

Context-anchored Asymmetric Conformal Prediction for Uncertainty-aware Reasoning over Knowledge Graphs

Boyang Fan¹ Ziyue Qiao^{1,*} Wei Ju^{2,*} Xiao Luo³

¹School of Computing and Information Technology, Great Bay University, Dongguan, China

²School of Computer Science, Peking University, Beijing, China

³University of Wisconsin–Madison, Madison, WI, USA

boyangfan01@gmail.com ziyuejoe@gmail.com

juwei@pku.edu.cn xiao.luo@wisc.edu

Abstract

This paper studies the problem of uncertainty-aware reasoning over knowledge graphs (KGs), which aims to incorporate uncertainty quantification into unseen knowledge predictions over KGs. Previous works typically incorporate uncertainty measurement tools such as conformal predictions into graph representation learning. Despite the progress, the performance of these approaches remains limited since they usually neglect both structural information in KGs and intrinsic asymmetry between over-estimation and under-estimation. Towards this end, we propose a novel approach named Context-anchored Asymmetric Conformal Prediction for uncertainty-aware Reasoning over KGs (**CARE**). The core of our **CARE** is to identify the context of each query using anchors, which guide the construction of triple-specific asymmetry prediction intervals. In particular, we first take a subset of the calibration set as a reference pool. For each triple query, we summarize the KG structural statistics into a vector, which guides the identification of its anchor samples from the pool. The confidence residuals between query and anchors are utilized to quantify over-estimation and under-estimation, producing contextual nonconformity scores. Finally, we utilize the unselected calibration dataset to determine cutoff levels for constructing asymmetric prediction intervals. Extensive experiments on benchmark datasets validate the effectiveness of **CARE**. The code can be found at <https://github.com/fanboyang/CARE>.

1 Introduction

Knowledge graphs (KGs) encode relational facts as triples with the three elements head entity, relation, and tail entity, representing heterogeneous data (Hogan et al., 2021; Nickel et al., 2016; Ji et al., 2022). They support applications such as biomedical knowledge integration, drug repur-

posing, polypharmacy side-effect modeling, protein drug target discovery, and recommendation (Nicholson and Greene, 2020; Himmelstein et al., 2017; Zitnik et al., 2018; Mohamed et al., 2020; Wang et al., 2019). Many of these facts are not cleanly true or false. Extraction noise and probabilistic phenomena like interactions between proteins (Szklarczyk et al., 2017) produce *uncertain knowledge graphs* (UnKGs), where each triple comes with a confidence score between 0 and 1 (Pujara et al., 2013; Chen et al., 2019; Zhu et al., 2025c). On an UnKG, the task is no longer link prediction but *confidence-score prediction*, where an unseen triple is mapped to a confidence score.

Existing uncertain knowledge graph embedding (UnKGE) models return a point estimate of that confidence as a single scalar (Chen et al., 2019, 2021a,b), and downstream pipelines use the scalar to rank candidates or flag triples for expert review. A scalar alone does not tell us how reliable the estimate is. Two triples with the same predicted confidence score can have different levels of reliability, and UnKGE models cannot quantify predictive uncertainty around the score itself (Safavi et al., 2020; Tabacof and Costabello, 2020). Each point estimate therefore needs a calibrated *prediction interval* that quantifies uncertainty around the confidence-score estimate while keeping marginal coverage at or above a target level. Conformal prediction provides such a route by turning point predictions into prediction intervals using a calibration set drawn from the same distribution as the test data (Vovk et al., 2005; Angelopoulos and Bates, 2023). UNKGCP (Zhu et al., 2025c) applies this route to UnKGs through the score’s binary-entropy-normalized absolute residual, so the interval width adapts to the score’s predictive uncertainty.

However, interval baselines for confidence-score prediction reduce uncertainty to global residual distributions or scalar score-dependent scales. Parametric intervals assume a fixed distributional shape,

*Corresponding authors.

standard CP uses one absolute-residual scale, and entropy-normalized CP such as UNKGCP depends only on score and entropy. This baseline family leaves two limitations for interval modeling. On the one hand, triples with the same score can receive the same width even when their local topology, relation frequency, and neighbor support imply different residual risks. On the other hand, the interval is symmetric around the backbone score, although the backbone can be systematically optimistic in some graph regions and conservative in others. Thus, confidence-score intervals should adapt to graph-local residual behavior and directional estimation bias while preserving conformal coverage.

To address these two gaps, we propose Context-anchored Asymmetric Conformal REasoning (**CARE**). Rather than treating confidence-score uncertainty as a function of the scalar prediction alone, the framework reframes interval construction as a context-sensitive and direction-aware calibration problem. Specifically, **CARE** anchors calibration in graph-local evidence, allowing the interval to reflect residual behavior from queries with similar structural and relational contexts. It further separates under-estimation from over-estimation, so the lower and upper sides of the interval can adapt to directional bias. These local asymmetric adjustments are then integrated into split conformal prediction, retaining the marginal coverage guarantee under exchangeability (Lei et al., 2018). This design turns structural context and estimation bias into calibrated interval widths, yielding sharper and more informative uncertainty estimates without retraining the underlying UnKGE backbone. On standard uncertain knowledge graph benchmarks and across multiple backbones, **CARE** produces sharper intervals than UNKGCP at the same coverage level, cutting sharpness by 30% on average.

We make three main contributions:

- ❶ **New Perspective.** We propose that the uncertainty of confidence scores in knowledge graph datasets is significantly shaped by local graph structure and is directionally asymmetric.
- ❷ **Novel Methodology.** We introduce **CARE**, a conformal predictor that estimates asymmetric scales from local graph context. It preserves the marginal coverage guarantee of split conformal prediction.
- ❸ **Extensive Experiments.** Ablation and sensitivity studies on standard uncertain knowledge graph benchmarks attribute the sharpness gain to graph context and directional asymmetry, not to tuning.

2 Preliminaries

2.1 Problem Definition

In this work, we learn a post-hoc interval predictor for confidence-score prediction on uncertain knowledge graphs. Let E and R be entity and relation sets. An uncertain knowledge graph contains weighted triples (q_i, c_i) , where $q_i = \langle h_i, r_i, t_i \rangle \in E \times R \times E$ and $c_i \in [0, 1]$ is the confidence label (Chen et al., 2019; Zhu et al., 2025c). A trained UnKGE backbone $M_{\mathcal{T}_{\text{train}}}$ maps a query to a fixed point score $s = M_{\mathcal{T}_{\text{train}}}(q) \in [0, 1]$ (Chen et al., 2019, 2021a,b). Given fixed scores and calibration triples, the goal is to map each unseen query q_* , with unknown confidence label c_* , to a clipped interval $\Gamma^\alpha(q_*) = [L(q_*), U(q_*)] \subseteq [0, 1]$ without retraining $M_{\mathcal{T}_{\text{train}}}$. For a target coverage level α , the interval should satisfy

$$\mathbb{P}(c_* \in \Gamma^\alpha(q_*)) \geq \alpha, \quad (1)$$

while remaining short.

We evaluate interval quality with two metrics, coverage and sharpness. Coverage is the fraction of contained test labels and measures reliability. Sharpness is the mean interval length and measures informativeness. At comparable coverage, lower sharpness means stronger modeling capability.

2.2 Conformal Prediction

Conformal prediction turns fixed predictors into prediction sets or intervals with distribution-free coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018; Xiao et al., 2026; Luo et al., 2026). In **CARE**, $\mathcal{T}_{\text{ref}}$ estimates query-specific lower/upper residual scales, and $\mathcal{T}_{\text{conf}}$ selects the final max-normalized multiplier. For a nonconformity score $A(q, c \mid M)$ and calibration scores $a_i = A(q_i, c_i \mid M)$ from the conformal holdout, split conformal prediction selects $\hat{m}_\alpha = a_{(\lceil \alpha(\ell+1) \rceil)}$, where $a_{(1)} \leq \dots \leq a_{(\ell)}$ are the ℓ sorted calibration scores. Labels with nonconformity no larger than $\hat{m}_\alpha$ form the prediction interval. Thus α controls final coverage, while the choice of A controls interval efficiency.

Existing task-specific conformal methods instantiate A with scalar residuals. Direct CP uses absolute residuals and gives a symmetric interval around $M_{\mathcal{T}_{\text{train}}}(q_*)$, while UNKGCP (Zhu et al., 2025c) normalizes them by score entropy. The resulting interval has entropy-scaled half-width $\hat{m}_\alpha H(M_{\mathcal{T}_{\text{train}}}(q_*))$. This adjusts width for each

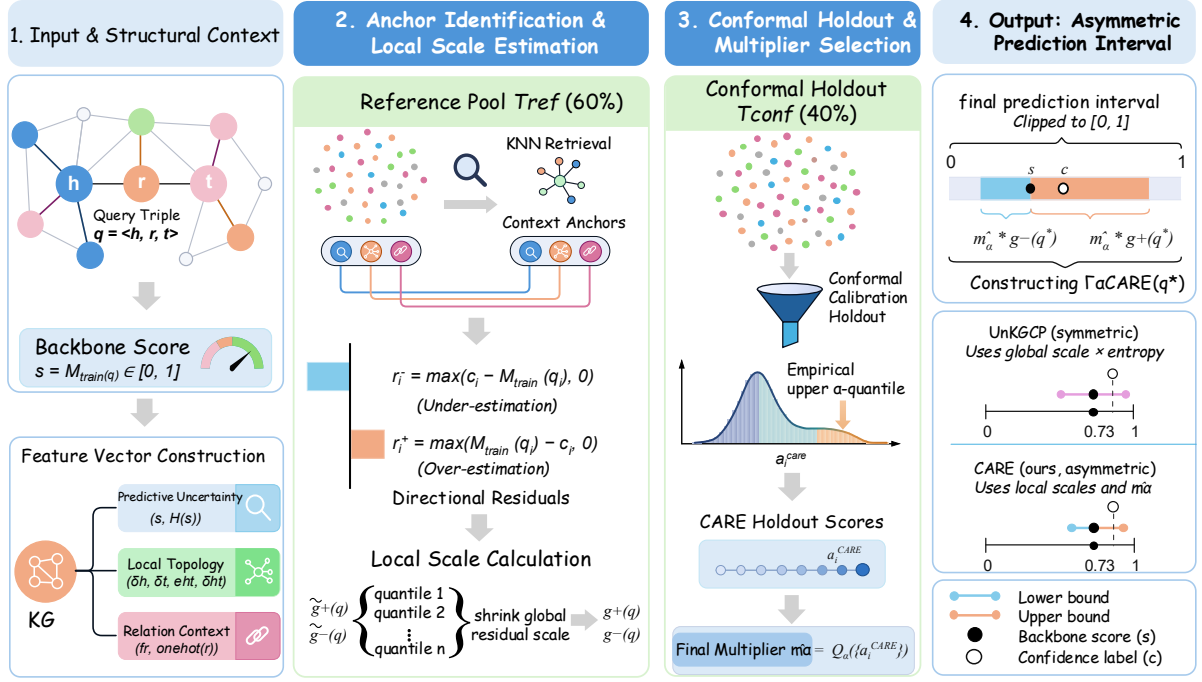


Figure 1: Overview of CARE. The main innovation is to integrate context-anchored calibration and asymmetric interval construction into the conformal framework. Specifically, the method uses fixed backbone scores and training-graph statistics to identify context anchors, estimate asymmetric residual scales from the reference pool, and apply a conformal holdout multiplier to produce calibrated lower/upper confidence-score intervals.

query by scalar score information but ignores graph context and directional error scales.

3 The Proposed CARE

3.1 Framework Overview

In this paper, we study confidence-score interval prediction for uncertain knowledge graphs and propose CARE from a context-anchored calibration perspective. Existing conformal methods attach symmetric intervals to fixed UnKGE scores, using absolute residuals or score entropy to control width (Vovk et al., 2005; Lei et al., 2018; Zhu et al., 2025c). Although they inherit split-conformal coverage, they do not use local graph context and cannot distinguish systematic over-estimation from under-estimation. CARE instead retrieves nearby calibration triples to estimate directional uncertainty, following localized and graph conformal ideas that use nearby examples or graph neighborhoods to improve efficiency (Guan, 2023; H. Zargarbashi et al., 2023; Zargarbashi and Bojchevski, 2024). The overall framework has four parts. Structural context acquisition represents each query with score, topology, and relation evidence. Calibration splitting separates anchor retrieval from coverage calibration. Anchor scale estimation turns nearby calibration residuals into lower and upper

scales. Final interval construction applies the held-out conformal multiplier to produce a clipped confidence interval. Figure 1 summarizes this workflow.

3.2 Structural Context Acquisition for Anchor Identification

Previous conformal baselines use the confidence score as a single indicator and ignore graph structure around a query (Zhu et al., 2025c). We build a structural-context vector for anchor identification in the reference pool: triples are locally comparable when predictive uncertainty, topology, and relation context are similar. The vector is computed from training-graph statistics and retrieves anchors in $\mathcal{T}_{\text{ref}}$, so residual estimates are grounded in local KG structure rather than scalar score alone.

For a query $q = \langle h, r, t \rangle$ with score $s = M_{\mathcal{T}_{\text{train}}}(q)$, the anchor feature vector is

$$\phi(q) = [\phi_{\text{pred}}(q), \phi_{\text{topo}}(q), \phi_{\text{rel}}(q)], \quad (2)$$

where $\phi_{\text{pred}} = (s, H(s))$ captures predictive evidence for anchor matching, $\phi_{\text{topo}} = (\delta_h, \delta_t, e_{ht}, \delta_{ht})$ encodes local topology, and $\phi_{\text{rel}} = (f_r, \text{onehot}(r))$ represents relation context. Here $H(s)$ is the binary entropy of the score. The terms δ_h and δ_t encode log degrees of h and t , and δ_{ht} encodes their common-neighbor support. The indicator $e_{ht} \in \{0, 1\}$ marks whether a training edge

between h and t exists, while f_r and $\text{onehot}(r)$ encode relation frequency and identity. All statistics come from $\mathcal{T}_{\text{train}}$ alone; calibration and test triples share the feature map but add no edges.

3.3 Asymmetric Residuals for Confidence-Score Nonconformity

Local KG structure can make the confidence score of a query systematically over-estimated or under-estimated, so local calibration should be directional. We therefore estimate asymmetric residual scales from context anchors. This follows the conformal-regression view that residual calibration can use side-specific and instance-dependent evidence rather than one global absolute-residual scale (Romano et al., 2019; Guan, 2023). Two design choices follow from this motivation. Residuals are split by direction so that systematic over-/under-estimation errors receive different scales, and local anchor estimates are smoothed by a global reference quantile to stay stable in sparse or noisy graph neighborhoods during scale estimation.

Directional Residuals. The first step records how the backbone errs on each calibration triple in a way that preserves the direction of the error. A single absolute residual $|M_{\mathcal{T}_{\text{train}}}(q_i) - c_i|$ collapses optimistic and pessimistic mistakes into one quantity, so any local scale derived from it cannot tell the two cases apart. The framework therefore decomposes each calibration error into two directional residuals with nonnegative values:

$$r_i^+ = \max(M_{\mathcal{T}_{\text{train}}}(q_i) - c_i, 0), \quad (3)$$

$$r_i^- = \max(c_i - M_{\mathcal{T}_{\text{train}}}(q_i), 0). \quad (4)$$

Here $r_i^+ > 0$ marks an over-estimation example whose evidence will be used to push the lower endpoint of a query downward, while $r_i^- > 0$ marks an under-estimation example that pushes the upper endpoint upward. Each calibration triple contributes to at most one side, so r^+ and r^- together cover the residual signal without double counting.

Context-Anchor Local Scales. The next step turns nearby calibration evidence into a directional scale estimate for an unseen query. Let $\mathcal{N}_K(q \mid \mathcal{T}_{\text{ref}})$ denote the K nearest neighbors of q in the standardized ϕ -space. The local lower and upper scales are the ρ -quantiles of the corresponding directional residuals over these anchors:

$$\tilde{g}^\pm(q) = \text{Quantile}_\rho\{r_i^\pm \mid i \in \mathcal{N}_K(q \mid \mathcal{T}_{\text{ref}})\}. \quad (5)$$

Using a high quantile rather than a sample mean asks how large the anchors’ directional residuals are in the conservative tail. This avoids letting a few low-residual neighbors collapse the width on a query that needs a wider interval.

Shrinkage Toward a Global Baseline. A purely local quantile is informative only when the K neighbors are themselves representative. In a sparse graph region the anchors can have small or noisy residuals, producing either an overconfident scale near zero or a single-anchor-driven spike. This makes shrinkage a safeguard against unreliable anchor evidence rather than a replacement for local calibration. Each local scale is therefore shrunk toward a global reference $g_0^\pm$ computed over all of $\mathcal{T}_{\text{ref}}$:

$$g^\pm(q) = \frac{K}{K + \kappa} \tilde{g}^\pm(q) + \frac{\kappa}{K + \kappa} g_0^\pm + \epsilon, \quad (6)$$

$$g_0^\pm = \text{Quantile}_\rho\{r_j^\pm \mid j \in \mathcal{T}_{\text{ref}}\}. \quad (7)$$

The parameter κ has the interpretation of κ virtual anchors carrying the global residual, so real evidence with weight K is averaged against a stable baseline with weight κ . The two limits $K \gg \kappa$ and $\kappa \gg K$ recover the purely local and purely global scales respectively. The resulting $g^+(q)$ and $g^-(q)$ are query-specific asymmetric scales, calibrated jointly with the conformal holdout in the next step.

3.4 Final Prediction Interval Construction

The final construction turns asymmetric local residual scales into a coverage-calibrated prediction interval. On the conformal holdout $\mathcal{T}_{\text{conf}}$, a joint max-normalized nonconformity score is computed for each calibration triple:

$$a_i = \max\left\{\frac{r_i^+}{g^+(q_i)}, \frac{r_i^-}{g^-(q_i)}\right\}. \quad (8)$$

The empirical upper α -quantile of $\{a_i \mid i \in \mathcal{T}_{\text{conf}}\}$ defines the final conformal multiplier $\hat{m}_\alpha$, which carries the joint coverage budget across both directions of the interval following the split conformal calibration principle (Papadopoulos et al., 2002; Lei et al., 2018). For a test query q_* with point score $s_* = M_{\mathcal{T}_{\text{train}}}(q_*)$, the same standardized KNN retrieval procedure on $\mathcal{T}_{\text{ref}}$ produces query-specific over-/under-estimation scales on the training graph, and these are combined with $\hat{m}_\alpha$ to form

the prediction interval:

$$L_\alpha(q_*) = s_* - \hat{m}_\alpha g^+(q_*), \quad (9)$$

$$U_\alpha(q_*) = s_* + \hat{m}_\alpha g^-(q_*), \quad (10)$$

$$\Gamma^\alpha(q_*) = [L_\alpha(q_*), U_\alpha(q_*)] \cap [0, 1]. \quad (11)$$

The final interval is clipped to $[0, 1]$.

Theoretical Guarantees. The construction above is supported by two coverage results and one conditional monotonicity result. Throughout the statements, write $Z_i = (q_i, c_i)$, $\ell = |\mathcal{T}_{\text{conf}}|$, and $k_\alpha = \lceil (\ell + 1)\alpha \rceil$. Theorem 3.1 gives the marginal coverage guarantee of the final interval. Theorem 3.1 gives the marginal guarantee, and Proposition 3.2 states sufficient conditions under which self-reuse contracts the interval. Theorem 3.3 justifies using one multiplier for the whole asymmetric interval rather than two separate side multipliers. Proofs are given in Appendices A.1–A.3.

Theorem 3.1. Coverage guarantee of CARE. *For this statement, fix the trained backbone, training graph, calibration split, and $\mathcal{T}_{\text{ref}}$. If $\mathcal{T}_{\text{conf}}$ and a future triple are exchangeable conditional on them, then for any α with $k_\alpha \leq \ell$,*

$$\mathbb{P}(c_{\ell+1} \in \Gamma^\alpha(q_{\ell+1})) \geq \alpha. \quad (12)$$

This is the standard split-conformal coverage guarantee for the final interval.

Proposition 3.2. Conditional interval containment under self-reuse. *Let $T_A(Z)$ denote the CARE non-conformity score in Eq. (8) with local scales estimated from reference set A . Set $\tilde{A} = \mathcal{T}_{\text{ref}} \cup \mathcal{T}_{\text{conf}}$, $S_i = T_{\mathcal{T}_{\text{ref}}}(Z_i)$, and $\tilde{S}_i = T_{\tilde{A}}(Z_i)$; let $\hat{m}_\alpha$ and $\tilde{m}_\alpha$ be the corresponding $k_\alpha = \lceil (\ell + 1)\alpha \rceil$ -th order statistics. If the future local scales do not increase, $g_A^\pm(q_{\ell+1}) \leq g_{\mathcal{T}_{\text{ref}}}^\pm(q_{\ell+1})$, and self-reuse contracts the holdout scores, $\tilde{S}_i \leq S_i$ for $i = 1, \dots, \ell$, then $\tilde{m}_\alpha \leq \hat{m}_\alpha$ and*

$$\Gamma_{\text{reuse}}^\alpha(q_{\ell+1}) \subseteq \Gamma_{\text{split}}^\alpha(q_{\ell+1}). \quad (13)$$

The score and scale inequalities are explicit assumptions. Under these conditions, reuse yields a contained interval. The disjoint construction in Theorem 3.1 supplies the marginal guarantee.

Theorem 3.3. Joint multiplier calibration. *Under the same exchangeability assumption as Theorem 3.1, the joint construction in Eq. (11) satisfies*

$$\mathbb{P}(c_{\ell+1} \in \Gamma^\alpha(q_{\ell+1})) \geq \alpha. \quad (14)$$

If lower and upper events are calibrated separately as two side-specific events at a common side level α' and combined by the union-bound certificate, then the same interval target requires the stricter condition

$$\alpha' \geq \frac{1 + \alpha}{2}. \quad (15)$$

Thus the joint multiplier calibrates the whole interval event directly in one step, without side splitting, instead of imposing a stricter side target.

3.5 Time Complexity Analysis

The computational overhead is restricted to the post-hoc interval layer, since the UnKGE backbone is fixed and never retrained. Per query, the layer adds feature lookup, exact L2 retrieval in $\mathcal{T}_{\text{ref}}$, and a ρ -quantile over K neighbors. Let $n_R = |\mathcal{T}_{\text{ref}}|$, $\ell = |\mathcal{T}_{\text{conf}}|$, n_T the test size, d_ϕ the feature dimension, and $C_{\text{loc}} = n_R d_\phi + K \log K$. Scales run on $n_E = \ell + n_T$ holdout and test points, one $\ell \log \ell$ sort picks the multiplier, and $|\mathcal{T}_{\text{train}}|$ is one-time graph preprocessing, yielding a total cost of

$$\mathcal{O}(|\mathcal{T}_{\text{train}}| + n_E C_{\text{loc}} + \ell \log \ell). \quad (16)$$

The cost is linear in n_T at fixed n_R, ℓ, K, d_ϕ .

4 Experiments

4.1 Experimental Setup

Datasets and Backbones. CARE is evaluated on three uncertain-KG benchmarks, CN15K, PPI5K, and NL27K. CN15K contains commonsense confidence statements from ConceptNet (Speer et al., 2017; Zhu et al., 2025c). PPI5K contains protein–protein interaction confidence scores from STRING (Szklarczyk et al., 2017; Zhu et al., 2025c). NL27K contains web-extracted relational facts from NELL (Mitchell et al., 2018; Zhu et al., 2025c). We use fixed score exports from three confidence predictors, UKGE, PASSLEAF, and BEURRE (Chen et al., 2019, 2021a,b). Dataset and backbone details are in Appendix B.1–B.2.

Baselines. The comparison includes FPI and two conformal prediction baselines. FPI is a Fisher-style calibration-score Gaussian prediction-interval baseline (FISHER, 1935). CP is standard split conformal prediction with absolute residuals (Vovk et al., 2005; Lei et al., 2018). UNKGCP normalizes conformal residuals by the binary entropy of the backbone score (Zhu et al., 2025c). Baseline details are in Appendix B.3.

Table 1: Horizontal comparison of four interval modeling methods. Baseline rows use the original conformal prediction method as the reference for sharpness changes. **Green** marks coverage at or above 0.90, **bold** marks best sharpness, and non-CP sharpness cells mark absolute changes compared with CP. **Blue downward** arrows indicate shorter intervals and **red upward** arrows indicate longer intervals.

BACKBONE	METHOD	CN15k		PPI5k		NL27k	
		Coverage	Sharpness ↓	Coverage	Sharpness ↓	Coverage	Sharpness ↓
UKGE	CP	0.90	0.845	0.90	0.188	0.90	0.443
	FPI	0.78	0.750 _{↓0.095}	0.89	0.683 _{↑0.495}	1.00	0.748 _{↑0.305}
	UNKGCP	0.90	0.833 _{↓0.012}	0.90	0.186 _{↓0.002}	0.90	0.362 _{↓0.082}
	CARE	0.90	0.690 _{↓0.155}	0.90	0.128 _{↓0.060}	0.90	0.259 _{↓0.184}
PASSLEAF	CP	0.90	0.862	0.90	0.207	0.90	0.544
	FPI	0.35	0.704 _{↓0.158}	0.89	0.684 _{↑0.478}	1.00	0.765 _{↑0.221}
	UNKGCP	0.90	0.845 _{↓0.017}	0.90	0.202 _{↓0.005}	0.90	0.440 _{↓0.104}
	CARE	0.90	0.673 _{↓0.189}	0.90	0.119 _{↓0.088}	0.90	0.231 _{↓0.313}
BEURRE	CP	0.90	0.864	0.90	0.261	0.90	0.418
	FPI	0.79	0.699 _{↓0.165}	0.89	0.701 _{↑0.440}	1.00	0.667 _{↑0.249}
	UNKGCP	0.90	0.807 _{↓0.057}	0.90	0.271 _{↑0.010}	0.90	0.384 _{↓0.035}
	CARE	0.90	0.604 _{↓0.260}	0.90	0.202 _{↓0.059}	0.90	0.263 _{↓0.155}

Implementation Details. We report coverage and sharpness, using target coverage $\alpha = 0.90$. The default CARE configuration uses 10 seeds for every backbone–dataset cell, a residual-sign-stratified 60/40 split into $\mathcal{T}_{\text{ref}}$ and $\mathcal{T}_{\text{conf}}$, exact L2 retrieval, and shared hyperparameters $K = 200$, $\rho = 0.90$, and $\kappa = 50$. All methods use the same fixed backbone scores and positive test triples for the main comparison. Other evaluation subsets and code-level details are in Appendix B.4.

4.2 Performance Comparison

Overall, CARE reduces sharpness by 29.76% on average across the nine positive cells at comparable coverage. The result is recorded in Table 1. The table leads to three observations. First, FPI shows coverage fluctuations because it builds a parametric Gaussian calibration-score interval rather than using empirical conformal ranks. This assumption is brittle when score-label errors vary across graph contexts. Second, CP, UNKGCP, and CARE follow the same split-calibration principle, so the conformal methods maintain stable coverage and differ in interval sharpness once the coverage target is matched. Third, CARE gives the sharpest intervals in all nine cells, and the sharpness reduction is most pronounced on NL27k. This gain comes from context-anchor retrieval and lower/upper residual scales, which replace one entropy-based symmetric scale with query-specific asymmetric evidence. At

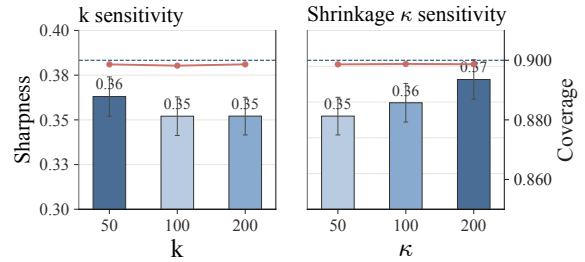


Figure 2: Sensitivity analyses for local calibration under the shared protocol. The two plots vary neighborhood size K and shrinkage strength κ .

the cell level, the advantage on NL27k suggests that anchor retrieval is especially helpful when relational evidence is heterogeneous and score entropy alone cannot separate local residual regimes. Coverage–sharpness analysis is in Appendix D.1.

4.3 Sensitivity Analyses

To test whether the gain depends on a narrow local-calibration setting, we vary the neighborhood size K and shrinkage strength κ under the same protocol, with results shown in Figure 2. These settings change the retrieved context and the amount of smoothing. From the results, we make two observations. ❶ As K decreases, the calibration neighborhood becomes narrower, but sharpness does not decrease and can even rise slightly. This suggests that simply shrinking the calibration domain is not always better, since too few anchors make the esti-

Table 2: Ablation studies of different variants, averaged over UKGE, PASSLEAF, and BEURRE backbones.

VARIANT	CN15k		PPI5k		NL27k		MEAN
	Coverage	Sharpness ↓	Coverage	Sharpness ↓	Coverage	Sharpness ↓	Sharpness inc.
CARE w/o AS	0.90	0.795 $\uparrow$ 0.139	0.90	0.168 $\uparrow$ 0.019	0.90	0.361 $\uparrow$ 0.110	25.4%
CARE w/o LS	0.90	0.745 $\uparrow$ 0.089	0.90	0.212 $\uparrow$ 0.062	0.90	0.317 $\uparrow$ 0.066	20.6%
CARE w/ RC	0.90	0.760 $\uparrow$ 0.104	0.90	0.430 $\uparrow$ 0.280	0.90	0.323 $\uparrow$ 0.073	43.3%
CARE w/ LT	0.90	0.718 $\uparrow$ 0.062	0.90	0.167 $\uparrow$ 0.017	0.90	0.264 $\uparrow$ 0.013	8.7%
CARE w/ PU	0.90	0.714 $\uparrow$ 0.058	0.90	0.181 $\uparrow$ 0.031	0.90	0.227 $\downarrow$ 0.024	6.2%
CARE (Full Model)	0.90	0.656	0.90	0.150	0.90	0.251	—

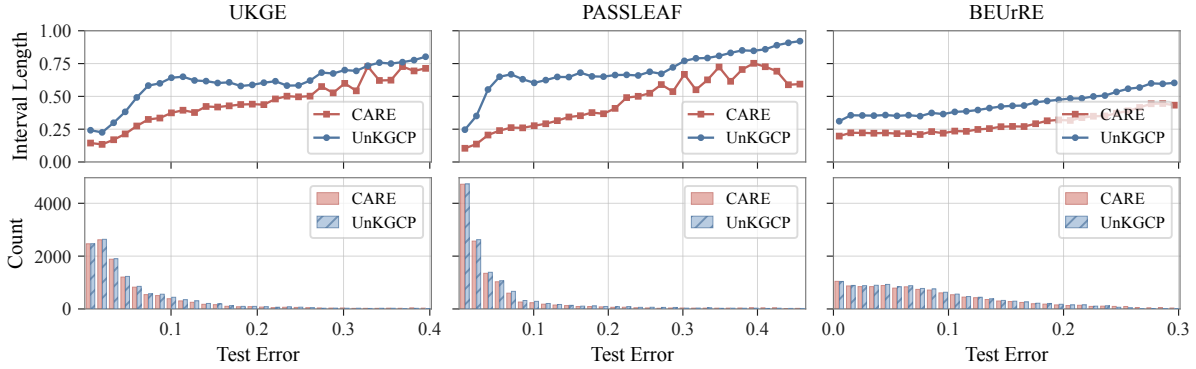


Figure 3: Difficulty-tracking analysis on NL27K test triples. Test examples are grouped by absolute prediction error to probe difficulty-aware interval allocation. The line plot reports mean interval length over covered examples in each bin, and the histogram reports the test-error distribution.

mate more sensitive to local graph noise. ② As κ increases, local calibration is gradually smoothed toward global calibration, and coverage rises accordingly. Together, the trends show that CARE is stable near the default regime under this shared experimental protocol. This robustness matters because the same setting is used across datasets and backbones. The result supports the role of neighborhood calibration in improving interval precision, matching the theoretical analysis in Section 3. Details are in Appendix C.

4.4 Ablation Studies

To evaluate the effectiveness of different components in CARE, we introduce five parallel variants in Table 2. ① CARE w/o AS pools the lower and upper residual sides into one absolute-residual scale. ② CARE w/o LS replaces the query-specific KNN residual scale with the global ρ -quantile on $\mathcal{T}_{\text{ref}}$. ③ CARE w/ Relation Context uses only relation frequency and relation identity. ④ CARE w/ Local Topology uses only local graph-support features. ⑤ CARE w/ Predictive Uncertainty uses only the score and score entropy. The results show that both context-anchor calibration and bidirectional

residual calibration improve interval precision. The three feature groups carry different amounts of information: predictive uncertainty is slightly more informative than local topology, and both are much more informative than relation context. Full ablation details are in Appendix C.

4.5 Further Analysis

Difficulty Tracking Analysis. Aggregate coverage and sharpness do not show whether an interval method reacts to query difficulty. We therefore group NL27K test triples by absolute prediction error and compare mean interval length within each bin in Figure 3. The figure leads to two observations. First, both UnKGCP and CARE assign longer intervals to harder triples, showing that conformal prediction methods can effectively identify more difficult examples and assign them wider prediction intervals. Second, CARE stays below UnKGCP across the displayed backbone plots, indicating that context anchors bring stronger interval modeling ability while preserving this difficulty-aware trend. This micro-level view complements Table 1 by checking whether shorter intervals come from useful allocation rather than uniform shrinkage.

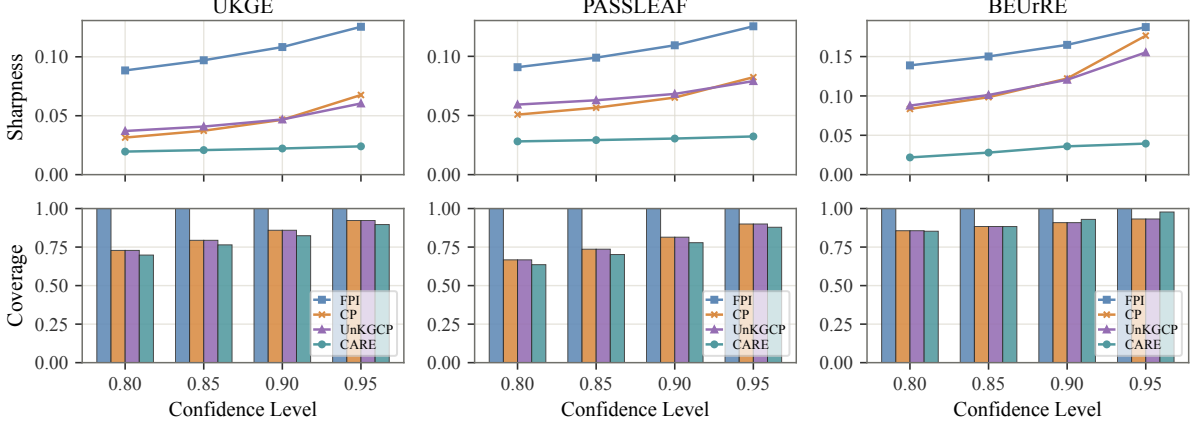


Figure 4: Distribution-Shift Scope Diagnostic. The top row reports sharpness curves over target coverage levels on negative subsets, and the bottom row reports empirical coverage.

Table 3: Conditional comparison on positive triples across the evaluation matrix. Each pair is coverage/sharpness; gap is deviation from 0.90.

Method	Overall cov./sharp.	Stable-rel. cov./sharp.	Gap
GLOBAL CP	0.9002/0.5144	0.8967/0.5092	0.0685
RELATION-MONDRIAN CP	0.9028/0.5052	0.9046/0.4914	0.0221
CARE	0.8990/0.3521	0.8937/0.3381	0.0405

Table 4: Directional residuals on positive triples. Frequency is the proportion in each direction; mean $q_{0.90}$ is the average directional residual quantile.

DIRECTION	Frequency	Mean $q_{0.90}$
OVER-ESTIMATION (r^+)	0.2402	0.0495
UNDER-ESTIMATION (r^-)	0.7597	0.3474

Distribution-Shift Scope Diagnostic. The main experiments use positive triples from the original datasets, while this diagnostic uses artificially constructed negatives to examine model behavior under distribution-shifted data. We use Figure 4 to probe this shifted setting. The figure gives two observations. First, negative triples make empirical coverage less stable for conformal methods, confirming that the exchangeability condition behind split conformal calibration is sensitive to the data source. Because the negative subsets alter entity support, the result is best read as a scope diagnostic for shifted inputs. Second, CARE keeps the shortest intervals in the sharpness plots across confidence levels, showing strong interval-modeling ability under distribution shift. Implementation details for both diagnostics are in Appendix B.4.

4.6 Additional Validation

Conditional comparison. In this task, relation identity is the predicate type of a query triple and provides a natural conditional grouping because different relations encode distinct semantics and residual patterns in uncertain KG reasoning. We therefore compare CARE with Relation-Mondrian CP, which calibrates absolute residuals separately for each relation; Table 3 reports overall and stable-relation summaries. Relation-Mondrian CP aligns stable-relation coverage more closely, whereas CARE yields sharper intervals overall and within stable relations. ❶ This coverage pattern indicates that relation-wise calibration captures residual structure shared at the predicate level. ❷ The sharpness advantage of CARE shows that relation identity remains too coarse to explain variation among triples of the same relation, supporting query-specific anchors as a finer basis for interval allocation.

Residual asymmetry. Over- and under-estimation affect opposite interval endpoints, so their frequency and tail scale indicate whether one absolute-residual distribution can represent both error directions. We compare these two statistics across positive runs, with Table 4 reporting direction frequencies and mean directional $q_{0.90}$. Under-estimation is both more frequent and broader in its tail scale. ❶ This imbalance indicates that the backbone errors follow a persistent directional pattern rather than random sign variation. ❷ Because the two directions require different corrections, separate scales preserve information that a pooled absolute-residual scale would erase, supporting the use of asymmetric calibration for endpoint-specific interval allocation.

5 Related Work

5.1 Knowledge Graph Embeddings

Knowledge graph embedding (KGE) methods map symbolic triples into continuous representations and score relational facts. Early deterministic link-prediction models include TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), and ConvE (Dettmers et al., 2018). Later graph neural variants include R-GCN (Schlichtkrull et al., 2018) and CompGCN (Vashishth et al., 2020). For uncertain KGs, UKGE (Chen et al., 2019), BEURRE (Chen et al., 2021a), and PASSLEAF (Chen et al., 2021b) predict continuous confidence scores, with related geometric representations in BoxE (Abboud et al., 2020). These models provide the fixed backbone score $M_{\mathcal{T}_{\text{train}}}(q)$ used by CARE, while calibration studies ask whether such scores behave like probabilities (Tabacof and Costabello, 2020; Safavi et al., 2020). CARE instead adds a post-hoc interval layer that makes residual uncertainty explicit without retraining the UnKGE backbone.

5.2 Uncertainty Quantification

Uncertainty quantification attaches reliability information to predictions, and split conformal prediction provides finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Shafer and Vovk, 2007; Lei et al., 2018; Angelopoulos and Bates, 2023). Adaptive variants include conformalized quantile regression and localized conformal prediction (Romano et al., 2019; Guan, 2023), with extensions to classification, distribution shift, and weaker exchangeability (Romano et al., 2020; Gibbs and Candès, 2021; Barber et al., 2023), alongside known limits for distribution-free conditional coverage (Barber et al., 2021). For structured data, conformal methods have been applied to GNNs and graph predictors (H. Zargarbashi et al., 2023; Zargarbashi and Bojchevski, 2024; Huang et al., 2023), and KG-specific work forms answer sets for link prediction (Zhu et al., 2025a,b). UN-KGCP (Zhu et al., 2025c) is closest because it constructs intervals for continuous uncertain-KG scores. CARE keeps the same holdout calibration route, but estimates lower and upper residual scales at the query level from graph-local anchors instead of relying on a single scalar.

6 Conclusion

In this paper, we propose CARE, a context-anchored asymmetric conformal prediction framework for uncertainty-aware reasoning over KGs without retraining the backbones. The framework builds on two findings. Context-anchored neighborhood calibration improves interval precision, and separate lower/upper calibration corrects directional output errors of scoring models. Across commonsense, biomedical, and web-extracted KG benchmarks, experiments show that CARE effectively uses context anchors, turns structural evidence in an UnKG into an uncertainty signal, and reaches state-of-the-art interval modeling for the studied task while preserving the same coverage target.

Limitations

Although CARE shows strong performance on the studied uncertain-KG benchmarks, the current evaluation is still limited to three datasets and fixed score exports from three UnKGE backbones. Future work should test the framework on broader KG domains, more data sources, and additional confidence-score predictors.

The guarantee of CARE is marginal split-conformal coverage under exchangeability. Per-relation and distribution-shift coverage require additional assumptions or calibration protocols. The final label-independent random split aligns the implementation with the stated theorem for the evaluated source distribution. On artificially constructed negatives, coverage can fall under shift, especially on PPI5k; we therefore reserve efficiency comparisons for settings with matched coverage.

Ethics Statement

This work studies uncertainty intervals for KG confidence scores, including settings related to biomedical prediction. Such intervals should be treated as evidence for expert inspection rather than as substitutes for domain validation. We take responsibility for the authenticity and reproducibility of the reported results.

Acknowledgments

The work of Ziyue Qiao is supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515140238, No. 2024A1515140114) and the National Natural Science Foundation of China (No. 62676075, No. 62406056).

References

- Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. BoxE: A box embedding model for knowledge base completion. *NeurIPS Proceedings*, 33.
- Anastasios N. Angelopoulos and Stephen Bates. 2023. [Conformal Prediction: A Gentle Introduction](#). *Foundations and Trends in Machine Learning*, 16(4):494–591.
- Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. 2021. [The limits of distribution-free conditional predictive inference](#). *Information and Inference: A Journal of the IMA*, 10(2):455–482.
- Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. [Conformal prediction beyond exchangeability](#). *The Annals of Statistics*, 51(2):816–845.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. [Translating Embeddings for Modeling Multi-relational Data](#). In *Neural Information Processing Systems (NIPS)*, pages 2787–2795, South Lake Tahoe, United States.
- X. Chen, M. Chen, W. Shi, Y. Sun, and C. Zaniolo. 2019. [Embedding Uncertain Knowledge Graphs](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3363–3370.
- Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew McCallum. 2021a. [Probabilistic Box Embeddings for Uncertain Knowledge Graph Reasoning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 882–893, Online. Association for Computational Linguistics.
- Z.-M. Chen, M.-Y. Yeh, and T.-W. Kuo. 2021b. [PASSLEAF: A Pool-bAsed Semi-Supervised LEArning Framework for Uncertain Knowledge Graph Embedding](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4019–4026.
- Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. [Convolutional 2D Knowledge Graph Embeddings](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):1811–1818.
- R. A. FISHER. 1935. [THE FIDUCIAL ARGUMENT IN STATISTICAL INFERENCE](#). *Annals of Eugenics*, 6(4):391–398.
- Isaac Gibbs and Emmanuel Candès. 2021. [Adaptive Conformal Inference Under Distribution Shift](#). *Preprint*, arXiv:2106.00170.
- Leying Guan. 2023. [Localized conformal prediction: a generalized inference framework for conformal prediction](#). *Biometrika*, 110(1):33–50.
- Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. 2023. [Conformal Prediction Sets for Graph Neural Networks](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12292–12318. PMLR.
- Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. [Systematic integration of biomedical knowledge prioritizes drugs for repurposing](#). *eLife*, 6:e26726.
- Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D’amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. [Knowledge Graphs](#). *ACM Comput. Surv.*, 54(4).
- Kexin Huang, Ying Jin, Emmanuel Candès, and Jure Leskovec. 2023. [Uncertainty Quantification over Graph with Conformalized Graph Neural Networks](#). *Preprint*, arXiv:2305.14535.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. [A Survey on Knowledge Graphs: Representation, Acquisition, and Applications](#). *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.
- J. Johnson, M. Douze, and H. Jégou. 2021. [Billion-Scale Similarity Search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Roger Koenker and Gilbert Bassett. 1978. [Regression Quantiles](#). *Econometrica*, 46(1):33–50.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. [Distribution-Free Predictive Inference for Regression](#). *Journal of the American Statistical Association*, 113(523):1094–1111.
- Junyu Luo, Jinsheng Huang, Yang Xu, Lutong Zou, Xiao Luo, Bohan Wu, Yifan Wang, Wei Ju, and Ming Zhang. 2026. Conformalized hierarchical calibration for uncertainty-aware adaptive hashing. In *International Conference on Learning Representations*, volume 2026, pages 105223–105246.
- T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, and 7 others. 2018. [Never-ending learning](#). *Commun. ACM*, 61(5):103–115.

- Sameh K Mohamed, Vít Nováček, and Aayah Nounu. 2020. [Discovering protein drug targets using knowledge graph embeddings](#). *Bioinformatics*, 36(2):603–610.
- David N. Nicholson and Casey S. Greene. 2020. [Constructing knowledge graphs and their biomedical applications](#). *Computational and Structural Biotechnology Journal*, 18:1414–1428.
- M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. 2016. [A Review of Relational Machine Learning for Knowledge Graphs](#). *Proceedings of the IEEE*, 104(1):11–33.
- Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammernan. 2002. Inductive Confidence Machines for Regression. In *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge Graph Identification. In *The Semantic Web – ISWC 2013*, pages 542–557, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. 2019. [Conformalized Quantile Regression](#). *Preprint*, arXiv:1905.03222.
- Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. 2020. [Classification with Valid and Adaptive Coverage](#). *Preprint*, arXiv:2006.02544.
- Tara Safavi, Danai Koutra, and Edgar Meij. 2020. [Evaluating the Calibration of Knowledge Graph Embeddings for Trustworthy Link Prediction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8308–8321, Online. Association for Computational Linguistics.
- Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*, pages 593–607, Cham. Springer International Publishing.
- Glenn Shafer and Vladimir Vovk. 2007. [A tutorial on conformal prediction](#). *Preprint*, arXiv:0706.3188.
- R. Speer, J. Chin, and C. Havasi. 2017. [ConceptNet 5.5: An Open Multilingual Graph of General Knowledge](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. [RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space](#). In *International Conference on Learning Representations*.
- Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. 2017. [The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible](#). *Nucleic Acids Research*, 45(D1):D362–D368.
- Pedro Tabacof and Luca Costabello. 2020. [Probability Calibration for Knowledge Graph Embedding Models](#). In *International Conference on Learning Representations*.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. [Complex Embeddings for Simple Link Prediction](#). *Preprint*, arXiv:1606.06357.
- Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. [Composition-based Multi-Relational Graph Convolutional Networks](#). In *International Conference on Learning Representations*.
- Vladimir Vovk, Alexander Gammernan, and Glenn Shafer. 2005. [Algorithmic Learning in a Random World](#). Springer New York, NY.
- Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. [KGAT: Knowledge Graph Attention Network for Recommendation](#). In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, pages 950–958, New York, NY, USA. Association for Computing Machinery.
- Zhiping Xiao, Yusheng Zhao, Qixin Zhang, Jiaye Xie, Wanjia Zhao, Weizhi Zhang, Xiao Luo, Philip Yu, and Ming Zhang. 2026. Sample lottery: Unsupervised discovery of critical instances for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pages 8919–8949.
- Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. [Embedding Entities and Relations for Learning and Inference in Knowledge Bases](#). *Preprint*, arXiv:1412.6575.
- Soroush H. Zargarbashi and Aleksandar Bojchevski. 2024. [Conformal Inductive Graph Neural Networks](#). In *The Twelfth International Conference on Learning Representations*.
- Yuqicheng Zhu, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Evgeny Kharlamov, and Steffen Staab. 2025a. [Predicate-Conditional Conformalized Answer Sets for Knowledge Graph Embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4145–4167, Vienna, Austria. Association for Computational Linguistics.
- Yuqicheng Zhu, Nico Potyka, Jiarong Pan, Bo Xiong, Yunjie He, Evgeny Kharlamov, and Steffen Staab. 2025b. [Conformalized Answer Set Prediction for Knowledge Graph Embedding](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 731–750, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuqicheng Zhu, Jingcheng Wu, Yizhen Wang, Hongkuan Zhou, Jiaoyan Chen, Evgeny Kharlamov, and Steffen Staab. 2025c. [Certainty in Uncertainty: Reasoning over Uncertain Knowledge Graphs with Statistical Guarantees](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8730–8752, Suzhou, China. Association for Computational Linguistics.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. [Modeling polypharmacy side effects with graph convolutional networks](#). *Bioinformatics*, 34(13):i457–i466.

A Theoretical Analysis

This appendix expands the two coverage theorems and the conditional self-reuse proposition used in Sec. 3. Appendix A.1 proves the marginal coverage guarantee of the final interval. Appendix A.2 proves the interval-containment conclusion under the explicit score- and scale-contraction assumptions.

Appendix A.3 explains why a single joint multiplier is preferable to two separately calibrated side multipliers for the asymmetric interval.

A.1 Expected Coverage

Proof of Theorem 3.1. All probabilities are conditional on the fixed objects in the validity statement in the main text. After conditioning on $\mathcal{T}_{\text{ref}}$, the trained backbone, and the realized split, the only random objects left are the conformal holdout triples and the future test triple. At this point the nearest-neighbor search, local residual quantiles, and shrinkage rule have already produced fixed positive scale functions g^+ and g^- . Hence the score map used below is fixed before the holdout and test labels are ranked. It is not re-estimated after observing which point is treated as the future test triple. Applying the same nonconformity-score map in Eq. 8 to each exchangeable holdout or test triple preserves exchangeability, so

$$A_i = A(q_i, c_i \mid \mathcal{T}_{\text{ref}}), \quad i = 1, \dots, \ell + 1$$

are exchangeable scores (Papadopoulos et al., 2002; Vovk et al., 2005). For the observed holdout triples, the scores used to compute the conformal multiplier are $a_i = A_i$ for $i = 1, \dots, \ell$. Equivalently, for every permutation π of $\{1, \dots, \ell + 1\}$,

$$(A_1, \dots, A_{\ell+1}) \stackrel{d}{=} (A_{\pi(1)}, \dots, A_{\pi(\ell+1)}).$$

Let $a_{(1)} \leq \dots \leq a_{(\ell)}$ be the sorted holdout scores and let $\hat{m}_\alpha = a_{(k_\alpha)}$ be the conformal multiplier. Let $R_{\ell+1}$ be the rank of $A_{\ell+1}$ among the pooled scores $A_1, \dots, A_{\ell+1}$, with smaller scores ranked earlier and ties broken uniformly at random. Exchangeability then makes

$$\mathbb{P}(R_{\ell+1} = r) = \frac{1}{\ell + 1}, \quad r = 1, \dots, \ell + 1.$$

The implemented event uses the non-strict comparison $A_{\ell+1} \leq \hat{m}_\alpha$, so it contains the randomized-rank event $R_{\ell+1} \leq k_\alpha$ and is conservative at ties.

Therefore the standard split-conformal rank argument gives

$$\begin{aligned} \mathbb{P}(A_{\ell+1} \leq \hat{m}_\alpha) &\geq \mathbb{P}(R_{\ell+1} \leq k_\alpha) \\ &= \frac{k_\alpha}{\ell + 1} \geq \alpha. \end{aligned} \quad (17)$$

By construction of the score in Eq. 8, the event $A_{\ell+1} \leq \hat{m}_\alpha$ means that the multiplier $\hat{m}_\alpha$ is large enough to cover the test label on both the lower and upper sides of the asymmetric interval in Eq. 11. Since confidence labels lie in $[0, 1]$, clipping the interval to $[0, 1]$ cannot remove a covered label. Therefore

$$\{A_{\ell+1} \leq \hat{m}_\alpha\} \subseteq \{c_{\ell+1} \in \Gamma^\alpha(q_{\ell+1})\}.$$

Combining the rank bound with this containment proves

$$\mathbb{P}(c_{\ell+1} \in \Gamma^\alpha(q_{\ell+1})) \geq \alpha. \quad (18)$$

□

A.2 Split Calibration

Proof of Proposition 3.2. The split between $\mathcal{T}_{\text{ref}}$ and $\mathcal{T}_{\text{conf}}$ is not only an implementation detail. The reference pool is used to estimate local scales, whereas the conformal holdout is used to rank nonconformity scores. If the same triple is used in both roles, it can become its own nearest neighbor during anchor retrieval, which may shrink its residual scale and make the conformal multiplier too optimistic. Proposition 3.2 states this finite-sample behavior as a conditional monotonic comparison.

For a reference set A , write $T_A(Z)$ for the CARE nonconformity score obtained when the local scale is estimated from A :

$$T_A(q, c) = \max \left\{ \frac{\max(M_{\mathcal{T}_{\text{train}}}(q) - c, 0)}{g_A^+(q)}, \frac{\max(c - M_{\mathcal{T}_{\text{train}}}(q), 0)}{g_A^-(q)} \right\}. \quad (19)$$

In the split construction, holdout scores are computed by the fixed map $T_{\mathcal{T}_{\text{ref}}}$. In the reused construction, each calibration point $Z_i \in \mathcal{T}_{\text{conf}}$ is also available to the KNN scale estimator because $\tilde{A} = \mathcal{T}_{\text{ref}} \cup \mathcal{T}_{\text{conf}}$. Thus Z_i can be a zero-distance anchor for itself:

$$\begin{aligned} g_A^\pm(q_i) &= G^\pm \left(\{r_j^\pm : q_j \in N_{\tilde{A}}(q_i)\} \right), \\ q_i &\in N_{\tilde{A}}(q_i), \end{aligned}$$

where $G^\pm(\cdot)$ denotes the local quantile followed by shrinkage. The statement isolates the resulting optimism by assuming the coordinatewise contraction $\tilde{S}_i \leq S_i$ for the calibration scores. Order statistics are monotone under coordinatewise domination, so

$$\tilde{m}_\alpha = \tilde{S}_{(k_\alpha)} \leq S_{(k_\alpha)} = \hat{m}_\alpha. \quad (20)$$

For a test query, the lower and upper half-widths are the multiplier times $g^+(\cdot)$ and $g^-(\cdot)$. If the reused test scales are no larger than the split test scales on both sides, the inequality $\tilde{m}_\alpha \leq \hat{m}_\alpha$ gives

$$\begin{aligned} \tilde{m}_\alpha g_A^+(q_{\ell+1}) &\leq \hat{m}_\alpha g_{\mathcal{T}_{\text{ref}}}^+(q_{\ell+1}), \\ \tilde{m}_\alpha g_A^-(q_{\ell+1}) &\leq \hat{m}_\alpha g_{\mathcal{T}_{\text{ref}}}^-(q_{\ell+1}). \end{aligned}$$

Both intervals are centered at the same fixed backbone score. Therefore the reused interval is contained in the split interval:

$$\Gamma_{\text{reuse}}^\alpha(q_{\ell+1}) \subseteq \Gamma_{\text{split}}^\alpha(q_{\ell+1}). \quad (21)$$

The corresponding coverage event for the reused interval is also contained in the split coverage event. This proves the subset relation in Proposition 3.2. It remains to state why the split interval has the conformal guarantee. Because $\mathcal{T}_{\text{ref}}$ is disjoint from $\mathcal{T}_{\text{conf}}$, no holdout triple contributes to its own scale estimate. Conditioning on $\mathcal{T}_{\text{ref}}$ fixes the map $T_{\mathcal{T}_{\text{ref}}}$ for every holdout and future triple. If $Z_1, \dots, Z_{\ell+1}$ are exchangeable conditional on $\mathcal{T}_{\text{ref}}$, then the scores $S_i = T_{\mathcal{T}_{\text{ref}}}(Z_i)$ obey

$$(S_1, \dots, S_{\ell+1}) \stackrel{d}{=} (S_{\pi(1)}, \dots, S_{\pi(\ell+1)})$$

for every permutation π . The rank argument gives

$$\mathbb{P}(S_{\ell+1} \leq \hat{m}_\alpha \mid \mathcal{T}_{\text{ref}}) \geq \frac{k_\alpha}{\ell+1} \geq \alpha. \quad (22)$$

The event $S_{\ell+1} \leq \hat{m}_\alpha$ is exactly the event that the split interval covers $c_{\ell+1}$, after clipping to $[0, 1]$. Thus

$$\mathbb{P}(c_{\ell+1} \in \Gamma_{\text{split}}^\alpha(q_{\ell+1}) \mid \mathcal{T}_{\text{ref}}) \geq \alpha. \quad (23)$$

□

A.3 Joint Multiplier Calibration

Proof of Theorem 3.3. The interval in Eq. 11 is asymmetric, but the conformal event is still one event: the label lies inside the whole interval. This is why CARE calibrates the max-normalized score with one multiplier rather than assigning separate

conformal budgets to the lower and upper sides. Although $g^+(q)$ and $g^-(q)$ can differ, both sides are evaluated on the same holdout triples and together determine whether the interval covers the label. A single score A_i therefore ranks interval-level failures directly, while two side scores would rank lower-side and upper-side failures as separate marginal events. Theorem 3.3 formalizes this point by comparing the joint event with a two-side alternative certified through a union bound.

The joint-multiplier coverage follows directly from the expected-coverage argument in Appendix A.1. Let $A_i = A(q_i, c_i \mid \mathcal{T}_{\text{ref}})$ for $i = 1, \dots, \ell+1$, with $a_i = A_i$ on the observed holdout triples. Under exchangeability of $\{A_i\}_{i=1}^{\ell+1}$ conditional on $\mathcal{T}_{\text{ref}}$ and the fixed backbone, the rank argument gives

$$\mathbb{P}(A_{\ell+1} \leq \hat{m}_\alpha) \geq \alpha. \quad (24)$$

By definition,

$$A_i = \max \left\{ \frac{r_i^+}{g^+(q_i)}, \frac{r_i^-}{g^-(q_i)} \right\}.$$

Thus the event $\{A_{\ell+1} \leq \hat{m}_\alpha\}$ is equivalent to

$$\begin{aligned} \{A_{\ell+1} \leq \hat{m}_\alpha\} &= \{r_{\ell+1}^+ \leq \hat{m}_\alpha g^+(q_{\ell+1})\} \\ &\quad \cap \{r_{\ell+1}^- \leq \hat{m}_\alpha g^-(q_{\ell+1})\}, \end{aligned} \quad (25)$$

which by Eq. 11 is exactly $\{c_{\ell+1} \in \Gamma^\alpha(q_{\ell+1})\}$ after clipping to $[0, 1]$.

For the two-side alternative, let $B_i = r_i^+ / g^+(q_i)$ and $C_i = r_i^- / g^-(q_i)$. Separate split-conformal calibration on $\mathcal{T}_{\text{conf}}$ produces

$$\begin{aligned} \hat{m}_{\alpha'}^+ &= B_{(\lceil(\ell+1)\alpha'\rceil)}, \\ \hat{m}_{\alpha'}^- &= C_{(\lceil(\ell+1)\alpha'\rceil)}, \end{aligned} \quad (26)$$

each guaranteeing $\mathbb{P}(B_{\ell+1} \leq \hat{m}_{\alpha'}^+) \geq \alpha'$ and $\mathbb{P}(C_{\ell+1} \leq \hat{m}_{\alpha'}^-) \geq \alpha'$ by the same rank argument. The two-side interval covers $c_{\ell+1}$ when both events hold. By the union bound,

$$\begin{aligned} \mathbb{P}(\{B_{\ell+1} \leq \hat{m}_{\alpha'}^+\} \cap \{C_{\ell+1} \leq \hat{m}_{\alpha'}^-\}) \\ \geq 2\alpha' - 1. \end{aligned} \quad (27)$$

Requiring $2\alpha' - 1 \geq \alpha$ yields $\alpha' \geq (1 + \alpha)/2$. Thus the two-side certificate must use a stricter side-specific target, whereas the joint multiplier calibrates the interval event directly at level α . □

B Details of Experimental Settings

B.1 Datasets

The three benchmark datasets differ in domain, relation vocabulary size, and confidence-score semantics.

- **CN15K.** CN15K is a commonsense uncertain-KG benchmark derived from ConceptNet (Speer et al., 2017; Zhu et al., 2025c). It contains 15,000 entities, 36 relations, and 241,158 positive triples in total, split into 204,984 training triples, 16,881 validation triples, and 19,293 test triples. Its confidence labels reflect subjective plausibility of commonsense statements.
- **PPI5K.** PPI5K is a protein–protein interaction benchmark derived from STRING (Szkarczyk et al., 2017; Zhu et al., 2025c). It contains 4,999 entities, 7 relations, and 271,666 positive triples in total, split into 230,929 training triples, 19,017 validation triples, and 21,720 test triples. Its confidence labels correspond to evidence-derived interaction probabilities.
- **NL27K.** NL27K is an automatically constructed web-knowledge benchmark derived from NELL (Mitchell et al., 2018; Zhu et al., 2025c). It contains 27,221 entities, 417 relations, and 175,412 positive triples in total, split into 149,100 training triples, 12,278 validation triples, and 14,034 test triples. Its confidence labels reflect the extraction system’s confidence in each statement.

The benchmark aliases and source KGs are introduced in Sec. 4. We follow the same benchmark protocol (Zhu et al., 2025c), and Table A.1 gives the split sizes used by the score exports (Speer et al., 2017; Mitchell et al., 2018; Szkarczyk et al., 2017). The original split, inherited from the UKGE dataset protocol, assigns 85% of triples to backbone training, 7% to validation/calibration, and 8% to testing (Chen et al., 2019; Zhu et al., 2025c). The positive evaluation subset uses the original validation and test triples. The negative calibration file is constructed from validation positives by replacing the head or tail entity under the same relation, assigning the synthetic triple a zero confidence label, and scoring it with the frozen backbone. The negative test file contains one zero-label synthetic triple for

each positive test triple; each synthetic triple is sampled from the entity and relation vocabularies and rejected if it appears in the train, validation, or test positives. The combined subset concatenates the corresponding positive and negative files for both calibration and testing. Positive triples are the main benchmark task. Negative and combined subsets are reported as evaluation-subset diagnostics.

B.2 Backbones

The interval layer is evaluated on the fixed score exports named in the main setup and summarized in Table A.2. CARE does not update entity embeddings, relation embeddings, box parameters, or scoring parameters. This keeps the comparison focused on how a fixed point score is converted into a calibrated interval.

UKGE. UKGE extends knowledge graph embeddings to uncertain triples by regressing continuous confidence labels (Chen et al., 2019). In the DistMult-style variant used here, the raw triple interaction is the trilinear product

$$z(q) = \langle \mathbf{e}_h, \mathbf{e}_r, \mathbf{e}_t \rangle = \sum_{j=1}^d e_{h,j} e_{r,j} e_{t,j}.$$

UKGE_{logi} then maps this interaction to a confidence score through

$$M_{\text{UKGE}}(q) = \sigma(wz(q) + b),$$

where w and b are learned scalar mapping parameters. With positive weighted triples $\mathcal{T}^+$ and negative triples $\mathcal{T}^-$, the regression objective is written schematically as

$$\begin{aligned} \mathcal{L}_{\text{UKGE}} &= \mathcal{L}_{\text{UKGE}}^+ + \lambda \mathcal{L}_{\text{UKGE}}^-, \\ \mathcal{L}_{\text{UKGE}}^+ &= \sum_{(q,c) \in \mathcal{T}^+} (M_{\mathcal{T}_{\text{train}}}(q) - c)^2, \\ \mathcal{L}_{\text{UKGE}}^- &= \sum_{q \in \mathcal{T}^-} M_{\mathcal{T}_{\text{train}}}(q)^2. \end{aligned}$$

PASSLEAF. PASSLEAF keeps the UKGE confidence-prediction task but adds semi-supervised pseudo-labeled triples to improve the use of uncertain evidence (Chen et al., 2021b). Its loss contains supervised, negative, and

Table A.1: Dataset statistics and split sizes used by the paper-protocol score exports.

DATASET	Entities	Relations	Train	Validation	Test	Main subset
CN15K	15000	36	204984	16881	19293	positive
PPI5K	4999	7	230929	19017	21720	positive
NL27K	27221	417	149100	12278	14034	positive

semi-supervised terms:

$$\begin{aligned}\mathcal{L}_{\text{PASS}} &= \mathcal{L}_{\text{pos}} + \frac{\mathcal{L}_{\text{neg}} + \mathcal{L}_{\text{semi}}}{|\mathcal{T}^-| + |\mathcal{T}^{\text{semi}}|}, \\ \mathcal{L}_{\text{pos}} &= \sum_{(q,c) \in \mathcal{T}} (M_{\mathcal{T}_{\text{train}}}(q) - c)^2, \\ \mathcal{L}_{\text{neg}} &= \sum_{q \in \mathcal{T}^-} M_{\mathcal{T}_{\text{train}}}(q)^2, \\ \mathcal{L}_{\text{semi}} &= \sum_{(q,\tilde{c}) \in \mathcal{T}^{\text{semi}}} (M_{\mathcal{T}_{\text{train}}}(q) - \tilde{c})^2.\end{aligned}$$

BEURRE. BEURRE represents entities and relations with probabilistic boxes and predicts confidence from expected box-volume overlap (Chen et al., 2021a). At a high level, the score of a triple $q = \langle h, r, t \rangle$ is an approximate conditional probability of a transformed tail box being compatible with the head box:

$$\begin{aligned}V_q &= \text{Vol}(\text{Box}_h \cap H_r(\text{Box}_t)), \\ M_{\text{BEURRE}}(q) &\approx \mathbb{E} \left[\frac{V_q}{\text{Vol}(\text{Box}_h)} \right],\end{aligned}$$

with relation-specific transformations $H_r(\cdot)$. The model is trained to regress the same continuous confidence labels as the other backbones.

B.3 Baselines

Let the generic calibration subset used by a baseline be $\mathcal{C} = \{(q_i, c_i)\}_{i=1}^n \subseteq \mathcal{T}_{\text{cal}}$, with exported backbone scores $s_i = M_{\mathcal{T}_{\text{train}}}(q_i)$. For a test query q , write $s = M_{\mathcal{T}_{\text{train}}}(q)$. All interval baselines are evaluated on the same exported backbone scores as CARE. The main comparison focuses on post-hoc interval predictors that reuse fixed backbone scores. QR trains an additional quantile model (Koenker and Bassett, 1978) and is outside this fixed-score comparison. All interval endpoints are clipped to $[0, 1]$ before reporting coverage and sharpness.

FPI. Fisher Prediction Intervals are a classical parametric baseline that assumes the calibration backbone scores are approximately Gaussian (Fisher, 1935). The calibration score mean and standard

deviation are

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i, \quad \hat{\sigma}_s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (s_i - \bar{s})^2}.$$

The implemented FPI baseline uses the same fixed interval for every test query in the evaluation subset:

$$\begin{aligned}L_{\text{FPI}}^\alpha &= \bar{s} - t_{(1+\alpha)/2, n-2} \hat{\sigma}_s \sqrt{\frac{n}{n-1}}, \\ U_{\text{FPI}}^\alpha &= \bar{s} + t_{(1+\alpha)/2, n-2} \hat{\sigma}_s \sqrt{\frac{n}{n-1}}, \\ \Gamma_{\text{FPI}}^\alpha &= [L_{\text{FPI}}^\alpha, U_{\text{FPI}}^\alpha] \cap [0, 1].\end{aligned}$$

CP. Split conformal prediction calibrates absolute residuals (Vovk et al., 2005; Lei et al., 2018):

$$a_i = |c_i - s_i|.$$

If $Q_\alpha(\cdot)$ denotes the finite-sample conformal quantile at index $\lceil (n+1)\alpha \rceil$, then

$$\hat{m}_\alpha = Q_\alpha(\{a_i\}_{i=1}^n).$$

The interval is

$$\Gamma_{\text{CP}}^\alpha(q) = [s - \hat{m}_\alpha, s + \hat{m}_\alpha] \cap [0, 1].$$

Entropy-Normalized CP. This baseline normalizes the absolute residual by binary entropy of the backbone score (Zhu et al., 2025c):

$$\begin{aligned}H(s) &= -s \log s - (1-s) \log(1-s), \\ a_i &= \frac{|c_i - s_i|}{H(s_i)}.\end{aligned}$$

The calibrated interval is symmetric around the backbone score, with method-specific endpoints:

$$\begin{aligned}L_{\text{UnKGCP}}^\alpha(q) &= s - \hat{m}_\alpha H(s), \\ U_{\text{UnKGCP}}^\alpha(q) &= s + \hat{m}_\alpha H(s),\end{aligned}$$

After clipping to $[0, 1]$, $\Gamma_{\text{UnKGCP}}^\alpha(q)$ is the interval with endpoints $L_{\text{UnKGCP}}^\alpha(q)$ and $U_{\text{UnKGCP}}^\alpha(q)$. Thus two triples with the same score receive the same interval length regardless of graph support.

Table A.2: Backbone score-export settings. We keep the original train/validation/test protocol and reuse fixed score exports. Per-run early-stopping provenance is retained with the score-export metadata where available.

BACKBONE	VARIANT	Epochs	lr	Dim.	Batch size	Neg./pos
UKGE	UKGE _{logi}	3000	0.001	128	128 (CN15K,NL27K), 256 (PPI5K)	10
PASSLEAF	UKGE _{logi} -m2	600	0.001	512	512	10
BEURRE	Bigumbelbox	1000	0.0001	64	4096 (CN15K), 2048 (NL27K,PPI5K)	30

B.4 Implementation Details

This subsection gives the implementation-level protocol used by the released evaluation code. It does not repeat the method definitions in Section 3. Instead, it records the concrete feature formulas, calibration routine, and aggregation rules used in the retained experiments.

Experiment-Specific Settings and Reproduction.

The main text reports the default coverage level, seed count, split ratio, retrieval metric, and **CARE** hyperparameters. We therefore only record code-level choices omitted there, including entropy inputs clipped into $(10^{-6}, 1 - 10^{-6})$, a feature scaler fitted only on $\mathcal{T}_{\text{ref}}$, finite-sample quantiles with the higher rule, and exact-neighbor indices stored in a hash-checked cache. Numerical tables and figures are regenerated from retained summary outputs. Configuration files and summary outputs may use a legacy internal method identifier, which maps to the paper method **CARE**.

Input Score Frames. Each backbone–dataset–seed directory provides positive calibration, negative calibration, positive test, negative test, and combined test score files. The main comparison reports the positive test file. Negative and combined files are used only for scope diagnostics. Every row contains (h, r, t) , a gold confidence value c , and an exported backbone score s .

KNN Feature Coordinates. The implementation instantiates the symbols in Section 3 with seven scalar quantities before appending the relation-identity one-hot vector. Let $p = s$ for PPI5K and NL27K, and $p = (s + 1)/2$ for CN15K. The value p is clipped into $(10^{-6}, 1 - 10^{-6})$ before entropy is evaluated. Let $E_{\mathcal{T}_{\text{train}}}^u$ be the undirected relation-agnostic training edge set and $\mathcal{N}_{\mathcal{T}_{\text{train}}}^u(v)$ be the corresponding neighbor set. For $q = \langle h, r, t \rangle$, the

scalar quantities are

$$\begin{aligned}
 s &= M_{\mathcal{T}_{\text{train}}}(q), \\
 H(s) &= -p \log p - (1 - p) \log(1 - p), \\
 \delta_h &= \log(1 + d_h), \\
 \delta_t &= \log(1 + d_t), \\
 f_r &= \log(1 + n_r), \\
 e_{ht} &= \mathbb{I}\{\{h, t\} \in E_{\mathcal{T}_{\text{train}}}^u\}, \\
 \delta_{ht} &= \log(1 + |\mathcal{N}_{\mathcal{T}_{\text{train}}}^u(h) \cap \mathcal{N}_{\mathcal{T}_{\text{train}}}^u(t)|).
 \end{aligned}$$

These quantities instantiate the main-text blocks $\phi_{\text{pred}} = (s, H(s))$, $\phi_{\text{topo}} = (\delta_h, \delta_t, e_{ht}, \delta_{ht})$, and $\phi_{\text{rel}} = (f_r, \text{onehot}(r))$. Here d_h and d_t are training-graph degrees, and n_r is the training count of relation r . Entities or relations absent from the training graph receive zero count features before the log transform. The relation one-hot block has dimension $|R|$ for the training-split relation vocabulary, and calibration/test triples are vectorized by the same map without being inserted as graph edges.

Feature Standardization and Exact Retrieval.

For each evaluation subset, the KNN reference pool defines the normalization statistics. If $\phi_j(q)$ is the j -th raw feature coordinate, the implemented standardized coordinate is

$$z_j(q) = \frac{\phi_j(q) - \mu_j(\mathcal{T}_{\text{ref}})}{\sigma_j(\mathcal{T}_{\text{ref}})}.$$

The same fitted scaler is then applied to conformal-holdout and test triples. Exact FAISS L2 retrieval (Johnson et al., 2021) returns neighbors according to

$$\mathcal{N}_K(q \mid \mathcal{T}_{\text{ref}}) = \underset{i \in \mathcal{T}_{\text{ref}}}{\text{smallest}} \arg \text{topK} \|z(q) - z(q_i)\|_2^2.$$

The released implementation stores exact neighbor indices in a hash-checked cache. The cache only avoids repeated distance computation and does not change the neighbor ranking.

Quantile Routine and Bounded Labels. The conformal baselines and **CARE** share the same finite-sample quantile helper. For n calibration scores

and target coverage α , the code evaluates

$$\tau_n(\alpha) = \min \left\{ 1, \frac{\lceil (n+1)\alpha \rceil}{n} \right\},$$

and returns

$$\hat{Q}_\alpha(\{v_i\}_{i=1}^n) = Q_{\tau_n(\alpha)}^{\text{higher}}(\{v_i\}_{i=1}^n).$$

All interval endpoints are clipped to the label range $[0, 1]$ before computing metrics. Entropy-based features and baselines use the same probability preprocessing as $H(s)$ above.

Metric Computation and Aggregation. For a test subset $\mathcal{T}$ and intervals $\Gamma(q_i) = [L_i, U_i]$, the reported metrics are

$$\text{Coverage}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{(q_i, c_i) \in \mathcal{T}} \mathbb{I}\{L_i \leq c_i \leq U_i\},$$

$$\text{Sharpness}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{(q_i, c_i) \in \mathcal{T}} (U_i - L_i).$$

For 10-seed supplementary tables, the aggregation code reports sample means and, where shown, sample variances:

$$\bar{x} = \frac{1}{m} \sum_{j=1}^m x_j,$$

$$\text{Var}(x) = \frac{1}{m-1} \sum_{j=1}^m (x_j - \bar{x})^2.$$

Further-Analysis Diagnostics. The difficulty-tracking diagnostic uses positive NL27K test triples and evaluates UKGE, PASSLEAF, and BEURRE at target coverage $\alpha = 0.90$. Test difficulty is the absolute score error $|s - c|$. For each backbone, the bin range is defined from covered UnKGCP test examples and divided into 30 equal-width bins. CARE uses the same bins for comparability. Each method then reports mean interval length and histogram counts only over the test examples covered by that method. Empty bins are recorded as zero, and examples outside the shared right boundary are not folded into the last bin.

The distribution-shift scope diagnostic uses the negative calibration and test files for all three datasets, all three backbones, and all 10 retained seeds. It evaluates FPI, CP, UNKGCP, and CARE at target coverage levels $\{0.80, 0.85, 0.90, 0.95\}$. The plotted points first average each backbone–dataset–method–level cell over seeds and then average across datasets within each backbone. CARE

uses the same final configuration as the main experiments, including exact L2 retrieval, $K = 200$, $\rho = 0.90$, $\kappa = 50$, and the label-independent random reference–holdout split.

B.5 Artifact and Reproducibility Details

Artifacts and Licenses. The experiments use existing uncertain-KG benchmarks (ConceptNet, STRING, NELL) under their original splits and the negative protocol of prior uncertain-KG conformal work (Zhu et al., 2025c). The review PDF redistributes no raw datasets or checkpoints. Any anonymous supplementary package contains only review-safe scripts, configuration files, processed summary tables, and reproduction instructions, with license notes for any redistributed code.

Software Packages. TensorFlow is used for the UKGE and PASSLEAF score exporters. PyTorch is used for BEURRE. Other packages include NumPy/Pandas, scikit-learn, and Matplotlib.

Compute Resources. Backbone score exports were produced on Linux CUDA workers with one GPU-visible process per job. Post-hoc interval construction, ablations, sensitivity sweeps, and figure generation run on CPU/FAISS from fixed score exports without retraining backbones. The retained 10-seed backbone export matrix takes approximately 167 hours on a single RTX 4090 GPU, and the retained score-export files occupy 536 MB.

C Supplementary Result Tables

Table A.3: Sensitivity analysis for the residual scale and the conformal level. For each evaluation subset, the table reports coverage and sharpness under different values of ρ and α .

ρ	α	POSITIVE		NEGATIVE		COMBINED	
		Coverage	Sharpness ↓	Coverage	Sharpness ↓	Coverage	Sharpness ↓
0.7	0.7	0.70	0.242	0.62	0.024	0.66	0.142
0.7	0.8	0.80	0.295	0.74	0.028	0.78	0.186
0.7	0.9	0.90	0.443	0.85	0.030	0.88	0.291
0.8	0.7	0.70	0.250	0.62	0.022	0.66	0.162
0.8	0.8	0.80	0.285	0.73	0.027	0.77	0.190
0.8	0.9	0.90	0.384	0.85	0.030	0.88	0.260
0.9	0.7	0.70	0.264	0.62	0.021	0.66	0.184
0.9	0.8	0.80	0.301	0.73	0.023	0.77	0.217
0.9	0.9	0.90	0.352	0.84	0.030	0.88	0.260

Table A.4: Diagnostics of query adaptivity on positive triples. We divide the examples within each run into degree buckets and relation frequency buckets, and report coverage and sharpness for **CARE** and **UNKGCP** in each group across the evaluation matrix.

BUCKET TYPE	BUCKET	CARE		UNKGCP	
		Coverage	Sharpness	Coverage	Sharpness
Degree	Low	0.9027	0.3450	0.8857	0.4740
Degree	Mid	0.9052	0.3521	0.9141	0.4871
Degree	High	0.8879	0.3592	0.8991	0.4821
Relation frequency	Low	0.8871	0.3308	0.8808	0.4624
Relation frequency	Mid	0.9038	0.3605	0.9123	0.4784
Relation frequency	High	0.9113	0.2793	0.9052	0.4616

D Additional Experimental Analysis

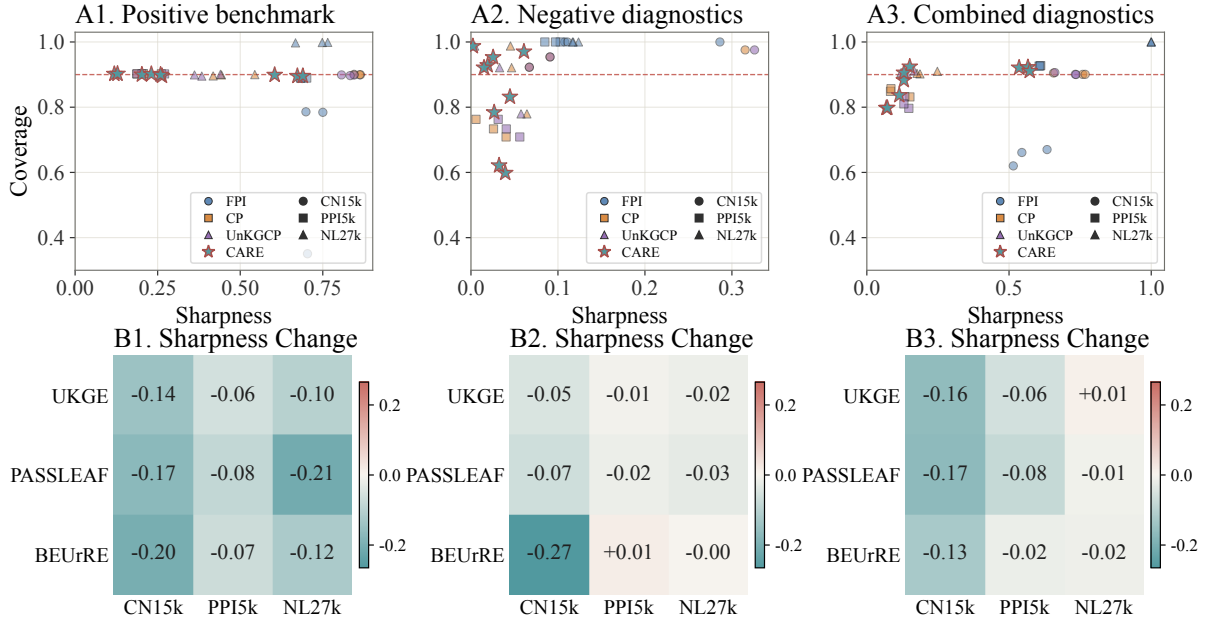


Figure A.1: Full result overview across evaluation subsets. Scatter plots show coverage–sharpness trade-offs, with **CARE** marked by stars. Heatmaps show **CARE** sharpness change relative to the entropy-normalized baseline, where negative values mean shorter intervals.

D.1 Full Coverage and Sharpness Overview

The main table compresses the benchmark results into aggregate cells, but it does not show whether the sharpness gain is systematic across evaluation sources. We therefore collect three evaluation sources in one coverage–sharpness overview in Figure A.1. The scatter plots place every method and benchmark cell in the same coverage–sharpness plane, while the heatmaps summarize the cell-wise sharpness change of **CARE** relative to the entropy-normalized baseline.

A1 shows the benchmark setting used for the main comparison. The conformal methods stay near the target coverage, with mean coverage 0.900 for CP, 0.900 for UnKGCP, and 0.899 for **CARE**, whereas FPI ranges from 0.351 to 0.999 because its parametric residual assumption is less stable. Under this matched-coverage regime, **CARE** moves left in the coverage–sharpness plane: its mean sharpness is 0.352, compared with 0.481 for UnKGCP. This supports the main benchmark claim that context anchors shorten intervals without changing the split-conformal operating point.

The middle setting deliberately changes the evaluation source. Coverage no longer concentrates as tightly around 0.90: **CARE** has mean coverage 0.844, while CP and UnKGCP both average 0.861.

The important point is therefore not a new validity claim but a scope diagnostic under distribution shift. In this shifted setting, **CARE** still gives the shortest intervals, with mean sharpness 0.030 compared with 0.079 for UnKGCP. B2 shows that most cells remain negative, especially BEUrRE–CN15k at -0.27 , although BEUrRE–PPI5k becomes slightly positive at $+0.01$. Thus, local residual evidence controls width in many corrupted neighborhoods, but the effect is more heterogeneous than on original triples.

The rightmost setting combines original and shifted examples to expose the trade-off between benchmark consistency and shifted inputs. The mean coverage of **CARE** is 0.878, close to UnKGCP at 0.874, while mean sharpness decreases from 0.330 to 0.260. B3 shows that the reductions remain strongest on CN15k, with changes between -0.13 and -0.17 , and are milder on PPI5k and NL27k. One UKGE–NL27k cell is slightly positive at $+0.01$, but the remaining cells are negative. Overall, distribution shift weakens the uniformity of the gain.

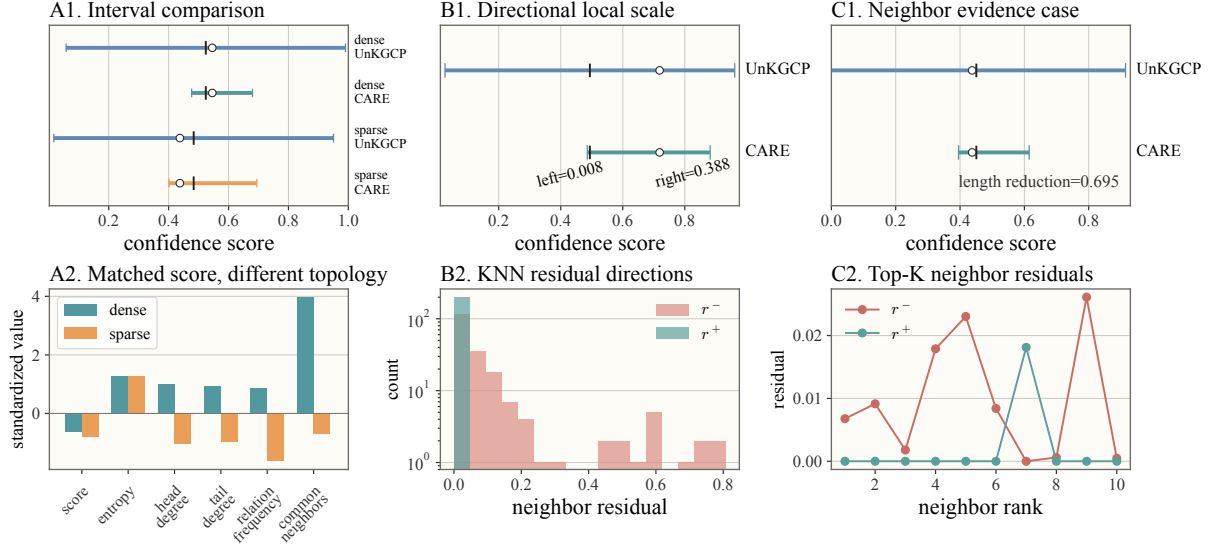


Figure A.2: Case studies for context-anchored calibration on PASSLEAF–NL27K. The top row contrasts queries with similar backbone scores but different graph support, leading to different interval lengths. The middle row shows lower and upper local residual scales for queries where the backbone is directionally biased. The bottom row traces interval shortening to the residuals of the retrieved calibration neighbors.

D.2 Case-Level Mechanism Study

Aggregate results show that **CARE** produces shorter intervals, but they do not explain how a specific query receives a sharper or more asymmetric interval. We therefore select three representative PASSLEAF–NL27K cases with fixed rules and make the context-anchored calibration mechanism inspectable in Figure A.2. The goal is to check whether the shorter intervals in the tables can be traced to concrete local calibration evidence.

The first case selection restricts candidates to triples with close backbone scores and entropy, then chooses a dense–sparse pair with the largest topology-support gap. A1 compares intervals for compact and sparse graph regions. UnKGCP cannot separate them because it mainly uses score and entropy, whereas **CARE** constructs a sharper interval for the compact query with stronger local evidence. A2 makes the mechanism explicit: score and entropy remain matched, but head degree, tail degree, relation frequency, and common neighbors clearly separate the two queries. The interval width therefore follows structural support from context anchors rather than scalar score uncertainty alone.

The second case selection chooses a covered query with little boundary clipping and strongly imbalanced lower/upper local scales. In B1, the backbone score lies below the true confidence label, so the main risk is upward correction: the interval needs a larger right/upper extension than left/lower extension. UnKGCP expands both sides

with one entropy-normalized width, whereas **CARE** keeps the left side tight and assigns most width to the right side. B2 explains this decision from the retrieved anchors: r^- residuals, which measure under-estimation and support upper-side expansion, are much heavier than r^+ residuals. The right-heavy interval in B1 therefore directly matches the r^- -heavy anchor evidence in B2, showing why side-specific calibration is needed.

The final case selection chooses a covered query from high-reduction candidates and keeps the one whose top anchors have the smallest mean residual. In C1, UnKGCP gives a broad interval because it lacks local graph-structural evidence beyond score entropy. **CARE** uses the same fixed backbone score but forms a sharper asymmetric interval, leaving a shorter left/lower side while giving the right/upper side more width. C2 exposes the cause: nearby anchors are not uniformly noisy, and their nonzero residuals are mostly r^- , pointing to under-estimation and right-side expansion. This is the advantage of context-anchor calibration: **CARE** identifies where width is needed instead of paying the broad UnKGCP width on both sides.