Conformal Path Reasoning

Trustworthy KGQA via Path-Level Calibration

Shuhang Lin[1] * Chuhao Zhou[2] * Xiao Lin[3] Zihan Dong[1]
Kuan Lu[2] Zhencan Peng[1] Jie Yin[4] † Dimitris N. Metaxas[1] †

ICML 2026

[1]Rutgers University [2]Independent Researcher [3]UIUC [4]University of Sydney
*Equal contribution †Corresponding authors

Outline

Motivation

Background & Problem

CPR: Conformal Path Reasoning

Experiments

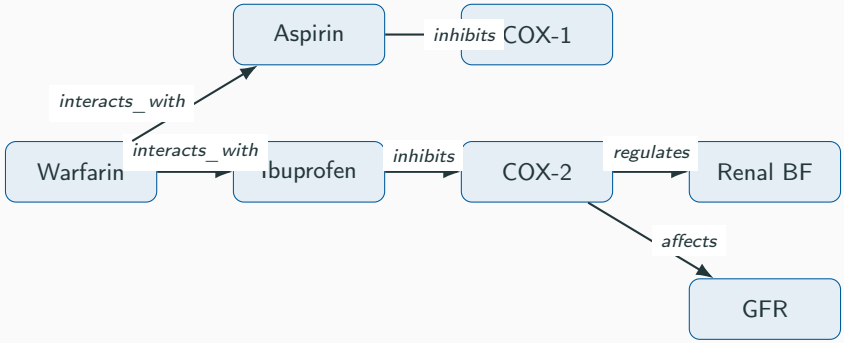
Future Work

Summary

Motivation

What is a Knowledge Graph?

A **Knowledge Graph (KG)** stores facts as **(head, relation, tail)** triples.



Where KGs power real systems:

Public & search: Wikidata, Google KG, OpenAlex
Science / health: drug discovery, biomedical QA
Industry: financial / supply-chain, recommendation
LLM era: RAG grounding, agent memory

Why Trustworthy KGQA?

Knowledge Graph Question Answering (KGQA) grounds reasoning in structured facts and produces interpretable multi-hop paths.

High-stakes applications need more than a single answer

- Medical diagnostics, financial risk assessment, scientific QA
- A single best-guess entity is **not enough**:
  - Ambiguous KGs (multiple valid answers)
  - No rigorous reliability statement for downstream use

Goal. Return a *set-valued prediction* that provably contains the ground-truth answer at a user-specified risk level α.

Running Example: A High-Stakes KGQA Query

Query (e.g., clinical decision support):

"What is the underlying mechanism of NSAID toxicity in chronic kidney disease (CKD) patients on Warfarin?"

- Hop 1: `Warfarin` $\xrightarrow{\text{interacts\_with}}$ {`Aspirin`, `Ibuprofen`, ...}
- Hop 2: each candidate $\xrightarrow{\text{inhibits}}$ target enzyme {`COX-1`, `COX-2`}
- Hop 3: `COX-2` $\xrightarrow{\text{regulates}}$ `Renal blood flow`, `GFR`, ...

Why a single answer is dangerous

Missing one unsafe interaction at hop 1 makes the correct mechanism unreachable at hop 2-3 – errors compound across hops.

We need an **answer set with a calibrated coverage guarantee**.

From Single Answer to a Set with Guarantees

In high-stakes KGQA, a single best-guess can be insufficient:

- Multiple plausible answers may exist (e.g., several unsafe drugs)
- One missed risk can mean a patient harmed or a fraud undetected

Shift the goal:

Single point	$\Longrightarrow$	**A set with a guarantee**
`COX-2 inhibition`		{`COX-2 inhibition`, `Vasoconstriction`, `Reduced GFR`}
"best guess"		"ground truth $\in$ set, prob. $\geq 1-\alpha$"

What we need: a tool that **provably** bounds the error rate on any new query, regardless of model or data.

$\Rightarrow$ This is exactly the promise of **Conformal Prediction**.

Background & Problem

Conformal Prediction: The Idea

Problem with point predictions. A model says *"diabetes"* with 70% confidence. Is that 70% calibrated? **No guarantee**.

Conformal Prediction (CP) returns a set, not a point.

Provably contains the truth with probability $\geq 1 - \alpha$.

Setting	**Point prediction**	**CP set ($\alpha = 0.1$)**
Diagnosis	"diabetes"	{diabetes, pre-diabetes}
Image	"cat"	{cat, lynx, fox}
KGQA	`COX-2 inhib.`	{`COX-2 inhib.`, `Vasoconstriction`, `Reduced GFR`}

Trade-off. Smaller set $\Leftrightarrow$ more useful; bigger set $\Leftrightarrow$ more reliable.
CP gives **the smallest set with the desired coverage**.

Background: Split Conformal Prediction

Given a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$ and a non-conformity score $S(x, y)$ (lower = better fit):

$$\tau_\alpha = \text{Quantile}\Big(\{S(x_i, y_i)\}_{i=1}^{n},\ \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\Big)$$

$$C(x) = \{y : S(x, y) \leq \tau_\alpha\}$$

Under **exchangeability** of calibration and test samples,

$$\mathbb{P}(y_{\text{test}} \in C(x_{\text{test}})) \geq 1 - \alpha.$$

Reading the recipe: score how "odd" each calibration point is, pick a threshold at the $(1-\alpha)$ quantile, then the test set is "everything not too odd".

Where Standard CP Hits Its Limits in KGQA

Two distinct issues – one breaks the *guarantee*, one breaks the *usefulness*.

Issue 1 – Validity (the guarantee breaks)
Multi-hop calibration: hop k's set depends on hop $k-1$'s threshold, so calibration scores get **entangled across calibration samples**.

Exchangeability is violated $\Rightarrow$ coverage guarantee no longer holds.

Issue 2 – Efficiency (the set blows up)
KGQA has no off-the-shelf *path-level* score; local-similarity scores can't tell correct paths from plausible-but-wrong ones.

Coverage still holds, but the prediction set is so large that it loses practical utility.

CPR addresses both:

- **QE-CP** (Query-Exchangeable CP) – query-level calibration $\Rightarrow$ fix validity
- **RCVNet** – learned discriminative path scorer $\Rightarrow$ fix efficiency

The Pitfall of Hop-Level Calibration

A natural strategy: calibrate at each hop k during multi-hop traversal (e.g. UAG).

Concrete failure (drug query example): if hop-1 drops `Ibuprofen`, the chain `Ibuprofen`→`COX-2`→`Renal blood flow` is unreachable at hop-2 and hop-3 – errors cascade and are unrecoverable by any later threshold.

What goes wrong (formal)
Let $S_k^{(i)}$ be the hop-k score for query i. Hop k depends on the prediction set of hop $k-1$:

$$S_k^{(i)} = f(x_i,\, C_{k-1}^{(i)}), \qquad S_1 \to C_1 \to S_2 \to C_2 \to \cdots$$

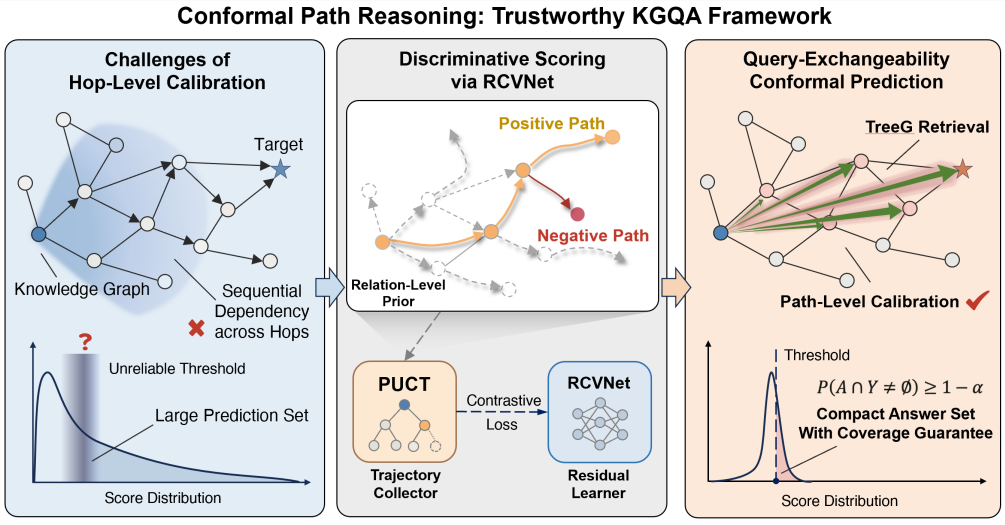
Because $C_{k-1}^{(i)}$ uses thresholds computed from *all* calibration samples, scores become **entangled across samples**.

$$\Rightarrow \text{ exchangeability is violated } \Rightarrow \text{ coverage guarantee fails.}$$

Key observation. *Paths within a query* are correlated, but *queries themselves* are i.i.d. samples. Exchangeability is restored at the **query level**.

CPR: Conformal Path Reasoning

CPR: Overview



(a) Hop-level pitfall (b) Discriminative scoring (RCVNet via PUCT) (c) Path-level CP

CPR: Two Innovations

1. **Theoretical** – **QE-CP**: query-level calibration over path scores (restores exchangeability $\Rightarrow$ valid coverage)
2. **Methodological** – a **Calibration–Inference paradigm**:
   - **Calibration** (offline): PUCT collects ⟨pos, neg⟩ paths $\rightarrow$ trains RCVNet (Residual FiLM-MLP value network)
   - **Inference** (online): TreeG parallel retrieval scored by RCVNet $\rightarrow$ CP threshold $\rightarrow$ answer set

Theoretical Component: Query-Exchangeable CP (QE-CP)

Idea. Move the calibration unit from *hops* to the *query as a whole*.

Hop-level (broken)

$S_1 \to C_1 \to S_2 \to C_2 \to \dots$

S_k depends on C_{k-1}

$\Rightarrow$ entangled *across queries*

$\Rightarrow$ **exchangeability violated**

Query-level (works)

One score per query: $S(q_i)$

Cal/test split is random

$\Rightarrow$ queries are i.i.d.

$\Rightarrow$ **exchangeable by construction**

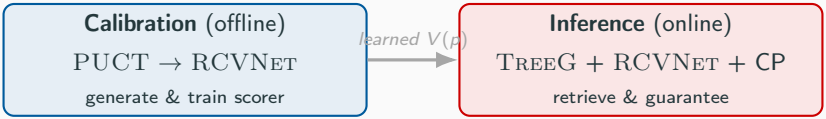
Per-query path score: $S(q_i) = \min_{p \in P^\star_{q_i}} V(p)$, V from RCVNET.

Theorem (Coverage validity)
If queries $\{q_i\}$ are exchangeable, then for any $\alpha \in (0, 1)$,

$$\mathbb{P}(\mathcal{A}_q \cap Y_q \neq \emptyset) \geq 1 - \alpha,$$

where Y_q is the (possibly multi-element) ground-truth answer set for query q.

Methodological Overview: A Two-Stage Paradigm



- **Expensive exploration is paid once** during calibration
- **Inference is fast**: no PUCT, just lightweight retrieval + CP
- Same RCVNET $V(p)$ is used both to *train* and to *score* at test

Calibration Phase: PUCT Trajectory Collection

Why PUCT? (Predictor + UCT, AlphaGo / AlphaZero)
Balances **exploit** (Q) and **explore** (visit-count bonus, prior P):

$$a^* = \arg\max_{a \in \mathcal{A}(s)} Q(s,a) + c_{\text{puct}} \cdot P(s,a) \cdot \frac{\sqrt{N(s)}}{1 + N(s,a)}$$

$Q(s,a)$: empirical mean return of action a at state s $P(s,a)$: action prior, softmax over (relation semantic similarity + mean of Beta posterior $\mu(r)$) $N(s,a)$: visit count of (s,a), $N(s) = \sum_{a'} N(s,a')$.

What it gives us. Walking the KG offline, PUCT produces:

- **Positive** (answer-reaching) + **negative** (dead-end) path pairs $\rightarrow$ supervision $\mathcal{D} = \{(p_{\text{pos}}, p_{\text{neg}})\}$ for RCVNET
- **Beta posteriors** $\pi(r) \sim \mathrm{Beta}(\alpha_r, \beta_r)$ on relations (rewarded on success, penalized on failure)

Calibration Phase: Training RCVNet

Goal. Learn a **discriminative path scorer** $V(p)$ from PUCT pairs.

Architecture. FiLM-modulated MLP – $\gamma(c), \beta(c)$ shape hidden activations:

$$h' = \gamma(c) \odot h + \beta(c)$$

Inputs.

- Scalar $x \in \mathbb{R}^3 = [\,\mathrm{sim}_{q\text{-}rel},\ \mathrm{sim}_{q\text{-}path},\ \mathrm{prior}_{rel}\,]$
- Conditioning $c = \mathrm{concat}(q_{\mathrm{emb}}, r_{\mathrm{emb}}, p_{\mathrm{emb}}) \in \mathbb{R}^{3D}$

Pairwise ranking loss on PUCT-collected pairs:

$$\mathcal{L} = \mathrm{softplus}\big(V(p_{\mathrm{pos}}) - V(p_{\mathrm{neg}})\big)$$

Inference Phase: TreeG Retrieval + Conformal Thresholding

Inference. Given a query, we walk the KG hop by hop:

- Start from each topic entity in the query.
- At every hop:
  - Look at all outgoing relations from the current frontier of paths
  - Score each candidate extension with RCVNET
  - Keep only the top-A most promising paths; discard the rest
- After H hops, every surviving path ends at a candidate answer.
- Apply the conformal threshold $\hat{q}$: keep paths with $V(p) \leq \hat{q}$.
  The terminal entities form the answer set, with coverage $\geq 1 - \alpha$.

Conformal threshold: $\hat{q} = \mathrm{Quantile}\left(S,\ \frac{(n+1)(1-\alpha)}{n}\right)$. Complexity: $\mathcal{O}(H \cdot B \cdot A)$, linear in depth H.

Experiments

Experimental Setup

- **Benchmarks:** WebQSP, CWQ, PathQuestion (PQ), PathQuestion-Large (PQL)
- **Baselines:** UAG, CLM, LoFreeCP
- **CPR variants:**
  - CPR (full): TreeG + RCVNet trained on PUCT-mined negatives
  - **Abl. 2** = w/o PUCT: RCVNet trained on *random* negatives
  - **Abl. 1** = w/o PUCT & RCVNet: TreeG with semantic scores only
- **Metrics:**
  - *ECR* – Empirical Coverage Rate (↑, must be $\geq 1 - \alpha$)
  - *APSS* – Average Prediction Set Size (↓)
  - *Coverage Efficiency* = ECR / APSS (↑)
- **LLM:** Qwen3-8B as soft relation hint generator

Main Results (4 Benchmarks)

Dataset	Method	ECR (%) ↑						APSS ↓						Coverage Efficiency (%) ↑					
		0.3	0.4	0.5	0.6	0.7	0.8	0.3	0.4	0.5	0.6	0.7	0.8	0.3	0.4	0.5	0.6	0.7	0.8
WebQSP	UaG	72.3	60.4	51.8	41.8	37.4	25.6	98.81	32.52	14.29	8.23	5.58	1.67	0.73	1.86	3.63	5.08	6.70	15.3
	CLM	–	60.2	51.0	40.1	35.6	24.6	–	46.69	14.37	5.74	**2.78**	1.82	–	1.29	3.55	6.99	12.8	13.5
	LoFreeCP	–	60.0	50.1	41.0	36.7	24.1	–	48.32	10.28	6.89	3.05	**1.05**	–	1.24	4.87	5.95	12.0	23.0
	CPR (Abl. 1)	76.0	65.5	57.3	53.4	49.8	43.2	22.05	14.84	10.10	8.26	3.80	2.16	3.45	4.41	5.67	6.46	13.1	20.0
	CPR (Abl. 2)	76.0	65.6	61.0	57.3	52.5	50.0	19.76	**9.89**	7.53	5.50	3.02	1.88	3.85	6.63	8.10	10.4	17.4	26.6
	CPR	**76.8**	**67.3**	**61.4**	**58.3**	**55.1**	**50.8**	**17.77**	10.10	**7.14**	**5.45**	3.01	1.62	**4.32**	**6.66**	**8.60**	**10.7**	**18.3**	**31.4**
CWQ	UaG	–	–	–	42.1	37.3	26.7	–	–	–	120.20	99.16	27.96	–	–	–	0.35	0.38	0.96
	CLM	–	–	–	–	32.2	22.5	–	–	–	–	10.48	5.33	–	–	–	–	3.07	4.22
	LoFreeCP	–	–	–	–	32.0	25.6	–	–	–	–	8.78	3.73	–	–	–	–	3.64	6.86
	CPR (Abl. 1)	72.2	65.5	57.8	50.0	42.2	27.6	33.43	27.84	20.50	14.41	10.96	4.61	2.16	2.35	2.82	3.47	3.85	5.99
	CPR (Abl. 2)	75.9	67.5	58.3	50.0	42.5	28.7	24.51	17.85	13.17	9.60	7.23	3.44	3.10	3.78	4.43	5.21	5.88	8.34
	CPR	**76.6**	**68.1**	**58.8**	**50.2**	**42.8**	**29.9**	**23.98**	**15.28**	**11.40**	**8.35**	**6.08**	**3.39**	**3.20**	**4.46**	**5.16**	**6.01**	**7.04**	**8.82**
PQ	UaG	–	–	54.6	41.3	37.5	33.1	–	–	3.79	1.56	**1.24**	**1.01**	–	–	14.4	26.5	30.2	32.8
	CLM	–	–	–	–	30.8	21.3	–	–	–	–	2.49	1.42	–	–	–	–	12.4	15.0
	LoFreeCP	–	–	–	–	31.2	22.4	–	–	–	–	1.98	1.21	–	–	–	–	15.8	18.5
	CPR (Abl. 1)	–	60.4	52.8	45.9	35.7	33.6	–	7.47	4.36	2.75	1.76	1.21	–	8.09	12.1	16.7	20.3	27.8
	CPR (Abl. 2)	**70.1**	60.8	54.2	46.3	38.5	36.4	22.5	6.92	3.11	1.64	1.32	1.19	3.12	8.79	17.4	28.2	29.2	30.6
	CPR	**70.1**	**61.3**	**55.1**	**47.1**	**40.8**	**37.5**	**21.8**	**6.80**	**2.95**	**1.49**	1.28	1.10	**3.22**	**9.01**	**18.7**	**32.3**	**31.9**	**34.1**
PQL	UaG	–	–	52.1	42.7	37.0	28.6	–	–	2.25	1.68	**1.05**	**0.79**	–	–	23.2	25.4	35.2	36.2
	CLM	–	–	–	–	33.3	27.6	–	–	–	–	8.61	6.18	–	–	–	–	3.88	4.47
	LoFreeCP	–	–	–	–	31.8	21.9	–	–	–	–	3.68	0.94	–	–	–	–	8.64	23.3
	CPR (Abl. 1)	–	60.6	51.4	42.1	35.1	31.9	–	7.25	3.47	2.46	1.62	1.12	–	8.36	14.8	17.1	21.7	28.5
	CPR (Abl. 2)	70.6	61.8	52.4	44.3	37.9	36.4	22.43	6.12	2.89	1.77	1.28	1.03	3.15	10.1	18.1	25.0	29.6	35.3
	CPR	**71.3**	**62.7**	**53.1**	**45.2**	**39.0**	**38.5**	**20.85**	**5.87**	**2.65**	**1.51**	1.09	0.95	**6.57**	**10.7**	**20.0**	**29.9**	**35.8**	**40.5**

"–" = coverage guarantee violated. CPR valid at *all* α.

Ablations & Sensitivity Studies

Component contributions (WebQSP, α=0.8).

- Full CPR: Coverage Efficiency = 0.314
- w/o PUCT (**Abl. 2**, RCVNET on random negatives): 0.266
- w/o PUCT *and* RCVNET (**Abl. 1**, semantic scores only): 0.200
- *Takeaway:* PUCT-mined negatives *and* learned scoring each contribute; together they raise coverage efficiency by ∼57%.

Robustness summary.

- LLM hint quality: CPR holds coverage even *w/o* LLM (ECR = 42%); UAG drops to 19.8%
- Scalability: 778k-node graph (∼600× larger), coverage holds, APSS +2.96
- Rollout budget: 32 rollouts is the sweet spot; random walk: coverage efficiency drops by 0.9 pp

Future Work

Future Work

- **Beyond single-graph KGQA**: dynamic / temporal KGs, cross-domain transfer, industrial-scale graphs (Freebase, Wikidata)
- **Stronger path scoring**: explore beyond FiLM-modulated MLP – e.g. graph-aware encoders for longer reasoning chains
- **Tighter conformal guarantees**: conditional / class-conditional coverage on KGs; query-difficulty-aware risk control
- **Reducing exploration cost**: tighter rollout-budget bounds for PUCT distillation; alternative trajectory collectors

Summary

Summary

- Identified a **theoretical flaw** in hop-level CP for KGQA
- Proposed **CPR**:
  - QE-CP– query-level calibration restores exchangeability
  - RCVNET (PUCT + FiLM) – learns discriminative path scores
  - TREEG– efficient online inference
- Empirical: +45% ECR, −52% APSS vs. conformal baselines; robust to LLM hint quality and 600× graph scale

Thank you.

Shuhang Lin • sl2148@cs.rutgers.edu

Code: `github.com/betterCallSherlock/CPR`

Backup: LLM Ablation (WebQSP)

LLM	CPR-ECR↑	CPR-APSS↓	UaG-ECR↑	UaG-APSS↓
Qwen3-8B	50.8	1.62	25.6	1.67
Qwen3-4B	45.6	2.45	22.6	2.70
Qwen3-1.7B	42.8	2.49	21.9	2.67
Qwen3-0.6B	42.3	2.50	20.4	2.75
w/o LLM	42.0	2.54	19.8	2.81

UaG's 19.8% falls below $1 - \alpha = 0.5$ – coverage guarantee fails.

Backup: Scalability (WebQSP, α=0.5)

Method	Graph scale	ECR (%)	APSS	Cov. Eff. (%)
UaG	Original ($\sim$1.3k nodes)	51.8	14.29	3.63
CPR	Original ($\sim$1.3k nodes)	61.4	7.14	8.60
CPR	Expanded (778k nodes)	55.3	10.10	5.48

Backup: PUCT Rollout Sensitivity (WebQSP, α=0.5)

#Rollout	ECR (%)	APSS	Cov. Eff. (%)
0 (random walk)	52.5	3.02	17.4
4	52.9	3.02	17.5
8	53.4	3.02	17.7
16	54.6	3.01	18.1
32	55.1	3.01	18.3
64	56.4	2.98	18.9