

Confidence over Time

Confidence Calibration with Temporal Logic
for Large Language Model Reasoning

Ivan Ruchkin

Malachowsky Family Endowed Rising Star Assistant Professor in ECE

Trustworthy Engineered Autonomy (TEA) Lab
University of Florida



CSCE 585: Machine Learning Systems | University of South Carolina | September 3, 2026



TEA Lab: Trustworthy $\exists$ ngineered $\forall$ tonomy

- **TEA Lab's mission:** make autonomous systems *safer* and *trustworthier* through rigorous computational methods with strong guarantees
 - Across many domains: autonomous driving, mobile robots, UAVs/UUVs
- *Research areas:* **formal methods, AI/ML, cyber-physical systems, robotics**
 - Theory $\rightarrow$ algorithms & tools $\rightarrow$ simulations+physical deployment



Research platform:
small-scale racing cars

tea.ece.ufl.edu



[Subscribe to us
on LinkedIn!](#)



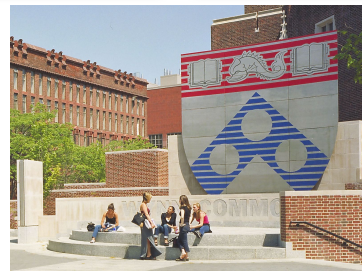
[Subscribe to us
on RedNote!](#)



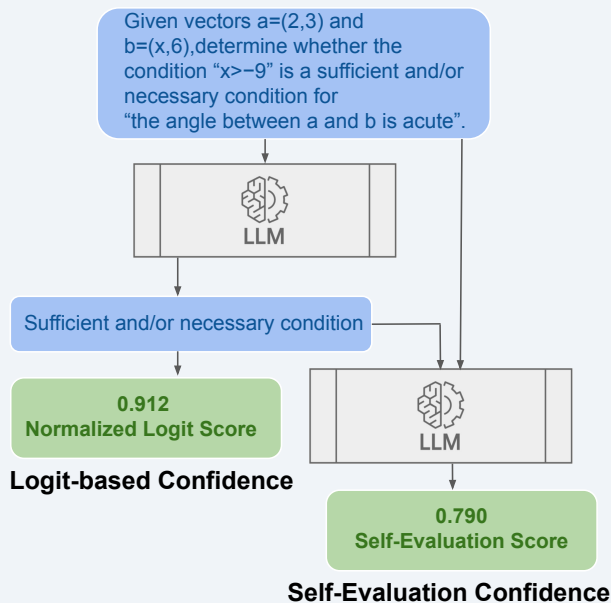
ivan.ece.ufl.edu
iruchkin@ece.ufl.edu

Ivan Ruchkin: brief bio

- 2006–2011: **Undergrad** @ Moscow State University 
 - Moscow, Russia — applied math & computer science
- 2011–2018: **PhD** @ Carnegie Mellon University 
 - Pittsburgh, PA — software engineering & formal methods
- 2018–2022: **Postdoc** @ University of Pennsylvania 
 - Philadelphia, PA — guarantees for learning-enabled systems
- 2022–present: **Assistant professor** @ University of Florida 
 - Gainesville, PA — trustworthy methods for autonomy
- **Industry and government roles**    
 - AFRL VFRP, NASA JPL, CMU SEI, Google Summer of Code, ...



How to Quantify LLM Confidence?



Logit-based

Cumulative token likelihoods, predictive entropy, max probability. Length-normalized variants help but conflate semantic & auxiliary tokens.

User: Could you give me one name of president?

LLM: Sure, here is a historical American president: ** ■

Next token probability distribution:

Barack Hussein Obama	0.377
Abraham Lincoln	0.202
John Fitzgerald Kennedy	0.178
George Washington	0.178
...	...

Entropy: 2.03

actual is certain, estimated is uncertain

Common Sense Question

User: Where will the future evolution of the observable universe ultimately lead?

LLM: The universe will eventually go to ** ■

Next token probability distribution:

heat death	0.828
Heat Death	0.136
singularity	0.003
the heat death	0.000
...	...

Entropy: 0.68

actual is uncertain, estimated is certain

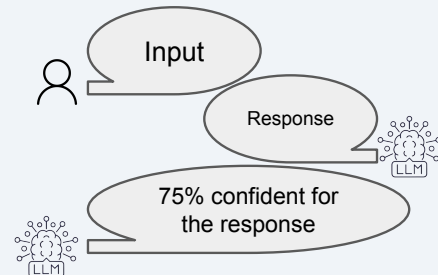
Unsolved Physics Problem

Malinin & Gales (2021) — *Uncertainty Estimation in Autoregressive Structured Prediction* (ICLR 2021)

Kuhn et al. (2023) — *Semantic Entropy: Linguistic Invariances for Uncertainty Estimation in NLG* (Nature / ICLR 2023).

Self-Evaluation

Prompt or fine-tune LLMs to express confidence in natural language. Often requires additional training or domain adaptation.

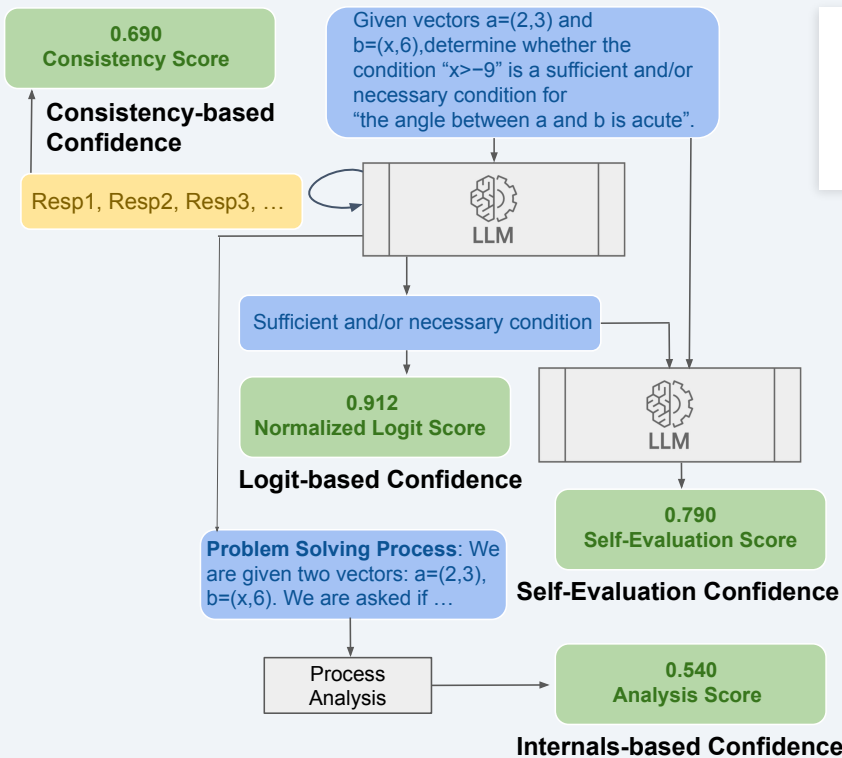


Lin et al. (2022) — *Teaching Models to Express Their Uncertainty in Words* (TMLR).

Tian et al. (2023) — *Just Ask for Calibration: Eliciting Calibrated Confidence from LLMs* (EMNLP 2023).

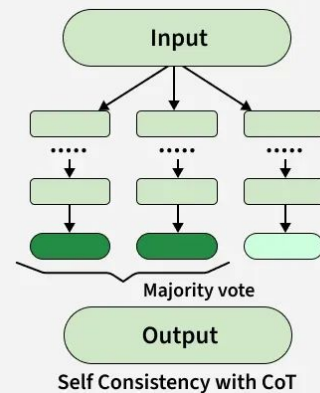
Xiong et al. (2024) — *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs* (ICLR 2024).

How to Quantify LLM Confidence?



Consistency-based

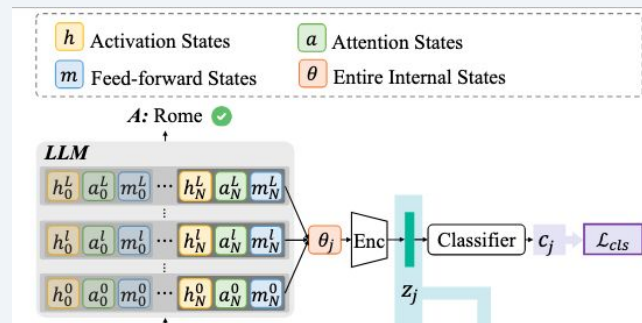
Self-Consistency, Semantic Entropy.
Uncertainty from multiple sampled outputs.
High inference overhead.



Wang et al. (2022) — *Self-Consistency Improves Chain of Thought Reasoning in Language Models* (ICLR 2023).
Farquhar et al. (2024) — *Detecting Hallucinations in LLMs Using Semantic Entropy* (Nature 2024).

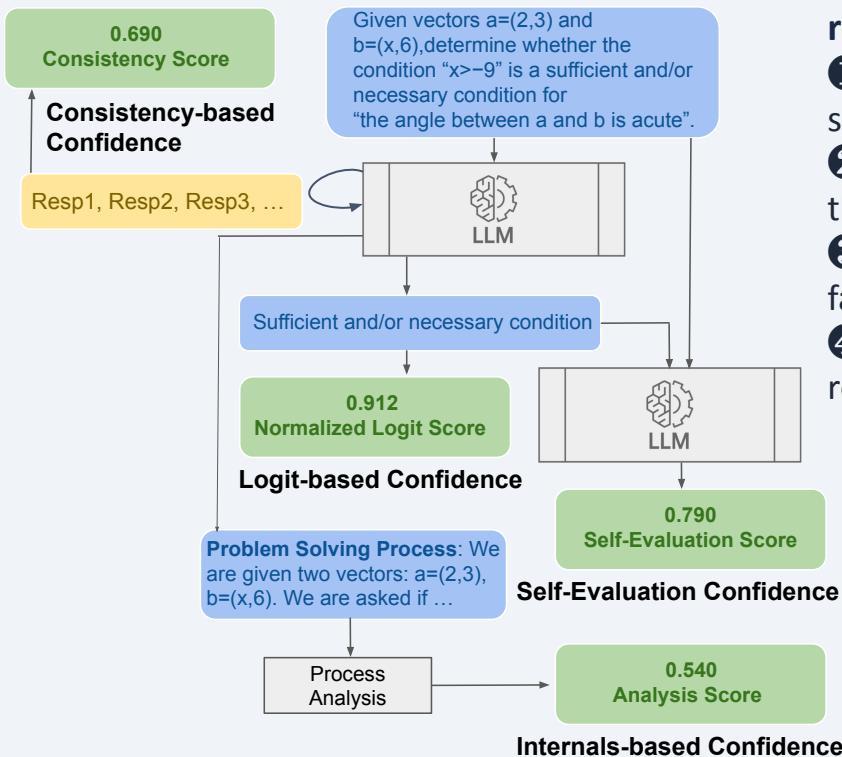
Internals-based

Hidden states / token attribution (e.g., SAR, InternalInspector). Can be expensive or require large validation sets.



Burns et al. (2022) — *Discovering Latent Knowledge in Language Models Without Supervision* (ICLR 2023).
Azaria & Mitchell (2023) — *The Internal State of an LLM Knows When It's Lying* (EMNLP Findings 2023).
Duan et al. (2024) — *SAR: Shifting Attention to Relevance* (ACL 2024).

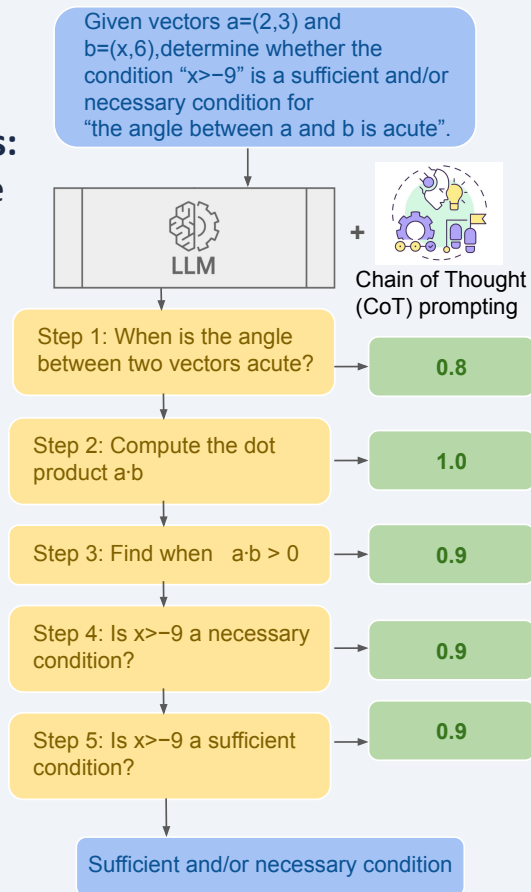
How to Quantify LLM Confidence?



LLMs excel at multi-step reasoning, but existing methods:

- ❶ Collapse reasoning into a single scalar value
- ❷ Ignore how confidence evolves throughout generation
- ❸ Are sensitive to superficial factors: length, verbosity, fillers
- ❹ Cannot distinguish correct reasoning from confident errors

Key Idea: Confidence should be viewed as a *temporal signal* evolving over reasoning steps, not a static scalar.



Problem: Confidence Estimation in LLMs



Divide 2 teachers and 4 students into 2 groups and arrange them to participate ...

Instruction: x



First, we need to divide 2 teachers and 4 students into 2 groups, each ...
Next, **we** select 2 students out of 4 for the first group. This can be done in ...
...

Reminder:

Softmax converts logits to probabilities
 $P(y_i) = e^{\text{logit}(y_i)} / (\sum_j e^{\text{logit}(y_j)})$

Response: y

top_logit: 26.5,
top_prob: 0.94,
top_k: {
 token: "we", logit: 26.5
 token: "from", logit: 22.9
 token: "for", logit: 12.4
 ...}

Confidence Estimation Objective

Estimate a confidence score approximating the posterior correctness:

$$C \approx P(c_T(x, y) = 1 \mid x),$$

where $c_T(x, y) \in \{0, 1\}$ is a binary correctness predicate for task T

Problem with Joint Probability

$$P(x, y_1, y_2, \dots, y_n) = P(x) \cdot \prod_{t=1}^n P(y_t \mid x, y_1, \dots, y_{t-1})$$

Suffers from *severe length bias*: assigns lower values to longer responses regardless of correctness; unsuitable for long-form reasoning.

Autoregressive Generation

Given input x , an LLM generates response $y = [y_1, \dots, y_t]$ autoregressively, by sampling per-step conditional probability: $P(y_t \mid y_{<t}, x)$

Problem: Confidence Estimation in LLMs

Confidence Estimation Objective

Estimate a confidence score approximating the posterior correctness:

$$C \approx P(c_T(x, y) = 1 \mid x),$$

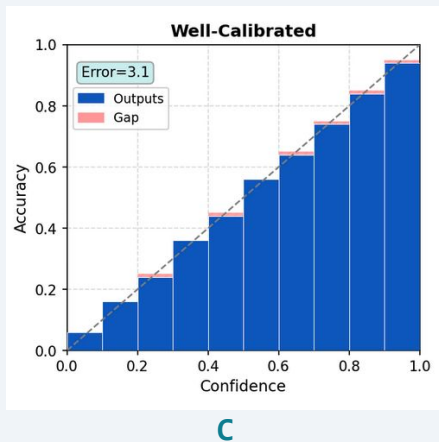
where $c_T(x, y) \in \{0, 1\}$ is a binary correctness predicate for task T

Evaluation Metric

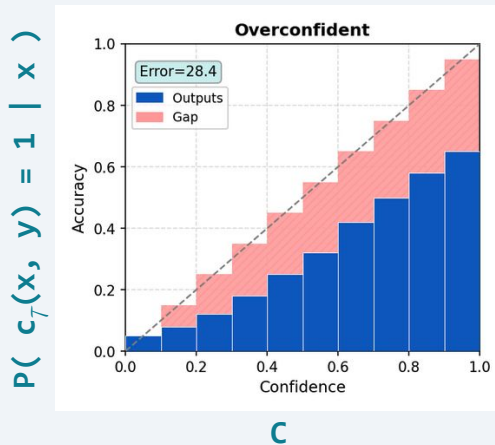
Expected Calibration Error (ECE) - weighted difference between binned accuracy & confidence

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

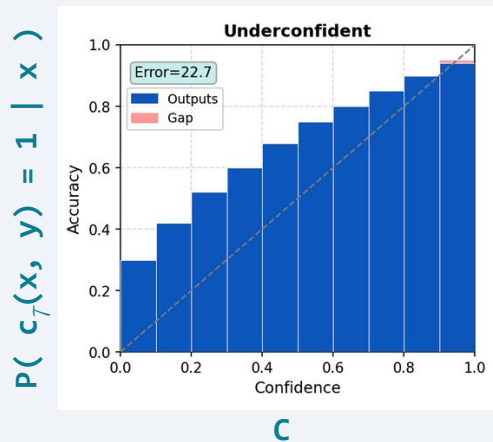
Perfect Calibration



Overconfidence



Underconfidence



Agenda

- Stepwise Confidence & Signal Temporal Logic
 - Research Questions
 - Question-Adaptive Hypernetworks
 - Experiments and Results

Stepwise Confidence Signal

Instruction: x

Divide 2 teachers and 4 students into 2 groups and arrange ...



Response: y

1. First, we need to divide 2 teachers and 4 students...

2. Next, we select 2 students out of 4 for the first group...

3. After, calculate ...

...



Confidence Signal:

$s = [0.88, 1.0, 0.94, \dots]$

Stepwise Confidence Signal

1. Segment response y into n consecutive steps

$$y = [u_1, \dots, u_n]$$

2. Define per-step confidences s_j as:

$$s_j = (1/L_j) \sum_k P(y(t_{j+k}) \mid y(<t_{j+k}), x)$$

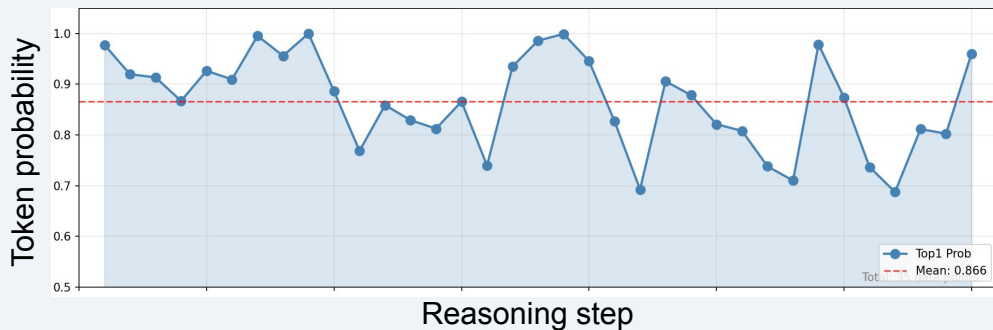
3. Aggregate into the confidence signal S :

$$S = [s_1, s_2, \dots, s_n]$$

Example confidence signals

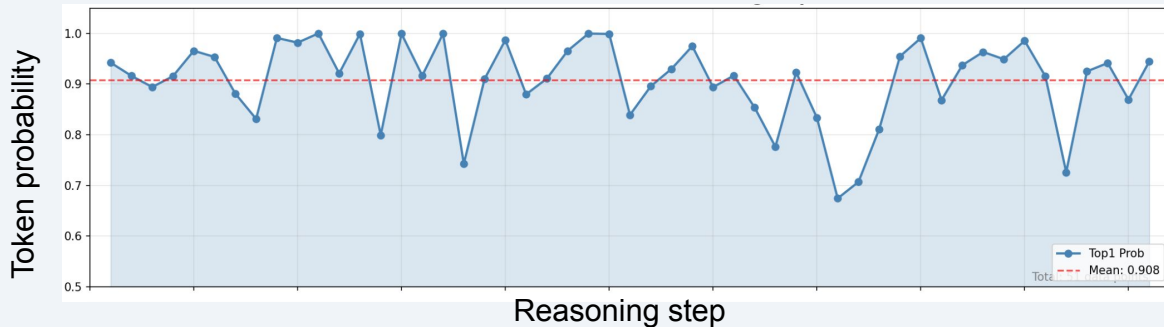
Scenario 1: given 2 vectors, select the angle acuteness criterion

→ *The LLM answer is incorrect*



Scenario 2: compute a vector product of parabola-line intersections

→ *The LLM answer is correct*



Tool: Signal Temporal Logic (STL)

Signal Temporal Logic (STL)

Syntax (fragment):

$\varphi ::= \mu \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \Box[a,b]\varphi \mid \Diamond[a,b]\varphi$

Robustness Semantics:

$\rho(\Box[a,b]\varphi, S, t) = \min_{k \in [t+a, t+b]} \rho(\varphi, S, k)$

$\rho(\Diamond[a,b]\varphi, S, t) = \max_{k \in [t+a, t+b]} \rho(\varphi, S, k)$

Positive → satisfied; **Negative** → violated

$$\rho(\varphi, S, 0) = 0.05$$

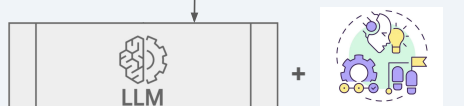
The signal is never above 3.5

$\varphi := \Box(x[t] < 3.5)$



STL-based CoT Confidence

Given vectors $a=(2,3)$ and $b=(x,6)$, determine whether the condition " $x>-9$ " is a sufficient and/or necessary condition for "the angle between a and b is acute".



Step 1: When is the angle between two vectors acute?

0.8

Step 2: Compute the dot product $a \cdot b$

1.0

Step 3: Find when $a \cdot b > 0$

0.9

Step 4: Is $x > -9$ a necessary condition?

0.9

Step 5: Is $x > -9$ a sufficient condition?

0.9

Sufficient and/or necessary condition

RQ1: Does *STL pattern structure* generalize across reasoning tasks?
RQ2: Given a structure, do *STL parameters* generalize across questions?

Stepwise Confidence Signal
 $S = [s_1, s_2, \dots, s_n]$



STL ϕ
 $\square[1, t] (s_t \geq \mu)$

Overall Confidence C

Discriminative Mining of Signal Temporal Logic

$\Phi+$ (Correct Patterns): Temporal signatures of correct reasoning.

$\Phi-$ (Incorrect Patterns): Patterns of hallucination / failure.

01

Base Template Instantiation

Start from *predefined base templates*, e.g. $\Box[a, b](s_t \geq \mu)$

Fit parameters to maximize discriminative power.

02

Robustness-Based Signal Augmentation

Signal lifting: compute robustness signals from base templates.

Build augmented signal S' for nested formulas.

03

Structural Composition

Temporal nesting & Boolean combination ($\wedge$ / $\vee$).

Generate more expressive STL formulas.

Primitive STL Templates

$\Phi+$ (Correct Patterns): Temporal signatures of correct reasoning.

$\Phi-$ (Incorrect Patterns): Patterns of hallucination / failure.

Positive Templates ($\Phi+$) for Correct Reasoning

Template	STL Formula
WeakestLink	$\Box[0, T](s_t \geq \mu)$
EndHigh	$\Box[T-k, T](s_t \geq \mu)$
StartHigh	$\Box[0, k](s_t \geq \mu)$
NeverSharpDrop	$\Box[0, T](\Delta t > -\epsilon)$
ConfidenceGain	$\bar{s}_{\text{end}} > \bar{s}_{\text{start}} + \epsilon$

Negative Templates ($\Phi-$) for Failure Modes

Template	STL Formula
EventuallyLow	$\Diamond[0, T](s_t \leq \mu)$
EndLow	$\Diamond[T-k, T](s_t \leq \mu)$
SharpDrop	$\Diamond[0, T](\Delta t \leq -\epsilon)$
FinalDecline	$\Box[T-k, T](\Delta t \leq 0)$
Recovery	$\Diamond(s_t \leq \mu_1) \wedge \Diamond(s_t \geq \mu_h)$

Reasoning Datasets

➤ **CLadder** **Background:** Let X = education level; Y = salary; $V2$ = proximity to a college; $V1$ = unobserved confounders.

Questions: If we intervene and set education level to high for everyone, does this increase expected salary?

Answer: No

➤ **SciQ** **Support:** Exothermic reactions release energy, often in the form of heat or light. Combustion is one of the most common exothermic reactions.

Question: What type of reaction releases energy in the form of heat or light?

- A. endothermic
- B. catalytic
- **C. exothermic** ✓
- D. isothermal

➤ **BIG-Bench Hard (BBH)** (23 subtasks)

① Boolean Expressions

Q: True and False and not True or False

A: False

② Date Understanding

Q: The concert was scheduled for 06/01/1943, but was delayed by one day to today. What is the date one week from today in MM/DD/YYYY?

A: 06/09/1943

③ Causal Judgment

Story A factory has a policy that workers must wear hard hats. Tom decides not to wear one. A beam falls and hits Tom on the head, causing injury. If Tom had worn a hard hat, would he have been injured?

Q: Did Tom's decision cause the injury?

A: Yes

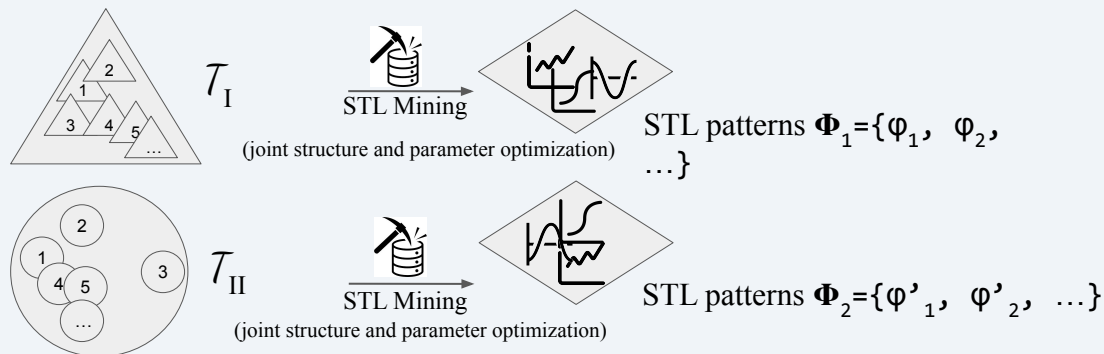
Agenda

- Stepwise Confidence & Signal Temporal Logic
- Research Questions
 - Question-Adaptive Hypernetworks
 - Experiments and Results

RQ1: Structural Generalization Across Tasks

RQ 1: Do STL pattern sets Φ_1 and Φ_2 overlap over tasks?

Key Findings

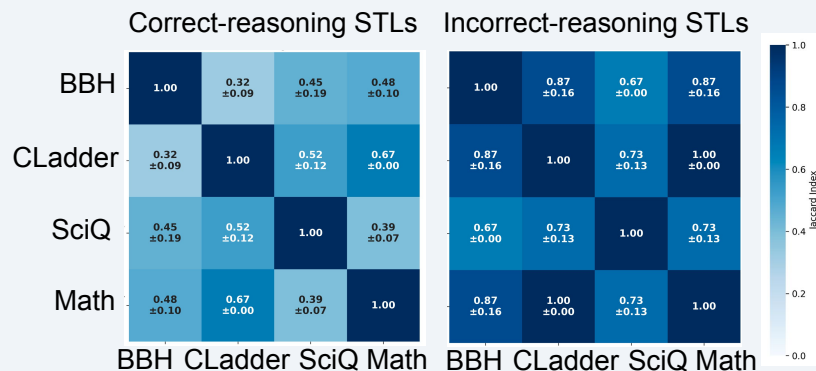


Asymmetric generalization:

- Incorrect-reasoning STL patterns Φ_- show *high structural similarity* across tasks (Jaccard/IoU ≈ 0.74 – 0.81)
- Correct-reasoning patterns Φ_+ are *task-dependent* (Jaccard/IoU ≈ 0.47 – 0.55)

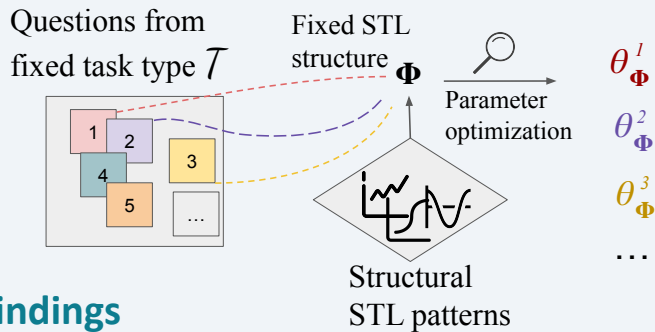
Implication:

Φ_- can serve as transferable structural templates, while Φ_+ should adapt to task-specific characteristics



RQ2: Parameter Sensitivity to Questions

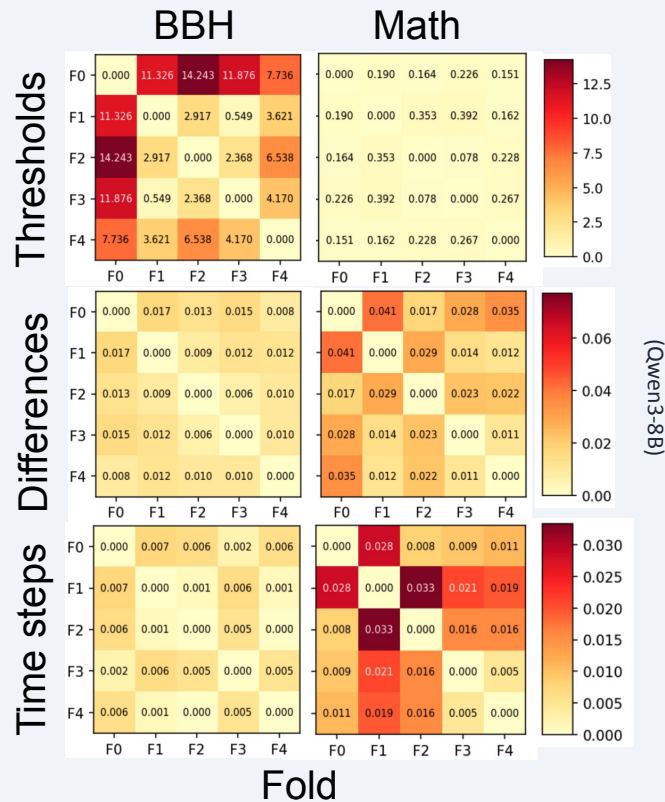
RQ 2: Are $\theta_{\phi 1}^1$ $\theta_{\phi 1}^2$ $\theta_{\phi 1}^3$... similar?



Key Findings

- **High threshold instability:** thresholds can vary drastically across folds (especially in BBH, up to 14.24)
- **Slightly variable difference parameters:** Difference parameters exhibit some cross-fold variation in both datasets
- **Modestly inconsistent time steps:** time step parameters are mostly stable in BBH (max 0.007), but show moderate variation in Math (max 0.033)

Implication: A single global STL parameterization is not enough. Question-adaptive estimation mechanism is needed.



Agenda

- Stepwise Confidence & Signal Temporal Logic
- Research Questions
- Question-Adaptive Hypernetworks
 - Experiments and Results

Question-Adaptive Hypernetworks

Motivation: RQ2 shows different questions need different parameters, but modifying STL structure would undermine interpretability.

Design Principle: Fixed STL structure + variable STL parameters (thresholds, differences, time steps)

Hypernetwork Architecture

Input: prompt x + confidence signal S

Output: STL parameters θ

$$\theta = H_{\psi}(x, S)$$

$x \rightarrow$ fixed text encoder

$S \rightarrow$ temporal features (mean, variance, slope)

Concatenate x and $S \rightarrow$ MLP $\rightarrow \theta$

Training Objective

End-to-end discriminative loss:

$$\min_{\psi} \frac{1}{N} \sum_{i=1}^N L(c_i, \hat{P}_i)$$

Confidence estimates are computed as:

θ_i from hypernetwork $\rightarrow$ fixed STL template $\varphi \rightarrow$
robustness $\rightarrow$ sigmoid

Confidence Estimation: STL Blocks

STL Block Design

Mapping robustness $\rightarrow$ confidence:

$$p = \sigma(\alpha \cdot \rho(\varphi_{\theta}, S) + \beta)$$

σ : sigmoid function

α : learnable scale (decision boundary sharpness)

β : learnable bias (boundary location)

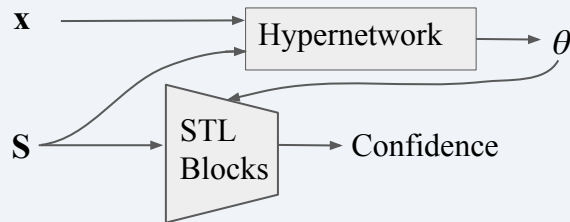
Aggregation of Temporal Patterns

Φ^+ scores are used as-is

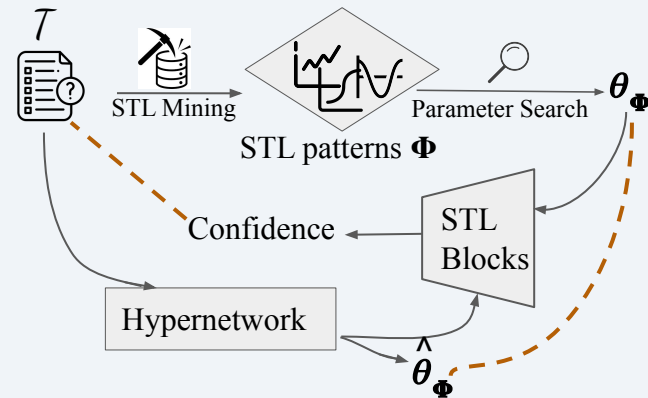
Φ^- scores are inverted via $1 - p$

Weighted aggregation $\rightarrow$ final confidence

Inference



Training



Agenda

- Stepwise Confidence & Signal Temporal Logic
- Research Questions
- Question-Adaptive Hypernetworks
- Experiments and Results

Experimental Setup

Datasets

- GAOKAO-Math (mathematical reasoning)
- CLadder (causal reasoning)
- SciQ (scientific QA)
- BigBench-Hard / BBH (multi-task)

Backbone Models

- Qwen3-8B
- Gemma-3-12B
- Llama-3-8B

Single-sample inference (except Self-Consistency)

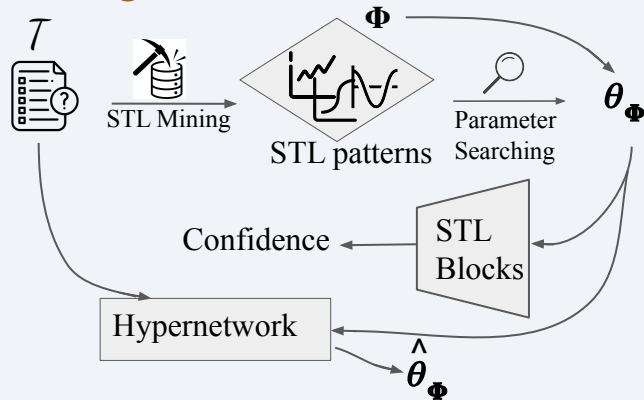
Baselines & Ablation Variants

Baselines: AveLogit, SAR, Self-Eval, Self-Consistency, InternalInspector

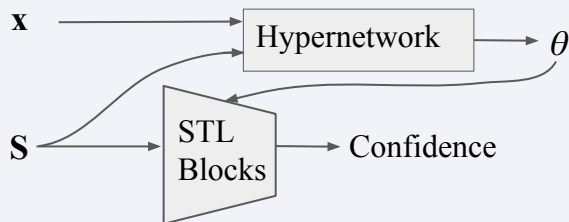
Ablations: A1: Cross-task reuse (all STL) | A2: Negative patterns only | A3: In-domain mining (fixed params) |
Ours: In-domain STL + Hypernetwork

Experimental Results: Ablations

Training



Inference

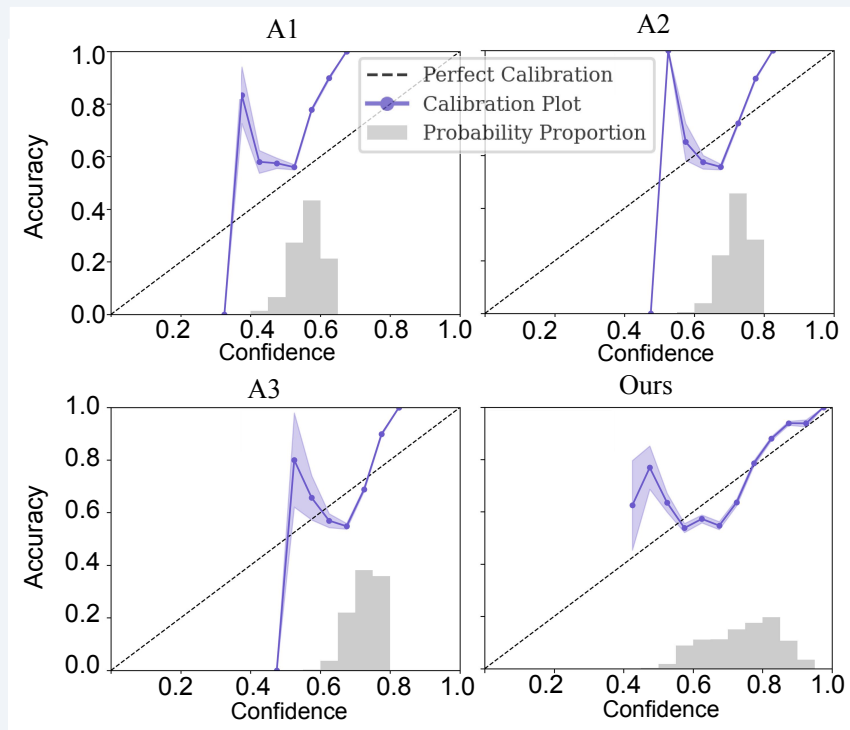


A1: cross-task reuse of both positive and negative STL patterns (mined on BBH)

A2: cross-task reuse of negative (failure-mode) patterns only

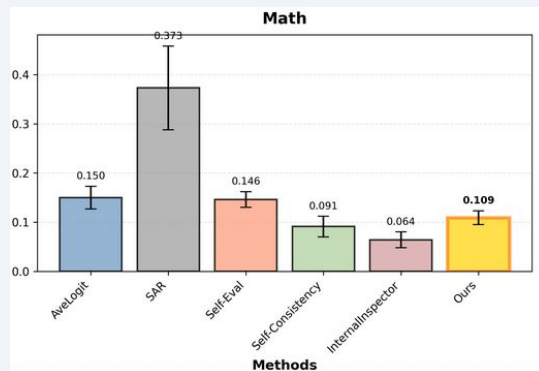
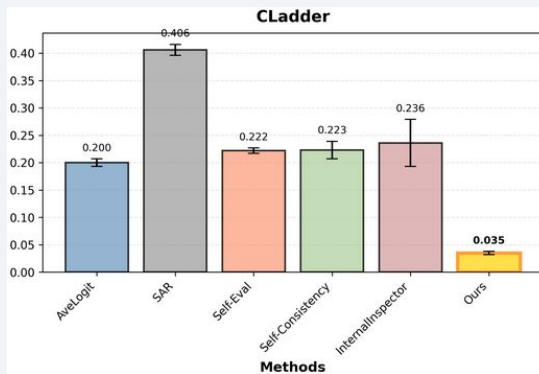
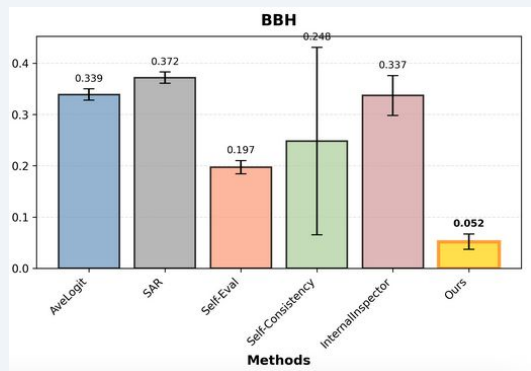
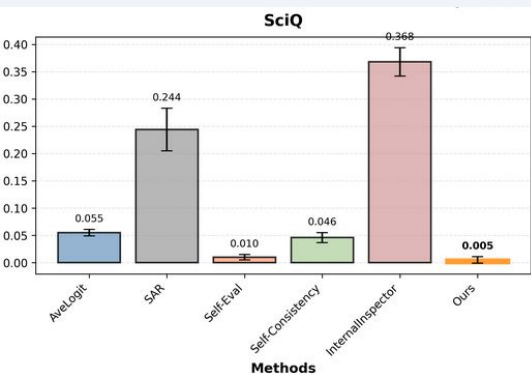
A3: in-domain STL mining with fixed parameters

Ours: in-domain STL with instance-adaptive parameters via a hypernetwork



Experimental Results: Comparison to Baselines

Expected Calibration Error (↓) Comparison (Qwen3-3B)



Takeaways

1. Our method shows consistently strong **calibration**, ranking first on three out of four datasets, with a clear margin over baselines.
2. **Math is a notable exception**, where our method underperforms InternallInspector and Self-Consistency, suggesting limitations on complex reasoning tasks.
3. **SAR systematically underperforms** across all datasets, indicating that token-relevance-weighted logits alone are insufficient for reliable calibration.
4. **InternallInspector is unreliable**, swinging between best and worst depending on the dataset.
5. **Task complexity matters**: harder reasoning tasks yield higher calibration error across all methods and shift the relative ranking.

Future Work

Methodological Extensions:

- Automatic segmentation strategies and multi-granularity fusion
- Richer signal sources: hidden states, attention patterns beyond token probabilities
- Open-ended generation tasks with ill-defined correctness

Applications to Agents & Collaboration:

- Agentic planning with confidence-triggered replanning
- Multi-agent collaboration: disagreement detection and consensus formation
- Confidence propagation in hierarchical task graphs

Theoretical Extensions:

- Why do failure patterns generalize better than success patterns?
- Multi-agent calibration guarantees

Conclusion

- 1 Discriminative STL mining from stepwise confidence signals in LLM reasoning
- 2 Temporal structures for incorrect reasoning generalize across tasks
- 3 Given a fixed STL structure, optimal discriminative parameters vary per question
- 4 STL Blocks + Hypernetwork architecture improves calibration in LLM reasoning

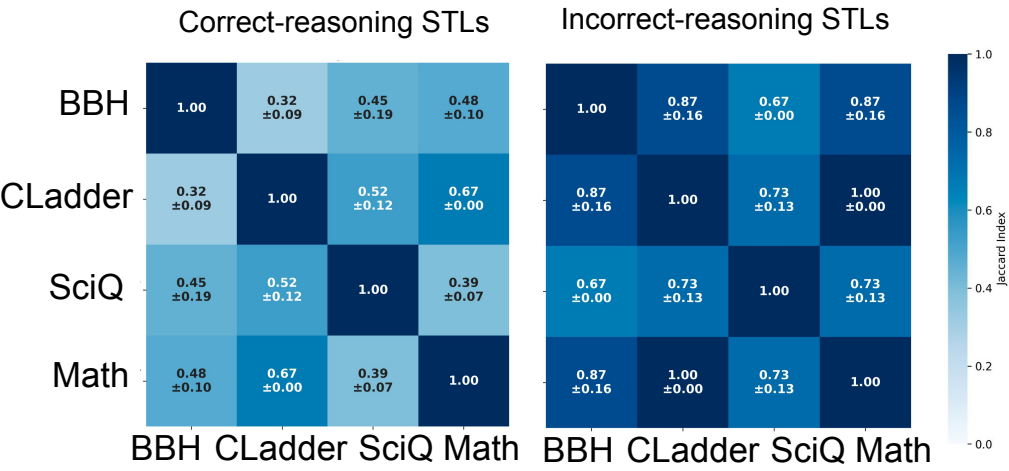
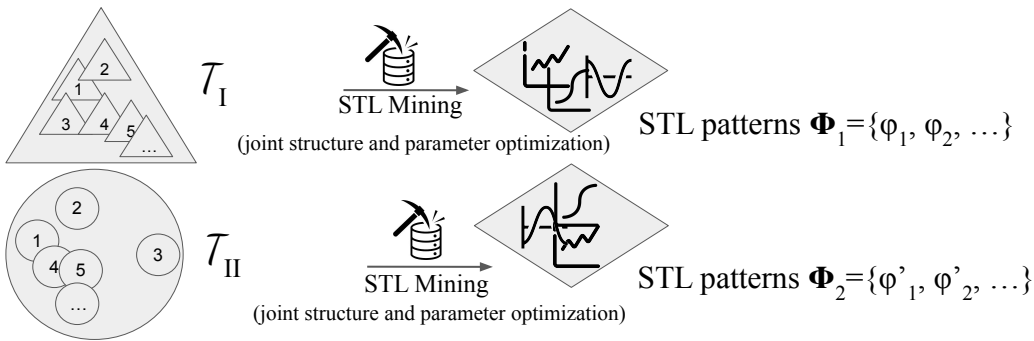
Limitations: Segmentation granularity affects signal quality | Effective calibration still requires in-task mining | Evaluated only on structured reasoning benchmarks

[Zhenjiang Mao*, Anirudhh Venkat*, Artem Bisliouk, Akshat Kothiyal, Sindhura Kumbakonam Subramanian, Saithej Singhu, Ivan Ruchkin. *Confidence over Time: Confidence Calibration with Temporal Logic for Large Language Model Reasoning*, in Findings of the ACL 2026]

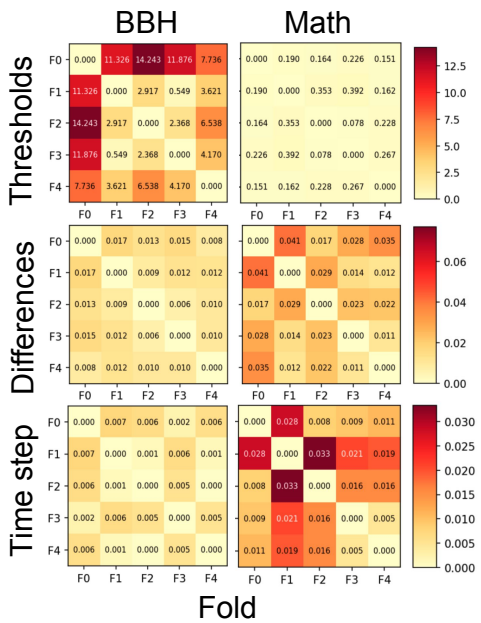
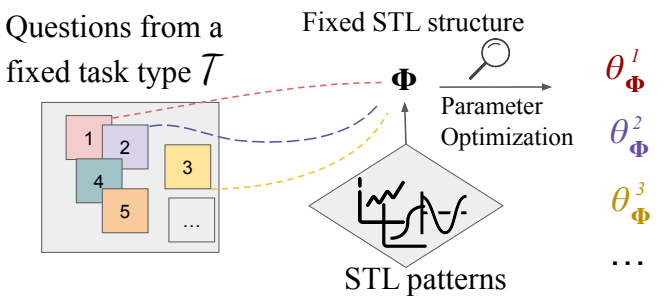
Backup

Research Questions

RQ 1: Do STL pattern sets Φ_1 and Φ_2 overlap?



RQ 2: Are $\theta^1_{\phi_1}$, $\theta^2_{\phi_1}$, $\theta^3_{\phi_1}$... similar?



Quantifying Confidence with STL

An STL formula φ is defined recursively as:

$$\varphi ::= \mu \mid \neg \varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathbf{G}_I \varphi \mid \mathbf{F}_I \varphi \mid \varphi_1 \mathbf{U}_I \varphi_2$$

μ denotes an atomic predicate,

G (Globally) specifies that a formula must hold at all times within the interval I

F (Finally) specifies that a formula must hold at some time within an interval

U (Until) specifies that φ_1 must hold continuously until φ_2 becomes true within the interval I

Instruction: x

 *Divide 2 teachers and 4 students into 2 groups and arrange them to participate ...*

Response: y

 *First, we need to divide 2 teachers and 4 students into 2 groups, each ... (step 1)
Next, we select 2 students out of 4 for the first group. This can be done in ... (step 2)
...
Finally, ... (step n)*

Step confidence signal:

$$S = [s_1, s_2, \dots, s_n]$$



$$\text{STL } \varphi := \square_{[1,t]} (s_t \geq \mu)$$



$$\text{Robustness score: } \rho(\varphi, S, t) \in \mathbb{R}$$