ISEE: Interactive Semantic Enrichment for Database Fields

Anonymous ACL submission

Abstract

LLM-based agents are increasingly being deployed for data-related tasks, including data sense-making, exploration, and retrieval. However, their performance heavily depends on the clarity and completeness of data semantics. Unfortunately, in practice, many field descriptions remain ambiguous or incomplete, as much of the essential context (e.g., the meaning of a customized field) originates from users' domain knowledge and is rarely documented publicly. This gap restricts the agents' task performance in downstream tasks, such as entity-linking. To bridge this gap, we introduce a novel and comprehensive **I**nteractive **SE**mantic **E**nrichment system (ISEE). Given a data field description, ISEE measures its quality through a scoring system, gathers domain knowledge, and collaboratively enriches the semantics with users. Through a user study, automated user simulation, quantitative evaluation, and case study, we demonstrate that ISEE significantly reduces cognitive load, improves description quality, and enhances downstream task performance.

1 Introduction

LLM-based agents are increasingly deployed in enterprise tasks such as entity linking (Shen et al., 2015; Kolitsas et al., 2018) and natural language querying of databases (Ning et al., 2023; Tian et al., 2023; Ning et al., 2024; Tian and Zhang, 2026). While academic environments assume well-documented or publicly available context, enterprise environments pose distinct challenges. Business databases often contain highly customized fields whose meaning is privately owned by customers. Consequently, when field descriptions are ambiguous or incomplete, LLMs face challenges in interpreting their meaning, leading to errors in downstream tasks. For example, when a user asks a natural language question, the entity-linking service is expected to accurately map the query to the corresponding mentioned fields. However, given a user-defined abbreviation name with unclear semantics, LLMs can only guess. This is known as out-of-distribution (Yang et al., 2024), where LLMs fail to respond when the required knowledge is outside their parametric space. In addition to training-based domain adaptation (Tian et al., 2025; Zhang et al., 2025b; Deotale et al., 2026), semantic enrichment (Belsky et al., 2016; Cai et al., 2025; Xue et al., 2021) can efficiently mitigate this issue by explicitly describing the missing semantics in documentation. However, they come with inherent limitations in practice. Existing semantic enrichment methods assume that the enrichment can be achieved using the general knowledge embedded in LLMs, while many rely on domain-specific expertise from end-users. Without corresponding context, it is infeasible for LLMs to automatically infer the missing information and enrich these fields. Thus, we argue that data field enrichment has to keep the domain experts in the loop and actively

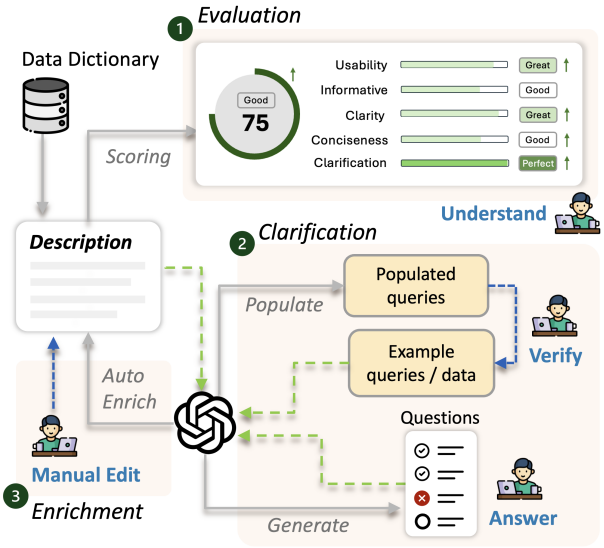

Figure 1: Pipeline of ISEE: (1) Users understand the quality of the current description by the scoring system. (2) Users provide feedback through clarification or query population. (3) Users refine the description.

solicit domain knowledge.

We present ISEE, a novel **I**nteractive **SE**mantic **E**nrichment system that automatically evaluates the quality of existing descriptions, efficiently elicits user knowledge, and collaboratively enhances descriptions with domain experts. ISEE introduces three key novel features. First, a *downstream-aware scoring system* comprehensively evaluates the quality of field descriptions across five dimensions, including usability, informativeness, clarity, conciseness, and readability. We develop unique algorithms for each dimension to measure a distinct aspect of the description quality. Second, a *conditional query population* component allows domain experts to enrich field descriptions by simply verifying LLM-populated natural queries related to this field. Third, a *taxonomy-guided clarification generator* leverages our proposed clarification taxonomy to generate specific, easy-to-answer questions to capture domain experts' knowledge.

We evaluated ISEE through a user study, an automated user simulation, a quantitative evaluation, and a case study. According to the user study, ISEE improved description enrichment accuracy by 119%, 32%, and 92% compared to manual editing, candidate selection, and ChatGPT, respectively, while significantly reducing cognitive load. According to the quantitative evaluation, the enriched descriptions further benefit downstream tasks, improving entity linking by 22% and text-to-SQL by 12% over automatic enrichment.

Our contributions include:

- A scoring system that evaluates the quality of a given description through five complementary aspects.
- A clarification question generator based on a newly proposed clarification taxonomy tailored for semantic enrichment.
- An interactive system that integrates scoring, clarification, and conditional query population into a unified workflow.
- A thorough evaluation including a user study, an automated user simulation, a quantitative evaluation, and a case study.

2 Related Work

2.1 Semantic Enrichment

Prior work has investigated automatically generating natural language descriptions for database schemas. Wretblad et al. (2024) generate column descriptions using a single LLM prompt and evaluate on text-to-SQL tasks. Several efforts extend this to both table and column levels (Zhang et al., 2025a; Anonymous, 2024; Singh et al., 2025; Gao and Luo, 2025): Zhang et al. (2025a) generate table-level descriptions to support readability, Anonymous (2024) employ an LLM-as-a-judge framework to select high-quality outputs, and Singh et al. (2025) leverage curated business glossaries for enrichment.

While these methods demonstrate the utility of schema descriptions, they rely on LLM general knowledge or static glossaries and generate descriptions in a single pass without incorporating domain expert feedback. In contrast, ISEE keeps the domain expert in the loop, iteratively eliciting and integrating user knowledge that cannot be inferred from public sources alone.

2.2 Human-AI Collaboration for Data

Human-AI collaborative systems combine human knowledge with AI capabilities to tackle tasks that neither can solve well alone (Amershi et al., 2019; Horvitz, 1999). In data management, recent systems adopt this paradigm for tasks such as SQL query construction (Tian et al., 2024), data exploration (Ning et al., 2023), and generation steering (Tian and Zhang, 2025), where users guide and validate AI-generated outputs. A key mechanism in such systems is clarification: asking targeted questions to resolve ambiguity and elicit missing information. Prior work on clarifying questions in information retrieval (Zamani et al., 2020a; Yang et al., 2025; Aliannejadi et al., 2019a) and NL interfaces (Purver, 2004; Aliannejadi et al., 2019b) has shown that well-structured questions significantly reduce misunderstanding. ISEE builds on these ideas by combining a clarification taxonomy with a multi-dimensional scoring system and query population, enabling a structured and iterative pipeline tailored for data field enrichment.

3 Method

3.1 System Overview

Figure 1 presents the pipeline of ISEE. Starting with field descriptions from a data dictionary (e.g., a relational database), ISEE includes three iterative stages—*Evaluation*, *Clarification*, and *Enrichment*. In the first stage (*Evaluation*), ISEE evaluates the quality of the current description. Users are pro-

vided with multiple evaluation scores, each measuring a distinct aspect of the description. Additionally, natural language (NL) explanations accompany each score to help users better understand the quality and current state of the description. This enables users to understand the situation and identify areas for improvement. In the second stage (*Clarification*), once users understand the current description, they can actively contribute related information (e.g., relevant NL queries or data records) using the query population feature. Alternatively, they can passively provide feedback by answering generated clarification questions. This stage facilitates users in efficiently providing missing semantic information. In the third stage, by incorporating user feedback, ISEE collaboratively suggests an updated field description, which users can further refine as needed.

This iterative loop, including three stages, continues until the description achieves a satisfactory level of quality. ISEE follows the design of Human-AI collaborative systems such as (Tian et al., 2024, 2025), where humans guide and validate the enrichment process while AI assists with automation and scalability. We use OpenAI's GPT-4o (OpenAI, 2024) as the base LLM and `text-embedding-3-small` [1] as the embedding model in ISEE. We discuss details of each component in the following sections.

3.2 Scoring System

ISEE provides a scoring system that quantitatively measures the given description's quality across five dimensions (0–100 for each). Unlike approaches that delegate scoring to an LLM (Chen et al., 2024; Zhuang et al., 2024; Liu et al., 2023), each metric uses a dedicated algorithm, ensuring greater stability than LLM-as-a-judge methods (Gu et al., 2025; Li et al., 2025). We summarize each dimension below; full details are in Appendix A.

3.2.1 Usability Score

Usability directly measures downstream task performance. For *entity linking*, it is the Mean Reciprocal Rank (MRR) over $|Q|$ evaluation queries:

$$S_{\text{usability}} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \times 100,$$

where rank_i is the rank of the correct field for query i. When downstream evaluation is unavailable, the remaining four metrics still provide meaningful assessment.

[1] https://platform.openai.com/docs/models/text-embedding-3-small

3.2.2 Informativeness Score

Informativeness captures the diversity of information in a description. We split the text into clause components, embed each with a text embedding model, and compute the average pairwise cosine similarity $\bar{s}$:

$$S_{\text{info}} = 100 \times (1 - \bar{s}), \quad S_{\text{info}} = 0 \text{ if } m < 2.$$

3.2.3 Clarity Score

Clarity measures interpretation consistency. Inspired by (Mu et al., 2024), we prompt an LLM to generate N independent interpretations, embed them, and compute mean pairwise similarity:

$$S_{\text{clarity}} = \frac{2}{N(N-1)} \sum_{i<j} \text{CosSim}(E_i, E_j) \times 100.$$

3.2.4 Conciseness Score

Conciseness detects redundancy via both syntactic (Jaccard (Levandowsky and Winter, 1971)) and semantic (embedding cosine) similarity between sentence pairs. Let $N_{\text{redundant}}$ be the number of sentences in at least one redundant pair out of N_{total}:

$$S_{\text{concise}} = 100 \times \left(1 - \frac{N_{\text{redundant}}}{N_{\text{total}}}\right).$$

3.2.5 Readability Score

We adopt the Flesch–Kincaid Reading Ease score (Flesch, 1943) as the readability metric.

3.3 Taxonomy-Guided Clarification

To add missing context and resolve ambiguity, we propose a clarification taxonomy to support clarification question generation for field enrichment, based on previous studies (Purver, 2004; Alianne-jadi et al., 2019b; Zamani et al., 2020b).

As shown in Table 1, each type targets a distinct cause of incompleteness or ambiguity: *Paraphrase* asks the user to restate the description, helping cross-validate and disambiguate meaning. *Narrow-down* resolves polysemy by restricting the scope of a term with multiple interpretations. *Attributes* collects finer-grained details when relevant subtypes or attributes are not explicitly specified. *Wh-clarification* elicits further context by asking who, when, where, or why questions. *Relationship* uncovers potential connections between the given field and other fields in the schema. *Logic* clarifies

Table 1: Clarification question taxonomy for semantic enrichment.

Type	Description	Example
Paraphrase	When the description is generally ambiguous, the system can ask the user to paraphrase it in another way. Paraphrasing can cross-validate the meaning and is also easy for users to perform.	*User:* "This is a field about Python." ISEE*:* "Can you paraphrase it in another way?"
Narrow-down	When there are different meanings of a certain mention, the system asks a clarification question to narrow down the meaning.	*User:* "Python." ISEE*:* "Are you talking about the programming language or the animal (snake)?"
Attributes	The mention may not be ambiguous, but there are multiple subtypes/attributes under it. The system can request more detail and identify which attributes the user cares about.	*User:* "Programming languages." ISEE*:* "Do you care more about Python, Java, or other languages? Or is this for all languages?"
Wh-clarification	The system asks *Who* (personal context), *When* (temporal context), *Where* (spatial context), or *Why* (purpose context) questions about the description.	*User:* "Laptop." ISEE*:* "Is there any purpose, such as gaming, work, or school?"
Relationship	There can be a relationship between this field and other fields in the schema/sandbox, but the user did not mention it. The system identifies a likely candidate and asks if such a relationship exists.	*User:* "This field collects user names who understand Python programming." ISEE*:* "Is there a relationship between this field and the other field *expertise_level*?"
Logic	The logic presented in the description can be interpreted in different ways.	*User:* "Python and Java." ISEE*:* "Do you mean 'either' or 'both' for the two conditions?"

the intended logical relationships between multiple elements or conditions in the description.

Given a field description and its context (e.g., schema information or previously answered questions), ISEE generates a clarification question in two steps. First, it prompts an LLM in an "LLM-as-a-Judge" role to select the most relevant clarification type from the taxonomy. Second, it builds a type-specific generation prompt containing the description, context, few-shot examples, and a short style template to produce a concise question along with candidate answers (multiple-choice or open-ended). This two-step pipeline ensures that each generated question is both consistent in form and targeted at the most pressing ambiguity.

3.4 Conditional Query Population

Without the additional instruction, ISEE generates diverse queries related to this field by default

Beyond passively answering clarification questions, ISEE allows users to proactively provide NL queries or example data that convey the purpose or usage scenario of a data field. Since manually creating such examples can be demanding, ISEE introduces a query population feature that automatically generates candidate queries for a given field. Our key insight is that verifying related queries is often easier than refining a textual description, so users only need to validate or refine populated candidates rather than author them from scratch. The process is iterative: previously verified or user-provided queries become context for subsequent rounds, progressively improving semantic coverage. Users can optionally steer generation with an NL instruction (e.g., "focus on time evaluation" to populate only chronological queries); without one, ISEE generates diverse queries by default.

3.5 User Interface

Figure 2 illustrates the UI of ISEE, which is organized around the three-stage pipeline described in Section 3. The interface is designed to minimize cognitive load while keeping the domain expert in control at every step.

The left panel (①) displays the current field description and a *Score* button that triggers the scoring system (Section 3.2). Results are shown as a radar chart of the five dimension scores, an overall quality score (0–100) with natural language explanations, and a *Downstream Results* dashboard that reveals the field's ranking performance on real evaluation queries. The middle-left area (②) presents taxonomy-guided clarification questions (Section 3.3) as clickable multiple-choice options or short free-text inputs, enabling users to provide feedback with minimal effort. The middle-right

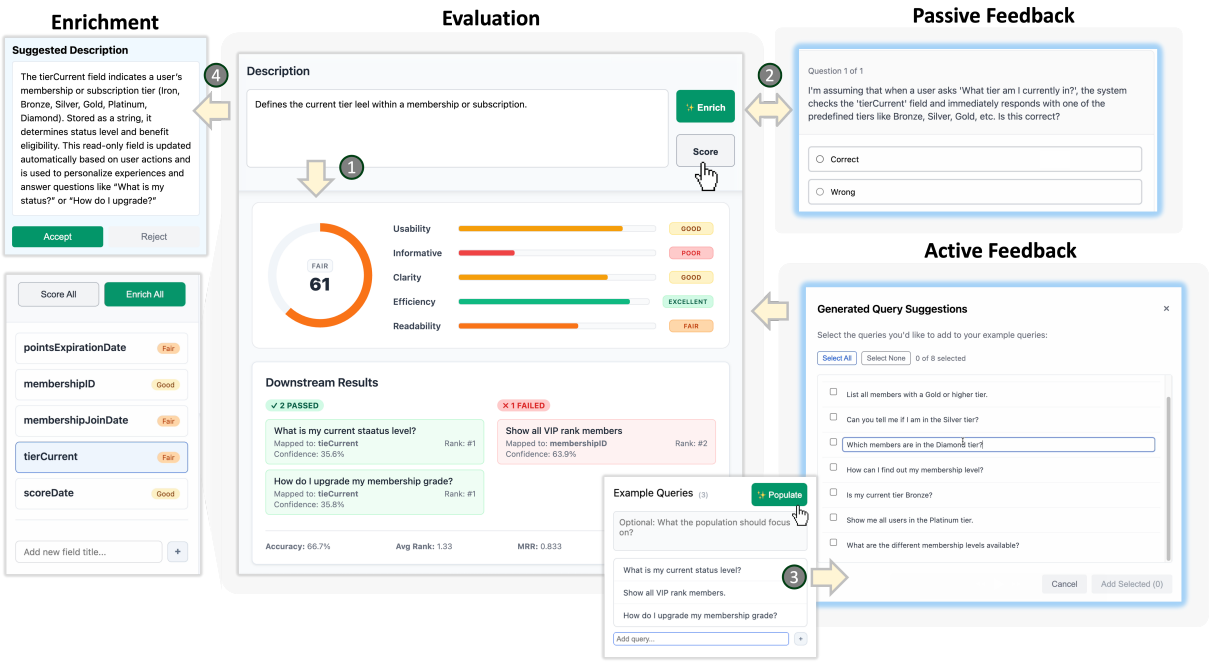

Figure 2: UI of ISEE. **(1) Evaluate:** Users click *Score* to obtain a 5-dimensional 0–100 scores and a *Downstream Results* dashboard to indicate the quality of the current description. **(2) Passive feedback:** The UI asks concise clarification questions (e.g., scope, attributes, logic); users answer with one click. **(3) Active feedback:** Users add example NL queries, or press *Populate* to review system-populated queries and select the relevant ones. **(4) Enrichment:** The system drafts a *Suggested Description* that the user can Accept/Reject or lightly edit.

area (③) provides the query population interface (Section 3.4), where users can type NL queries directly or press *Populate* to generate a batch of candidates displayed as selectable chips. An optional instruction field lets users steer the generation (e.g., "focus on time-related queries"), and selected queries accumulate across rounds. The right panel (④) shows a *Suggested Description* synthesized from all collected feedback, which users can accept, reject, or edit in place. A description history viewer lets users compare successive versions. The entire workflow is iterative: after accepting or editing a suggestion, users can re-score the updated description and continue refining until the quality meets their expectations.

4 Evaluation

4.1 User Study

As an interactive tool, we evaluate the usability and effectiveness of ISEE. We first conducted a within-subjects user study with 8 participants. All of them engage with data field descriptions as part of their day-to-day work. In the study, we compare ISEE to three baselines, including manually editing the description, selecting enriched description candidates, and using a conversational AI assistant. Following the user study, we conducted an additional qualitative evaluation in which human reviewers rated the enriched descriptions against the ground truth. The ratings were conducted with the reviewers unaware of the origin of the descriptions, as descriptions were shuffled across all users and sessions. This approach ensured an unbiased evaluation of whether ISEE enhances enrichment accuracy.

4.1.1 User Study Protocol

Each study began with a two-minute introduction that outlined the motivation and background of the study. Next, participants were asked to complete a pre-task survey to capture their background information and better understand user needs. To simulate a scenario that elicits domain experts' knowledge, participants were given three minutes to memorize a printed sheet containing the ground truth descriptions of four data fields, which would later serve as the study tasks. The sheet was then collected to simulate a scenario in which participants hold knowledge in their minds without documented information. Following this, participants watched a three-minute tutorial video explaining the study tasks and demonstrating how to use ISEE and the other tools. Participants proceeded to complete four task sessions, each lasting three minutes and

utilizing a different tool. Participants were asked to make their best effort to provide a good description within the allotted time for each session. Both the order of tools and their assigned tasks were shuffled to negate learning effects. Finally, after completing all tasks, participants filled out a post-task survey to compare their experiences across the different sessions.

4.1.2 Comparison Baselines

To evaluate the effectiveness of ISEE, we compared it against three baselines that reflect common practices for enriching data field descriptions:

- **Manual Editing.** Participants directly wrote or revised the field description without system support. This baseline simulates the most common practice, where users rely solely on their own domain knowledge.
- **Candidate Selection.** Participants were presented with several enriched descriptions generated by LLM prompting. They could select one of the candidates and manually edit it.
- **Conversational AI Assistant.** Participants interacted with a general-purpose conversational AI assistant, ChatGPT. Users iteratively refine the description through natural chatting, a process with which they are already familiar.

4.1.3 Pre-task Survey

Before starting the main tasks, participants completed a short pre-task survey to capture their prior experience and perceptions of field descriptions. All eight participants reported that they were not satisfied with the quality of the field descriptions that appeared in their work. In terms of usage frequency, six participants indicated that they interact with database fields on a daily basis, while the remaining two reported doing so on a weekly basis.

Participants also identified field description issues in their work. As shown in Table 2, the most frequent issue was that descriptions are too short or missing (100%). Other common limitations included a lack of concrete examples (71.4%), missing information about relationships between fields (71.4%), and insufficient business context or purpose (57.1%). Ambiguity in language (42.9%) and outdated or incorrect descriptions (28.6%) were also noted. These findings confirm that field descriptions often fail to meet practitioners' needs and motivate the design of ISEE to provide more complete, contextualized, and useful enrichment.

Table 2: Participants reported limitations of field descriptions (7 participants responded, multiple choices allowed).

Challenge	Responses (%)
Too short or missing	7 (100%)
Ambiguous or unclear language	3 (42.9%)
Outdated or incorrect	2 (28.6%)
Lacks business context or purpose	4 (57.1%)
Lacks concrete examples	5 (71.4%)
Lacks relationship information	5 (71.4%)
Other	0 (0%)

Table 3: Average participant rankings of contents of field descriptions (lower score = higher importance).

Content	Avg. Rank
Data Type and Format	1.57
Field Purpose	2.00
Usage Example	2.86
Field Relationship	3.14

Furthermore, we asked participants to reflect on the important aspects of field descriptions. Results are summarized in Table 3. *Field purpose* and *Data type/format* were consistently ranked as the most important content of a description.

4.1.4 User Study Results

We first evaluated ISEE against three baselines (manual editing, candidate selection, and conversational assistant) using the NASA TLX questionnaire (Hart and Staveland, 1988). As shown in Figure 3, ISEE significantly reduced perceived mental demand, effort, and frustration compared to the baselines. The reductions in mental demand and effort were statistically significant ($p < 0.05$). Participants reported that enriching field descriptions with ISEE required less cognitive load and felt smoother than manually editing or relying on a conversational assistant. Importantly, the highest ratings on performance indicate that participants felt more successful in completing the enrichment task when using ISEE.

To further assess enrichment quality, we conducted a blind review in which independent human raters scored enriched descriptions against ground-truth references without knowing their source (Table 4). Descriptions generated with ISEE received significantly higher accuracy ratings than all baselines, and this difference was statistically significant ($p < 0.05$), which is consistent with participants' self-perception.

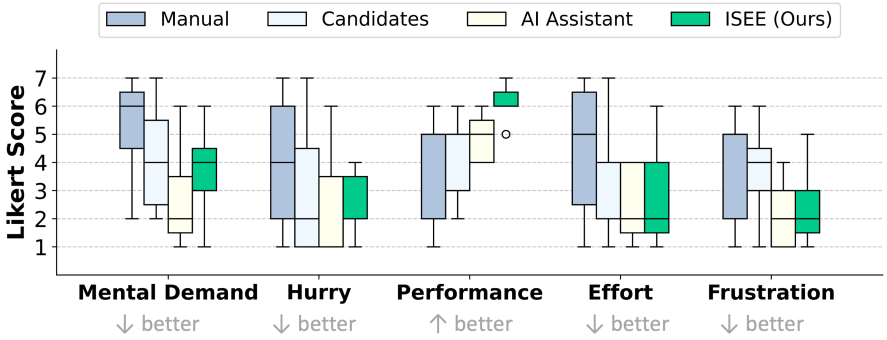

Figure 3: NASA Task Load Index Ratings.

Additionally, in the post-study survey, we gathered additional feedback. The majority of participants rated the quality score as useful or very useful, with 85.8% assigning a score of 4 or 5 on a 5-point scale. We also asked participants to rate the usefulness of three key features of our system on a 1–5 scale (higher is better). Query population received the highest rating ($M = 4.29$), with participants emphasizing that the automatically generated queries were "*diverse in nature, covering many possible questions a user may have about the field,*" which "*made it much easier to write a good description.*" Description scoring was also rated highly ($M = 3.86$), with one participant noting that "*the scoring of the interactive enrichment system is extremely helpful as it can guide me to improve my description.*" Finally, clarification questions received a relatively lower rating ($M = 2.71$). While generally considered useful, participants mentioned that "*sometimes [they] felt it less efficient to answer many questions.*" Taken together, these results suggest that all three features contribute to the effectiveness of semantic enrichment, with query population and description scoring being particularly impactful.

4.2 Automated User Simulation

To evaluate ISEE at scale beyond the user study, we follow prior work (Tian et al., 2023; Yao et al., 2019) and design an user simulation experiment where an LLM simulates a domain expert interacting with ISEE. We experiment on the BIRD benchmark (Li et al., 2024), a public text-to-SQL dataset containing real-world databases with column descriptions. We evaluate enrichment quality through two downstream tasks: *entity linking*, where we rank all columns by relevance to a given NL question and calculate Mean Reciprocal Rank (MRR) of the gold columns; and *text-to-SQL*, where we generate SQL using GPT-4o with column descriptions as context and report execution accuracy (EX), which checks whether the predicted SQL produces the same result as the gold query, and Valid Efficiency Score (VES), which combines execution accuracy with query efficiency by penalizing slow queries.

Table 4: Average reviewer ratings of enriched descriptions (1 = lowest quality, 7 = highest quality).

Method	Rating (1–7)
Manual Editing	3.2
Candidate Selection	4.0
Conversational AI Assistant	4.5
ISEE (Ours)	**6.2**

We construct a *reference description* for each column by combining its existing BIRD description, schema (table, columns, foreign keys), sampled database records, and evidence from gold SQL queries and their associated NL questions. This reference represents the domain knowledge an expert would hold and is used only to drive the simulated user's responses, which is *never* provided to ISEE. The simulated user (a persona prompt containing the reference context) can only interact through ISEE's UI features: answering clarification questions and accepting or rejecting populated queries. It cannot directly write or paste the description.

We run the simulation on the entire BIRD dev set (1534 queries, 11 databases) and compare among three conditions: (1) *Original*: the existing BIRD description; (2) *Auto-enrich*: a single-pass LLM baseline that improves the description using only schema context, without interactive feedback or the reference; and (3) ISEE *(simulated)*: enrichment through the simulated user interacting with ISEE. As shown in Table 5, ISEE with simulated users significantly outperforms both baselines. Compared

Table 5: User simulation results on full BIRD.

Method	Entity Linking	Text-to-SQL	
	MRR	EX	VES
Original	42.31	52.22	56.99
Auto-enrich	55.25	54.95	60.18
ISEE (simulated)	64.72	61.40	66.54

Table 6: Quantitative evaluation on sampled BIRD.

Method	Entity Linking	Text-to-SQL	
	MRR	EX	VES
Original	45.83	55.00	58.77
Auto-enrich	55.01	57.00	63.80
ISEE	66.99	64.00	64.65

to Auto-enrich, ISEE improves entity-linking MRR by 9.47 and text-to-SQL EX by 6.45, demonstrating that interactive clarification and evaluation can elicit domain knowledge that single generation cannot cover. Regarding token cost per column, Auto-enrich requires on average 150 output tokens, while ISEE consumes on average 380 output tokens.

4.3 Quantitative Evaluation

To complement the user simulation experiment, we further conducted a quantitative experiment where one author interactively enriches descriptions using ISEE. We randomly sampled 100 NL questions across 18 databases covering 135 columns from BIRD. The experimenter worked on each column description within a 3-minute time limit. Since the experimenter is fully familiar with the system, these results can be interpreted as an upper bound of user performance with ISEE.

Table 6 reports the results. Enrichment with ISEE improves entity-linking MRR by 21.16, text-to-SQL execution accuracy by 9.00, and VES by 5.88 over the original descriptions. ISEE outperforms Auto-enrich across all metrics, confirming that interactive enrichment yields higher-quality descriptions than automated generation.

4.4 Case Study

We conduct a case study to understand how enrichment with ISEE affects downstream task performance in a realistic setting. In our study, one participant was asked to enrich the field `purchases.value`, whose original description was simply *"Value."* The database also contains fields such as `order.priceTotal` (*"Total price of the order."*) and `purchases.id` (*"Purchase identifier."*). We examine how a user query, *"What is the total revenue from purchases last quarter?"*, is handled before and after enrichment.

Before enrichment. The entity-linking system cannot determine what *"value"* refers to. It ranks `order.priceTotal` first (since its description explicitly mentions *"price"*) and `purchases.value` third. Consequently, a text-to-SQL system generates an incorrect query: `SELECT SUM(order.priceTotal) FROM order WHERE order_date >= '2025-10-01'`. This returns the sum of all order totals, not the transaction revenue.

Enrichment with ISEE. Through clarification, the participant specifies that the field refers to a *"monetary amount"* at the *"per-transaction"* level. Through query population, the participant verifies relevant queries such as *"average purchase amount per customer."* The description becomes: *"The monetary amount (in base currency) of each completed purchase transaction. Used for revenue aggregation. Distinct from order-level totals such as order.priceTotal."* The score rises from 18 to 82.

After enrichment. With the enriched description explicitly mentioning *"monetary amount"* and *"revenue aggregation,"* entity linking now correctly ranks `purchases.value` first, as its description directly matches the user's intent. The text-to-SQL system accordingly generates the correct query: `SELECT SUM(purchases.value) FROM purchases WHERE purchase_date >= '2025-10-01'`, which correctly computes revenue from purchase transactions rather than order totals. This example illustrates how a single enrichment session with ISEE can resolve ambiguity that otherwise propagates through the downstream pipeline.

5 Conclusion

We present ISEE, an interactive semantic enrichment system that automatically scores field descriptions, and improve it through query population and clarification questions. A user study shows that ISEE significantly reduces cognitive load while improving performance. A user simulation experiment and a quantitative evaluation shows that data descriptions enriched by ISEE can significantly benefit downstream tasks. This work demonstrates that interactive enrichment is a promising approach for capturing missing domain knowledge and benefiting data-related applications.

Limitations

Like any interactive system that relies on human expertise, ISEE assumes that the user possesses domain knowledge about the fields to enrich. Regarding practical use, while users are domain experts who defined or work with these fields daily, their knowledge may occasionally be incomplete or outdated. Future work could incorporate confidence estimation or cross-validation mechanisms to further safeguard enrichment quality.

On the evaluation side, we designed three complementary studies to balance ecological validity with scale. Our user study recruited eight professional industry engineers and researchers with hands-on data management experience, ensuring high-quality feedback, though the sample size is naturally constrained by the cost of expert recruitment. To scale beyond this, we introduced an automated user simulation over the entire BIRD dev set; while this enables comprehensive coverage, simulated interactions inevitably simplify real human behavior such as partial or ambiguous responses. The quantitative evaluation, where the first author manually enriched sampled databases, bridges these two settings with controlled but smaller-scale evidence. We believe these three evaluations together offer a thorough and balanced assessment of ISEE.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019a. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, page 475–484. ACM.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019b. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 475–484, New York, NY, USA. Association for Computing Machinery.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Anonymous. 2024. Tabmeta: Table metadata generation with LLM-curated dataset and LLM-judges. In *Submitted to ACL Rolling Review - June 2024*. Under review.

Michael Belsky, Rafael Sacks, and Ioannis Brilakis. 2016. Semantic enrichment for building information modeling. *Computer-Aided Civil and Infrastructure Engineering*, 31(4):261–274.

Ting Cai, Stephen Sheen, and AnHai Doan. 2025. Columbo: Expanding abbreviated column names for tabular data using large language models. *Preprint*, arXiv:2508.09403.

Yi-Pei Chen, KuanChao Chu, and Hideki Nakayama. 2024. Llm as a scorer: The impact of output order on dialogue evaluation. *Preprint*, arXiv:2406.02863.

Rushikesh Deotale, Adithya Srinivasan, Yuan Tian, Tianyi Zhang, Pavlos Vlachos, and Hector Gomez. 2026. All-fem: Agentic large language models fine-tuned for finite element methods. *Preprint*, arXiv:2603.21011.

Rudolf Franz Flesch. 1943. *Marks of Readable Style: A Study in Adult Education*. Teachers College, Columbia University. Accessed: 30 January 2025, via National Library of Australia.

Yingqi Gao and Zhiling Luo. 2025. Automatic database description generation for text-to-sql. *Preprint*, arXiv:2502.20657.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 159–166, New York, NY, USA. Association for Computing Machinery.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. *Preprint*, arXiv:1808.07699.

Michael Levandowsky and David Winter. 1971. Distance between sets. *Nature*, 234(5323):34–35.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, and 1 others. 2024. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Calibrating llm-based evaluator. *Preprint*, arXiv:2309.13308.

Fangwen Mu, Lin Shi, Song Wang, Zhuohao Yu, Binquan Zhang, ChenXue Wang, Shichao Liu, and Qing Wang. 2024. Clarifygpt: A framework for enhancing llm-based code generation via requirements clarification. *Proc. ACM Softw. Eng.*, 1(FSE).

Zheng Ning, Yuan Tian, Zheng Zhang, Tianyi Zhang, and Toby Jia-Jun Li. 2024. Insights into natural language database query errors: from attention misalignment to user handling strategies. *ACM Trans. Interact. Intell. Syst.*, 14(4).

Zheng Ning, Zheng Zhang, Tianyi Sun, Yuan Tian, Tianyi Zhang, and Toby Jia-Jun Li. 2023. An empirical study of model errors and user error discovery and repair strategies in natural language database queries. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 633–649, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

J. Purver, Matthew R.2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College London. PhD thesis.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Mayank Singh, Abhijeet Kumar, Sasidhar Donaparthi, and Gayatri Karambelkar. 2025. Leveraging retrieval augmented generative llms for automated metadata description generation to enhance data catalogs. *Preprint*, arXiv:2503.09003.

Yuan Tian, Jonathan K. Kummerfeld, Toby Jia-Jun Li, and Tianyi Zhang. 2024. Sqlucid: Grounding natural language database queries with interactive explanations. *Preprint*, arXiv:2409.06178.

Yuan Tian, Daniel Lee, Fei Wu, Tung Mai, Kun Qian, Siddhartha Sahai, Tianyi Zhang, and Yunyao Li. 2025. Text-to-sql domain adaptation via human-llm collaborative data annotation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1398–1425, New York, NY, USA. Association for Computing Machinery.

Yuan Tian and Tianyi Zhang. 2025. Selective prompt anchoring for code generation. *Preprint*, arXiv:2408.09121.

Yuan Tian and Tianyi Zhang. 2026. Pv-sql: Synergizing database probing and rule-based verification for text-to-sql agents. *Preprint*, arXiv:2604.17653.

Yuan Tian, Zheng Zhang, Zheng Ning, Toby Jia-Jun Li, Jonathan K. Kummerfeld, and Tianyi Zhang. 2023. Interactive text-to-SQL generation via editable step-by-step explanations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16149–16166, Singapore. Association for Computational Linguistics.

Niklas Wretblad, Oskar Holmström, Erik Larsson, Axel Wiksäter, Oscar Söderlund, Hjalmar Öhman, Ture Pontén, Martin Forsberg, Martin Sörme, and Fredrik Heintz. 2024. Synthetic sql column descriptions and their impact on text-to-sql performance. *arXiv preprint arXiv:2408.04691*.

Fan Xue, Liupengfei Wu, and Weisheng Lu. 2021. Semantic enrichment of building and city information models: A ten-year review. *Advanced Engineering Informatics*, 47:101245.

Fan Yang, Yuan Tian, and Jiansong Zhang. 2025. Supporting construction worker well-being with a multi-agent conversational ai system. *Preprint*, arXiv:2506.07997.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *Preprint*, arXiv:2110.11334.

Ziyu Yao, Yu Su, Huan Sun, and Wen-tau Yih. 2019. Model-based interactive semantic parsing: A unified framework and a text-to-SQL case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5447–5458, Hong Kong, China. Association for Computational Linguistics.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020a. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 418–428, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020b. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 418–428, New York, NY, USA. Association for Computing Machinery.

Haoxiang Zhang, Yurong Liu, Aécio Santos, Juliana Freire, and 1 others. 2025a. Autoddg: Automated dataset description generation using large language models. *arXiv preprint arXiv:2502.01050*.

Tianshu Zhang, Kun Qian, Siddhartha Sahai, Yuan Tian, Shaddy Garg, Huan Sun, and Yunyao Li. 2025b. Evoschema: Towards text-to-sql robustness against schema evolution. *Proc. VLDB Endow.*, 18(10):3655–3668.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *Preprint*, arXiv:2310.14122.

A Scoring System Details

This appendix provides full descriptions of the five scoring metrics used in ISEE.

A.1 Usability Score

The usability score measures how well a field description supports downstream tasks. We consider this the most important metric, as it directly reflects the practical utility of the description in real-world applications. For the *entity linking* task, the score is based on the Mean Reciprocal Rank (MRR), which captures both top-rank accuracy and fine-grained ranking performance. Given a set of $|Q|$ evaluation queries:

$$S_{\text{usability}} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \times 100$$

where rank_i is the position of the correct field in the ranked list of retrieval results for query i. Higher scores indicate that the description enables more accurate and precise retrieval. As more downstream tasks are incorporated, this score becomes increasingly representative. When downstream evaluation is unavailable (e.g., due to privacy or insufficient data), the remaining four metrics still provide meaningful assessment.

A.2 Informativeness Score

Informativeness captures the richness and diversity of the information conveyed in the description. A more informative description provides multiple, non-overlapping aspects of meaning, offering more context and greater semantics.

We split the text into clause components by periods and commas. Each component is embedded with OpenAI's `text-embedding-3-small` and L2-normalized. Let m be the number of components and $f(\cdot)$ the embedding function. We compute the average pairwise cosine similarity:

$$\bar{s} = \frac{2}{m(m-1)} \sum_{i<j} \cos\big(f(c_i),\, f(c_j)\big),$$

and define the informativeness score:

$$S_{\text{info}} = 100\,(1 - \bar{s}), \quad \text{with } S_{\text{info}} = 0 \text{ if } m < 2.$$

The score is clipped to $[0, 100]$ to stabilize it and mitigate the impact of outliers.

A.3 Clarity Score

Clarity measures the extent to which a description can be interpreted in different ways. A high clarity score indicates that different readers or models are likely to interpret the text consistently.

Inspired by prior work (Mu et al., 2024) on ambiguity detection for code generation, we prompt an LLM (GPT-4o) to generate N independent interpretations of the input description, then encode each into a semantic embedding. We compute the mean pairwise cosine similarity:

$$S_{\text{clarity}} = \frac{2}{N(N-1)} \sum_{i<j} \text{CosSim}(E_i, E_j) \times 100$$

Lower agreement indicates higher ambiguity, resulting in a lower clarity score.

A.4 Conciseness Score

Conciseness measures whether a description conveys its intended meaning without unnecessary redundancy. Redundancy can distract LLMs (Tian and Zhang, 2025), thereby reducing downstream performance. We quantify redundancy using both syntactic and semantic similarity:

- **Syntactic redundancy**: Pairwise Jaccard similarity (Levandowsky and Winter, 1971) between word sets of two sentences: $J(A, B) = |A \cap B| / |A \cup B|$. Pairs exceeding a threshold are flagged as redundant.
- **Semantic redundancy**: Cosine similarity between sentence embeddings. High similarity indicates semantically equivalent sentences.

The final redundancy set is the union of pairs flagged by either method. Let N_{total} be the total number of sentences and $N_{\text{redundant}}$ the number in at least one redundant pair:

$$S_{\text{concise}} = 100 \times \Big(1 - \frac{N_{\text{redundant}}}{N_{\text{total}}}\Big)$$

A.5 Readability Score

Readability evaluates whether the structure of a description is easy to comprehend. Although originally designed for humans, poor readability indicates that the description is poorly organized and may also hinder LLM understanding. We use the Flesch–Kincaid Reading Ease score (Flesch, 1943) directly as our readability metric.

A.5.1 Overall Score

The overall score is a weighted sum of the five metrics:

$$S_{\text{total}} = \sum_{m=1}^{5} w_m \cdot S_m$$

The weights are hyperparameters that can be adjusted to match different enterprise priorities or domain requirements. In our setting, we use weights 0.50, 0.20, 0.15, 0.10, and 0.05 for usability, informativeness, clarity, conciseness, and readability, respectively. These values were calibrated in consultation with multiple engineers in our working environment, with usability weighted highest as it directly reflects downstream task performance. The primary purpose of the overall score is to provide a quick signal for identifying low-quality field descriptions that need attention, rather than serving as a precise ranking metric; the exact weight values are therefore less critical than their relative ordering.