Argument Graph Uncertainty

Quantifying Uncertainty from the Logical Structure of Reasoning Chains

Simon Baur [1] Sumire Honda [1] Wojciech Samek [1] Jackie Ma [1]

[1]Fraunhofer Heinrich-Hertz-Institute – Fraunhofer HHI, Berlin, Germany

arXiv:2610.20328

NEURAL INFORMATION PROCESSING SYSTEMS

Fraunhofer HHI

TL;DR

AGU turns a *single* chain-of-thought into an argument graph and reads the model's uncertainty off its logical structure: which steps support or contradict one another, and which answer they commit to. Post hoc, interpretable per step, and with no extra calls to the reasoning model.

8 / 9 models where AGU has the best PRR on GPQA Diamond

+0.10 PRR when AGU is blended with consistency (MMLU-Pro)

0 extra calls to the evaluated model

Reasoning models break today's UQ signals

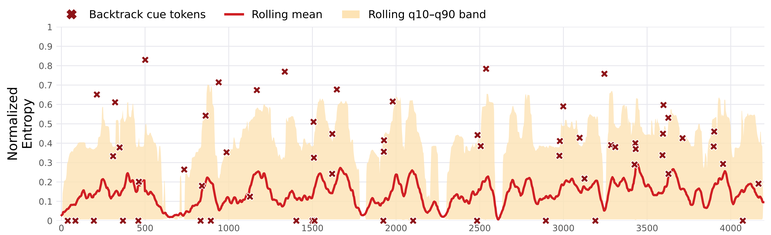


Figure 1. Token entropy along one GPQA reasoning trace. The rolling mean stays between about 0.1 and 0.2, and backtracking cues (×) do not line up with its spikes.

- **Token entropy** is cheap, but deep context inflates token confidence regardless of reasoning quality.
- **Consistency** samples S full traces and compares the final answers: S times the inference cost, and the reasoning itself is thrown away.
- **Trace heuristics** (length, backtracking cues) are confounded by problem difficulty and RL training artifacts.

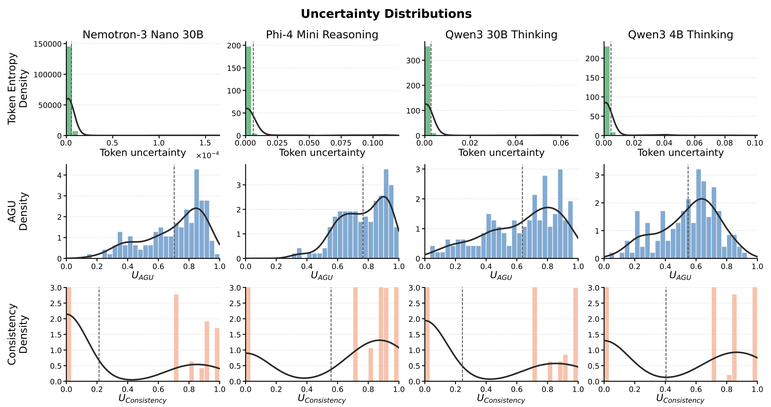
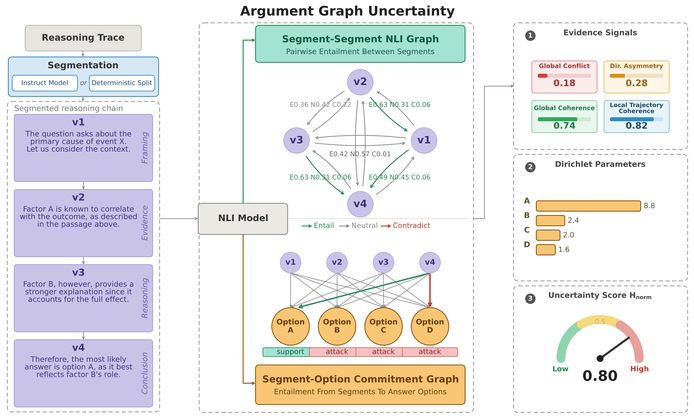


Figure 2. Score distributions for four models. Token entropy (top) collapses to near zero and consistency (bottom) is bimodal; AGU (middle) spreads across $[0, 1]$.

None of the existing signals looks at what decides whether a chain can be trusted: its logical structure.

Argument Graph Uncertainty (AGU)

Steps of the trace are nodes; NLI entailment and contradiction act as **support** and **attack** relations (Dung, 1995). The evidence feeds a Dirichlet posterior over the answer options (Sensoy et al., 2018).



NLI edge scores

$$f_{\mathrm{NLI}}(p, h) = (p_E,\, p_N,\, p_C), \qquad p_E + p_N + p_C = 1$$

Scored for segment pairs (v_i, v_j) and segment–option pairs (v_i, h_k), with h_k = "The answer is Option k: ...". Net entailment: $s_{ij} = p_E(v_i, v_j) - p_C(v_i, v_j)$.

1 Evidence signals

$$\mathrm{Conflict} = \frac{1}{n(n-1)} \sum_{i \neq j} p_C(v_i, v_j)$$

$$\mathrm{Asym} = \frac{2}{n(n-1)} \sum_{i<j} \left| p_E(v_i, v_j) - p_E(v_j, v_i) \right|$$

$$\mathrm{coh}_i = \frac{1}{2}\left(1 + \frac{1}{n-1} \sum_{j \neq i} s_{ij}\right)$$

$$\mathrm{traj}_i = \frac{1}{2}\left(1 + \frac{1}{|T(i)|} \sum_{(a,b) \in T(i)} s_{ab}\right)$$

$T(i)$: the edges from the previous step into v_i and from v_i into the next.

Option support and step reliability

$$e_{i,k} = \max\left(0,\; p_E(v_i, h_k) - p_C(v_i, h_k)\right)$$

$$r_i = \lambda\, \mathrm{coh}_i + (1 - \lambda)\, \mathrm{traj}_i$$

A step is evidence for an option only if it entails it more than it contradicts it, weighted by how well it fits the rest of the trace.

2 Dirichlet parameters

$$\kappa_{\mathrm{eff}} = \kappa\, \exp\left(-\gamma_c\, \mathrm{Conflict} - \gamma_a\, \mathrm{Asym}\right)$$

$$\alpha_k = \alpha_0 + \kappa_{\mathrm{eff}} \sum_{i=1}^{n} r_i\, e_{i,k}$$

Traces that contradict themselves get less evidence overall.

3 Uncertainty score

$$\hat{p}_k = \frac{\alpha_k}{\sum_j \alpha_j}, \qquad U_{\mathrm{AGU}} = -\frac{1}{\log K} \sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k$$

Defaults for all models: $\alpha_0 = 1, \kappa = 12, \gamma_c = 1.2, \gamma_a = 1, \lambda = 0.5$.

AGU on one real trace

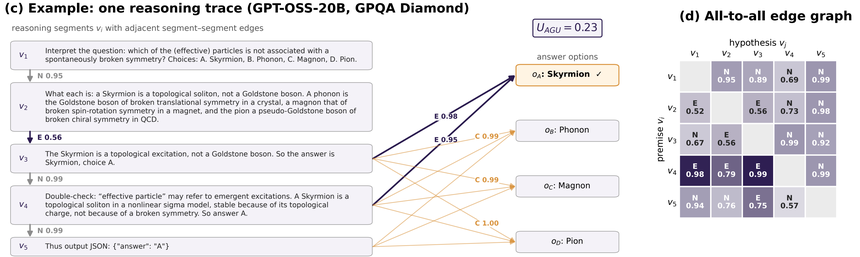


Figure 3. The concluding segments entail the correct option ($p \geq 0.95$) and contradict the others ($p \geq 0.99$).

Results

Setup. Nine open reasoning models on GPQA Diamond, MMLU-Pro and TruthfulQA, with one fixed set of hyperparameters. Metric: Prediction Rejection Ratio (PRR; 1 = oracle rejection, 0 = random).

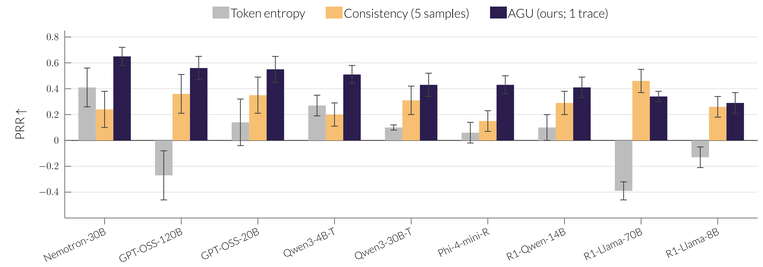
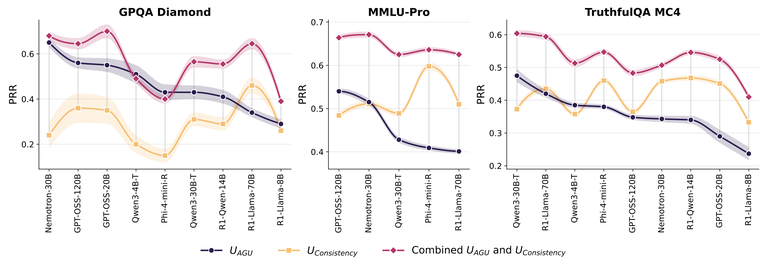


Figure 4. GPQA Diamond, PRR (mean ± std): AGU is the best signal on 8 of 9 models.

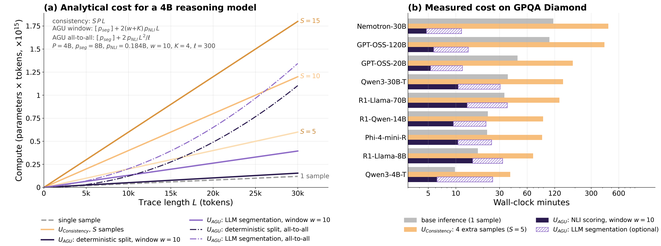
AGU can be combined with consistency: +0.10 PRR

Equal-weight blend vs. the stronger of the two on MMLU-Pro (+0.09 on TruthfulQA). Best signal for every model on both, and for 7 of 9 on GPQA Diamond.



AGU is about 10× faster than consistency

Median over nine models of the measured wall-clock overhead on GPQA Diamond, against four extra samples (range 5× to 97×).



Limitations and next steps

- Multiple-choice only: open-ended generation needs a commitment graph without a fixed answer set.
- General-domain NLI can misjudge specialised scientific language; a scorer trained on reasoning text is the next step.

Supported by the Senate of Berlin and the European Commission's Digital Europe Programme (DIGITAL), grant TEF-Health (101100700). Contact: simon.baur@hhi.fraunhofer.de