Scaling Trust: Veracity-Driven Defect Detection in Entity Search

Ornella Irrera*
ornella.irrera@unipd.it
University of Padua
Padua, Italy

Stefano Marchesin*
stefano.marchesin@unipd.it
University of Padua
Padua, Italy

Gianmaria Silvello
gianmaria.silvello@unipd.it
University of Padua
Padua, Italy

Omar Alonso†
omralon@amazon.com
Amazon
Palo Alto, California, USA

Abstract

Veracity is a critical dimension of data quality that directly impacts a wide range of tasks. In entity search scenarios, Knowledge Graphs (KGs) such as DBpedia and Wikidata serve as core resources for accessing factual content. The veracity of these KGs is therefore essential for ensuring the reliability and trustworthiness of retrieved entities – factors that directly influence user confidence in the search system. However, ensuring the truthfulness of entities remains a major challenge due to the complexities associated with the scale, development, and maintenance of KGs.

This paper critically analyzes the impact of veracity in entity search, using DBpedia as the underlying KG. To this end, we introduce *e*Rank, a veracity-driven re-ranking strategy that enhances entities' trustworthiness without sacrificing the ranking's overall relevance. Furthermore, we propose the Active Learning-based verAcity-Driven Defect IdentificatioN (ALADDIN) system, a lightweight and scalable framework for veracity-driven defect detection. ALADDIN identifies incorrect KG facts and exhibits high effectiveness in downstream entity-centric tasks, such as entity summarization, entity card generation, and defect recommendation.

CCS Concepts

• **Information systems** → **Information systems applications**; **Information retrieval**; **Evaluation of retrieval results**.

Keywords

Entity Search, Knowledge Graph, Defect Detection, Active Learning

ACM Reference Format:
Ornella Irrera, Stefano Marchesin, Gianmaria Silvello, and Omar Alonso. 2025. Scaling Trust: Veracity-Driven Defect Detection in Entity Search. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3746252.3761208

*These authors contributed equally to this work.
†Work does not relate to the author's position at Amazon.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761208

1 Introduction

In today's digital landscape, the widespread reuse and redistribution of online content has raised concerns about information reliability [30]. Users rely on search engine results and structured data from Knowledge Graphs (KGs), which not only summarizes key facts but also generates *entity cards* to provide concise information and organize entity details, enabling context-aware user query responses that are, ideally, also accurate and reliable [12, 70]. As data from KGs is increasingly repurposed across various platforms, ensuring its veracity – comprising accuracy, truthfulness, and consistency – becomes essential. Indeed, *data veracity* is considered a key metric in quality management, complementing the other four V's of Big Data (volume, velocity, variety, and value) [8, 14, 78], and plays a crucial role in guaranteeing information trustworthiness [53]. Hence, guaranteeing KGs veracity is crucial for ensuring the reliability of information. Still, it is also challenging due to the noisy, large-scale, and constantly evolving nature of KGs [82].

Entity search, the target of this work, focuses on structuring information around entities, their attributes, and the relationships between them [4], usually represented in KGs. Notable examples of KGs include Wikidata [81] and DBpedia [3]. However, these KGs are (semi-)automatically constructed and updated, which include potentially incorrect or unreliable information [65, 82].

A key aspect of veracity estimation in KGs is identifying incorrect or unreliable triples – potential defects – compromising data quality. Detecting such triples is crucial for improving KG veracity and supporting downstream tasks. Despite its importance, veracity is often addressed through manual or ad hoc methods that do not scale with the growing size and complexity of KGs [36].

Existing efforts have proposed sampling-based approaches to KG veracity estimation [28, 56, 57, 69], and recent work [59] introduced a utility model based on entity popularity, showing promising but not yet scalable results. The literature shows limited work on systematic solutions for assessing and leveraging veracity in entity-centric systems. This gap limits our ability to guarantee the reliability of retrieved entities and negatively impacts user experience in entity search or related entity-centric downstream tasks.

We address these challenges by exploring two research questions:

RQ1: How does integrating veracity estimation affect the effectiveness and scalability of entity search systems?

RQ2: How can veracity signals be operationalized to support decision-making and veracity control in entity-centric tasks?

To address RQ1, we propose *e*Rank, a novel veracity-estimation metric, and apply it to large-scale KGs to examine how veracity influences entity search. Our experimental evaluation shows that integrating veracity estimates into entity search systems effectively prioritizes entities with higher veracity while maintaining relevance to the query. This approach can enhance the user experience by providing more reliable and credible results. *e*Rank is centered on enhancing veracity at the entity level; however, it does not discern high-quality facts or individual triples within a KG at a more granular level.

Hence, in addressing RQ2, we introduce the Active Learning-based verAcity-Driven Defect IdentificatioN (ALADDIN) system: a lightweight system engineered for defect detection. ALADDIN integrates an offline-trained linear regression model with an online active learning framework incorporating human feedback. In this online phase, ALADDIN presents entity cards showcasing top-ranked facts for a given entity, enabling users to assess their veracity in context. This annotation approach using entity cards allows ALADDIN to collect user feedback on the facts considered most pertinent by the ranking model, enhancing its defect detection abilities where user validation is most crucial.

We evaluate the proposed method on four downstream tasks: *defect detection*, *entity summarization*, *entity card generation*, and *defect recommendation*. Each task leverages the predicted veracity scores to serve a distinct purpose: *defect detection* targets the detection of incorrect triples; *entity summarization* ranks facts based on their estimated veracity; *entity card generation* constructs entity cards guided by veracity scores and evaluates their quality through human annotation; and *defect recommendation* selects potentially incorrect triples under a budget constraint to prioritize corrections. Our experiments demonstrate that ALADDIN progressively enhances the accuracy of veracity predictions over time, while requiring only a limited amount of labeled data.

The main **contributions** of this work are:

(1) A novel metric, *e*Rank, for estimating entity-level veracity, and extensive analyses to understand its impact on large-scale entity search.
(2) ALADDIN, a veracity-driven, fact-level defect detector.
(3) An in-depth evaluation of ALADDIN on four downstream tasks demonstrating its effectiveness in predicting veracity and supporting diverse entity-centric tasks at scale.

Outline. The rest of the paper is as follows. Section 2 provides the necessary background to develop *e*Rank (Section 3) and ALADDIN (Section 4). Sections 5 and 6 present the experiments, their setup, and the results. Section 7 reports on related work. Finally, Section 8 concludes the paper and outlines possible future work directions.

2 Background

We explore the role of data veracity in entity-centric tasks from a utilitarian view and the related evaluation challenges. Next, we outline a utility-oriented framework for KG veracity assessment, forming the basis for the development of *e*Rank and ALADDIN.

The Role of Data Veracity in Utilitarian Analysis. In the evolving landscape of web search, there is a growing consensus that traditional evaluation metrics focused primarily on ranking quality are insufficient for capturing the full spectrum of user satisfaction and task success [13]. This shift has led to the emergence of a broader evaluation framework known as *utilitarian analysis*, which emphasizes the holistic nature of search experiences. Rather than solely measuring relevance, utilitarian analysis considers the user's overall journey, including contextual factors and implicit costs such as time, cognitive effort, and required interaction. This paradigm applies across various search modalities, from explicit queries and recommendation systems to content feeds and entity search.

Within this framework, the notion of *Delphic costs and benefits* captures search operations' intangible, non-monetary consequences. These include impairments like misinformation, disinformation, and misrepresentation, which can distort the utility users derive from search results [13]. In recent years, particular attention has turned to the role of data veracity on these Delphic dimensions [58], especially in the context of entity search tasks [59]. KGs – in particular those generated through semi- or fully-automated methods [84] – are inherently prone to inaccuracies [24, 25, 68], thereby amplifying the costs associated with misinformation. As a result, tasks such as entity cards generation become vulnerable; when these information capsules contain erroneous data, they not only diminish their intended benefits but also impose additional cognitive and trust-related burdens on users.

Practical Challenges in Veracity Assessment for KGs. Ensuring data veracity is crucial for user experiences, which requires a focused evaluation of KGs, emphasizing the importance of verification in downstream tasks. A foundational step in this process involves manually verifying the correctness of KG contents – i.e., assessing the accuracy of its facts, primarily represented as triples in the form of subject-predicate-object (s, p, o) relationships. However, real-world KGs such as Wikidata [81], DBpedia [3], and YAGO [76] contain hundreds of millions, or even billions, of facts, making exhaustive manual annotation impractical and prohibitively costly.

To overcome this limitation, recent work has explored the use of sampling and estimation techniques [28, 56–58]. These methods offer a cost-effective and statistically robust solution to veracity assessment, especially with contained labeling budgets. Notably, the utility of different KG parts may vary with respect to the downstream task. For example, entities with higher popularity or query volume typically have a greater impact on user experience [29, 37]. As such, prioritizing verifying facts associated with these high-utility entities can trigger the deployment of specific filtering or correction mechanisms when data quality is compromised [17, 25].

Utility-Oriented Framework for KG Veracity Assessment. To operationalize the connection between data veracity and task usefulness in entity search, we extend the utility-oriented KG evaluation framework introduced by Marchesin et al. [58]. This framework underpins our methodology for veracity-driven defect identification. The key innovation lies in explicitly modeling the utility of individual facts, segmenting the KG according to that utility, and assessing veracity in a scalable and statistically robust way.

Framework input. The framework models a KG as a directed, edge-labeled multi-graph $G = (V, R, \eta)$, where $V = \{E \cup A\}$ is the set of nodes – comprising entities E and attributes A; R is the set of labeled relationships; and $\eta : R \rightarrow E \times (E \cup A)$ assigns ordered node pairs to relationships. This representation yields the ternary relation $T \subseteq E \times R \times (E \cup A)$, whose elements – the (s, p, o) triples

– are the KG facts. Facts sharing the same subject $e \in E$ define an entity cluster $G[e] = \{(s, p, o) \in T \mid s = e\}$, whose size is denoted by $M_e = |G[e]|$. The total number of facts is denoted by $M = |T|$.

Utility and partitions. A popularity measure is used as a proxy for the utility on downstream tasks. We employ SPARQL query logs to assign popularity scores to KG facts, using query frequency as a proxy for task relevance. Then, the Cumulative Square Root of Frequency (CSRF) method [22], that has been shown to be effective in similar settings [28, 54, 55], is adopted to create a partition family $\mathcal{P} = \{P_i\}_{i=1}^{k}$, with each partition representing a popularity stratum.

Sampling and estimation. To assess veracity within each partition, the Two-stage Weighted Cluster Sampling (TWCS) method [28] is applied. In the first stage, n_i entity clusters are sampled from a partition P_i with probability $\pi_{ij} = M_{ij}/M_i$, where M_{ij} is the size of the jth cluster from P_i and $M_i = |P_i|$. In the second stage, up to m facts are sampled from each selected cluster via simple random sampling without replacement. Once fact annotations are gathered for a partition, they are fed to an estimator $\hat{\mu}$ to gauge the partition veracity. By computing the estimated veracity $\hat{\mu}_{ij}$ of the jth sampled cluster as the mean veracity of its sampled facts, the estimator of $\mu(P_i)$ can be defined as $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{\mu}_{ij}$, which is known to be unbiased [20].

Cost minimization. The framework employs an iterative estimation procedure to minimize the annotation cost. In a nutshell, the framework iteratively samples facts from a partition P_i using TWCS, gathers manual annotations, and computes an unbiased veracity estimate $\hat{\mu}_i$ for the target partition P_i. The framework evaluates whether the current estimate meets a specified confidence requirement at each iteration by constructing a $1-\alpha$ Confidence Interval (CI). The process terminates once the Margin of Error (MoE), defined as half the CI width, falls below a user-defined threshold ε_i, ensuring that the partition estimate is both cost-efficient and statistically robust.

Formally, the iterative procedure aims to solve the following constrained optimization problem:

$$\begin{aligned} \underset{\mathcal{S}}{\text{minimize}} \quad & \text{cost}(\mathcal{S}(P_i)) \\ \text{subject to} \quad & \text{MoE}(\hat{\mu}_i, \alpha) \leq \varepsilon_i \vee |\mathcal{S}(P_i)| = b_i \end{aligned}$$

where $\mathcal{S}(P_i)$ denotes the sample drawn from partition P_i under the given sampling strategy $\mathcal{S}$; $\text{cost}(\mathcal{S}(P_i))$ represents the effort required to annotate the sampled facts; and b_i is the portion of the overall annotation budget b allocated to P_i. The dual condition in the constraint ensures that the process halts either when sufficient precision is achieved (i.e., MoE $\leq \varepsilon_i$) or when the partition budget is exhausted (i.e., $|\mathcal{S}(P_i)| = b_i$).

Framework output. Once the estimation process is complete, each partition P_i is associated with an estimated veracity score $\hat{\mu}_i$. These partition-level scores can be propagated to the individual facts within the partition – i.e., for any fact $t \in P_i$, its estimated veracity is denoted as $v(t) = \hat{\mu}_i$. The final output is a ranking of all the facts in the KG associated with a veracity score defined at the partition level. The annotated samples and partition-level estimates produced by this framework serve as *foundational blocks* for the veracity-driven strategies we develop in this work.

3 Entity-Level Veracity-Driven Re-Ranking

Recent work introduced vRank [59], a fact-level veracity-driven re-ranking method that combines veracity signals with relevance scores from entity summarization models, to account for lower-veracity facts in the rankings. Formally, given a fact t, it first applies min-max normalization to the summarization model's raw score to obtain an fScore(t), and then adds the veracity estimate $v(t)$ to produce: $v\text{Rank}(t) = \text{fScore}(t) + v(t)$. The facts are then re-ranked by descending vRank to favor those deemed both relevant and more likely true [59].

However, entity search requires ranking whole entities, not individual facts, in response to a user query [4]. Hence, inspired by vRank, we compute each entity's overall veracity as the mean of its associated fact veracity scores to bring veracity into this setting. Then, we combine this with the entity search model's normalized relevance score. Formally, for entity e with fact cluster $G[e]$ of size M_e, we define:

$$e\text{Rank}(e) = \text{eScore}(e) + \frac{1}{M_e} \sum_{t \in G[e]} v(t),$$

where eScore(e) is the normalized relevance assigned by the entity search model, and $\frac{1}{M_e} \sum_t v(t)$ is the entity's mean veracity.

Regarding granularity, vRank operates at the fact level – boosting or demoting individual facts by adding their veracity scores to normalized relevance – whereas eRank aggregates these fact-level veracity estimates into a single mean score to rank entire entities. Moreover, vRank applies veracity as a final re-ranking adjustment for facts, while eRank integrates veracity directly into the core scoring function, blending relevance and truthfulness in one step.

4 The ALADDIN System

In this section, we first motivate the need for operational, fine-grained veracity inference by analyzing the limitations of existing fact-level approaches. Then, we present the ALADDIN system, a pipeline designed to address these challenges.

The need for operational fine-grained veracity inference. Although vRank has been proposed to improve ranking by integrating fact-level veracity with relevance, its effectiveness in real-world, noisy settings is lacking. To motivate this, Figure 1 presents the distribution of facts for 100 widely recognized DBpedia entities commonly used in entity search tasks [33, 35]. The figure indicates that facts are primarily concentrated in one partition for most entities. A bar with a uniform color represents this concentration, while a bar with multiple colors signifies scattered facts across partitions with varying popularity and veracity. As a result, the veracity scores, which remain largely consistent for most facts of a given entity, act as a constant additive factor. Therefore, they do not alter the relative ranking of facts, making vRank ineffective. This reveals a flaw in vRank: its dependency on partition-level veracity scores lacks the necessary detail to differentiate between true and false facts within the same entity. This discovery highlights the requirement for a finer-grained approach to assess veracity at the fact level. Our proposed system, ALADDIN, addresses this limitation.

ALADDIN. Below, we describe the main steps of ALADDIN pipeline illustrated in Figure 2.

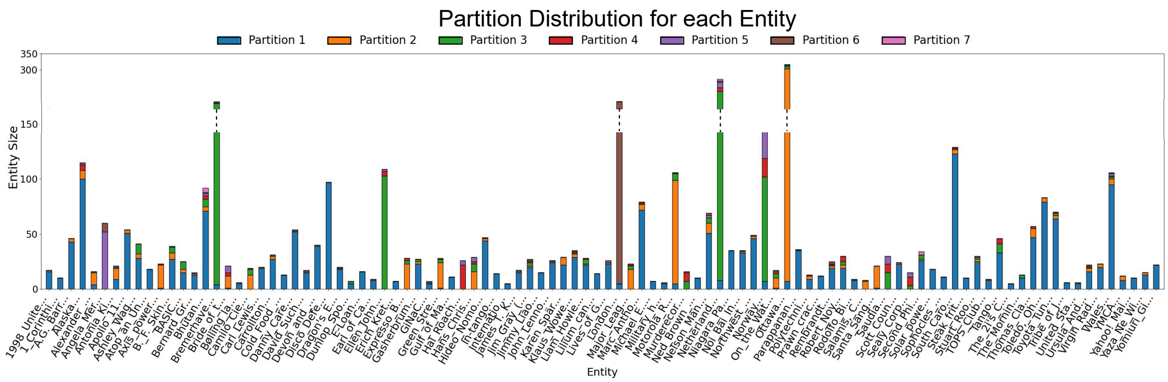

Figure 1: Partition distribution across the 100 entities widely used in entity-centric tasks. Partition 7 is the most popular, down to Partition 1, which is the least popular. Veracity does not correlate with popularity.

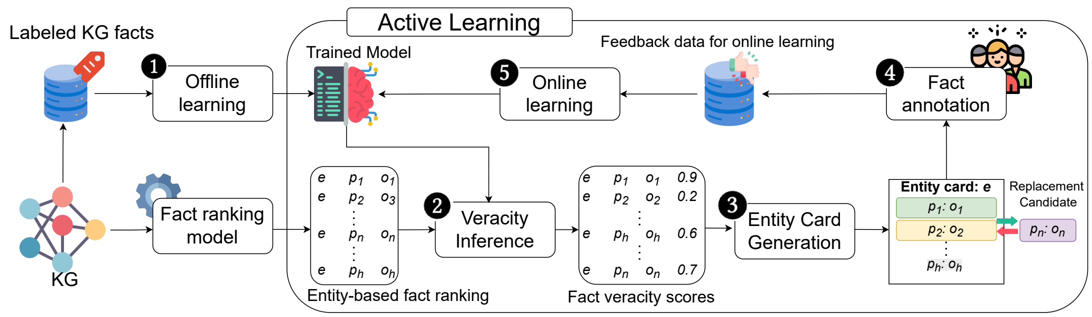

Figure 2: The ALADDIN pipeline: (1) Offline training of a linear regression model on KG fact embeddings, (2) veracity inference on ranked facts, (3) entity card generation with threshold-based defect detection and replacement suggestions, (4) human fact annotation with candidate evaluation, (5) online active learning updates.

❶ **Offline learning**. The pipeline begins with the offline training of a linear regression model to estimate the veracity of facts within the KG. Due to its simplicity and computational efficiency, the linear regressor is particularly well-suited for online learning scenarios, where rapid and incremental updates are essential [75]. The model is trained on a dataset of facts labeled with binary correctness annotations using the Squared Error loss function and optimized via stochastic gradient descent [10].

Although the training labels are discrete, using a regression model allows for more nuanced predictions. This design encourages the model to produce outputs that tend to polarize toward 0 or 1, while retaining the flexibility to assign intermediate veracity scores to more uncertain facts, avoiding strict binary classification.

Each KG fact is embedded with TransE [11] to enrich the model input. TransE graph embeddings capture both the structural (topological) and semantic (textual) aspects of the KG, enabling an expressive representation. Once trained, the model assigns to each fact t a real-valued veracity score $v_{\mathcal{A}}(t) \in \mathbb{R}$, optimized to fall within the $[0, 1]$ interval. Scores closer to 0 suggest a defective fact, while those closer to 1 a plausible fact.

❷ **Veracity inference.** The trained regression model is then applied to the ranked list of facts representing an entity (cluster) generated by a fact ranking system. The model infers a veracity score for each fact, indicating its plausibility. These scores are then analyzed to identify low-quality but top-ranked facts, thereby flagging them as potential and influential defects.

❸ **Entity card generation.** Once veracity scores have been computed for an entity's facts, ALADDIN presents human annotators with an *entity card*: a concise, familiar layout that lists the top-h ranked facts for that entity. Annotators review and label these facts, reliable or not, which in turn triggers the system's active learning loop. This structured, user-friendly annotation format is meant to make the annotation process faster and more intuitive [33, 71].

Each fact in the entity is associated with an inferred veracity score, allowing us to set a threshold θ_v to distinguish between reliable ($> \theta_v$) and unreliable ($\leq \theta_v$) facts. For any fact t deemed as unreliable, a replacement candidate t' is suggested. This candidate is selected from the lower-ranked portion of the fact list (i.e., positions beyond h), and is the highest-ranked fact with estimated veracity above the threshold.

This entity card-based annotation strategy allows ALADDIN to focus user feedback on the most critical and relevant facts by prioritizing identifying defective content in the most visible and high-traffic parts of the KG, thereby maximizing utility for entity-centric downstream tasks.

❹ **Fact annotation.** Human annotators evaluate the facts once an entity card is generated. For each fact deemed reliable, annotators are asked to assess whether the fact is correct or not. For unreliable facts, annotators are presented with a candidate replacement and asked to choose among four options: (i) the replacement is correct, (ii) the original fact is correct, (iii) both are correct, or (iv) both are incorrect. Options (iii) and (iv) are included to mitigate annotation bias and accommodate cases where both facts are equal.

After collecting annotations, each fact is mapped to a binary label. When a candidate replacement is marked as correct, it receives a 1 while the original fact gets a 0; conversely, if the original is marked correct, the reverse labeling is applied. When both facts are correct (or incorrect), both are assigned a 1 (or a 0), respectively.

❺ **Online learning.** Annotations are then used incrementally as labeled data to refine the linear regression model through online active learning. The model is updated in batches, where each batch is the set of annotations derived from a single entity card.

The linear regressor makes the model susceptible to overfitting when the incoming annotation stream is heavily biased, e.g., consisting predominantly of correct or incorrect facts [7]. We propose a *residual-based selection strategy* that filters for the most informative and representative training examples. Specifically, we exclude annotations where the model predictions closely align with human evaluations, as such cases are less likely to contribute a meaningful learning signal and may reinforce bias.

Formally, let e be an entity with annotated facts $t_1, \ldots, t_h$. We define the batch of selected annotations for model update as:

$$\Omega_e = \left\{ \langle t_i, \mathbb{1}(t_i) \rangle : |v_{\mathcal{A}}(t_i) - \mathbb{1}(t_i)| > \theta_{\mathcal{S}}, i \in [1..h] \right\}$$

where $\mathbb{1}(t_i)$ denotes the binary correctness label and $\theta_{\mathcal{S}}$ the residual threshold. We compute the residuals between predicted veracity scores and annotated correctness labels. We retain only those annotations where the residual is greater than $\theta_{\mathcal{S}} = 0.5$, representing the decision boundary between predicting correctness and incorrectness. A residual of > 0.5 indicates that the model prediction leans towards the opposite correctness label compared to the human annotation, making the discrepancy informative for model updates. Note that $\theta_{\mathcal{S}}$ can be adjusted within the $[0, 1]$ interval to impose more stringent or lenient filtering.

This selection mechanism ensures that the model focuses on learning from high-disagreement cases, where human judgment and model output diverge significantly. As a result, it mitigates the risk of overfitting caused by imbalanced annotation distributions and strengthens the model's generalization capabilities.

Active learning. The steps from ❷ to ❺ collectively represent the active learning strategy enforced by ALADDIN to enhance defect detection in a utility-aware, cost-effective, and scalable manner. These steps are repeated until a predefined allocation budget is exhausted or a custom stopping criterion is met.

ALADDIN can handle KGs through the support for incremental updates and integrating new annotations. It is inherently lightweight, relying on a simple linear regression model that is computationally inexpensive, transparent, and explainable, allowing for downstream error analysis. Additionally, since the input graph embeddings can be pre-computed offline, the system further benefits from improved efficiency. Most importantly, ALADDIN is both *cost-effective* and *utility-aware*, as it aims at minimizing the need for manual labeling by directing annotation efforts toward the most informative and critical examples relevant to downstream tasks.

5 Experimental Setup

We present the KG and collections, along with the experimental design and task setup developed to address the research questions.[1]

The reference KG is the English version of **DBpedia 2015-10** [43], which comprises 6.2 million entities, 1.1 billion facts, and 739 ontology types. As documented in [58], DBpedia 2015-10 is the reference also for the utility-oriented framework for evaluating the veracity of KGs. [58] created seven popularity-oriented partitions, each reflecting a distinct level of veracity. The reported veracity levels for these partitions (with their CIs), are as follows: the Partition 1 has a veracity of 0.83 ± 0.02; partitions 2 and 3 a veracity of 0.85 ± 0.02; Partitions 4 and 5 a veracity of 0.89 ± 0.02; Partition 6 a veracity of 0.90 ± 0.03; and, Partition 7 a veracity of 0.87 ± 0.03. Partition 1 is the least popular up to Partition 7, which is the most popular (cf. Section 2). This indicates that veracity and popularity are independent, or orthogonal, assessment metrics.

While DBpedia 2015-10 overall is of high quality, the observed variation in veracity across partitions highlights the importance of methods that can effectively prioritize high-veracity data to maximize utility in downstream tasks. At the same time, the narrow CIs underscore the robustness of the underlying estimation procedure.

We consider three collections in our experiments.

DBpedia-Entity v2. Introduced in [35] for evaluating entity search systems over DBpedia 2015-10. This collection retains only DBpedia entities with both title and abstract, resulting in a total of 4.6 million entities. It comprises 485 queries with relevance judgments on a three-point graded scale, ranging from 0 (irrelevant) to 2 (highly relevant). In addition to the queries and judgments, the collection includes the official runs of 12 systems.

DBpedia-Defect. Derived from [58] within the utility-oriented framework for KG veracity assessment. This collection comprises 9,930 facts from DBpedia 2015-10, spanning various topics including entertainment, news, history, sports, business, and science. At least three crowdworkers annotated each triple using binary correctness labels. Final labels were determined through a quality-weighted majority voting scheme, where annotator reliability was estimated using a separate validation set annotated by domain experts. There are 7,949 correct and 1,395 incorrect triples.

DBpedia-Fact. Introduced in [33] to evaluate query-dependent entity summarization methods over DBpedia 2015-10. The collection includes only DBpedia entities with a title, an abstract, and at least five valid predicates, ensuring minimum descriptive richness. It comprises 100 query-entity pairs, representing a subset of those available in DBpedia-Entity v2. Each query q_i is associated with a target entity e_i, and the corresponding set of facts from the entity cluster $G[e_i]$ forms the input for summarization. Overall, the collection contains 4,069 facts, with an average of 41 facts per entity. Judgments are provided across three dimensions: (i) fact importance to the entity (3-point scale); (ii) fact relevance to the query

[1] Code and data: https://github.com/KGAccuracyEval/defect-detection4entity-search

(3-point scale); and (iii) a combined utility score integrating the two (5-point scale). In addition, the collection includes official runs for each of these dimensions using four fact ranking approaches: RELIN [16], a PageRank-based approach, and three variants of DynES [33], a learning-to-rank approach, trained on importance (DynES/i), relevance (DynES/r), and combined utility (DynES/u) judgments, respectively.

Task 1: Entity search. In this first experiment, we investigate the impact of veracity on entity search systems (**RQ1**) on the DBpedia-Entity v2 collection. We consider the 12 official baseline runs and apply eRank, our veracity-driven re-ranking strategy, to re-rank their results. We adopt min-max normalization for the eScore$(\cdot)$ component of eRank and, to be consistent with the original evaluation [35], we report nDCG@10 and nDCG@100.

Task 2: Defect detection. In this set of experiments, we evaluate the performance of ALADDIN in predicting fact veracity (**RQ2**). The experiments use the DBpedia-Defect collection for offline training and offline/online testing, and a portion of DBpedia-Fact for online training. Specifically, we randomly sample 1,000 facts from DBpedia-Defect as the test set, using the rest for offline training. For online training, we consider the four official runs from DBpedia-Fact, each comprising a ranked list of facts for 100 entities, resulting in a total of 400 fact rankings. From these, 50 rankings are randomly selected and held out for evaluating subsequent tasks, while the remaining 350 are split into 300 for training and 50 for validation.

The KG facts used as input to ALADDIN are encoded using TransE [11], an energy-based model for learning KG embeddings, trained for 100 epochs with a learning rate of 0.1 to optimize Focal Loss [48], which effectively handles imbalanced data like DBpedia-Defect. Each fact is embedded by concatenating the embeddings of the subject, predicate, and object – each represented with 384 features. The linear regressor is trained offline for 1,000 epochs. For online training, entity cards are generated from DBpedia-Fact rankings following the procedure described in [33]. We set the veracity threshold $\theta_v = 0.5$ to distinguish between reliable and unreliable facts, considering a fact *correct* if its predicted veracity exceeds this threshold. We adopt an *invscaling* learning rate schedule, which gradually reduces the learning rate as training progresses [9]. This approach is well-suited to online learning, as it enables larger parameter updates in the early stages and more conservative updates later, reducing the risk of oscillations as new annotations are provided [75]. To further prevent overfitting, we also apply early stopping based on validation performance: after each annotation round, we evaluate the model on the validation set and stop training if performance degrades for more than five consecutive rounds, retaining the best-performing model. This process converged after 130 rounds, with 300 annotated facts selected for training by the residual-based selection strategy. Entity card annotations were performed by one expert annotator using a custom interface, as shown in Figure 3.

Performance is reported on the DBpedia-Defect test set for offline and online settings, using {binary, weighted} F1 and balanced accuracy to assess class imbalance.

Task 3: Entity summarization. In this experiment, we evaluate the impact of the veracity scores produced by ALADDIN on fact ranking (**RQ2**). To this end, we resort to the DBpedia-Fact collection. We take the official baseline runs from the four fact

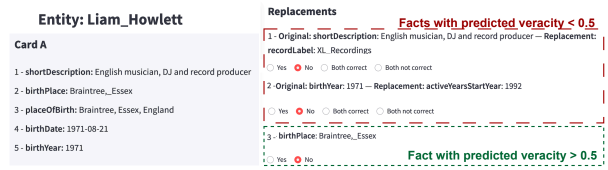

Figure 3: Interface for the online active learning stage. Candidate replacements (highlighted in the red box) are shown only when the veracity predicted by ALADDIN is below 0.5; otherwise, no replacement is suggested (in the green box).

ranking approaches and re-rank their results using the veracity scores predicted by ALADDIN. Specifically, for each fact t in the ranking, we compute the ALADDIN-enhanced score as

$$\mathcal{A}\,\text{Rank}(t) = \text{fScore}(t) + v_{\mathcal{A}}(t),$$

where fScore(t) is the min-max normalized score produced by the original fact ranking model, and $v_{\mathcal{A}}(t)$ is the predicted veracity score from ALADDIN. We compare the performance of these re-ranked outputs with those of the original methods and with the vRank-based re-ranking proposed in [59]. Results are reported using nDCG@5 and nDCG@10 for consistency with prior work [33, 59].

Task 4: Entity card generation. The goal of this experiment is to assess how users perceive the quality of entity cards generated from ALADDIN-enhanced rankings compared to those derived from the original rankings (**RQ2**). For this evaluation, we use the 50 rankings previously held out from DBpedia-Fact during the setup of *Task 2*. Each ranking is re-ranked using ALADDIN, and entity cards are generated for both the original and re-ranked versions following the procedure outlined in [33].

For each pair of competing cards, four annotators were asked to indicate which version – original or ALADDIN re-ranked – presented more accurate factual content, or to choose "no preference" if both are equally accurate. To prevent bias, the position of the two cards (left or right) was randomized when presented to annotators.

Evaluation is based on annotator preferences across all card pairs. Specifically, we measure the number of preferences expressed in favor of the original cards, the ALADDIN-based cards, and the cases where no preference was indicated.

Task 5: Defect recommendation. In this experiment, we assess the effectiveness of ALADDIN in recommending facts that are most likely to be incorrect (**RQ2**). The goal of defect recommendation is twofold: (i) to flag low-veracity facts that are likely incorrect, and (ii) to enable a budget-aware correction strategy that prioritizes the most critical issues – ensuring that limited resources are allocated efficiently to address the most impactful errors first.

We perform this evaluation on the DBpedia-Defect test set, where ALADDIN is used to predict fact veracity scores. The facts are then ranked in ascending order of predicted veracity, allowing identification and prioritization of the least trustworthy ones. Performance is measured in terms of precision and recall.

6 Experimental Results

RQ1: Veracity impact at scale. The first experiment addresses how veracity estimation impacts *entity search* (Task 1). Table 1

Table 1: Performances of entity-search systems comparing the original ranking (Orig) to the eRank-based ranking.

	nDCG@10		nDCG@100	
Model	Orig	eRank	Orig	eRank
BM25	0.26	0.26	0.36	0.37
PRMS	0.39	0.39	0.47	0.47
MLM-All	0.40	0.41	0.49	0.49
LM	0.42	0.42	0.50	0.50
SDM	0.42	0.43	0.51	0.52
LM+ELR	0.42	0.42	0.51	0.51
SDM+ELR	0.44	0.43	0.52	0.52
MLM-CA	0.44	0.44	0.51	0.52
BM25-CA	0.44	0.45	0.53	0.54
FSDM	0.45	0.46	0.53	0.54
BM25F-CA	0.46	0.47	0.55	0.56
FSDM+ELR	0.46	0.46	0.54	0.54

Table 2: τ and τ_{AP} metrics for DBpedia-entity-v2 retrieval methods: original ranking vs eRank-based ranking.

	cutoff@10		cutoff@100	
Model	τ	τ_{AP}	τ	τ_{AP}
BM25	0.77	0.75	0.39	0.36
PRMS	0.69	0.68	0.28	0.24
MLM-ALL	0.70	0.68	0.28	0.25
LM	0.73	0.71	0.30	0.26
SDM	0.79	0.77	0.34	0.30
LM+ELR	0.71	0.69	0.29	0.26
SDM-ELR	0.79	0.78	0.34	0.30
MLM-CA	0.75	0.73	0.32	0.28
BM25-CA	0.76	0.74	0.32	0.29
FSDM	0.73	0.71	0.32	0.28
BM25F-CA	0.74	0.72	0.33	0.29
FSDM-ELR	0.74	0.71	0.32	0.28

compares the performance of the original search systems (Orig column) with that of the eRank-based re-ranking strategy (eRank column), using nDCG@10 and nDCG@100 as evaluation metrics. The results show that applying eRank does not degrade retrieval effectiveness, while contributing to improved veracity. In some cases, eRank even yields modest gains in nDCG.

To confirm the retrieval stability of eRank, we assessed whether it introduces statistically significant performance differences compared to the original rankings. The goal is to verify that eRank preserves effectiveness while promoting higher-veracity entities. To this end, we computed Cohen's d to measure effect size and performed a paired t-test. Across all systems, effect sizes remained well below 0.2 and p-values exceeded 0.01 in nearly all cases, indicating no significant difference in retrieval performance between the original and eRank-based rankings.

To verify that the ranking changes introduced by eRank are substantial, we computed Kendall's Tau (τ) and Tau Average Precision (τ_{AP}) between the original and eRank-based rankings. As shown in Table 2, all runs yielded low τ and τ_{AP} values. Low τ values indicate that eRank significantly alters the original ranking order, while low τ_{AP} values highlight that these changes affect the top-ranked positions. Combined with stable nDCG scores, these results suggest that eRank generates meaningfully different rankings that maintain relevance to the query while promoting higher-veracity entities.

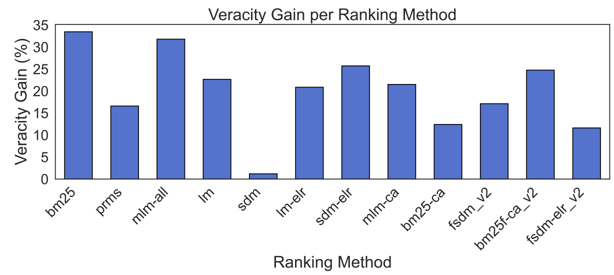

Figure 4: Veracity gain of entity search systems on the DBpedia-Entity-v2 collection after eRank re-ranking.

Table 3: ALADDIN performance for defect detection. $\mathcal{A}_{\text{off}}$ refers to ALADDIN trained only offline, while $\mathcal{A}_{\text{on}}$ consider also the online training.

	binary F1	weighted F1	balanced accuracy
$\mathcal{A}_{\text{off}}$	0.79	0.70	0.63
$\mathcal{A}_{\text{on}}$	0.83	0.72	0.64

Finally, we conduct a fact-level analysis to examine how the entity-level re-ranking introduced by eRank affects the positioning of higher-veracity facts. Specifically, using the partition-based veracity estimates computed for DBpedia 2015-10 by [58], we quantify veracity gain as the relative improvement in rank positions of facts from higher-veracity partitions after applying eRank, compared to their original positions. This allows us to assess whether prioritizing higher-veracity entities also promotes more trustworthy facts. As shown in Figure 4, all tested entity search systems benefit from eRank, with improvements reaching up to 34%. These results confirm that eRank not only enhances entity-level veracity, but also boosts higher-veracity facts, thus promoting trust at scale.

RQ1: Take-home message

eRank promotes entities with higher-veracity facts to top ranks, providing a re-ranking strategy that enhances ranking quality while maintaining the overall relevance to the query untouched.

RQ2: Veracity effects in action. To explore how veracity signals can support decision-making and veracity control in entity-centric tasks, we begin with *defect detection* (Task 2). For this task, we assess the effect of online active learning and human-in-the-loop supervision by comparing ALADDIN trained with online feedback to its offline counterpart. Table 3 reports performance on the DBpedia-Defect test set, evaluated using {binary, weighted} F1 and balanced accuracy. All metrics indicate improved performance with active online training. The binary F1 increases from 0.79 to 0.83, showing that user feedback helps the model refine its predictions. Both the weighted F1 and balanced accuracy also rise, reflecting more effective handling of class imbalance.

We next examine the effects of ALADDIN's veracity scores in *entity summarization* (Task 3), where they are used to re-rank the outputs of RELIN and DynES on the DBpedia-Fact collection. Table 4 compares the original, vRank-based, and ALADDIN-based rankings. Differences in nDCG@5 and nDCG@10 are minimal,

Table 4: Analysis of the DynES and RELIN runs in terms of importance, relevance, and utility. Original, vRank and ALADDIN ($\mathcal{A}$Rank) are the three ranking approaches considered.

	Importance						Relevance						Utility					
	nDCG@5			nDCG@10			nDCG@5			nDCG@10			nDCG@5			nDCG@10		
Model	Orig	vRank	$\mathcal{A}$Rank	Orig	vRank	$\mathcal{A}$Rank	Orig	vRank	$\mathcal{A}$Rank	Orig	vRank	$\mathcal{A}$Rank	Orig	vRank	$\mathcal{A}$Rank	Orig	vRank	$\mathcal{A}$Rank
RELIN	0.48	0.52	0.49	0.53	0.56	0.54	0.36	0.37	0.32	0.43	0.43	0.40	0.47	0.50	0.46	0.54	0.56	0.53
DynES/i	0.79	0.79	0.78	0.80	0.81	0.79	0.47	0.48	0.46	0.54	0.55	0.53	0.72	0.73	0.70	0.76	0.76	0.75
DyneES/r	0.58	0.58	0.56	0.62	0.62	0.60	0.53	0.53	0.51	0.58	0.58	0.56	0.62	0.62	0.60	0.66	0.66	0.64
DynES/u	0.77	0.77	0.76	0.78	0.77	0.79	0.58	0.59	0.56	0.65	0.65	0.62	0.76	0.76	0.74	0.79	0.79	0.77

Table 5: Ranking correlations for the entity summarization models, between the original ranking and vRank, and between the original ranking and ALADDIN ($\mathcal{A}$Rank).

	cutoff@5				cutoff@10			
	vRank		$\mathcal{A}$Rank		vRank		$\mathcal{A}$Rank	
Model	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}	τ	τ_{AP}
RELIN	0.98	0.94	0.06	0.05	0.98	0.95	0.10	0.05
DynES/i	0.80	0.79	0.28	0.26	0.81	0.79	0.27	0.25
DynES/r	0.89	0.90	0.24	0.25	0.79	0.80	0.20	0.25
DynES/u	0.84	0.77	0.27	0.24	0.84	0.77	0.27	0.17

with no statistically significant variation (Cohen's $d < 0.2$, p-value > 0.01), indicating that veracity-driven re-ranking does not compromise effectiveness. Notably, this finding holds across all considered evaluation dimensions – i.e., importance, relevance , and utility.

While both vRank and ALADDIN maintain performance, correlation metrics (τ and τ_{AP} in Table 5) reveal a key distinction: vRank largely preserves the original ranking order, offering only marginal veracity gains, whereas ALADDIN introduces substantial reordering. This confirms the limitations of vRank (cf. Section 3) and underscores ALADDIN's ability to prioritize higher-veracity facts without compromising ranking effectiveness.

Given the substantial ordering differences introduced by ALADDIN-based rankings, we assess their impact on *entity card generation* (Task 4). We analyzed annotations from four experts who assessed 50 pairs of entity cards generated from DBpedia-Fact rankings. Based on majority voting for each pair, ALADDIN-generated cards were preferred in 20 cases (40%), the originals in 13 (26%), and no preference was expressed in 17 cases (34%). These results indicate that ALADDIN improves or preserves user perception without detriment in 74% of the cases.

Finally, given the inexpensive, transparent, and explainable nature of ALADDIN, we evaluate its utility for *defect recommendation* (Task 5) – a critical first step for downstream error analysis. Specifically, we use ALADDIN to compute the veracity scores of facts and assess how effectively these scores guide the identification of potentially defective or low-veracity parts of the KG.

To test ALADDIN under varying resource constraints, we evaluate its performance on the DBpedia-Defect test set by measuring precision and recall at different selection budgets – specifically, the top 10%, 20%, 25%, and 50% of facts ranked from lowest to highest predicted veracity. Table 6 reports results for both the standard (online) ALADDIN model and its offline variant. Following the DBpedia partitioning defined in [58], evaluation is conducted over three subsets: the full test set (*All*), 253 facts from low-veracity partitions (*Low*), and 258 facts from high-veracity partitions (*High*).

Table 6: Precision and recall for ALADDIN before ($\mathcal{A}_{\text{off}}$) and after online training ($\mathcal{A}_{\text{on}}$) across different budget levels. We report performance for all facts (*All*), and the least (*Low*) and most popular partitions (*High*). The % gain column shows the improvement from offline to online.

		Precision			Recall		
Budget	Subset	$\mathcal{A}_{\text{off}}$	$\mathcal{A}_{\text{on}}$	% gain	$\mathcal{A}_{\text{off}}$	$\mathcal{A}_{\text{on}}$	% gain
10%	All	0.42	0.42	–	0.65	0.65	–
	Low	0.60	0.66	+10.00	0.57	0.65	+14.00
	High	0.54	0.60	+11.10	0.50	0.59	+18.00
20%	All	0.34	0.77	+126.50	0.58	0.61	+5.00
	Low	0.64	0.67	+4.68	0.64	0.68	+6.25
	High	0.70	0.71	+1.43	0.68	0.73	+7.35
25%	All	0.29	0.66	+127.58	0.54	0.63	+16.66
	Low	0.66	0.69	+4.54	0.67	0.70	+4.47
	High	0.73	0.73	–	0.72	0.75	+4.16
50%	All	0.61	0.66	+8.19	0.58	0.66	+13.80
	Low	0.67	0.69	+2.98	0.67	0.71	+5.97
	High	0.71	0.72	+1.40	0.71	0.75	+5.63

Results show that online ALADDIN consistently outperforms its offline counterpart, particularly in identifying low-veracity facts under constrained budgets. At 10% budget, both versions perform similarly on the full test set, but the online variant achieves higher precision and recall for both *Low* and *High* subsets. At 20%, it exhibits a sharp increase in *All* precision (from 0.34 to 0.77), along with gains in recall. Improvements occur also across veracity-specific subsets. As the budget increases to 25% and 50%, the advantages of online learning persist, confirming its effectiveness.

These findings underscore ALADDIN's effectiveness in defect recommendation, even under strict budget constraints. As such, prioritizing the verification of its recommended facts can trigger the deployment of filtering or correction mechanisms when data veracity issues are detected.

RQ2: Take-home message

ALADDIN operationalizes veracity signals in entity search, enabling defect detection across entity-centric tasks and enhancing data veracity without compromising search effectiveness.

7 Related Work

KG and Data Veracity. KGs support information integration and semantic organization, playing a pivotal role in applications such as information retrieval, question answering, recommender systems, and so on. Prominent examples include DBpedia [3], YAGO [77], and Wikidata [81], all comprising millions of entities and from hundreds of millions to billions of facts. Due to their large-scale and often automated construction, KGs frequently lack meticulous curation, which leads to issues such as incompleteness, inconsistencies, and noise [86]. As a result, the reliability of the information they contain cannot always be guaranteed, making data veracity assessment a critical aspect [1].

Veracity assessment in KGs is essential for ensuring their reliability, yet it remains relatively underexplored. The standard approach relies on manually auditing the veracity of KG facts – a process that is infeasible for large and dynamically evolving KGs. To overcome this limitation, a range of efficient and cost-effective techniques have emerged. These include sampling methods [28, 56, 57], inference-based crowdsourcing approaches [63], cross-KG fact validation techniques [51], human-machine collaborative systems [69], and utility-oriented estimation strategies [58, 59].

Among its applications, veracity assessment plays a crucial role in mitigating misinformation within information retrieval tasks [60]. This line of work is often framed within the broader paradigm of *credible retrieval*, which focuses on retrieving accurate and trustworthy information from sources that meet established reliability standards [30]. Notable efforts in this direction include the ROMCIR workshop series [41, 66, 67, 72], the TREC Health Misinformation tracks [18, 19], and the CLEF eHealth CHS tasks [31, 40]. Yet, despite their relevance, these efforts have not directly examined the contribution of KGs or their veracity to credible retrieval.

Beyond veracity assessment, error detection mechanisms are also essential for supporting entity-centric tasks in KGs. Existing approaches to error detection can be categorized into four groups: (i) outlier detection using statistical and machine learning techniques [64, 65, 85]; (ii) external validation through comparison with web data or other KGs [27, 44, 46]; (iii) graph exploration and structural analysis [38]; and (iv) error diagnosis via ontology reasoning and formal constraints [47]. Despite their utility, these methods often suffer from scalability limitations, high computational costs, and heavy dependence on ontology reasoning, resulting impractical for modern, large-scale KGs. Even LLM-based solutions are not mature yet to address defect detection effectively – both when used as standalone approaches [58] and within RAG-based pipelines [73].

Entity-centric tasks. KG entities play a crucial role in web search, providing the structured semantic data essential for entity search, summarization, and card generation [35].

The **entity search** task has witnessed a growing range of methods, from traditional text-based approaches [5, 34] to learning-based methods [15, 23, 52]. In this context, various datasets have been introduced to support benchmarking and comparison of entity search approaches. The most notable are the DBpedia-Entity v1 [6] and DBpedia-Entity v2 [35]. More recently, Arabzadeh et al. [2] released LaQuE, a framework for entity search with real-user queries. LaQuE builds on queries from the ORCAS dataset [21] and maps them to DBpedia 2015-10 entities. However, in contrast to DBpedia-Entity v2, which features an average of 36 relevant entities per query, the LaQuE dataset contains only an average of 1.08 relevant entities per query. This makes DBpedia-Entity v2 better suited to investigate the effects of veracity-driven re-ranking strategies, as changes in ranking can have a tangible impact on the ordering of relevant entities. Therefore, we adopt DBpedia-Entity v2 as the reference collection for our experiments, offering a more robust basis for evaluating the impact of veracity on entity search.

Another key task is **entity summarization**, which involves identifying the most salient facts about an entity to generate an optimal, size-constrained summary comprising a subset of relevant information [49]. A variety of approaches have been proposed, leveraging strategies such as PageRank-based ranking [79, 80], hierarchical clustering [32], and deep neural networks [26, 42, 45, 50, 61, 83]. Given our focus on the role of veracity in entity search and its downstream applications, we restrict our attention to a specific subtask: *query-dependent entity summarization* [33], which entails ranking the facts of an entity based on both their importance for the entity and relevance to the query. In this setting, two methods represent the state-of-the-art: RELIN [16], a PageRank-inspired approach, and DynES [33], a learning-to-rank method.

Building on entity summaries, **entity cards** have been shown to influence search behavior. Bota et al. [12] found out that users engage more with relevant cards, boosting interactions with organic search results. Navalpakkam et al. [62] showed that entity cards reduce information-seeking time as they contain relevant content. Shokouhi and Guo [74] analyzed how proactive card usage is affected by time and local factors. Hasibi et al. [33] further showed that users favor query-dependent summaries over static ones. In healthcare, Jimmy et al. [39] observed that users prioritize the first entity card when seeking condition-related information.

8 Conclusions and Future Work

In this work, we analyzed the impact of KG veracity on entity search. We introduced *e*Rank, an entity-level, veracity-driven re-ranking strategy that prioritizes high-quality entities in search results without compromising retrieval effectiveness.

To support veracity assessment at scale, we proposed ALADDIN – a lightweight system for defect detection. ALADDIN combines a linear regression model trained offline with an online active learning mechanism powered by human feedback. During the online phase, the system presents users with entity cards featuring the top-ranked facts about an entity, enabling fast and intuitive accuracy assessment. By prioritizing feedback on the most relevant facts, this strategy guides the learning process toward correcting the most significant errors, making user input both targeted and efficient.

Extensive experiments confirmed the effectiveness of ALADDIN across entity-centric tasks. The use of online active learning improved both defect detection and recommendation, even under tight annotation budgets. For entity summarization and card generation, ALADDIN showed greater sensitivity to fact quality and enhanced the perceived content relevance. Together, these results highlight the value of fine-grained veracity estimation and the benefits of lightweight, feedback-driven systems for enhancing KG reliability.

As future work, we aim to integrate ALADDIN into KG error correction pipelines, enabling precise error detection, quantification, and feedback-driven corrections.

Acknowledgments

The work was supported by the HEREDITARY project, as part of the EU Horizon Europe program under Grant Agreement 101137074.

GenAI Usage Disclosure

In accordance with ACM's guidelines on the use of generative AI tools, we disclose that generative AI technologies were used solely to assist in grammar checking and sentence rephrasing during the preparation of this paper.

References

[1] Z. Abedjan, X. Chu, D. Deng, R. Castro Fernandez, I. F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, and N. Tang. 2016. Detecting Data Errors: Where are we and what needs to be done? *Proc. VLDB Endow.* 9, 12 (2016), 993–1004.

[2] N. Arabzadeh, A. Bigdeli, and E. Bagheri. 2024. LaQuE: Enabling Entity Search at Scale. In *Proc. of the 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024 (Lecture Notes in Computer Science, Vol. 14609).* Springer, 270–285. https://doi.org/10.1007/978-3-031-56060-6_18

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007 (Lecture Notes in Computer Science, Vol. 4825).* Springer, 722–735. https://doi.org/10.1007/978-3-540-76298-0_52

[4] K. Balog. 2018. *Entity-Oriented Search.* The Information Retrieval Series, Vol. 39. Springer. https://doi.org/10.1007/978-3-319-93935-3

[5] K. Balog, E. Meij, and M. de Rijke. 2010. Entity search: building bridges between two worlds. In *Proceedings of the 3rd International Semantic Search Workshop, SEMSEARCH '10, Raleigh, North Carolina, USA, April 26, 2010.* ACM, 9:1–9:5. https://doi.org/10.1145/1863879.1863888

[6] K. Balog and R. Neumayer. 2013. A test collection for entity search in DBpedia. In *Proc. of SIGIR.* ACM, 737–740. https://doi.org/10.1145/2484028.2484165

[7] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. 2020. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30063–30070. https://doi.org/10.1073/pnas.1907378117

[8] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi. 2015. From Data Quality to Big Data Quality. *J. Database Manag.* 26, 1 (2015), 60–82. https://doi.org/10.4018/JDM.2015010103

[9] A. G. Baydin, R. Cornish, D. Martínez-Rubio, M. Schmidt, and F. D. Wood. 2017. Online Learning Rate Adaptation with Hypergradient Descent. *CoRR* abs/1703.04782 (2017). arXiv:1703.04782 http://arxiv.org/abs/1703.04782

[10] M. Belkin, D. Hsu, S. Ma, and S. Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854. https://doi.org/10.1073/pnas.1903070116

[11] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* 2787–2795. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

[12] H. S. Bota, K. Zhou, and J. M. Jose. 2016. Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload. In *Proc. of CHIIR.* ACM, 131–140. https://doi.org/10.1145/2854946.2854967

[13] A.Z. Broder and P. McAfee. 2023. Delphic Costs and Benefits in Web Search: A utilitarian and historical analysis. *CoRR* abs/2308.07525 (2023). https://doi.org/10.48550/ARXIV.2308.07525 arXiv:2308.07525

[14] L. Cai and Y. Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* 14 (2015), 2. https://doi.org/10.5334/DSJ-2015-002

[15] S. Chatterjee and L. Dietz. 2019. Why does this Entity matter?: Support Passage Retrieval for Entity Retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019.* ACM, 221–224. https://doi.org/10.1145/3341981.3344243

[16] G. Cheng, T. Tran, and Y. Qu. 2011. RELIN: Relatedness and Informativeness-Based Centrality for Entity Summarization. In *Proc. of ISWC (Lecture Notes in Computer Science, Vol. 7031).* Springer, 114–129. https://doi.org/10.1007/978-3-642-25073-6_8

[17] P. Cimiano and H. Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web* 8, 3 (Jan. 2017), 489–508. https://doi.org/10.3233/SW-160218

[18] C. L. A. Clarke, M. Maistro, and M. D. Smucker. 2021. Overview of the TREC 2021 Health Misinformation Track. In *Proc. of TREC (NIST Special Publication, Vol. 500-335).* National Institute of Standards and Technology (NIST).

[19] C. L. A. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, and G. Zuccon. 2020. Overview of the TREC 2020 Health Misinformation Track. In *Proc of TREC (NIST Special Publication, Vol. 1266).* National Institute of Standards and Technology (NIST).

[20] W. G. Cochran. 1977. *Sampling Techniques, 3rd Edition.* John Wiley. https://doi.org/10.1017/S0013091500025724

[21] N. Craswell, D. Campos, B. Mitra, E. Yilmaz, and B. Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. *CoRR* abs/2006.05324 (2020). arXiv:2006.05324 https://arxiv.org/abs/2006.05324

[22] T. Dalenius and J. L. Hodges. 1959. Minimum Variance Stratification. *J. Am. Stat. Assoc.* 54, 285 (1959), 88–101. https://doi.org/10.1080/01621459.1959.10501501

[23] J. Dalton, L. Dietz, and J. Allan. 2014. Entity query feature expansion using knowledge base links. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014.* ACM, 365–374. https://doi.org/10.1145/2600428.2609628

[24] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. 2013. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proc. of SIGMOD.* ACM, 1209–1220. https://doi.org/10.1145/2463676.2465297

[25] S. Faralli, A. Lenzi, and P. Velardi. 2023. A Benchmark Study on Knowledge Graphs Enrichment and Pruning Methods in the Presence of Noisy Relationships. *J. Artif. Intell. Res.* 78 (2023), 37–68. https://doi.org/10.1613/JAIR.1.14494

[26] A. F. Firmansyah, D. Moussallem, and A.-C. Ngonga Ngomo. 2021. GATES: Using Graph Attention Networks for Entity Summarization. In *K-CAP '21: Knowledge Capture Conference, Virtual Event, USA, December 2-3, 2021.* ACM, 73–80. https://doi.org/10.1145/3460210.3493574

[27] M. H. Gad-Elrab, D. Stepanova, J. Urbani, and G. Weikum. 2019. Tracy: Tracing Facts over Knowledge Graphs and Text. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3516–3520. https://doi.org/10.1145/3308558.3314126

[28] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, and J. Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (2019), 1679–1691.

[29] D. Garigliotti, D. Albakour, M. Martinez, and K. Balog. 2019. Unsupervised Context Retrieval for Long-tail Entities. In *Proc. of ICTIR.* ACM, 225–228. https://doi.org/10.1145/3341981.3344244

[30] A. L. Gînsca, A. Popescu, and M. Lupu. 2015. Credibility in Information Retrieval. *Found. Trends Inf. Retr.* 9, 5 (2015), 355–475. https://doi.org/10.1561/1500000046

[31] L. Goeuriot, H. Suominen, G. Pasi, E. Bassani, N. Brew-Sam, G. N. González Sáez, L. Kelly, P. Mulhem, S. Seneviratne, R. Upadhyay, M. Viviani, and C. Xu. 2021. Consumer Health Search at CLEF eHealth 2021. In *Proc. of the Working Notes of CLEF (CEUR Workshop Proceedings, Vol. 2936).* CEUR-WS.org, 751–769.

[32] K. Gunaratna, K. Thirunarayan, and A. P. Sheth. 2015. FACES: Diversity-Aware Entity Summarization Using Incremental Hierarchical Conceptual Clustering. In *Proc. of AAAI.* AAAI Press, 116–122. https://doi.org/10.1609/AAAI.V29I1.9180

[33] F. Hasibi, K. Balog, and S. E. Bratsberg. 2017. Dynamic Factual Summaries for Entity Cards. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017.* ACM, 773–782. https://doi.org/10.1145/3077136.3080810

[34] F. Hasibi, K. Balog, D. Garigliotti, and S. Zhang. 2017. Nordlys: A Toolkit for Entity-Oriented and Semantic Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017.* ACM, 1289–1292. https://doi.org/10.1145/3077136.3084149

[35] F. Hasibi, F. Nikolaev, C. Xiong, K. Balog, S. E. Bratsberg, A. Kotov, and J. Callan. 2017. DBpedia-Entity v2: A Test Collection for Entity Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017.* ACM, 1265–1268. https://doi.org/10.1145/3077136.3080751

[36] A. Hur, N. Janjua, and M. Ahmed. 2021. A Survey on State-of-the-art Techniques for Knowledge Graphs Construction and Challenges ahead. In *Fourth IEEE International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2021, Laguna Hills, CA, USA, December 1-3, 2021.* IEEE, 99–103. https://doi.org/10.1109/AIKE52691.2021.00021

[37] F. Ilievski, P. Vossen, and S. Schlobach. 2018. Systematic Study of Long Tail Phenomena in Entity Linking. In *Proc. of COLING.* Association for Computational Linguistics, 664–674. https://aclanthology.org/C18-1056/

[38] S. Jia, Y. Xiang, X. Chen, K. Wang, and S. E. 2019. Triple Trustworthiness Measurement for Knowledge Graph. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019.* ACM, 2865–2871. https://doi.org/10.1145/3308558.3313586

[39] Jimmy, G. Zuccon, B. Koopman, and G. Demartini. 2019. Health Cards to Assist Decision Making in Consumer Health Search. In *Proc. of AMIA.* AMIA. https://knowledge.amia.org/69862-amia-1.4570936/t005-1.4574828/t005-1.4574829/3201885-1.4574890/3201686-1.4574887

[40] Jimmy, G. Zuccon, J. R. M. Palotti, L. Goeuriot, and L. Kelly. 2018. Overview of the CLEF 2018 Consumer Health Search Task. In *Working Notes of CLEF (CEUR Workshop Proceedings, Vol. 2125).* CEUR-WS.org.

[41] U. Kruschwitz, M. Petrocchi, and M. Viviani. 2025. ROMCIR 2025: Overview of the 5th Workshop on Reducing Online Misinformation Through Credible Information Retrieval. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 15576).* Springer, 339–344. https://doi.org/10.1007/978-3-031-88720-8_52

[42] A. Kushk and K. Kochut. 2021. ESDL: Entity Summarization with Deep Learning. In *IJCKG'21: The 10th International Joint Conference on Knowledge Graphs, Virtual Event, Thailand, December 6 - 8, 2021*. ACM, 186–190. https://doi.org/10.1145/3502223.3502743

[43] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195. https://doi.org/10.3233/SW-140134

[44] F. Li, X. L. Dong, A. Langen, and Y. Li. 2017. Knowledge Verification for Long-Tail Verticals. *Proc. VLDB Endow.* 10, 11 (2017), 1370–1381.

[45] J. Li, G. Cheng, Q. Liu, W. Zhang, E. Kharlamov, K. Gunaratna, and H. Chen. 2020. Neural Entity Summarization with Joint Encoding and Weak Supervision. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 1644–1650. https://doi.org/10.24963/IJCAI.2020/228

[46] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. 2015. Truth Finding on the Deep Web: Is the Problem Solved? *CoRR* abs/1503.00303 (2015). arXiv:1503.00303 http://arxiv.org/abs/1503.00303

[47] J. Liang, Y. Xiao, Y. Zhang, S. Hwang, and H. Wang. 2017. Graph-Based Wrong IsA Relation Detection in a Large-Scale Lexical Taxonomy. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder Singh and Shaul Markovitch (Eds.). AAAI Press, 1178–1184. https://doi.org/10.1609/AAAI.V31I1.10676

[48] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988. https://doi.org/10.1109/ICCV.2017.324

[49] Q. Liu, G. Cheng, K. Gunaratna, and Y. Qu. 2021. Entity summarization: State of the art and future challenges. *J. Web Semant.* 69 (2021), 100647. https://doi.org/10.1016/J.WEBSEM.2021.100647

[50] Q. Liu, G. Cheng, and Y. Qu. 2020. DeepLENS: Deep Learning for Entity Summarization. In *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2020) co-located with the 17th Extended Semantic Web Conference 2020 (ESWC 2020), Heraklion, Greece, June 02, 2020 - moved online (CEUR Workshop Proceedings, Vol. 2635)*. CEUR-WS.org. https://ceur-ws.org/Vol-2635/paper2.pdf

[51] S. Liu, M. d'Aquin, and E. Motta. 2017. Measuring Accuracy of Triples in Knowledge Graphs. In *Proc. of LDK (LNCS, Vol. 10318)*. Springer, 343–357. https://doi.org/10.1007/978-3-319-59888-8_29

[52] X. Liu and H. Fang. 2015. Latent entity space: a novel retrieval approach for entity-bearing queries. *Inf. Retr. J.* 18, 6 (2015), 473–503. https://doi.org/10.1007/S10791-015-9267-X

[53] T. Lukoianova and V. L. Rubin. 2014. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances in Classification Research Online* 24, 1 (2014), 4–15. https://doi.org/10.7152/acro.v24i1.14671

[54] N. G. Marchant and B. I. P. Rubinstein. 2017. In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling. *Proc. VLDB Endow.* 10, 11 (2017), 1322–1333.

[55] N. G. Marchant and B. I. P. Rubinstein. 2021. Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance. In *Proc. of KDD*. ACM, 1180–1190. https://doi.org/10.1145/3447548.3467435

[56] S. Marchesin and G. Silvello. 2024. Efficient and Reliable Estimation of Knowledge Graph Accuracy. *Proc. VLDB Endow.* 17, 9 (2024), 2392–2404. https://doi.org/10.14778/3665844.3665865

[57] S. Marchesin and G. Silvello. 2025. Credible Intervals for Knowledge Graph Accuracy Estimation. *Proc. ACM Manag. Data* 3, 3 (2025), 142:1–142:26. https://doi.org/10.1145/3725279

[58] S. Marchesin, G. Silvello, and O. Alonso. 2024. Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 12. 105–114. https://doi.org/10.1609/hcomp.v12i1.31605

[59] S. Marchesin, G. Silvello, and O. Alonso. 2024. Veracity Estimation for Entity-Oriented Search with Knowledge Graphs. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*. ACM, 1649–1659. https://doi.org/10.1145/3627673.3679561

[60] G. C. Milanese, G. Peikos, G. Pasi, and M. Viviani. 2025. Fact-Driven Health Information Retrieval: Integrating LLMs and Knowledge Graphs to Combat Misinformation. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 15574)*. Springer, 192–200. https://doi.org/10.1007/978-3-031-88714-7_17

[61] A. Moradan, M. Sorkhpar, A. Miyauchi, D. Mottin, and I. Assent. 2025. Untapping the Power of Indirect Relationships in Entity Summarization. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM 2025, Hannover, Germany, March 10-14, 2025*. ACM, 820–828. https://doi.org/10.1145/3701551.3703566

[62] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. J. Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proc. of WWW*. International World Wide Web Conferences Steering Committee / ACM, 953–964. https://doi.org/10.1145/2488388.2488471

[63] P. Ojha and P. P. Talukdar. 2017. KGEval: Accuracy Estimation of Automatically Constructed Knowledge Graphs. In *Proc. of EMNLP*. ACL, 1741–1750. https://doi.org/10.18653/v1/d17-1183

[64] H. Paulheim. 2014. Identifying Wrong Links between Datasets by Multi-dimensional Outlier Detection. In *Proceedings of the Third International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2014, co-located with 11th Extended Semantic Web Conference (ESWC 2014), Anissaras/Hersonissou, Greece, May 26, 2014 (CEUR Workshop Proceedings, Vol. 1162)*. CEUR-WS.org, 27–38. https://ceur-ws.org/Vol-1162/paper3.pdf

[65] H. Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8, 3 (2017), 489–508. https://doi.org/10.3233/SW-160218

[66] M. Petrocchi and M. Viviani. 2023. ROMCIR 2023: Overview of the 3rd Workshop on Reducing Online Misinformation Through Credible Information Retrieval. In *Proc. of ECIR (Lecture Notes in Computer Science, Vol. 13982)*. Springer, 405–411. https://doi.org/10.1007/978-3-031-28241-6_45

[67] M. Petrocchi and M. Viviani. 2024. ROMCIR 2024: Overview of the 4th Workshop on Reducing Online Misinformation Through Credible Information Retrieval. In *Proc. of ECIR (Lecture Notes in Computer Science, Vol. 14612)*. Springer, 403–408. https://doi.org/10.1007/978-3-031-56069-9_54

[68] J. Pujara, E. Augustine, and L. Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *Proc. of EMNLP*. ACL, 1751–1756. https://doi.org/10.18653/v1/d17-1184

[69] Y. Qi, W. Zheng, L. Hong, and L. Zou. 2022. Evaluating Knowledge Graph Accuracy Powered by Optimized Human-Machine Collaboration. In *Proc. of SIGKDD 2022*. ACM, 1368–1378. https://doi.org/10.1145/3534678.3539233

[70] R. Reinanda, E. Meij, and M. de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Found. Trends Inf. Retr.* 14, 4 (2020), 289–444. https://doi.org/10.1561/1500000063

[71] S. Salimzadeh, D. Maxwell, and C. Hauff. 2021. The Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proc. of ICTIR*. ACM, 63–72. https://doi.org/10.1145/3471158.3472255

[72] F. Saracco and M. Viviani. 2021. ROMCIR 2021: Reducing Online Misinformation through Credible Information Retrieval. In *Proc. of ECIR (Lecture Notes in Computer Science, Vol. 12657)*. Springer, 714–717. https://doi.org/10.1007/978-3-030-72240-1_87

[73] F. Shami, S. Marchesin, and G. Silvello. 2025. Fact Verification in Knowledge Graphs Using LLMs. In *Proc. of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*. ACM, 3985–3989. https://doi.org/10.1145/3726302.3730142

[74] M. Shokouhi and Q. Guo. 2015. From Queries to Cards: Re-ranking Proactive Card Recommendations Based on Reactive Search History. In *Proc. of SIGIR*. ACM, 695–704. https://doi.org/10.1145/2766462.2767705

[75] A. Strehl and M. Littman. 2007. Online linear regression and its application to model-based reinforcement learning. *Advances in Neural Information Processing Systems* 20 (2007), 737–744. https://doi.org/10.5555/2981562.2981740

[76] F. Suchanek, M. Alam, T. Bonald, L. Chen, P.-H. Paris, and J. Soria. 2024. YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy. In *Proc. of SIGIR 2024*. ACM.

[77] F. M. Suchanek, M. Alam, T. Bonald, L. Chen, P.-H. Paris, and J. Soria. 2024. YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*. ACM, 131–140. https://doi.org/10.1145/3626772.3657876

[78] I. Taleb, M. A. Serhani, and R. Dssouli. 2018. Big Data Quality: A Survey. In *2018 IEEE International Congress on Big Data, BigData Congress 2018*. IEEE Computer Society, 166–173. https://doi.org/10.1109/BIGDATACONGRESS.2018.00029

[79] A. Thalhammer, N. Lasierra, and A. Rettinger. 2016. LinkSUM: Using Link Analysis to Summarize Entity Data. In *Proc. of ICWE (Lecture Notes in Computer Science, Vol. 9671)*. Springer, 244–261. https://doi.org/10.1007/978-3-319-38791-8_14

[80] A. Thalhammer and A. Rettinger. 2014. Browsing DBpedia Entities with Summaries. In *Proc. of ESWC Satellite Events (Lecture Notes in Computer Science, Vol. 8798)*. Springer, 511–515. https://doi.org/10.1007/978-3-319-11955-7_76

[81] D. Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489

[82] Y. Wang, F. Ma, and J. Gao. 2020. Efficient Knowledge Graph Validation via Cross-Graph Representation Learning. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1595–1604. https://doi.org/10.1145/3340531.3411902

[83] D. Wei, Y. Liu, F. Zhu, L. Zang, W. Zhou, J. Han, and S. Hu. 2019. ESA: Entity Summarization with Attention. In *Proceedings of the 2nd International Workshop on EntitY REtrieval co-located with 28th ACM International Conference on Information and Knowledge Management (CIKM 2019), Beijing, China, November 3, 2019 (CEUR Workshop Proceedings, Vol. 2446)*. CEUR-WS.org, 40–44. http://ceur-ws.org/Vol-2446/paper6.pdf

[84] G. Weikum, X. L. Dong, S. Razniewski, and F. M. Suchanek. 2021. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Found. Trends Databases* 10, 2-4 (2021), 108–490. https://doi.org/10.1561/1900000064

[85] D. Wienand and H. Paulheim. 2014. Detecting Incorrect Numerical Data in DBpedia. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings (Lecture Notes in Computer Science, Vol. 8465)*, Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai (Eds.). Springer, 504–518. https://doi.org/10.1007/978-3-319-07443-6_34

[86] B. Xue and L. Zou. 2023. Knowledge Graph Quality Management: A Comprehensive Survey. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4969–4988. https://doi.org/10.1109/TKDE.2022.3150080