DoG: Sniffing Out Overconfidence in LLM Agents

via Post-hoc Trajectory Restructuring

Hyunjun Jeon* Dongha Lim* Kunwoong Kim Daewon Choi Jinwoo Shin

Korea Advanced Institute of Science and Technology (KAIST)

* equal contribution

NeurIPS 2026

LLM Agents Are Overconfident

- Agents **plan**, **call tools**, and solve multi-step tasks …

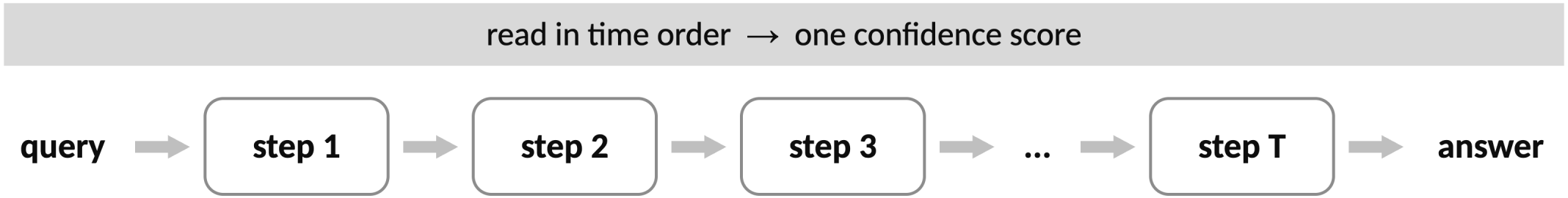


"I'm confident." — even when the answer is **wrong**

- … but often report **high confidence** on **incorrect** answers
- Sampling more rollouts to check agreement is **impractical**
  - every rollout = many tool calls, real API cost, possible side effects

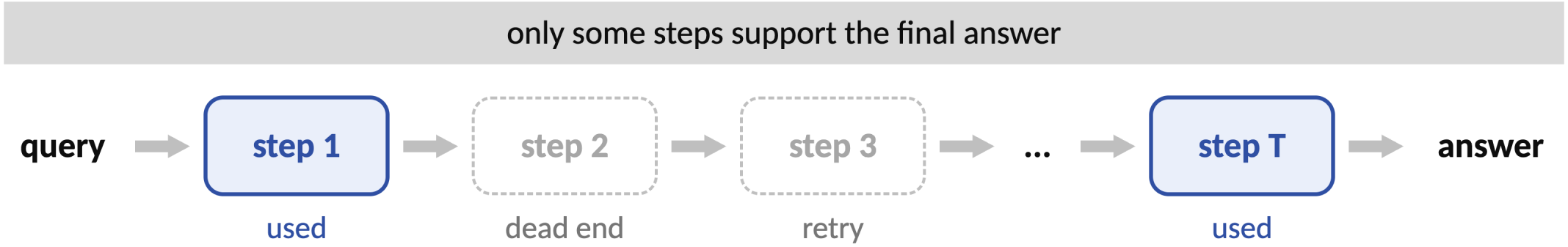
Trajectory: Sequence or Graph?

- Existing estimators read the trajectory as a **flat, time-ordered sequence**
  - Static LLM: P(True), Verbalized, Hedging [Kadavath et al, 2022; Tian et al, 2023; Tao et al, 2025]
  - Agent-specific: TRACER, HTC [Tayebati et al, 2026; Zhang et al, 2026] — need log-probs or an embedding model

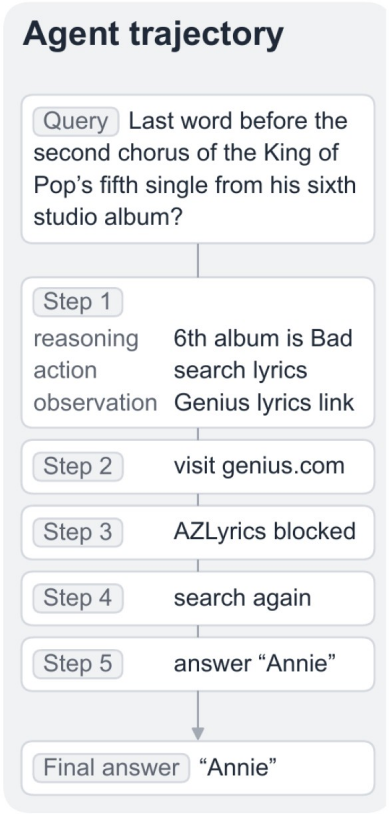


Trajectory: Sequence or Graph?

- Existing estimators read the trajectory as a **flat, time-ordered sequence**
  - Static LLM: P(True), Verbalized, Hedging [Kadavath et al, 2022; Tian et al, 2023; Tao et al, 2025]
  - Agent-specific: TRACER, HTC [Tayebati et al, 2026; Zhang et al, 2026] — need log-probs or an embedding model
- But real agents **branch**, **retry**, and hit **dead ends** — which steps matter?

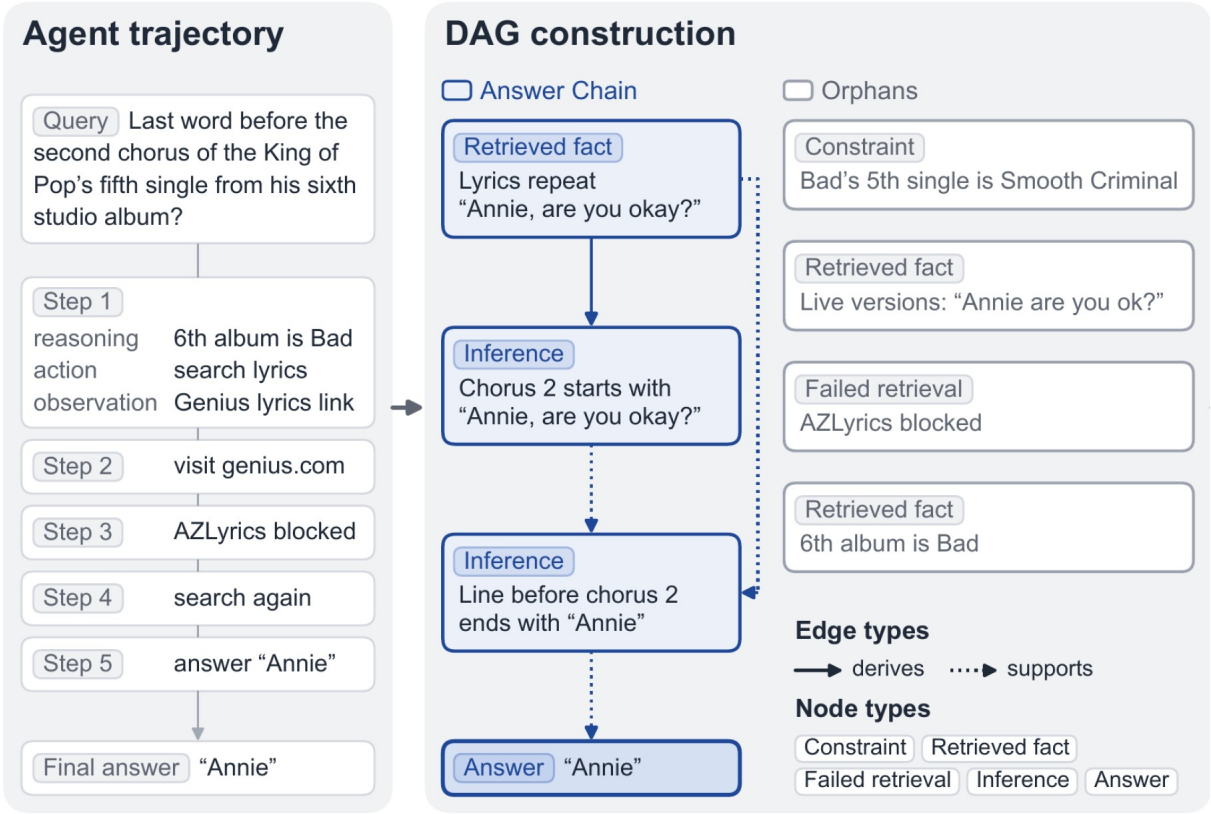


DAG of Grounds (DoG)



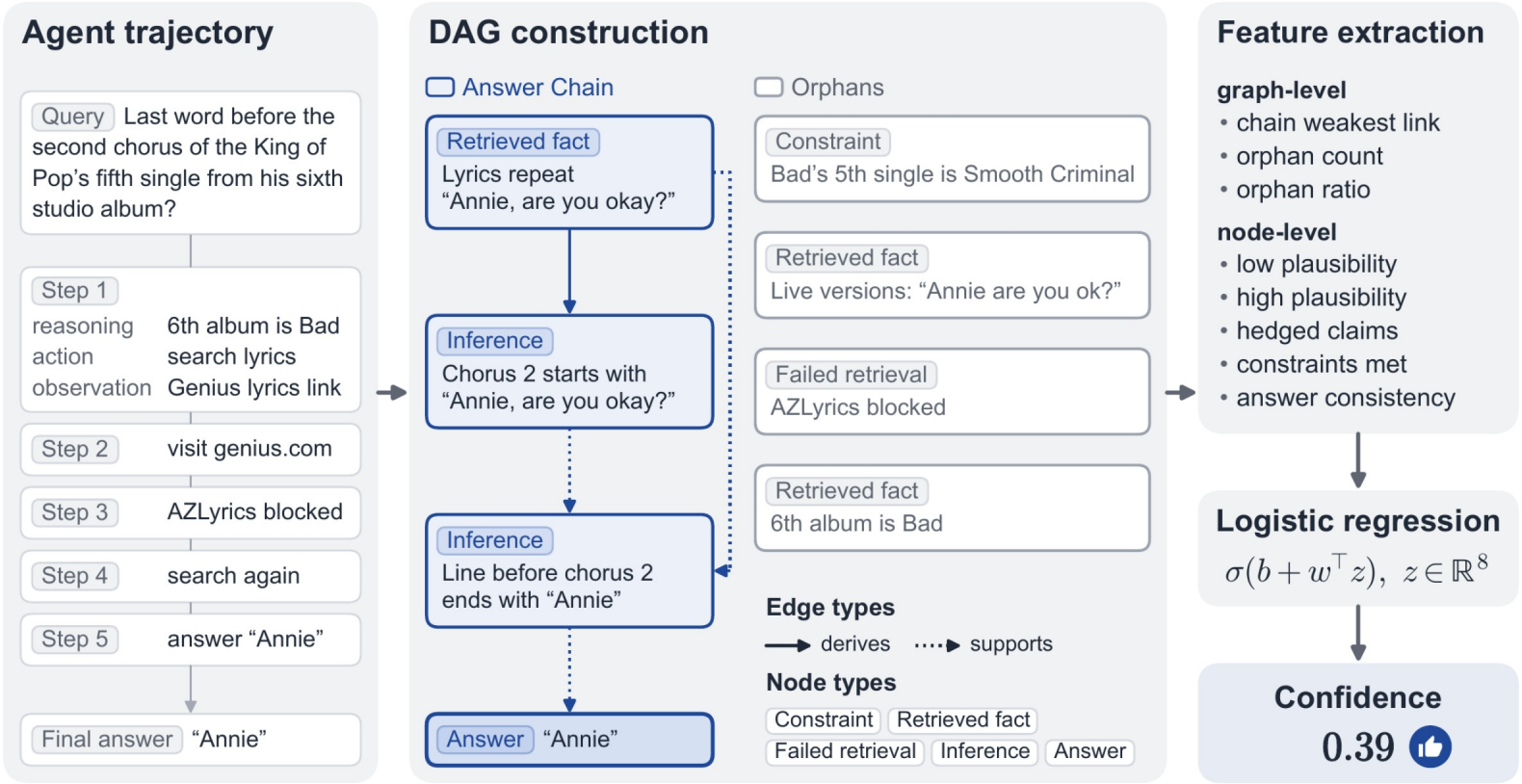
Input: a **finished** trajectory — a real GAIA run of GPT-5.4-nano, wrong answer

DAG of Grounds (DoG)



One post-hoc LLM call → a DAG: **Answer Chain** vs. **orphans**

DAG of Grounds (DoG)



8 features → **9-parameter** logistic regression → confidence **0.39** (no further LLM call)

From DAG to Confidence

graph-level

- chain-restricted weakest link
- orphan count
- orphan ratio

node-level

- low-plausibility inferences
- high-plausibility inferences
- hedged claims
- constraint satisfaction
- answer consistency

$$c = \sigma(b + w^T z)$$

z: 8 standardized features 9 learnable parameters

Properties:

- **black-box**: no logits, no embedding model
- **no extra rollouts**; agent loop unchanged
- **one** post-hoc LLM call (≈ $0.0023 per task)
- calibrator fit on **labeled** target trajectories

Example: Fluent but Ungrounded

Query (GAIA): “In April of 1977, who was the Prime Minister of the first place mentioned by name in the Book of Esther?”

Answer Chain

Retrieved fact
Hoveyda: PM of Iran, Jan 1965 – Aug 1977

↓ derives

Inference
so Hoveyda was PM in April 1977

↓ supports

Answer
“Amir-Abbas Hoveyda”

Orphans

Retrieved fact
first place named in Esther: “the citadel of Susa”

the key premise is **never linked** to the answer

Failed retrieval
web result on Esther 1

Verbalized confidence

0.70

DoG

0.39

Results: Calibration

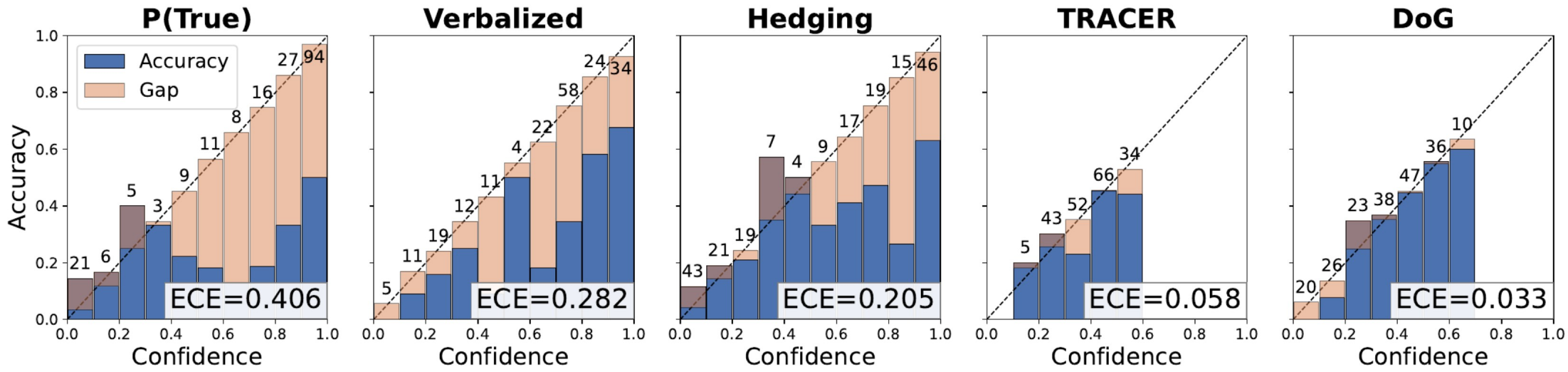
	GPQA			GAIA			HLE		
Method	**ECE ↓**	**Brier ↓**	**AUROC ↑**	**ECE ↓**	**Brier ↓**	**AUROC ↑**	**ECE ↓**	**Brier ↓**	**AUROC ↑**
P(True)	0.406	0.394	0.666	0.173	0.223	0.715	0.453	0.434	0.547
Verbalized	0.282	0.287	0.729	0.150	0.235	0.655	0.332	0.290	0.637
Hedging	0.205	0.254	0.728	0.297	0.318	0.661	0.182	0.223	0.626
TRACER	0.058	0.221	0.616	0.054	0.230	0.547	0.088	0.167	0.632
HTC†	0.102	0.213	0.706	0.142	0.233	0.686	0.031	0.090	0.644
DoG (ours)	0.045	0.195	0.731	0.041	0.194	0.727	0.056	0.156	0.709

DoG: **1st or 2nd** in all 9 (benchmark, metric) cells

blue = best in column. GPT-5.4-nano agent with full tools; base accuracy 35.5% (GPQA), 64.2% (GAIA), 22.0% (HLE); DoG = mean of 5 runs. †HTC numbers are copied from its original paper (different base LLM and base accuracy).

Results: Calibration

- Reliability diagrams on GPQA: **accuracy** vs. **predicted confidence**



zero-shot baselines: **overconfident**

never predicts above ~0.6

spread out and on the diagonal

Results: Which Signals Matter?

- **Leave-one-out** feature ablation on GAIA

drop **low-plausibility inferences**

+0.066

ECE gets worse

largest calibration effect (≈ 8σ of run-to-run variation)

drop **chain-restricted weakest link**

−0.060

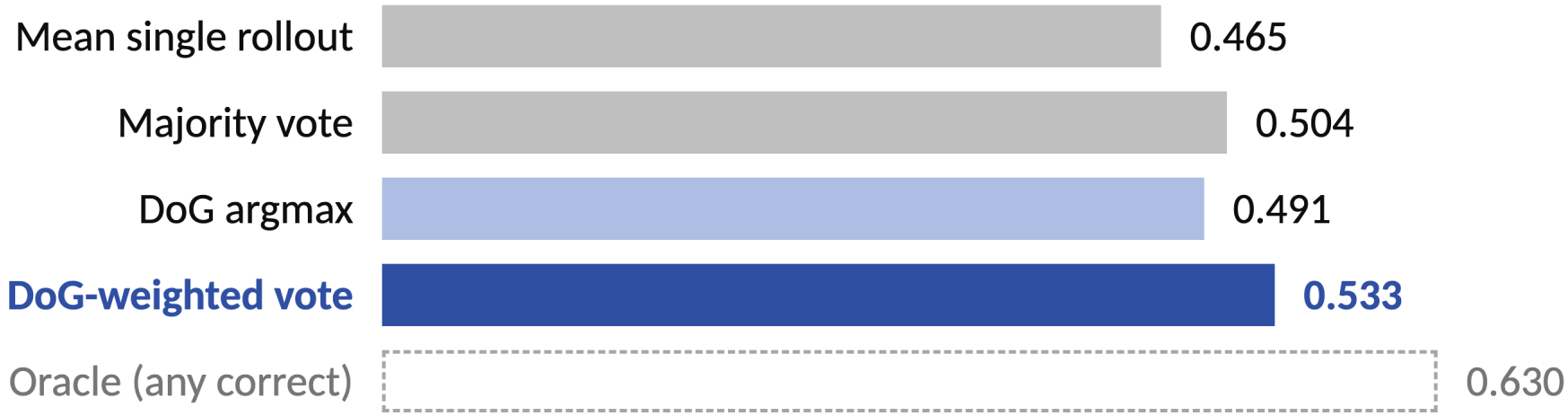
AUROC drops

3–4× larger than any other feature

node-level judgments → calibration · **graph** structure → discrimination

Application: Test-time Scaling

- **Best-of-4** rollouts on all 165 GAIA tasks: which answer to pick?



+3.0 points over majority voting · recovers **41%** of the oracle gap

one extra LLM call per rollout; no labels from the new rollouts; DoG argmax alone trails majority voting

Robustness Across Agent Environments

- **7** base LLMs, **3** tool regimes, **3** benchmarks — three levels of robustness

✓ **Per-setting calibration**

across 7 backbones: **ECE < 0.09**, **AUROC > 0.67**

calibrator fit in each setting (GAIA)

✓ **Zero-shot ranking**

mean AUROC change: **−0.007 / −0.017 / −0.008**

applied unchanged across benchmarks / backbones / tool regimes

---

✗ **Zero-shot calibration**

ECE rises to **0.198 / 0.121 / 0.247**

refitting only the intercept b closes 93% / 79% / 79% of the gap, but needs target labels

Conclusion

- Agent trajectories are **graphs**, not flat sequences
- DoG: **one post-hoc call** → DAG of grounds → **9-parameter** calibrator
  - black-box: no logits, no external models, no extra rollouts
- **1st or 2nd** on all 9 (benchmark, metric) cells; ranking **transfers**
- Limitations: LLM-built DAGs contain **errors**; calibration needs **target labels**
  - 8.0% of randomly sampled edges unsupported (single annotator)
- Interactive diagnostic tool: calibration-playground.vercel.app

Thank you!