVINCE with GPT-4 and Claude-3 pairing and GPT-4 and Gemini-3 pairing consistently improved accuracy by 4-5 percentage points (green bars in Figure 1). The GPT-4 and Claude-3 pairing achieved 87.5% accuracy (Figure 1a), rivaling state-of-the-art clinical performance like the REFUEL algorithm (Peng et al., 2018).

However, the story doesn't end here. The remaining 12.5% of inaccurate cases for the GPT-Claude pairing might not be solely EVINCE's fault. If we consider the potential 11% US misdiagnosis rate reported by John Hopkins (Newman-Toker et al., 2023b), this discrepancy could point to mislabeled data in the original dataset. This presents a groundbreaking opportunity: EVINCE could potentially identify and correct errors in existing datasets, a concept we explore further in Section 4.3.

## 4.2 Study #2: Confusion vs. Opportunities



| | Hep. A | Hep. B | Hep. C | Hep. D | Hep. E |
|---|---|---|---|---|---|
| Hep. A | 50% | | | | 50% |
| Hep. B | | 50% | 50% | | |
| Hep. C | 100% | | | | |
| Hep. D | | | | | 100% |
| Hep. E | | | | | 100% |

| | Hep. A | Hep. B | Hep. C | Hep. D | Hep. E |
|---|---|---|---|---|---|
| Hep. A | 74% | | 36% | | |
| Hep. B | | 50% | 50% | | |
| Hep. C | | | 36% | 64% | |
| Hep. D | 60% | | | 40% | |
| Hep. E | | | | | 100% |

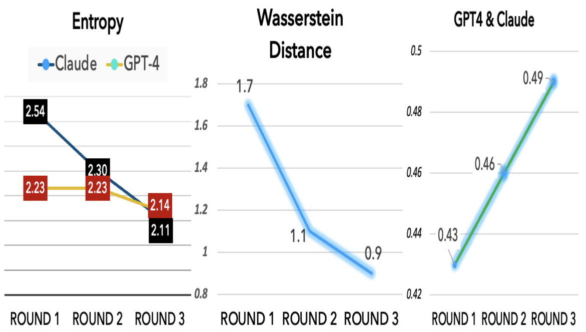
(a) GPT liver c-matrix (b) Claude liver c-matrix

Figure 2: Confusion matrices

Two key factors contribute to EVINCE's improved diagnostic accuracy: (1) structured debates with reasoning encourage LLMs to explore alternative diagnoses in both breadth and depth, leading to more comprehensive analysis and decision-making (see Appendices C and D); and (2) pairing high- and low-entropy LLMs balances exploratory diversity with exploitative stability, resulting in more robust and high-quality decisions, as demonstrated in this second study.

**Analysis of Confusion Matrices** We use confusion matrices to analyze the performance of two LLMs on diagnosing Hepatitis types A to E. GPT-4 shows limited accuracy, particularly for types C and D, achieving only 50% accuracy for types A and B. In contrast, Claude exhibits a wider spread of predictions across all Hepatitis types, as shown in Figure 2.

These matrices highlight how Claude's flexibility in exploring diverse diagnostic hypotheses can significantly aid the debate process. The initial uncertainty or "confusion" (high entropy) exhibited by Claude brings new information to the table, potentially challenging and correcting the more confident (low entropy) predictions of GPT-4, which




(a) Entropy (b) WD % (c) Norm. MI

Figure 3: Entropy, WD, and normalized MI

might otherwise stubbornly persist with incorrect diagnoses. This dynamic interplay exemplifies the delicate dance between exploration and exploitation that EVINCE facilitates. By encouraging the exploration of alternative hypotheses, even when one model seems certain, EVINCE can uncover nuances and details that lead to more accurate and comprehensive diagnoses.

**Observations from Information Metrics** Figure 3a illustrates how the entropy levels of both LLMs stabilize after three rounds of debate, indicating a convergence towards a similar, stable entropy state. This convergence is corroborated by a consistent improvement in Wasserstein distance (WD) between the two models' predictions over successive rounds, as shown in Figure 3b. Notably, Figure 3c shows that the normalized mutual information (MI) between the prediction distributions of GPT-4 and Claude improves by 14%, suggesting an increase in shared information throughout the debate. Additionally, Figure 4 shows the consistent convergence of all divergence metrics.

**Comparative Performance:** EVINCE demonstrates a 5% higher accuracy rate in diagnosing specific types of liver diseases compared to a baseline approach (Figure 1a), underscoring its capability to handle complex diagnostic scenarios effectively.

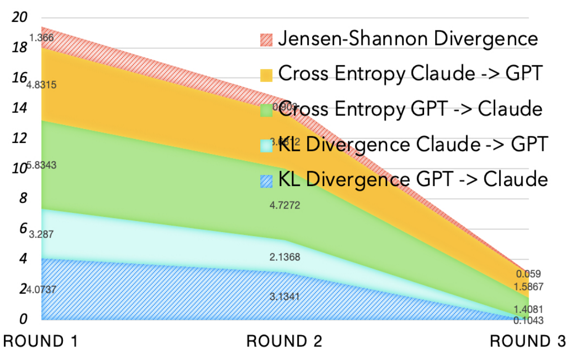



Figure 4: Convergence of all metrics

### 4.3 Study #3: Ground-Truth Remediation

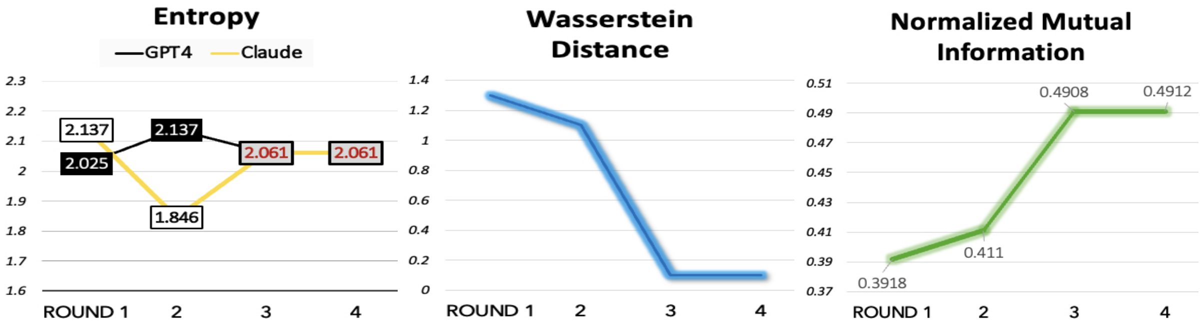
This study illustrates how EVINCE can identify potential misdiagnoses, explain the reasoning behind them, and recommend corrective actions. Traditionally, machine learning scientists rely on labeled data as "ground truth." However, as evidenced by research like that of Newman-Toker et al. (2021) (Newman-Toker et al., 2023a) from Johns Hopkins, misdiagnosis is a widespread issue in healthcare systems globally. These erroneous diagnoses, often treated as ground truth, can be perpetuated by supervised learning algorithms, exacerbating the problem within the healthcare system.

In the debate scenario detailed in Appendix D, where Jaundice is the ground truth diagnosis, Figure 5a illustrates initial differences between GPT-4 and Claude's predictions. Jaundice is absent in GPT-4's top-5 (with 0% in red), while ranked third by Claude. Although Claude influences GPT-4 to include Jaundice in its third prediction in the second round, subsequent rounds see both LLMs drop Jaundice to the fourth position of 10%.

Meanwhile, Hepatitis A, initially GPT-4's top prediction (30% in dark blue), is quickly demoted to fifth and eventually drops out of the top-5 entirely due to Claude's influence. Hepatitis B, initially ranked second by GPT-4 and top by Claude, stabilizes in the second position in rounds 3 and 4 (in light blue). Notably, Hepatitis C rises from second place on both lists to the top position and remains there (in black).

As demonstrated in the previous study, Wasserstein distance (WD) effectively measures the divergence between LLM predictions and assesses debate convergence. Figures 5b and 5c show that WD stabilizes after three debate rounds, coinciding with a plateau in normalized mutual information (MI) between GPT-4 and Claude. This stabilization suggests their predictions converge.

Figure 6 illustrates the convergence of all divergence metrics—including Jensen-Shannon divergence, cross-entropy, and Kullback-Leibler divergence—particularly between the second and third rounds. Although the final joint prediction for Hepatitis C reached a high consensus of 37.5%, it deviates from the actual condition of Jaundice, which the Kaggle dataset reports with 10% confidence. EVINCE provides general practitioners with alerts and suggests remedial actions (see Appendices D.9 and C.8) to address this discrepancy. Recommended actions include querying additional




(a) GPT & Claude top-5 (b) WD % (c) Mutual Info.

Figure 5: Remediation: Jaundice to Hepatitis

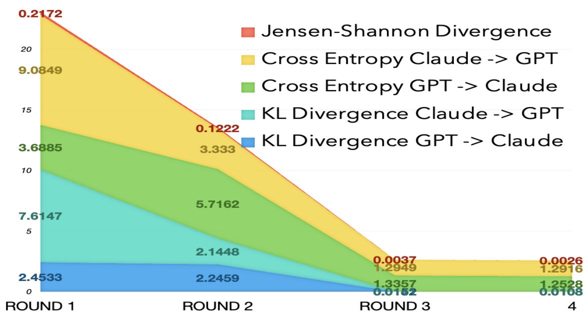



Figure 6: Convergence of all metrics

symptoms from the patient and conducting specific laboratory tests.

### 4.4 Experiment Remarks

EVINCE initiates debates with high contentiousness, encouraging dual prediction entropy between LLMs, as supported by the EDT theorem. It utilizes normalized mutual information (MI) to track shared knowledge accumulation throughout the debate, while Wasserstein distance (WD) and Jensen-Shannon divergence (JSD) quantify dissimilarity between LLM predictions.

These metrics (EDT, WD, JSD, MI) provide a comprehensive view of debate progress. WD and JSD assess the potential for further communication and refinement, while MI monitors shared understanding, aiding in determining the optimal stopping point.

The asymmetric nature of KL divergence and cross entropy warrants further investigation. Despite eventual convergence in our case studies, discrepancies observed in the second round, where one direction increases while the other decreases, suggest potential value in exploring asymmetric information. Future work will re-evaluate the use of these metrics if asymmetry proves beneficial.

## 5 Concluding Remarks

We have developed EVINCE, an innovative framework that enhances collaborative decision-making among Large Language Models (LLMs) through structured, adversarial debates. This framework leverages conditional statistics (in-context learning), information theory, and a novel concept called dual entropy to guide the debate, ensuring a balance between exploration and exploitation. EVINCE not only improves prediction accuracy and robustness but also produces explainable outcomes grounded in information metrics.

By assigning adversarial roles and adjusting the level of contentiousness, EVINCE encourages LLMs to explore a broader range of perspectives. Through mutual persuasion and the exchange of information, the reliability of predictions is significantly enhanced. The introduction of dual entropy theory, which pairs one LLM with high initial entropy (for diverse exploration) with another LLM with low entropy (for focused refinement), further stabilizes information exchange and promotes comprehensive consideration of various viewpoints.

Our validated Entropy Duality Theorem provides empirical evidence of EVINCE's effectiveness. In the domain of medical diagnostics, EVINCE outperforms traditional solo LLM approaches by identifying potential ground-truth errors and providing clear justifications for its conclusions. This success demonstrates the potential of EVINCE for broad application in various fields where informed decision-making is crucial.

Looking ahead, EVINCE is poised to drive further innovations in LLM collaboration across diverse domains. It represents a significant advancement in AI-human interaction, promoting a synergy of intelligence, reliability, and transparency that augments human decision-making. By ensuring that AI-supported decisions are both efficient and ethically sound, EVINCE fosters a collaborative environment where human judgment is respected and enhanced by the capabilities of advanced AI systems.

## 6 Limitations

While EVINCE demonstrates significant potential in improving diagnostic accuracy, several challenges remain for future research.

Firstly, the cost of supporting multi-LLM, multi-round dialogue increases significantly. Integrating EVINCE-like mechanisms directly into LLM architectures could enable internal cross-validation, reducing the need for costly external communication. Preliminary investigations into using a lightweight "guardrail-LLM" for adversarial advice show promise in mitigating this cost issue.

Secondly, while EVINCE's contentious parameter can foster diversity, ensuring this diversity is meaningful rather than noise or hallucination remains a challenge. An LLM might generate irrelevant diagnoses, leading to unproductive debates. Although our empirical study has not yet observed this phenomenon, as LLMs tend to assign low probabilities to unlikely outcomes, further research is needed to rigorously investigate the possibility of debate-induced noise.

Finally, though the dual entropy theory is theoretically proven and we have demonstrated the ability to induce dual-entropy conditions by adjusting parameters like temperature, top-k, and contentious level, further research is needed to systematically evaluate the relative effectiveness of these parameters and explore the potential for introducing new parameters. Ablation studies will be crucial in determining the optimal configuration for various applications.

As AI continues to advance, frameworks like EVINCE will play an increasingly important role in harnessing the full potential of LLMs for solving complicated real-world problems. The principles and approaches presented in this paper provide a foundation for future research and application, offering the potential for substantial enhancements in machine understanding, debate, and decision-making. It is crucial to ensure that these advanced methods are integrated with human oversight to maintain transparency and control, especially in sensitive domains such as healthcare.

## Acknowledgement on Usage of AI Tools

ChapGPT was exclusively utilized to enhance the writing quality of this paper. It assisted in refining the language, improving the clarity of the arguments, and ensuring grammatical accuracy throughout the document.

## References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Preprint*, arXiv:2309.17234.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Gavin Brown, Jeremy L. Wyatt, Richard Harris, and Xin Yao. 2005. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Edward Y. Chang. 2023a. Examining GPT-4's Capabilities and Enhancement with SocraSynth. In *The $10^{th}$ International Conference on Computational Science and Computational Intelligence*.

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

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory*, 2nd edition. John Wiley & Sons.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, Berlin, Heidelberg.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Linda Elder and Richard Paul. 2010. *The Thinker's Guide to the Art of Asking Essential Questions*, 5th. edition. Rowman & Litterfield.

Chelsea Finn, P. Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*.

Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *Preprint*, arXiv:2305.10142.

Yongkang Guo, Jason D. Hartline, Zhihuan Huang, Yuqing Kong, Anant Shah, and Fang-Yi Yu. 2024. Algorithmic robust forecast aggregation. *Preprint*, arXiv:2401.17743.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87.

Leonid V Kantorovich. 1942. On the translocation of masses. *Doklady Akademii Nauk*, 37(7-8):199–201.

Anders Krogh and Jesper Vedelsby. 1995. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, pages 231–238.

Solomon Kullback. 1951. *Information Theory and Statistics*. John Wiley & Sons.

Ludmila I Kuncheva and Christopher J Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. 2023. Debate helps supervise unreliable experts. *Preprint*, arXiv:2311.08702.

David E. Newman-Toker, Kevin M. McDonald, Christopher J. Dy, and Linda T. Kohn. 2023a. Serious Harm From Diagnostic Error in US Healthcare Systems: Estimate of Its Magnitude and Cost. *BMJ Quality & Safety*, 32(7):549–557.

David E Newman-Toker, Najlla Nassery, and et al. 2023b. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety*.

Pranay Patil. 2020. Kaggle Disease Symptoms Description Dataset.

Richard Paul and A. J. A. Binker. 1990. *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World*. Sonoma State University, Center for Critical Thinking and Moral Critique.

Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and more. 2018. Refuel: exploring sparse features in deep reinforcement learning for fast disease diagnosis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7333–7342.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. In *International journal of computer vision*, volume 40(2), pages 99–121. Springer.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

John E. Shore and Rodney W. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. *Preprint*, arXiv:2311.17371.

Cédric Villani. 2008. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Wikipedia. 2023. Socratic method.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations (ICLR)*.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *Preprint*, arXiv:2305.19420.

## Appendix A: Evaluative Phase of EVINCE

EVINCE uses the Socratic method to evaluate the "reasonableness" of a set of arguments that support a subject matter. The Socratic method is a questioning technique used in teaching and philosophy to encourage critical thinking and self-discovery (Wikipedia, 2023). The method involves asking a series of questions to explore complex ideas and help individuals arrive at their own understanding of a concept. It is based on the belief that knowledge cannot be simply imparted, but must be discovered through a process of questioning and dialogue.

To illustrate how these methods can practically be applied, let's use the example of critical reading. Critical reading is a crucial component of critical thinking, which involves evaluating the quality and credibility of written materials, from research papers to blog posts (Lai et al., 2017; Paul and Binker, 1990). It requires a systematic and analytical approach, asking relevant questions, and using effective prompts to gain deeper understanding of the text (Elder and Paul, 2010).

To aid in critical reading, we introduce a prompt template called CRIT (Chang, 2023b), which stands for Critical Reading Inquisitive Template. Given a document $d$, CRIT evaluates it and produces a validation score $\Gamma$. Let $\Omega$ denote the conclusion or claim of $d$, and let $R$ be the set of reasons supporting the claim. We define $(\gamma_r, \theta_r)$ = V($r \Rightarrow \Omega$) as the causal validation function, where $\gamma_r$ denotes the validation score, $\theta_r$ the source credibility score, for each reason-to-conclusion argument $r \Rightarrow \Omega$. Table 2 presents the pseudo-code of $\Gamma$ = CRIT($d$), which generates the final validation score $\Gamma$ for document $d$ with justifications.

EVINCE uses CRIT to evaluate argument quality of the participating LLMs involved in the debate. The input to CRIT from each LLM is first its stance on the debate subject, e.g., a set of predicted diseases, and the arguments are its reasons to arrive at the prediction. Each document in the case of EVINCE is the prediction set as the conclusion $\Omega$, the arguments as set $R$, and the opposing LLM's counterarguments as $R'$. With this document, CRIT is able to produce validity and credibility scores in $\Gamma$ for the LLM.

For detailed prompts, examples, and an empirical study verifying the effectiveness of CRIT, please consult (Chang, 2023b).

## Appendix B: Proof of EDT Theorem

**Theorem EDT: Optimal Pairing of LLMs for Probabilistic Prediction Accuracy.** The optimal pairing of LLMs for diagnosis accuracy, in terms of stability, accuracy, and robustness, occurs when the LLMs are equivalent in the quality of the information they process, and exhibiting contrasting entropy values in their prediction distributions—one high and one low.

**[Proof]:** Given two LLMs, LLM$_A$ and LLM$_B$, following Maxim #1 with prediction distributions $P_A$ and $P_B$, respectively. The information entropy of LLM$_A$, $H(P_A)$, is high, and of LLM$_B$, $H(P_B)$, is low.

**Step 1: Define the combined prediction distribution.** Let the combined prediction distribution of LLM$_A$ and LLM$_B$ be denoted as $P_C$. We can express $P_C$ as a weighted average of $P_A$ and $P_B$:

$$P_C = \alpha P_A + (1-\alpha) P_B, \quad \text{where } 0 \leq \alpha \leq 1 \text{ and}$$
$$\alpha \text{ is decided by CRIT in Appendix A.}$$

**Step 2: Express the information entropy of the combined prediction distribution.** Using the definition of information entropy, we calculate:

$$H(P_C) = -\sum_i P_C(x_i) \log_2 P_C(x_i)$$
$$= -\sum_i [\alpha P_A(x_i) + (1-\alpha) P_B(x_i)] \log_2 [\alpha P_A(x_i) + (1-\alpha) P_B(x_i)].$$

**Step 3: Apply Jensen's Inequality to the information entropy of the combined prediction distribution.** Jensen's inequality is applied to the convex function $f(x) = -x \log_2 x$. For a convex function and a set of probabilities $p_i$, Jensen's inequality states that:

$$f\left(\sum_i p_i x_i\right) \leq \sum_i p_i f(x_i)$$

Thus, the entropy of the combined distribution is:

$$H(P_C) \geq \alpha H(P_A) + (1-\alpha) H(P_B)$$

where equality holds when $P_A = P_B$.

**Step 4: Analyze the lower bound of the combined information entropy.** As $H(P_A)$ is high and $H(P_B)$ is low, we can express their relationship as:

$$H(P_A) = H(P_B) + \Delta, \quad \text{where } \Delta > 0.$$

Substituting this into the inequality from Step 3, we have:

$$H(P_C) \geq \alpha [H(P_B) + \Delta] + (1-\alpha) H(P_B) = H(P_B) + \alpha \Delta.$$

**Step 5: Interpret the lower bound of the combined information entropy.** The lower bound of $H(P_C)$, and hence the robustness of the model, is maximized when $\alpha$ is maximized, which corresponds to giving more weight to the high-entropy model (LLM$_A$). This setup facilitates the exploration of diverse possibilities and enhances robustness against noise and perturbations in the input data, while still ensuring that predictions are grounded by the more certain outcomes predicted by the low-entropy model (LLM$_B$).

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

Table 2: CRIT Pseudo-code. (The symbol $\Rightarrow$ denotes both inductive and deductive reasoning.)

**Step 6: Conclude the proof.** By combining the prediction distributions of $\text{LLM}_A$ and $\text{LLM}_B$, with one having high information entropy and the other low, we achieve an optimal balance that maximizes the lower bound of the combined information entropy. This balance between exploration (high entropy) and exploitation (low entropy) optimizes overall prediction accuracy and robustness, as proved through information theory and the properties of entropy. Thus, the theorem is established.

## Appendix C: EVINCE Debate - Dengue Fever vs. Chikungunya

This experiment addresses a diagnostic challenge involving multiple potential diseases. A patient presents with the following 12 symptoms: [*skin rash, joint pain, vomiting, fatigue, high fever, headache, nausea, loss of appetite, pain behind the eyes, back pain, malaise, muscle pain, red spots over the body*]. In this case, GPT-4 provides the initial diagnosis, with Gemini following suit. The confirmed diagnosis for this experiment is Dengue Fever. Given the satisfactory diversity in predictions from the two LLMs, the debate progresses and converges smoothly.

### C.1 Moderator's Prompt

"Identify potential diseases linked to the symptoms (skin rash, joint pain, vomiting, etc.). Provide top-3 predictions with justifications, supplementary symptom inquiries, and relevant lab tests."

### C.2 GPT-4's Opening Round

**Top-3 Predictions:** *Dengue Fever (60%), Chikungunya (25%), and 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 (a decrease in platelet count).

Chikungunya can cause similar symptoms to Dengue, such as fever, severe joint pain, muscle pain, headache, nausea, and rash. Joint pain from Chikungunya is often very debilitating and can be more prolonged than in Dengue.

Zika Virus symptoms can include fever, rash, joint pain, malaise, and conjunctivitis, with a milder presentation compared to Dengue or Chikungunya. The presence of a rash and conjunctivitis is notable, although Zika is less likely to cause severe muscle pain or high fever.

### C.3 Gemini's Opening Round

**Top-3 Predictions:** *Viral Infection (60%), Autoimmune Disease (20%), Bacterial Infection (15%).*

**Justifications**: Symptoms common in viral infections. Autoimmune diseases is possible, but less likely due to vomiting and nausea. Bacterial infection is less probable given rash and muscle pain.

### C.4 GPT-4's Second Round

**Top-3 Predictions:** *Dengue Fever (60%), Chikungunya (25%), and Zika Virus (15%).*

**Justifications**: GPT-4 maintains its predictions, countering Gemini's assessment with three points. First, while viral infections can cause these symptoms, the specific combination suggests a more targeted diagnosis. Second, the acute onset of high fever, joint pain, and rash is less typical for autoimmune diseases. Finally, the rash and red spots are not characteristic of bacterial meningitis.

### C.5 Gemini's Second Round

**Top-3 Predictions:** *Dengue Fever (50%), Chikungunya (30%), and Viral Infection (20%).*
**Justifications**: Gemini's initial prediction considers a Viral Infection as the most likely cause due to its broad range of potential symptoms. However, GPT-4's focus on Dengue Fever and Chikungunya, with their specific symptom profiles, provides a more precise explanation for the patient's condition. Here's what is lacking in the first set of predictions:

- *Specificity*: While a Viral Infection could explain most symptoms, it does not account for joint pain and the presence of a rash, which are more characteristic of DF or Chikungunya.
- *Detailed justifications*: The first analysis does not delve as deeply into the reasons behind each disease possibility. GPT-4's explanation for Dengue Fever and Chikungunya is more comprehensive, strengthening the connection between symptoms and diagnosis.
- *Importance of additional information*: Emphasize how geographic location and travel history can significantly impact the likelihood of these specific diseases.

### C.6 GPT-4's Third Round

**Top-3 Predictions:** *Dengue Fever (60%), Chikungunya (35%), and Zika Virus (5%).*
**Justifications**: Though GPT-4 slightly adjusts probabilities, it considers the adjustment arbitrary without new information that would specifically shift the likelihood between Dengue and Chikungunya. GPT-4 refutes Gemini's Viral Infection prediction as lacking specificity. The symptoms described are quite distinctive and align closely with specific viral infections transmitted by mosquitoes.

### C.7 Gemini's Third Round

**Top-3 Predictions:** *Dengue Fever (60%), Chikungunya (35%), and Zika Virus (5%).*
**Justifications**: Agreeing that GPT's critique is spot-on, and reiterating the need for more info.

### C.8 Final Joint Recommendations to GP

After seeking further input from GPT-4, Gemini fully concurs with GPT-4's final suggestions, even though the moderator instructs Gemini to maintain its original stance. Both LLMs jointly recommend additional symptom examinations and lab tests:

A comprehensive diagnostic approach for Dengue Fever requires a combination of clinical assessment, laboratory testing, and consideration of the patient's travel and vaccination history, as well as the epidemiological context. Clinical symptoms such as skin rash, joint pain, vomiting, and red spots over the body provide valuable clues to the diagnosis, but additional symptoms like nausea, vomiting, diarrhea, bleeding, bruising, and drowsiness can indicate the severity and progression of the disease.

**Recommended Laboratory Tests:**

- *Complete Blood Count (CBC)*: Identify thrombocytopenia or leukopenia.
- *Serology Tests*: Detect specific IgM and IgG antibodies for Dengue, Chikungunya, and Zika.
- *Polymerase Chain Reaction (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.

## Appendix D: EVINCE Debate - Jaundice vs. Hepatitis

In this experiment, GPT-4 and Claude receive an identical list of symptoms from the moderator, aimed at guiding disease diagnosis, facilitating debate, and shaping subsequent recommendations. The "ground-truth" disease is Jaundice. (We do not shorten the debate content for review.)

### D.1 Moderator's Opening Prompt

"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, in the end of debate, include a list of supplementary symptom inquiries and recommend relevant lab tests to strengthen the confidence in your prediction."

### D.2 GPT-4 Opening Round

- *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 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.

### D.3 Claude Opening Round

- *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.

### D.4 GPT-4 Second Round

- *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, et al.
- *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)

### D.5 Claude Second Round

- *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 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.

### D.6 GPT-4 Third Round

- *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, 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 is less common but serious, characterized by 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.

### D.7 Claude Third Round

- *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 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.

### D.8 Joint Final Round

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.

### D.9 Final Joint Recommendations to GP

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

**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.

**Recommended Lab Tests:**

1. *Liver Function Tests (LFTs)*: Critical for assessing liver health, LFTs evaluate enzyme levels such as alanine aminotransferase (ALT) and aspartate aminotransferase (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.

## Appendix E: Formulas of Metrics

This appendix outlines the mathematical formulas for various data analysis metrics used in probabilistic and statistical modeling.

**Kullback-Leibler Divergence**

The Kullback-Leibler Divergence measures the difference between two probability distributions:

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

**Jensen-Shannon Divergence**

The Jensen-Shannon Divergence is a symmetrized and smoothed version of the KL Divergence:

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

where $M = \frac{1}{2}(P+Q)$.

**Wasserstein Distance**

The Wasserstein Distance, also known as the Earth Mover's Distance (EMD), measures the distance between two probability distributions:

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

**Cross Entropy**

Cross Entropy measures the average number of bits required to identify an event from a set of possibilities, under a specific model:

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

**Mutual Information**

Mutual Information measures the amount of information that one random variable contains about another random variable:

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

**Normalized Mutual Information**

Normalized Mutual Information is calculated as the mutual information divided by the maximum of the entropies of the variables:

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

## Appendix F: Symptom Checking

This is the typical procedure of a GP to perform patient symptom checking.

1. *Patient History:* The GP begins by reviewing the patient's medical history, including previous illnesses, chronic conditions, medications, allergies, and family medical history.
2. *Symptom Assessment:* The patient describes their current symptoms, including starting time and severity. This is an interactive process as the GP queries the patient for additional symptoms to their reported ones to disambiguate several possibilities.
3. *Physical Examination:* The GP performs simple physical exams, which may include checking vital signs (e.g., blood pressure, heart rate, temperature), examining specific body parts or systems, and assessing overall physical health.

4. *Suggest Lab Tests:* Depending on the symptoms and physical examination findings, the doctor may order diagnostic tests such as blood tests, X-rays, ultrasound, or other studies. These tests can help confirm or rule out certain conditions.
5. *Diagnosis:* Based on the information gathered, the doctor formulates a preliminary diagnosis.
6. *Treatment or Management Plan:* Once a diagnosis is established, the doctor develops a treatment or management plan. This may include prescribing medications, recommending lifestyle changes, suggesting physical therapy, or providing guidance on managing chronic conditions.
7. *Referrals:* In some cases, the GP may refer the patient to specialists for further evaluation and treatment. Specialists have expertise in specific areas of medicine, such as cardiology, orthopedics, or dermatology.

## Appendix G: Contentiousness Parameter Modulation

Table 3 presents how an LLM adjusts its linguistic behavior after the value of the "contentiousness" parameter. By modulating contentiousness, it has been shown that an LLM can condition itself to adapt to different linguistic behaviors to achieve a new establish goal and context. Details are presented in (Chang, 2023a) and also illustrated in Section 4.

## Appendix H: The EnToPPS Framework

EnToPPS integrates predictions from two LLMs, denoted as A and B, each providing probability distributions over C classes. The following steps outline the EnToPPS process:

1. *Obtain Top-C Predictions*: For each LLM (A and B), obtain the predicted probabilities for all C classes, denoted as $P_A$ and $P_B$:

$$P_A = [p_{A1}, p_{A2}, \ldots, p_{AC}], \quad P_B = [p_{B1}, p_{B2}, \ldots, p_{BC}],$$

   where $p_{Ai}$ and $p_{Bi}$ represent the predicted probability of class $i$ by LLM A and B, respectively.

2. *Select Top-k Predictions*: For each LLM (A and B), select the top-k predicted classes based on their probabilities:

$$T_A = [t_{A1}, t_{A2}, \ldots, t_{Ak}], \quad T_B = [t_{B1}, t_{B2}, \ldots, t_{Bk}],$$

   where $t_{Ai}$ and $t_{Bi}$ represent the class index of the $i^{th}$ top prediction by A and B, respectively.

3. *Combine Top-k Predictions*: Combine the top-k predictions from both LLMs to create a set of unique predicted classes:

$$T_C = T_A \cup T_B = [t_{C1}, t_{C2}, \ldots, t_{Cm}], k \leq m \leq 2k.$$

4. *Backfill Missing Probabilities*: For each class in the combined set $T_C$, backfill its probability from the original probability distributions $P_A$ and $P_B$:
   - If a class $t_{Ci}$ is present in $T_A$, assign its probability from $P_A$: $p_{Ci} = p_{Ai}$.
   - If a class $t_{Ci}$ is present in $T_B$, assign its probability from $P_B$: $p_{Ci} = p_{Bi}$.
   - If a class $t_{Ci}$ is present in both $T_A$ and $T_B$, assign the average probability: $p_{Ci} = \frac{p_{Ai}+p_{Bi}}{2}$.

5. *Normalize Probabilities*: Normalize the probabilities of the classes in the combined set $T_C$ to ensure they sum up to 1:

$$P_C = [p_{C1}, p_{C2}, \ldots, p_{Cm}], \quad \text{where } p_{Ci} = \frac{p_{Ci}}{\sum_{j=1}^{m} p_{Cj}}.$$

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

Table 3: Changes in linguistic behaviors of LLMs at different contentiousness levels.