ISEE: Interactive Semantic Enrichment for Data Field Description

Yuan Tian[1], Yiru Chen[2], Rakesh R. Menon[2], Zifan Liu[2], Ting Cai[3], Fei Wu[2], Anudeep Chimakurthi[2], Prashanthi Ramamurthy[2], Sridevi Aishwariya Ganesan[2], Kun Qian[2], Yunyao Li[2]

[1]Purdue University
[2]Adobe
[3]University of Wisconsin-Madison
tian211@purdue.edu,
{yiruc, rakeshra, zifanl, feiw, achimakurthi, pramamur, sridevi, kunq, yunyaol}@adobe.com,
tingcai@cs.wisc.edu

Abstract

LLM-powered agents are increasingly being deployed for data-related tasks, including data import, data exploration, data visualization, and data analytics. However, their performance heavily depends on the clarity and completeness of data field semantics. Unfortunately, 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 effectiveness of LLM-based agents in downstream tasks, such as entity linking. To bridge this gap, we introduce a novel **I**nteractive **SE**mantic **E**nrichment system (ISEE). Given a field description, ISEE assesses its quality using a novel scoring system, efficiently gathers user knowledge, and collaboratively enriches the semantics with users. Both a user study and a qualitative evaluation demonstrate that our approach significantly enhances the accuracy of field descriptions while imposing less cognitive load on users compared to baselines.

Introduction

LLM-powered agents are increasingly deployed in enterprise tasks such as entity linking (Shen, Wang, and Han 2015; Kolitsas, Ganea, and Hofmann 2018) and natural language querying of databases (Ning et al. 2023; Tian et al. 2023; Ning et al. 2024). 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 domain experts. 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 the unclear semantics, LLMs can only guess and make mistakes. While automatic enrichment methods (Belsky, Sacks, and Brilakis 2016; Cai, Sheen, and Doan 2025; Xue, Wu, and Lu 2021) can mitigate this issue, they come with inherent limitations. This is primarily because these works utilize the broad, general knowledge embedded in LLMs, rather than relying on the domain-specific expertise of users. Without sufficient context, it is nearly impossible for LLMs to automatically infer the missing information and enrich these fields. Thus, field enrichment has to keep the domain experts in the loop and actively solicit domain knowledge.

To bridge this gap, 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, an *instruction-guided query population* module that allows domain experts to enrich field descriptions by simply verifying potentially LLM-suggested natural queries that can be answered using the 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 within-subject user study and a qualitative assessment. In the user study, we recruited eight participants with relevant professional backgrounds and tasked them with enriching field descriptions using ISEE as well as three baseline systems: manually editing descriptions, selecting from pre-enriched candidates, and working with a conversational AI Assistant (i.e., ChatGPT). According to participants' self-reported experiences, ISEE significantly reduced their cognitive load and increased their confidence in producing enriched descriptions. Following the study, an independent human evaluation of the enriched descriptions showed that ISEE significantly improved description enrichment accuracy by 119%, 32%, and 92% compared to manual editing, candidate selection, and ChatGPT, respectively.

Method

System Overview

Figure 1 presents the pipeline of ISEE. Starting with field descriptions from a data dictionary (Adobe Experience Data

Model[1]), ISEE includes three iterative stages—*Evaluation*, *Clarification*, and *Enrichment*. In the first stage (*Evaluation*), ISEE evaluates the quality of the current description. Users are provided 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` [2] as the embedding model in ISEE. We discuss details of each component in the following sections.

Scoring System

The first key feature of ISEE is a multi-dimensional scoring system that provides a quantitative and interpretable assessment of a field description's quality. This evaluation enables users to gauge the current quality of a description and identify specific areas for improvement. The overall score ranges from 0 to 100 and is computed from five distinct metrics. Unlike existing approaches (Chen, Chu, and Nakayama 2024; Zhuang et al. 2024; Liu et al. 2023) that rely on prompting an LLM to directly generate a score, our method does not delegate the scoring decision to the model. Instead, the calculation of each metric is based on a dedicated algorithm targeting a different aspect of quality. This ensures greater stability and consistency compared to early LLM-as-a-judge methods (Gu et al. 2025; Li et al. 2025). We discuss details for each score below.

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 example, in 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, the MRR is computed as:

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

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 in downstream tasks.

As more different downstream tasks are considered into the evaluation, this score can become increasingly accurate and representative of real-world performance. However, sometimes the downstream evaluation may be unavailable due to privacy or insufficient data. In such cases, our scoring system still functions effectively based on the other four downstream-independent metrics.

Informativeness Score Informativeness score 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.

To achieve this goal, we measure the diversity among component embeddings within the description. Specifically, we split the text into clause components by periods and commas. Each component is embedded with OpenAI's `text-embedding-3-small` and then normalized (L2 normalization). Let m be the number of components, and let $f(\cdot)$ maps each component to its normalized embeddings. 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

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

To ensure the score remains easy to interpret, it is clipped between 0 and 100: values below 0 are adjusted to 0, while values above 100 are adjusted to 100. This clipping process stabilizes the score and mitigates the impact of outliers.

Clarity Score Clarity measures the extent to which a description can be interpreted in different ways. Even a comprehensive description can include vague, ambiguous terms, leading to random interpretation by LLMs. A high clarity score indicates that different human readers or models are likely to interpret the text in consistent ways.

Inspired by prior work (Mu et al. 2024) on ambiguity detection for code generation, we extend the approach to handle arbitrary natural language descriptions. To be more specific, we prompt an LLM (GPT-4o) to generate N independent interpretations of the input description, then encode each interpretation into a semantic embedding (OpenAI's `text-embedding-3-small`). We compute the mean pairwise cosine similarity between these embeddings and scale the result to a 0–100 range:

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

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

[1] https://github.com/adobe/xdm

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

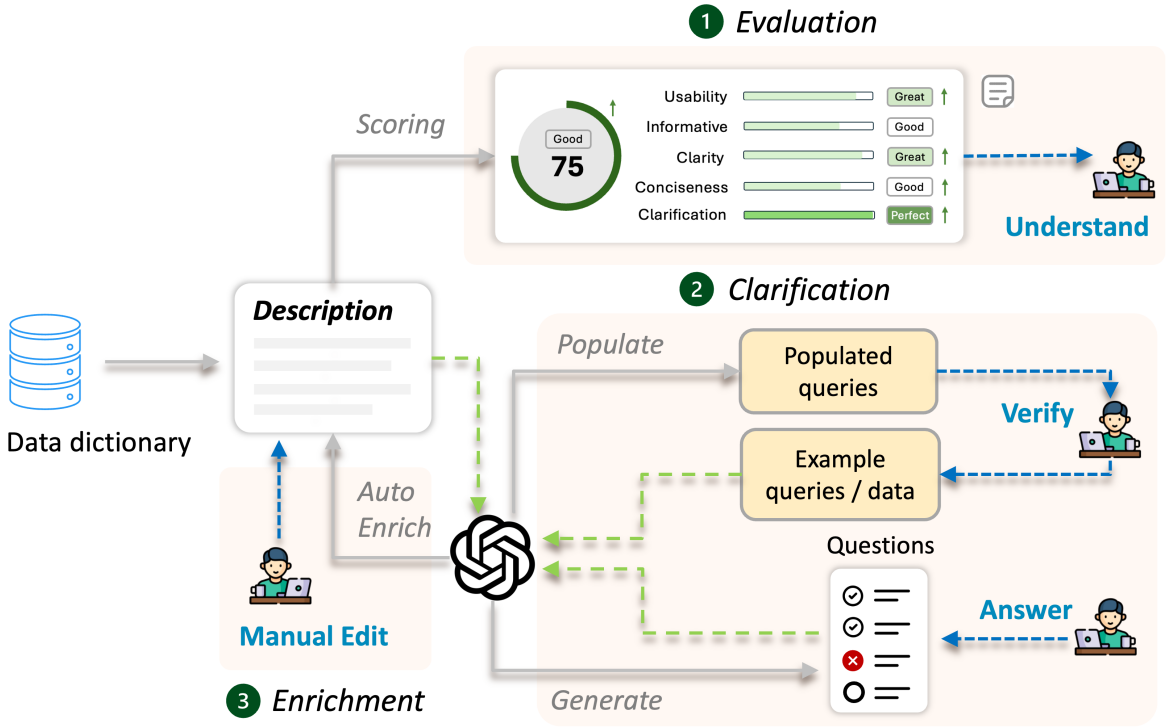

Figure 1: Pipeline of ISEE: (1) Users understand the quality of the current description by a multi-dimensional scoring system. (2) Users provide feedback through query population or clarification questions. (3) Users refine the suggested description.

Conciseness Score The conciseness score measures whether a description conveys its intended meaning without unnecessary redundancy or repetition.

A description can be complete and unambiguous, but contain overlapping or semantically equivalent sentences that restate the same information. Such redundancy can make the description unnecessarily lengthy and distract LLMs (Tian and Zhang 2025), thereby reducing downstream performance. We quantify redundancy using both syntactic and semantic similarity:

- **Syntactic redundancy**: We compute the pairwise Jaccard similarity (Levandowsky and Winter 1971) between the sets of words in two sentences:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

  where A and B are the sets of words in the respective sentences. If the similarity exceeds a predefined threshold, the sentences are marked as redundant.
- **Semantic redundancy**: Following the approach used in the *informativeness score*, we compute the cosine similarity between sentence embeddings. High similarity scores indicate semantically equivalent sentences, even if their wording differs.

To obtain the final redundancy set, we take the union of all sentence pairs flagged as redundant by either syntactic or semantic detection. This ensures that sentences identified as duplicates in either form are only counted once.

Let N_{total} be the total number of sentences in the description, and $N_{\text{redundant}}$ be the number of unique sentences appearing in at least one redundant pair. The redundancy ratio is then defined as:

$$\text{Redundancy ratio} = \frac{N_{\text{redundant}}}{N_{\text{total}}}.$$

The conciseness score is calculated as:

$$\text{Conciseness} = 100 \times \big(1 - \text{Redundancy ratio}\big),$$

so that higher scores indicate more concise descriptions with less repetition.

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

Overall Score The overall quality score is computed as a weighted sum of the five individual metric scores:

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

where S_m denotes the score for the m-th metric and w_m is its corresponding weight.

The metrics are prioritized in the following order of importance: *usability*, *informativeness*, *clarity*, *conciseness*, and *readability*. In our scenario, we manually tuned their weights to 0.50, 0.20, 0.15, 0.10, and 0.05, respectively. The weights can be dynamically adjusted to match enterprise-specific priorities, such as lowering the *usability* weight when downstream evaluation is not reliable.

Taxonomy-Guided Clarification

To add missing context and resolve ambiguity, ISEE elicits users' domain knowledge by generating specific, easy-to-answer clarification questions. To make clarification questions generation more consistent and useful, we first propose a clarification taxonomy based on careful literature review (Purver 2004; Aliannejadi et al. 2019; Zamani et al. 2020) and adapt it to our scenario. Leveraging this clarification taxonomy, ISEE prompts LLMs to select the most relevant question type based on LLM-as-a-Judge and generate the clarification question based on few-shot learning. We discuss more details in the sections below.

Clarification Taxonomy We propose a clarification taxonomy (Table 1) designed to support the generation of clarification questions for field enrichment. The taxonomy includes six distinct types of clarification questions, with each addressing a specific cause of incompleteness or ambiguity.

Paraphrase prompts the user to restate the description in another way, helping to validate and disambiguate meaning. *Narrow-down* resolves polysemy by asking targeted questions to restrict the scope of a term with multiple interpretations. *Attributes* collects finer-grained details when a mention has relevant subtypes or attributes not explicitly specified. *Wh-clarification* aims to elicit further context by asking who, when, where, or why questions. *Relationship* uncovers potential relevance between the given field and other fields in the schema that the user may not have mentioned. Finally, *Logic* clarifies the intended logical relationships between multiple elements or conditions in the description, which are often obscured by natural language.

Question Generation Given an input field description and the context (e.g., schema information or previously answered questions), ISEE generates a clarification question in two steps. First, it selects the most relevant clarification type from our six-type taxonomy by prompting a LLM in an "LLM-as-a-Judge" role. This step enables the LLM to identify the clarification type most likely to resolve the current key ambiguity. Second, using the selected type, ISEE builds a type-specific generation prompt that contains the description, the context, few-shot examples, and a short template that encodes style. This prompt directs the model to generate a specific and concise question along with its possible answers. The question may take the form of a multiple-choice question with straightforward options or an open-ended question. This taxonomy-guided clarification generation pipeline ensures consistent question generation while effectively addressing a key issue.

Instruction-Guided Query Population

In addition to passively answering clarification questions to give feedback, ISEE also allows users to proactively provide NL queries or example data related to this field, which are important context as they make the system understand the purpose or usage scenario of this field.

However, manually creating these examples can be demanding. To reduce effort, ISEE introduces a query population feature that automatically generates many queries for a given field. Our key insight is that while refining a textual description can be challenging, it is often easier for users to verify related queries. Thus, the system populates candidate queries, and the user only needs to validate or refine them rather than create them from scratch. The process is iterative: previously verified or user-provided queries are incorporated as context, enabling subsequent query population to better capture the intended semantics.

To make the query population more controllable, users can further guide this process through an optional NL instruction. For example, a user might specify "focus on time evaluation," so ISEE only populates chronological queries. Without the additional instruction, ISEE generates diverse queries related to this field by default.

Evaluation

As an interactive semantic enrichment service, 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.

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

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." *System:* "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." *System:* "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." *System:* "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." *System:* "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." *System:* "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." *System:* "Do you mean 'either' or 'both' for the two conditions?"

Table 1: Clarification question taxonomy for semantic enrichment

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.

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 pre-generated enriched descriptions generated by LLM prompting. This baseline reflects a basic industrial setting where systems provide static suggestions. Participants could either select one of the candidates or refine them, but no further interaction was supported.
- **Conversational AI Assistant.** Participants interacted with a general-purpose conversational AI assistant (ChatGPT) to iteratively refine the description through natural dialogue. This baseline reflects a widely recognized interactive modality that users are already familiar with.

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.

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.

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 2, ISEE significantly reduced perceived mental demand, effort, and frustration compared to the baselines. 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 perfor-

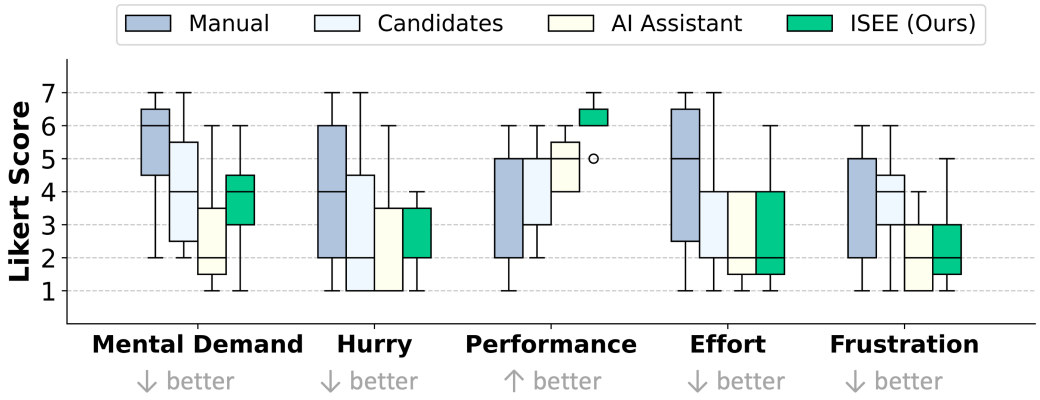

Figure 2: NASA Task Load Index Ratings.

Challenge	Responses (%)
Description is too short or missing	7 (100%)
Language is ambiguous or unclear	3 (42.9%)
Description is outdated or incorrect	2 (28.6%)
Lacks business context or purpose	4 (57.1%)
Lacks concrete examples of values/queries	5 (71.4%)
Lacks information about relationships	5 (71.4%)
Other	0 (0%)

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

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

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

mance 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 rating, which is consistent with participants' self-perception in the study.

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.

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

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

Conclusion

We present ISEE, an interactive semantic enrichment system that enhances enterprise field descriptions through description scoring, query population, and clarification questions. A user study and a qualitative evaluation show that ISEE significantly reduces cognitive load while improving performance. User study participants believed query population and description scoring as especially useful, highlighting their value in guiding and accelerating the enrichment process. This work demonstrates that interactive enrichment is an effective approach for capturing missing domain knowledge and benefiting enterprise data-related applications.

Acknowledgments

We appreciate all the participants in the user study for their valuable comments. This work was supported by Adobe during the first author's internship.

References

Aliannejadi, M.; Zamani, H.; Crestani, F.; and Croft, W. B. 2019. 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, 475–484. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361729.

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

Cai, T.; Sheen, S.; and Doan, A. 2025. Columbo: Expanding Abbreviated Column Names for Tabular Data Using Large Language Models. arXiv:2508.09403.

Chen, Y.-P.; Chu, K.; and Nakayama, H. 2024. LLM as a Scorer: The Impact of Output Order on Dialogue Evaluation. arXiv:2406.02863.

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

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Wang, Y.; Gao, W.; Ni, L.; and Guo, J. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Hart, S. G.; and Staveland, L. E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, P. A.; and Meshkati, N., eds., *Human Mental Workload*, volume 52 of *Advances in Psychology*, 139–183. North-Holland.

Kolitsas, N.; Ganea, O.-E.; and Hofmann, T. 2018. End-to-End Neural Entity Linking. arXiv:1808.07699.

Levandowsky, M.; and Winter, D. 1971. Distance between sets. *Nature*, 234(5323): 34–35.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594.

Liu, Y.; Yang, T.; Huang, S.; Zhang, Z.; Huang, H.; Wei, F.; Deng, W.; Sun, F.; and Zhang, Q. 2023. Calibrating LLM-Based Evaluator. arXiv:2309.13308.

Mu, F.; Shi, L.; Wang, S.; Yu, Z.; Zhang, B.; Wang, C.; Liu, S.; and Wang, Q. 2024. ClarifyGPT: A Framework for Enhancing LLM-Based Code Generation via Requirements Clarification. *Proc. ACM Softw. Eng.*, 1(FSE).

Ning, Z.; Tian, Y.; Zhang, Z.; Zhang, T.; and Li, T. J.-J. 2024. Insights into Natural Language Database Query Errors: from Attention Misalignment to User Handling Strategies. *ACM Trans. Interact. Intell. Syst.*, 14(4).

Ning, Z.; Zhang, Z.; Sun, T.; Tian, Y.; Zhang, T.; and Li, T. J.-J. 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, 633–649. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701061.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

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

Shen, W.; Wang, J.; and Han, J. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2): 443–460.

Tian, Y.; Kummerfeld, J. K.; Li, T. J.-J.; and Zhang, T. 2024. SQLucid: Grounding Natural Language Database Queries with Interactive Explanations. arXiv:2409.06178.

Tian, Y.; Lee, D.; Wu, F.; Mai, T.; Qian, K.; Sahai, S.; Zhang, T.; and Li, Y. 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, 1398–1425. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713064.

Tian, Y.; and Zhang, T. 2025. Selective Prompt Anchoring for Code Generation. arXiv:2408.09121.

Tian, Y.; Zhang, Z.; Ning, Z.; Li, T. J.-J.; Kummerfeld, J. K.; and Zhang, T. 2023. Interactive Text-to-SQL Generation via Editable Step-by-Step Explanations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16149–16166. Singapore: Association for Computational Linguistics.

Xue, F.; Wu, L.; and Lu, W. 2021. Semantic enrichment of building and city information models: A ten-year review. *Advanced Engineering Informatics*, 47: 101245.

Zamani, H.; Dumais, S.; Craswell, N.; Bennett, P.; and Lueck, G. 2020. Generating Clarifying Questions for Information Retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, 418–428. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370233.

Zhuang, H.; Qin, Z.; Hui, K.; Wu, J.; Yan, L.; Wang, X.; and Bendersky, M. 2024. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels. arXiv:2310.14122.