Trinity College
HARTFORD • CONNECTICUT

Uncertainty Quantification for Multi-Agent Debate via Logical Credal Networks

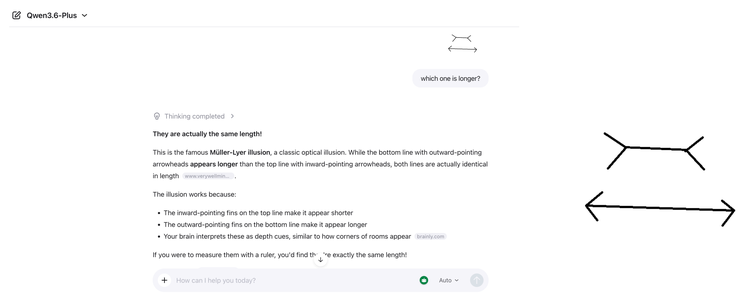
Hyun Lee*, **Christopher Yi***, Biruktayit Habtegiorgis, Shivang Mathur, Tanvir Ahmed Joy
Trinity College ▪ Advisors: Dr. Maminur Islam, Dr. Deepak Venugopal ▪ * Equal contribution

Working Paper
Submission to **Empirical Methods in Natural Language Processing (EMNLP 2026)**

Multi-agent debate
Interval uncertainty
Credal inference

1. Motivation

State-of-the-art LLMs can fail by relying too heavily on learned concepts rather than the actual evidence in front of them. In the example above, the lower line is in fact longer, yet the model is biased by the familiar Müller-Lyer pattern and fails to identify the correct answer. This illustrates a broader challenge: strong prior patterns can override direct evidence and lead to overconfident but incorrect judgments. Our goal is to mitigate such errors through a multi-agent debate framework, where agents challenge one another's interpretations and revise misleading initial judgments. We then use a logical credal network to quantify and bound the remaining uncertainty in the final answer.



A single model (Qwen 3.6 Plus) may be misled by learned visual patterns, while multi-agent debate combined with credal uncertainty bounds helps produce a more reliable final decision.

2. Cross-Calibrated Scoring

Each agent π_i scores every candidate answer Y_j generated by any agent. The **cross-calibrated score** aggregates these evaluations via teacher-forcing log-likelihood:

$$\text{Score}(Y_j) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\log P_{\pi_i}(y_t^{(j)} \mid p, y_{<t}^{(j)})$$

where p is the debate history (prompt + all prior rounds), N is the number of agents, and $T = |Y_j|$ is the answer length in tokens.

Why Cross-Calibration?

- Length normalization ($1/T$) and model averaging ($1/N$) remove scale effects, ensuring comparable scores across heterogeneous models with different tokenizers and vocabulary sizes.
- The score is equivalent to a geometric mixture: $\arg\max_Y \text{Score}(Y) \equiv \arg\max_Y \prod_i P_{\pi_i}(Y \mid p)^{1/N}$.
- **Computational advantage:** Teacher-forcing evaluation costs $O(N^2)$, compared to $O(N^2L)$ for autoregressive generation, enabling efficient calibration at scale.

From Scores to Interval Bounds

A single cross-calibrated score gives a point estimate. To obtain **interval bounds**, we exploit the variance across agents:

- Let $s_i(Y_j) = \frac{1}{T}\sum_t \log P_{\pi_i}(y_t^{(j)} \mid p, y_{<t}^{(j)})$ be agent i's individual score.
- The interval $[\min_i s_i, \max_i s_i]$ captures inter-model disagreement on the quality of answer Y_j.
- After sigmoid normalization, these intervals become the conditional probability bounds $[\ell, u]$ fed into the LCN (Section 3).

High agreement ($u - \ell \approx 0$) ⇒ tight credal constraint; high disagreement ⇒ wide interval reflecting epistemic uncertainty.

3. Logical Credal Networks (LCN)

Unlike standard PGMs that require exact point probabilities, an LCN operates on a **credal set** $K(\Phi)$—the set of all valid probability distributions P over possible worlds $\Omega = \{\omega\}$ satisfying interval constraints.
Given extracted claims and relational rules $\Phi = \{(\phi_i, \psi_i, \ell_i, u_i)\}_{i=1}^m$:

$$K(\Phi) = \{P : \ell_i \le P(\phi_i \mid \psi_i) \le u_i, \;\; i = 1, \ldots, m\}$$

For n atomic formulas there are $N = 2^n$ possible interpretations with probability vector $\vec{p} = (p_1, \ldots, p_N)$. Bounds on final answer $Y = y$ are obtained by solving two **non-linear programs (NLP)**:

$$\underline{P}(Y{=}y) = \min_{\vec{p}} \vec{A}_{Y=y} \odot \vec{p}, \qquad \overline{P}(Y{=}y) = \max_{\vec{p}} \vec{A}_{Y=y} \odot \vec{p}$$

Subject to:

$$\begin{aligned} \vec{A}_{\phi_i \wedge \psi_i} \odot \vec{p} - \ell_i \cdot \vec{A}_{\psi_i} \odot \vec{p} \ge 0 \quad &\text{(lower bounds)} \\ \vec{A}_{\phi_i \wedge \psi_i} \odot \vec{p} - u_i \cdot \vec{A}_{\psi_i} \odot \vec{p} \le 0 \quad &\text{(upper bounds)} \\ \textstyle\sum_j p_j = 1, \;\; p_j \ge 0 & \end{aligned}$$

Plus **quadratic independence constraints** from the Markov condition (Definition 5 in Marinescu et al., 2022):

$$(\vec{A}_\alpha \odot \vec{p})(\vec{A}_\beta \odot \vec{p}) - (\vec{A}_\gamma \odot \vec{p})(\vec{A}_\delta \odot \vec{p}) = 0 \quad (\forall x_i)$$

where $\alpha, \beta, \gamma, \delta$ encode the independence $x_i \perp\!\!\!\perp \text{ndnp}(x_i) \mid \text{pa}(x_i)$.

The quadratic constraints make exact inference NP-hard in general (non-convex NLP). The Generalized Markov Condition prevents vacuous bounds $[0, 1]$ and is identical to the Markov condition of Bayesian networks when the LCN represents a BN (all bounds collapse to point values).

4. Decision Rule Framework

From LCN inference we extract the midpoint $m(y)$ (expected belief) and width $w(y)$ (epistemic uncertainty):

$$m(y) = \tfrac{1}{2}\big(\underline{P}(Y{=}y) + \overline{P}(Y{=}y)\big), \qquad w(y) = \overline{P}(Y{=}y) - \underline{P}(Y{=}y)$$

$$y^* = \arg\max_y \big\{m(y) - \beta \cdot w(y)\big\}$$

Interpretation: Wide intervals indicate genuine unresolved disagreement (defer/review); narrow intervals show stable consensus (accept automatically). The hyperparameter β controls the trade-off between expected accuracy and risk aversion toward uncertain answers.

Role of β in Practice

- $\beta = 0$: Pure midpoint selection->equivalent to choosing the answer with the highest average belief, ignoring uncertainty entirely.
- $\beta > 0$: Risk-averse selection—>penalizes wide intervals, preferring answers with high confidence even if the midpoint is slightly lower.
- $\beta \to \infty$: Maximally conservative—>selects the answer with the narrowest interval (highest consensus), regardless of its midpoint value.

In practice, β is tuned on a held-out validation set. Across GSM8K, ARC, and MMLU, we find $\beta \in [0.3, 0.7]$ consistently outperforms $\beta = 0$, confirming that penalizing epistemic uncertainty improves answer selection.

5. Claim Extraction & Rule Translation

Toy problem: Sarah has 3 bags with 4 apples each. She buys 2 more apples. How many apples?

Agent 1 → Answer 14
C_1: $3 \times 4 = 12$
C_2: $12 + 2 = 14$

Agent 2 → Answer 9
C_3: $3 + 4 + 2 = 9$

Free-form reasoning becomes reusable proposition nodes $\{C_1, C_2, C_3, A_1, A_2\}$. Relations are tagged as support, contradict, or refine.

From Scores to LCN Constraints

Cross-calibrated scores are mapped to interval-valued conditional probability constraints. For a claim C_i supporting answer A_k, if the normalized score falls in $[s_\ell, s_u]$, we set:

$$\ell_{A_k|C_i} \le P(A_k \mid C_i) \le u_{A_k|C_i}$$

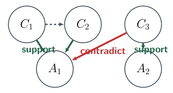
The width $u - \ell$ reflects inter-model disagreement on the reasoning step C_i.

6. Credal Belief Propagation

Because exact LCN inference involves non-convex NLPs that scale as $\mathcal{O}(\exp(N))$, we adopt a **Credal Loopy Belief Propagation (LBP)** scheme for approximate inference—an extension proposed as future work in Marinescu et al. (2022). Messages from node i to neighbor j:

$$\underline{\mu}_{i\to j}(x_j) = \min_{P \in K(\Phi_i)} \sum_{x_i} P(x_i \mid U_i) \prod_{k \in N(i)\setminus j} \mu_{k\to i}(x_i)$$

$$\overline{\mu}_{i\to j}(x_j) = \max_{P \in K(\Phi_i)} \sum_{x_i} P(x_i \mid U_i) \prod_{k \in N(i)\setminus j} \mu_{k\to i}(x_i)$$



Support from C_1 and C_2 tightens bounds around A_1; the conflicting C_3 branch widens uncertainty and supports A_2. Generalized Noisy-Or aggregates converging messages efficiently.

Noisy-Or Aggregation (Multi-Parent)
For answer A_k with parents $\{C_1, \ldots, C_r\}$:

$$P(A_k{=}1 \mid C_1, \ldots, C_r) = 1 - \prod_{i=1}^{r}\big(1 - P(A_k{=}1 \mid C_i)\big)$$

In the credal setting $\underline{P}, \overline{P}$ are obtained from extremal parent parameters, preserving causal independence semantics in polynomial time.

Credal Sets in Our Debate Graph
The credal set $K(\Phi)$ for the toy problem imposes these constraints simultaneously:

$0.80 \le P(A_1|C_1) \le 0.95$ C_1 supports A_1
$0.85 \le P(A_1|C_2) \le 0.98$ C_2 supports A_1
$0.75 \le P(A_2|C_3) \le 0.90$ C_3 supports A_2
$0.60 \le P(\neg A_1|C_3) \le 0.85$ C_3 contradicts A_1
$0.05 \le P(C_2|C_1) \le 0.20$ dependency

Inference yields: $\underline{P}(A_1) \in [0.20, 0.67]$, $\underline{P}(A_2) \in [0.18, 0.55]$. Wide interval on A_1 ⇒ **flagged for review**.

7. Theoretical Guarantees

Theorem 1 (Credal Bound Consistency)

Let G' be obtained from G by pruning only edges with $J(e) \le \epsilon$. Then

$$W_{G'}(y^*) \le W_G(y^*) + \mathcal{O}\left(\frac{\epsilon}{\lambda_2}\right).$$

Proof. Each pruned edge removes relational or independence constraints, so the feasible credal region can only expand: $\mathcal{P}_G \subseteq \mathcal{P}_{G'}$. Therefore lower probabilities cannot increase and upper probabilities cannot decrease, implying that pruning can only widen intervals. Since $J(e) = D_{\text{JS}} + \lambda_1 \max(0, \Delta S_e) + \lambda_2 \max(0, \Delta W_e)$ has nonnegative terms, $J(e) \le \epsilon$ implies $\Delta W_e \le \epsilon/\lambda_2$. Summing over accepted pruning steps yields the stated bound. ■

Theorem 2 (Convergence of Credal LBP)

Assume each local credal set $K(\Phi_i)$ is nonempty and compact. Define

$$(\mathcal{T}^{\pm}\mu)_{i\to j}(x_j) = \operatorname*{ext}_{P \in K(\Phi_i)} \sum_{x_i} P(x_i \mid U_i) \prod_{k \in N(i)\setminus j} \mu_{k\to i}(x_i),$$

where ext denotes min for $\mathcal{T}^-$ and max for $\mathcal{T}^+$. With $\underline{\mu}^{(0)} = \mathbf{0}$ and $\overline{\mu}^{(0)} = \mathbf{1}$,

$$\underline{\mu}^{(t+1)} = \mathcal{T}^-\underline{\mu}^{(t)}, \qquad \overline{\mu}^{(t+1)} = \mathcal{T}^+\overline{\mu}^{(t)}.$$

Proof. If $\mu \le \nu$ componentwise, then

$$\prod_k \mu_{k\to i}(x_i) \le \prod_k \nu_{k\to i}(x_i),$$

and since $P(x_i \mid U_i) \ge 0$, both $\mathcal{T}^-$ and $\mathcal{T}^+$ are isotone. Therefore

$$\mathbf{0} \le \underline{\mu}^{(0)} \le \underline{\mu}^{(1)} \le \cdots \le \mathbf{1}, \qquad \mathbf{1} \ge \overline{\mu}^{(0)} \ge \overline{\mu}^{(1)} \ge \cdots \ge \mathbf{0}.$$

Both sequences are monotone and bounded, hence converge to fixed points $\underline{\mu}^*$ and $\overline{\mu}^*$. Since LBP enforces local credal constraints while relaxing global consistency, the limit interval is conservative:

$$\underline{\mu}^* \le P_{\text{LCN}}(\cdot) \le \overline{\mu}^*.$$

■

8. Edge Pruning

Large debate graphs contain redundant or noisy edges. We prune edge e by minimizing the cost $J(e)$:

$$J(e) = D_{\text{JS}}\big(q_G(Y) \| q_{G_{-e}}(Y)\big) + \lambda_1 \max(0, \Delta S_e) + \lambda_2 \max(0, \Delta W_e)$$

Noisy-Or → JS Divergence Pipeline
The answer distribution $q_G(Y)$ is computed via the **Noisy-Or aggregation** (Section 6):

$$q_G(Y{=}y_k) = 1 - \prod_{i=1}^{r}\big(1 - P(A_k{=}1 \mid C_i)\big)$$

where the conditional probabilities come from converged Credal LBP messages.
For each candidate edge e:

1. Remove e from G to obtain G_{-e}.
2. Re-run Credal LBP on G_{-e}; recompute Noisy-Or marginals → $q_{G_{-e}}(Y)$.
3. Evaluate $D_{\text{JS}}(q_G \| q_{G_{-e}})$ using the log-likelihoods from both distributions:

$$D_{\text{JS}} = \tfrac{1}{2}\sum_y q_G(y)\log\frac{q_G(y)}{M(y)} + \tfrac{1}{2}\sum_y q_{G_{-e}}(y)\log\frac{q_{G_{-e}}(y)}{M(y)}, \quad M = \tfrac{1}{2}(q_G + q_{G_{-e}})$$

Edges with small D_{JS} (minimal distributional shift) and bounded width increase ($\Delta W_e \le \delta$) are safe to remove.

Algorithm 1: Budgeted Greedy Pruning

```
Input: Graph G^(0)=(V,E), thresholds ε,δ, cost threshold τ
Output: Pruned graph G^(m)
m ← 0
while |E^(m)| > 0 do
  Compute Noisy-Or marginals q_{G^(m)}(Y) via Credal LBP
  for each e ∈ E^(m): recompute q_{G^(m)_{-e}}(Y); evaluate J_m(e)
  e* ← arg min_e J_m(e)
  if J_m(e*) ≤ τ and D_JS(q_{G^(0)}, q_{G^(m)_{-e*}}) ≤ ε and ΔS_e* ≤ δ
    E^(m+1) ← E^(m) \ {e*};  m ← m + 1
  else break
return G^(m)
```

9. Experimental Results

Models: Qwen2.5-7B, Ministral-8B, Llama-3.1-8B (same total LLM calls for all methods).

GSM8K and ARC-Challenge Accuracy (%)

Method	GSM BoN	GSM Debate	ARC BoN
Random	76.18	81.00	83.90
Self-Certainty	–	–	88.91
Entropy	–	84.34	89.00
Gini Impurity	–	84.57	89.00
Agrawal et al.	77.16	84.88	89.00
LCN (Ours)	**79.50**	**87.31**	**91.20**
Upper Bound	89.51	90.00	94.62

MMLU Accuracy by Subset (%)

Method	FL	HSM	EM	CM	PHI	AA	Avg.
Random	47.2	46.4	79.0	40.8	69.8	43.9	54.5
Agrawal	49.6	50.6	82.2	43.9	70.1	42.9	57.9
LCN (BoN)	**51.8**	**53.1**	**84.5**	**46.2**	**72.5**	**45.3**	**60.5**
Agrawal (Deb.)	53.2	51.9	82.8	37.0	72.3	51.0	59.0
LCN (Deb.)	**55.4**	**54.3**	**85.2**	**39.5**	**74.7**	**53.5**	**61.4**
Upper Bound	72.0	65.9	93.0	64.0	81.0	57.0	72.2

Key Results

- **LCN outperforms all baselines** on every dataset and setting.
- GSM8K: **+2.3pp** (BoN), **+2.4pp** (Debate) over Agrawal et al.
- ARC-Challenge: **+2.2pp** over Agrawal et al.
- MMLU Average: **+2.5pp** (BoN), **+2.4pp** (Debate).
- LCN Debate (**87.31%**) surpasses single-model Best-of-9 (82.55%).

10. Discussion & Limitations

While LCN consistently outperforms point-estimate baselines, we identified two failure modes:

- **Cascading Hallucinations:** If all agents confidently agree on a flawed premise early, the credal set remains artificially narrow, producing false confidence.
- **Relational Extraction Errors:** Misclassifying a `refine` relation as `contradict` can unnecessarily widen the uncertainty, leading to over-flagging.

This highlights the need for robust relation-extraction mechanisms and targeted prompt tuning to ensure the debate graph accurately reflects semantic intent.

11. Conclusion & Future Work

- LCN provides **interval-valued uncertainty** instead of overconfident point scores.
- Cross-calibrated scores act as probabilistic constraints over interrelated claims.
- Lower/upper bounds expose genuine disagreement, improving answer selection.
- **Future:** Scalable approximate inference, single-agent settings, open-ended reasoning tasks, multi-modal extensions.

Wide interval → Flag for human review **Narrow interval → Accept automatically**

References: [1] Marinescu et al., "Logical Credal Networks," NeurIPS 2022. [2] Agrawal et al., "Uncertainty-Aware Answer Selection in Multi-LLM Systems," EMNLP 2025. [3] Srinivas, "A Generalization of the Noisy-Or Model," UAI 1993. [4] Du et al., "Improving Factuality and Reasoning via Multi-Agent Debate," ICML 2024.

Contact: Electrical Engineering Building, 465 Northwestern Ave, West Lafayette, IN 47907 · hlee6350@purdue.edu