Towards Reliable Conversational Data Analytics

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes
ameryahs@univ-grenoble-alpes.fr

Jasmina Bogojeska
Zurich U. of Applied Sciences
jasmina.bogojeska@zhaw.ch

Roberta Facchinetti
University of Verona
roberta.facchinetti@univr.it

Valeria Franceschi
University of Verona
valeria.franceschi@univr.it

Aristides Gionis
KTH Royal Institute of Technology
argioni@kth.se

Katja Hose
TU Wien
katja.hose@tuwien.ac.at

Georgia Koutrika
Athena Research Center
georgia@athenarc.gr

Roger Kouyos
University of Zurich
roger.kouyos@uzh.ch

Matteo Lissandrini
University of Verona
matteo.lissandrini@univr.it

Silviu Maniu
Univ. Grenoble Alpes
manius@univ-grenoble-alpes.fr

Katsiaryna Mirylenka
IBM Research
kmi@zurich.ibm.com

Davide Mottin
Aarhus University
davide@cs.au.dk

Themis Palpanas
Université Paris Cité
themis@mi.parisdescartes.fr

Mattia Rigotti
IBM Research
mrg@zurich.ibm.com

Yannis Velegrakis
Utrecht University
i.velegrakis@uu.nl

ABSTRACT

Conversational AI systems for data analytics aim to enable the extraction of analytical insights by means of conversational interfaces. Such interfaces are powered by a mix of query modalities and machine learning methods for analytics, and are relying on Large Language Models (LLMs) for natural language generation. However, critical challenges hinder their adoption. The question we discuss is how to devise *reliable Conversational Data Analytics (CDA) systems* producing *timely, consistent, and verifiable answers.* To reach this goal, we identify five properties that impose a paradigm shift in the way systems are built and in the way they interact with users. To illustrate that shift, we describe a prototypical CDA system. Realizing these properties involves either extending existing components, or redesigning components from scratch; both solutions require overcoming data management challenges and conducting a tight integration with advanced data management and machine learning techniques.

1 INTRODUCTION

Current trends in data analytics encourage users to *converse with data* by combining traditional OLAP-style queries with natural language (NL) interfaces [36, 41]. As a result, an analytics pipeline may start with an NL expression, which is sent to a large language model (LLM) that in turn produces and executes SQL queries and in doing so switches back and forth between SQL and NL. While very appealing for its expressivity and ease of use, this paradigm raises new challenges that require thinking deeply about how *reliable Conversational Data Analytics (CDA) systems* are built.

Reliability is essential for the long-term adoption of a CDA paradigm. A conversational interface benefits from a combination of structured languages such as SQL and SPARQL and generative models such as LLMs. SQL and SPARQL are expressive, deterministic, and efficient. LLMs offer an intuitive interface to non-technical users to tap into the vast amount of information on which an LLM is trained using natural language. In practice, LLMs already offer innovative solutions to traditional data management problems, such as entity resolution, schema matching, data discovery, and query synthesis [16, 36]. However, they heavily rely on unsupervised autoregressive pretraining and a probabilistic generation process that render them notoriously prone to "hallucinating" responses that are merely "statistically" related to the training data, as opposed to being factually supported by it. Despite their impressive abilities in language understanding, it is still challenging to ensure correct and faithful domain adaptation. Hence, relying on LLMs alone is not sufficient. A reliable CDA system needs to produce timely, consistent, and verifiable answers, and provide expert guidance to the user when in doubt or when missing information. Thus, we isolate 5 important properties that a reliable system needs to possess: *Efficiency, Grounding, Explainability, Soundness*, and *Guidance.* To develop a successful system, these properties should not be pursued in isolation but their interplay needs to be considered. In pursuing these reliability properties, our vision identifies three layers in a Conversational Data Analytics (CDA) system that need our attention: the conversational data exploration layer that works as a gateway to the data sources, the computational infrastructure that is indexing, processing and enabling data access and the NL model layer tasked with translating user commands and translating the output back to the user. In line with the recent need to build "compound AI systems" [65], we propose to design a compositional framework to encompass all five reliability properties. This will require significant technical and community-level efforts to develop: (*a*) components that synergistically achieve reliability, (*b*) integration mechanisms that preserve reliability under composition, and (*c*) metrics to assess component and system reliability. When building a reliable CDA system, we acknowledge that a natural approach to enhance reliability would be to extend existing components. However, achieving a high level of reliability may necessitate a complete redesign of existing solutions. For instance, the debate in explainable AI tend to favor *inherently interpretable models* over *post-hoc explanations* of opaque-box

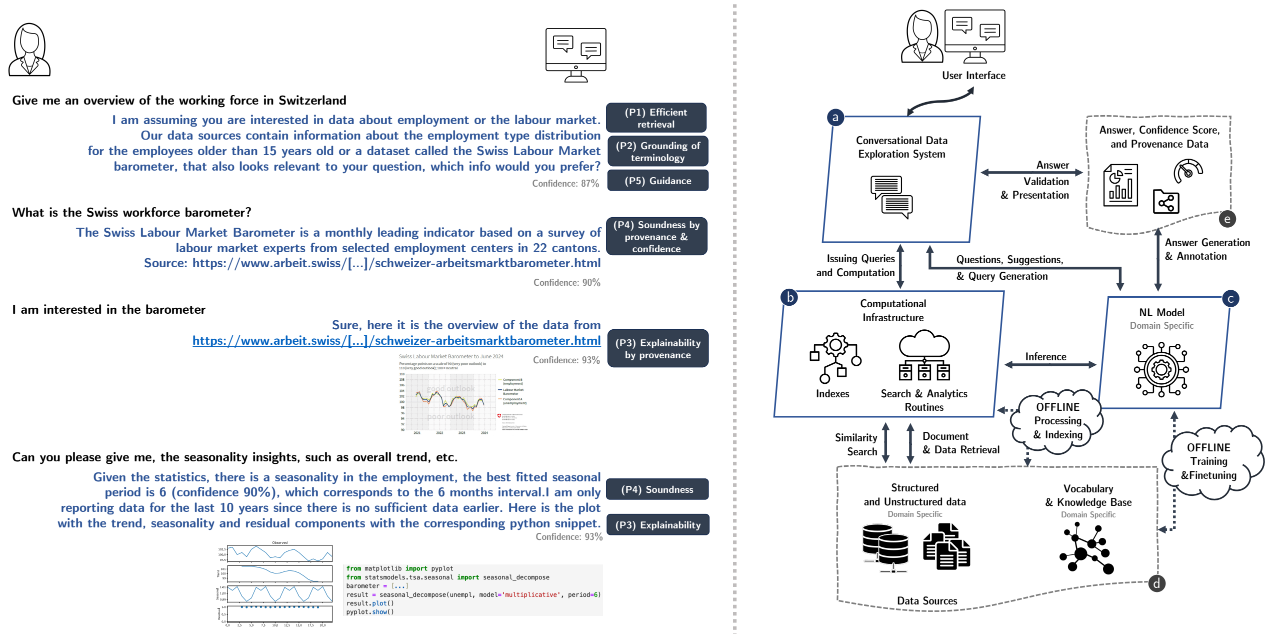

Figure 1: Left: Example use case of CDA system. Right: Architecture proposal

models, particularly in high-stakes decisions [46]. Inherently interpretable models are designed to operate in a specific domain, thereby enhancing transparency, trust, and accountability.

We describe the reliability properties and the data management challenges that arise when implementing a reliable CDA system in Section 2. Then, we discuss a prototypical CDA system and the path towards tackling those challenges in Section 3.

2 CDA: DESIDERATA AND CHALLENGES

2.1 Desiderata

The task of a CDA system is to produce analytical insights by engaging in a conversation with the user as illustrated in Figure 1.

Example. The user asks for an overview of the workforce in Switzerland. The system efficiently locates relevant datasets (*P1: Efficient Retrieval*). While the request is ambiguous, given the context, the system understood that the user is most likely interested in data about employment or the labor market (*P2: Grounding* of terminology) and it explains that it is basing the next interactions on this assumption (*P3: Explainability*). It provides succinct descriptions of the results and asks the user a follow-up question to capture their need (*P5: Guidance*). It also tells the user how confident it is for its computed answer by providing a confidence score (*P4: Soundness* by confidence). Then, the user inquires about one of the datasets, the barometer. The system provides a concise summary of the dataset coupled with the source where the answer was found (*P4: Soundness* by provenance). Finally, the user focuses on seasonality insights. The system provides a set of plots along with an acknowledgment that they were computed only where enough data was present (*P4: Soundness*) and with the code that produced them (*P3: Explainability*). In all these interactions, the system has deep knowledge of the domain and generates domain-specific computations and insights that adhere to the required standards (*P2: Grounding*). ■

As we see, a **Conversational Data Analytics (CDA) system** offers a user-friendly interface for data exploration and analysis. The system can handle a wide range of data types and analytical tasks, from simple data retrieval to more complex statistical analyses and visualizations, all driven by conversational prompts. It is aware of domain-specific vocabulary and is able to disambiguate or ask for clarification. Users describe the insights they seek, and the system responds by accessing relevant data sources, clarifying ambiguous requests, and suggesting refinements to improve query precision. For instance, it may offer proactive suggestions for additional data sources and analyses based on the user's initial query, helping to uncover unexpected patterns or relationships in the data. Further, it provides evidence for its answers, for instance, citing specific data sources or computations that produced the analysis and describing its confidence. Throughout the interaction, the system maintains context, allowing for follow-up questions and iterative refinement of analyses. As a result, users without advanced technical skills can confidently derive meaningful insights from complex datasets.

Properties of Reliable CDA. In the interactions between users and the system, as well as in the internals of a CDA system, we identify five key properties –summarized in Figure 2– that are *interwoven in the system fabric to synergistically achieve reliability*:

P1: Efficiency. Both computing an answer and providing an assessment of reliability should be fast and energy-efficient. The system should reliably retrieve the relevant data in a reasonable time, while also providing guarantees on the quality of approximation. Efficiency is central to any system, underpinning the other reliability properties. Answers should be delivered promptly and within bounded resource consumption, while maintaining a balance that minimizes sacrifices to accuracy.

P2: Grounding. Grounding is the process of connecting the system to real-world knowledge, contexts, and data, with reference to the domain of the user. This ensures that answers are relevant and factually consistent [17]. Grounding is a dynamic process that solicits user feedback when needed to enhance certainty.

P3: Explainability. To build trust in the system's responses it is essential that the system can explain its actions and outputs in understandable terms. An explanation needs to be concise while containing all relevant information about the generated

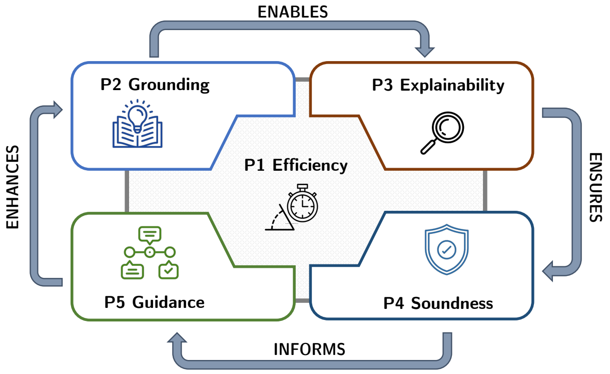

Figure 2: Properties of Reliable CDA and their interplay.

outputs. Explainability builds on the successful grounding of user intentions concerning relevant data and vocabulary.

P4: Soundness. The ability of the system to produce and evaluate an answer by connecting it to relevant data sources. The system should be able to judge whether an answer is, with sufficiently high probability, correct or not, and provide evidence of it to mitigate the model's hallucinations. At the same time, the system should be able to refrain from producing answers when unable to produce any answer with sufficient certainty. To achieve soundness, the system should be able to verify how answers are generated via explainability and provenance.

P5: Guidance. The ability to support users in pursuing their analytical goals by actively guiding them towards correct answers and desired insights more efficiently. The system should identify the steps to deliver complete, correct, and relevant answers by monitoring the soundness of the current answer and its alternative formulations to assist the user achieve their goal.

The limits of existing tools. Existing conversational tools (e.g., ChatGPT and Gemini [52]) can both process and generate human language, but lack the aforementioned properties.

W.r.t. Efficiency, current systems incur huge training and inference costs. While multiple pre-trained models are available, fine-tuning becomes necessary when domain knowledge is insufficient, potentially causing efficiency bottlenecks. Additionally, many approaches require searching for relevant content first, and in such cases, when searching in huge repositories, they adopt approximate methods that do not provide any guarantee on the quality of the information produced.

W.r.t. Grounding, existing models are trained as general-purpose models by processing vast amounts of data from heterogeneous data sources, primarily crawled from the Web. Thanks to this, they acquire essential abilities in language understanding and generation. To adapt to a new domain, they can leverage in-context learning [10]. Yet, some domains require precise grounding and alignment, necessitating fine-tuning or other forms of domain adaptation [8]. In these cases, current models require additional heavy training to adapt to evolving domains and datasets.

W.r.t. Explainability, existing approaches act as opaque boxes that generate answers, but cannot explain the process that produced them, due to the nature of their internal models that work as next-token predictors.

W.r.t. Soundness, existing systems are statistical generators that may hallucinate and cannot explicitly verify their answers. Current Retrieval Augmented techniques function as "prompt-enhancements" that try to provide contextual information from which the model can extract information, but do not change the fundamental generative nature of these architectures [24].

W.r.t. Guidance, current systems either wait for the user to provide a prompt, or the system automatically generates predetermined prompts without knowledge of the user's preferences.

2.2 Challenges

Designing a reliable CDA system leads to a wide spectrum of challenges at the intersection of data management and machine learning. From the outset, reliability properties need to be enforced within each system component but also in the interaction between the various components. It may not be sufficient to combine two sound components or two explainable components to ensure the result of their integration is still sound and explainable. This needs to be guaranteed formally: all components should have the formal properties that allow composability, i.e., individual properties (e.g., soundness) contribute to system-level formal guarantees. Finally, new metrics are needed to assess component and system reliability, in addition to accuracy, taking into account how those aspects might trade off [4, 37, 44].

Efficiency. A CDA system that operates at interactive speeds must navigate the trade-off between query execution time and answer quality. This is an elusive goal for most techniques used for information retrieval, also in dense representation spaces, since they are either fast and do not provide guarantees [3, 49], or provide quality guarantees and are relatively slow [9, 12, 15, 40, 60, 61]. To obtain methods that provide reliable answers under time constraints, would require adding quality guarantees to methods that do not support them yet, rendering methods with quality guarantees significantly faster, and devising a new generation of methods that meet all these requirements.

Grounding. We need to ensure access to the relevant terms and definitions specific to a domain. Further, the system needs access to rules and policies describing, in structured or semi-structured terms, the allowed processes and outputs. Hence, since this information is usually encoded in Knowledge Graphs and similar complex taxonomies and ontologies, the system should be able to query and perform reasoning over these resources. Important descriptive and prescriptive information may also be present in natural language. This additionally requires information extraction, data integration and alignment between different data models: structured, semi-structured, and unstructured. Further, the system should identify when an answer should be based on the result of specific computations and ensure that such computations are correctly carried out. This implies also a verification step for each answer. When the domain information changes, the system should then be able to refer to the most up-to-date model, and this cannot be always delegated to continuous retraining and fine-tuning. Hence, the challenge is how to keep the system up-to-date in its understanding of the domain and of the shifting semantics of the requests it serves.

Explainability. The system must address the problem of provenance computation, i.e., for every answer it should be possible to explain how the answer was computed. Moreover, new properties for explanation will need to be defined, in particular, *losslessness* and the ability to *mitigate errors*. Losslessness aims to make sure that an answer explanation is indeed representative of the calculations and source data used to generate it. Making an architecture lossless can rely on extending the data model of each component with expressive metadata. A stronger property is *invertibility*, i.e., to be able to recover individual calculations from an explanation. Explaining conversations must allow for expanding the scale and complexity of patterns recommended to a human. Furthermore, the system must be able to explain its actions and outputs in understandable terms, using code, natural language, graphs, or other means. This raises new challenges, such as hallucinated explanations when LLMs are used [50], high computational costs, and choosing the appropriate form of explanation. Error mitigation is the ability to re-calibrate provided explanations. This can be done for specific tasks such as entity recognition [48]. Thus, the challenge is how to combine LLM-based and data-based explanations to achieve those properties.

Soundness. Accurately quantifying the confidence of responses requires the system to be able to evaluate when it is competent or has sufficient and relevant data to provide a correct answer. When relying solely on an LLM, confidence scores may not accurately reflect the true probability of correctness. Additionally, confidence measures may be biased towards certain types of questions or topics, leading to inaccurate assessments for specific domains or tasks. An additional challenge arises when trying to closely guide the output of the system. This is achieved today by implementing reinforcement learning techniques. In these cases, the utmost challenge is the large action space. For instance, when applied naively, the actions are the number of tokens generated by an LLM, potentially going into the hundreds of thousands. Designing effective reward functions that capture the desired behavior is difficult, and collecting sufficient high-quality training data can be time-consuming and expensive. Moreover, balancing the exploration of new actions with the exploitation of known good ones is crucial for effective learning but can be difficult to achieve in practice. Finally, it is unclear how to encode safety and ethics requirements in these policies and reward functions.

Guidance. The biggest challenge is the ability to carry enough information to provide users with alternative options as opposed to the traditional single-answer approach. To do that, a graph enhanced with metadata in the form of external knowledge, enriched input, and human feedback, could be used. This raises the challenge of representing and composing such metadata in a seamless fashion as well as to perform the necessary speculative planning on top of this data structure. When planning is performed by NL models, the system must then mitigate potential hallucinations. Enabling the system to combine and switch between multiple guidance modes that also integrate user feedback is then an additional challenge.

3 BUILDING A CDA SYSTEM

We present the components required for CDA reliability, and discuss how they could be integrated *on top* of existing components.

3.1 A CDA Architecture

In pursuing reliability properties, we identify core layers in a CDA system (Figure 1 right): a the Conversational Data Exploration layer, b the Computational Infrastructure, and c the NL Model layer. These components continuously interact with two data layers: d various data and meta-data sources and e produced answers with their annotations. Indeed, the system must access both the data to be analyzed, spanning various modalities, and its accompanying metadata, such as domain-specific terminology, data models, and schemas. This data will be captured within knowledge bases, such as knowledge graphs [38]. The system will access documents and text, which may include past conversations between the user and the system, and query logs. A paradigm shift in this layer is a data model able to effectively interlink data and metadata and expose their connections uniformly to the system. Further, the system needs to be able to manage the continuous evolution of such data sources. Thus, this layer is not static: its content (data and metadata) is evolving. Central to that is an effective mechanism to cope with data rotting [25], i.e., the ability to identify and discard parts of the data that are outdated or obsolete. Consequently, the system needs to understand the intended use of the data and the ultimate user goals. Finally, the system should allow composability of the properties; this can be achieved by understanding holistically the requirements and the roles of each component within the pipeline.

The **Conversational Data Exploration layer** acts as a mediator between the user and the data sources and orchestrates all parts of a CDA system to implement the data exploration and analytical functionalities. This layer goes beyond acting as a proxy by guiding users to specific answers. It enables interactive dialogue, allowing users to ask follow-up questions to verify or explore answers. In return, the system connects responses to data, evaluates certainty, and poses questions to understand user intentions and preferences. Similar to current conversational recommender systems [34, 43, 58], it establishes when and how the system should proactively offer recommendations and ask questions. For example, a system can propose related data sources or additional computations and ask for the user's judgment. This user interaction not only provides relevant insights but also helps improve the system in formulating answers with higher confidence. *With that, conversations augmented with certainty levels become the new paradigm of interaction with a CDA system.*

The **NL Model layer** has the role of enabling user-system interactions in natural language. In this layer, different components, e.g., structured queries, exploratory analysis routines, or natural language prompts, serve different goals. For instance, this layer requires a component able to generate queries for the datastores (e.g., Knowledge Bases and Data Lakes) and snippets of formal code to process the data. These queries and snippets can be generated both in response to the user request or to the tasks inferred by the interactive conversational interface, e.g., when elaborating suggestions for the next steps. Another component, instead, will be responsible for generating natural language explanations of results or summaries of data sources. Moreover, the NL Model layer must be aware of domain-specific terminology, data models, and schemas. It also needs to be able to access and reason over multi-modal data, where modalities are not only images, text and video but also, for example, SQL and natural language questions. Finally, it should be able to interact with the computational infrastructure to delegate computations to appropriate domain-dependent reasoning and computation routines. *With that, multiple modalities will be seamlessly combined to facilitate expressing user needs.*

The **Computational Infrastructure** implements data preprocessing and indexing, as well as analytical computations including training the NL models and any other ancillary model. This layer has the primary role of enabling fast retrieval of data and documents that are needed by any querying modality to infer its answers. These operations should take place in real time. Different data may be differently significant for different tasks and this component must identify the best fit for a task at hand. Thus, the computational infrastructure should support (among others) efficient operations in high-dimensional spaces, such as vector-based operations. It also has the role of handling the entire data lifecycle, from ingestion to cleaning, de-duplication, disambiguation, and indexing. *With that, multiple data access modalities* *will be seamlessly combined with novel data structures and algorithms to facilitate achieving user needs providing fast approximate retrieval with bounded and certain quality guarantees.*

3.2 Implementation

The identified principles can address the challenges in Section 2.2 by either extending existing systems or designing new ones from scratch. While starting anew is likely unavoidable, we initially describe how to leverage existing components.

Efficiency. Both computing an answer and providing an assessment of its reliability should be fast and energy efficient by optimizing similarity search operations [39] that are at the core of many data-retrieval tasks. In this context, we see the need to design novel high-dimensional vector similarity search indexes that are able to provide a precise bound to the quality of approximation of their produced answers, while achieving shorter query answering times than existing techniques not offering quality guarantees [59]. These new solutions should also be able to return an empty set, when no answer exists with a given expected relevance. In addition, we focus our attention on a new class of algorithms, the "learning-augmented algorithms" [14, 57], in which machine learning models help similarity search algorithms make smart pruning decisions [13, 32] in order to produce accurate answers using the lowest execution time possible. However, to make conversational interfaces truly interactive and address efficiency concerns, one needs to redesign the full system architecture to treat interactivity as a first-class citizen. That is, the entire data-processing pipeline, from the user prompt to the answer computation, along with all other analytic processes taking place in parallel, should be accessible by a holistic optimizer, which identifies optimization opportunities, such as caching, batched computations, and sharing of computation and intermediate data.

Grounding. A grounded CDA system understands, disambiguates, and manipulates the correct vocabularies and factual knowledge about the domain and the data. Since such knowledge is often best encoded in knowledge graphs, we envision the computational infrastructure to include property graph and RDF DBMSes. However, not every piece of knowledge is relevant to every scenario. The right data can lead to better predictions and quicker goal achievement, while irrelevant or misplaced data, even if accurate, can cause hallucinations or erroneous conclusions. Although hallucinations cannot be entirely avoided, grounding the system to trustworthy sources effectively helps quantifying the model's uncertainty and provide explanations to the user. To this end, we propose a module that dynamically identifies the most relevant knowledge for the task at hand. This capability is crucial yet challenging in conversational systems, as data selection often occurs in real-time during user interactions.

Grounding is also achieved by exposing data-access APIs, accompanied by rich documentation, which in turn could also be generated semi-automatically by an LLM. Data access is achieved also by generating queries in the appropriate structured query languages or even data manipulation scripts. Thus, the model should be able to access a description of the schema of the data sources. Currently, this information is presented in textual form to the model [54]. Instead, we propose to encode this form of domain information in appropriate knowledge bases and enable the system to query and reason on these structures. Finally, entity extraction and entity linking processes will enrich a KG representation of both the schema and the contents of the data, i.e., by describing semantically the content of the data sources.

Real-world data often spans multiple modalities, such as text, tabular data, time series data, images, videos, and graphs. To fully leverage this diverse information, LLMs should be extended to access and integrate the complementary information embedded across these different data types. This can be addressed by creating effective dense representations of the different modalities in a unified space, forming a multimodal index. Additionally, developing a domain-specific multimodal knowledge graph to be utilized by the conversational system can significantly enhance its ability to model all the relevant data in the domain. Thus, we envision the necessity to redesign the KG data management systems, such that they can store, query, and manipulate directly a dense representation of entities and concepts for data representation of different modalities along the original plain data. Thus, this new data platform will offer a uniform and effective entry point to the data for the other modules.

Finally, since conversation logs with real users are part of the data sources used both during training and inference, the system needs to counteract the effect of any bias present in these logs, at processing time. The goal is to avoid that the model uses connoted or discriminatory language that perpetuates a distorted view of marginalized groups. We propose identifying such cases using approaches such as CADS (Corpus Assisted Discourse Analysis) [2] and sentiment analysis [51]. Human involvement will be fundamental, as it will be necessary to combine quantitative and qualitative analysis. Nonetheless, we see the need for new automatic methods for, at least partial, output evaluation to improve both effectiveness and accuracy in bias identification.

Explainability. Explainations of the various interactions and their possible alternative outcomes must be provided *succinctly*. This can be achieved by adding annotations to the pathways of question-answer in the natural language model layer. These kinds of annotation must allow concise summarization of primary sources, so as not to cognitively overload the user. The explanations should also be *consistent*: there should be contradictory explanations for the same outcome, and explanations of equivalent outcomes should also be equivalent. The annotation process should consist of adding metadata that is dynamically generated at conversation time. Those annotations must capture both *where-from* explanations, but also allow *where-to* analysis to support also the guidance ability of the system. To this end, we envision extending techniques already applied in query provenance computations [20]. Thus, to ensure explainability, provenance will be tracked across components, while the NL model needs to preserve this information received by the computational infrastructure when producing answers to the conversational data exploration system. Thus, in the provenance information, it must include data sources, query provenance, and code and APIs involved. Thanks to grounding, it should explain why the given sources are indeed appropriate. Furthermore, using provenance metadata captured for the sources, it should be possible to provide either a confidence score for the entire answer or for parts of the answer with differing scores and provenance information that can be used to trace back an answer to the pieces of information in the original sources that together led to it.

When LLMs are adopted, LLM interpretation (both interpreting LLMs and using LLMs for explanation) has the opportunity to redefine interpretability with a more ambitious scope across many applications, including in auditing LLMs themselves. Data-based interpretations will apply sequence summarization algorithms to a set of conversations and return a single summarized pattern enabling a better understanding of input conversations.

One viable approach is the fine-tuning of an LLM to take into account the propagation of annotations, in line with the current RAG-enabled explainability [45].

Soundness. A key aspect of delivering reliable answers and ensuring the correctness to a user's query is making sure that the response addresses the query as intended. Multi-turn dialogues are then a natural means to clarify user intent in the case of ambiguous queries. Enabling LLMs to conduct multi-turn dialogues also naturally provides a way of decomposing complex search queries in sequences of simpler steps, a process that was recently proposed as a way of replicating reasoning of data science and domain expertise [23, 53]. We envision "ask-and-refine" dialogues that, by actively integrating user inputs in each reasoning stage, go beyond the current LLM step-by-step reasoning approaches (e.g., Chain-of-Thoughts [62] and Tree-of-Thoughts [64]), and in addition combine database approaches of query reformulation with LLMs. At an implementation level, an active learning or active search component [28] could be in charge of eliciting feedback from users and actively probe the next question to ask with the goal of improving the answer certainty. To achieve soundness, the system should feature effective control methods for multi-modal generation that can produce high-quality, domain-aligned responses specifically tailored to conversational interfaces. The system will ensure correctness and answer quality by including direct control methods, e.g., offline reinforcement learning (ORL) [31], behavior cloning [30], and reward-augmented decoding [27]. Structured outputs can also be obtained through a combination of rejection sampling [26], constrained decoding and parsing [6]. The combination of these approaches offer enough flexibility to explore ways of optimizing the generation and ensure accurate and relevant results. We envision making the model generate robust and controllable outputs by re-purposing preference learning techniques borrowed from the alignment literature such as Reinforcement Learning from Human Feedback (RLHF) using reward models and policy optimization algorithms like PPO [47], or direct alignment algorithms such as Direct Preference Optimization (DPO) [42], combined supervised fine-tuning, and preference learning (ORPO) [21], and Alignment via Optimal Transport (AOT) [35].

Additional control components should be included in the system in tandem with uncertainty quantification methods [7, 29, 33] to quantitatively assess uncertainty. The user should control how the model defines correct and incorrect answers. Correct answers should form dense, separable n-dimensional shapes, while incorrect answers should be sparse, dissimilar to each other, and distinct from correct ones, with their frequency minimized during training. To introduce control during inference, the constrained decoding techniques should be redesigned to account for different types of data and depending on the input user-system interaction data; if the data has sufficient variability but a smaller vocabulary, imitation learning is preferred, otherwise ORL is more robust. Based on the results of previous reinforced active learning methods [63], we foresee that by exploring combinations of these methods the system can attain superior results.

Guidance. Our view of guidance is akin to adding planning to LLMs [22], but without necessarily knowing the objective beforehand. Interactions between the user and the system should be modeled as a first-class citizen. Systems like LangChain enable planning [22] and are designed to manage complex tasks. LangChain uses LangGraph, a specialized type of map, to connect and coordinate multiple "LLM agents". Each LLM agent can be customized with its own instructions and access to specific tools. This is illustrated in the AutoDev framework [55] that uses a combination of agents and tools to facilitate automated software development tasks. Our proposal is to generalize the previously proposed frameworks for Agentic AI [1, 11] to capture multi-modal human-AI and AI-AI interactivity. We propose to develop a new graph-based data model that captures the intricacies of relying on a mix of structured queries, LLMs, and human interactions to guide conversations. Nodes in the graph will represent LLMs or humans enabling to capture, from the ground up, the logic of recommending the next steps in interactive conversations. This would require every component of the architecture to "understand and manipulate" the proposed graph model.

The graph model runs alongside algorithms that guide the prompting strategy based on either previously successful tasks (as judged by human annotators), or by self-reasoning LLM agents via LLM reasoning algorithms [5, 19, 56, 62, 64] and running alternative scenarios behind the scenes. There are multiple libraries that can be adopted for LLM reasoning [18]. Finally, the systems, through profiling, should determine the level of expertise of the user and interact differently according to the inferred expertise.

Evaluation. Each system component will be evaluated with appropriate evaluation metrics. In general, common performance measures are still relevant, these include Precision, Recall, F1-score, Area Under the ROC Curve (AOC), Accuracy, Mean Least Square Error (MLSE) and perplexity for prediction tasks, MRR and NDCG for ranking tasks, as well as system measures, such as time, number of operations and memory consumption for efficiency. In addition, grounding will still require qualitative and manual assessment in tasks that analyze the language in the training data. Instead, new metrics may be needed to evaluate the quality of explainability or the probabilistic interpretation of any correctness estimation. There is further the need for new, end-to-end, benchmarks, that includes also user-interface considerations, for CDA system based on real-world multi-modal use cases that can span different domains, from healthcare, to education, e-commerce and business intelligence. Therefore, close collaboration with domain experts is vital to comprehensively evaluate all relevant aspects and requirements.

4 CONCLUSIONS

In this vision paper, we raise the question of reliability of CDA systems. Thus, we propose a first definition based on five cornerstone properties: efficiency in providing and assessing the quality of an answer, grounding to the terminology and the data values describing the domain as well as of the appropriate processes, explainability of any result of the interactions with the system, soundness of the answer produced and of the decisional processes, and guidance for the user towards the correct answers. To provide these properties, data management solutions will be central. We argue that current systems that rely almost completely on natural language processing cannot provide these properties. Thus, we propose a more comprehensive system for achieving reliable CDA. While entirely revamped architectures, where reliability is a first-class citizen, is a promising direction, in this paper, we proposed as a first step a solution that extends existing components to enforce reliability properties.

ACKNOWLEDGMENTS

This work has been partially supported by ARMADA, funded by the European Union's Horizon Europe Marie Skłodowska-Curie Actions (MSCA), under grant agreement No. 101168951.

REFERENCES

[1] Sihem Amer-Yahia. 2024. Intelligent Agents for Data Exploration. *Proc. VLDB Endow.* 17, 12 (2024).

[2] Ancarno, Clyde. 2020. *Corpus-Assisted Discourse Studies*. Cambridge University Press, 165–185.

[3] Ilias Azizi, Karima Echihabi, and Themis Palpanas. 2023. Elpis: Graph-Based Similarity Search for Scalable Data Science. *Proc. VLDB Endow.* 16, 6 (2023), 1548–1559.

[4] Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navratil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. 2024. Auditing and generating synthetic data with controllable trust trade-offs. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* (2024).

[5] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *AAAI Conference on Artificial Intelligence* 38, 16 (2024). http://arxiv.org/abs/2308.09687

[6] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2024. Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation. In *Forty-first International Conference on Machine Learning*.

[7] Debarun Bhattacharjya, Balaji Ganesan, Michael Glass, Junkyu Lee, Radu Marinescu, Katya Mirylenka, and Xiao Shou. 2024. Consistency-based Black-box Uncertainty Quantification for Text-to-SQL. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

[8] Robin Chan, Katsiaryna Mirylenka, Thomas Gschwind, Christoph Miksovic, Paolo Scotton, Enrico Toniato, and Abdel Labbi. 2024. Adapting LLMs for Structured Natural Language API Integration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 991–1000.

[9] Manos Chatzakis, Panagiota Fatourou, Eleftherios Kosmas, Themis Palpanas, and Botao Peng. 2023. Odyssey: A Journey in the Land of Distributed Data Series Similarity Search. *Proc. VLDB Endow.* 16, 5 (2023), 1140–1153.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 1107–1128. https://doi.org/10.18653/v1/2024.emnlp-main.64

[11] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568* (2024).

[12] Karima Echihabi, Panagiota Fatourou, Kostas Zoumpatianos, Themis Palpanas, and Houda Benbrahim. 2022. Hercules Against Data Series Similarity Search. *Proc. VLDB Endow.* 15, 10 (2022), 2005–2018.

[13] Karima Echihabi, Theophanis Tsandilas, Anna Gogolou, Anastasia Bezerianos, and Themis Palpanas. 2023. ProS: data series progressive k-NN similarity search and classification with probabilistic quality guarantees. *VLDB J.* 32, 4 (2023), 763–789.

[14] Talya Eden, Piotr Indyk, Shyam Narayanan, Ronitt Rubinfeld, Sandeep Silwal, and Tal Wagner. 2021. Learning-based Support Estimation in Sublinear Time. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=tilovEHA3YS

[15] Panagiota Fatourou, Eleftherios Kosmas, Themis Palpanas, and George Paterakis. 2023. FreSh: A Lock-Free Data Series Index. In *42nd International Symposium on Reliable Distributed Systems, SRDS 2023, Marrakesh, Morocco, September 25-29, 2023*. IEEE, 209–220.

[16] Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan. 2023. How Large Language Models Will Disrupt Data Management. *Proc. VLDB Endow.* 16, 11 (jul 2023), 3302–3309. https://doi.org/10.14778/3611479.3611527

[17] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Yoav Artzi and Jacob Eisenstein (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2–7. https://doi.org/10.18653/v1/P18-5002

[18] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. 2024. LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models. *arXiv preprint arXiv:2404.05221* (2024). https://www.llm-reasoners.net/

[19] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. In *Conference on Empirical Methods in Natural Language Processing*. 8154–8173.

[20] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A survey on provenance: What for? What form? What from? *VLDB J.* 26, 6 (2017), 881–906. https://doi.org/10.1007/S00778-017-0486-1

[21] Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic Preference Optimization without Reference Model. *arXiv preprint arXiv:2403.07691* (2024).

[22] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716* (2024).

[23] Madelon Hulsebos, Wenjing Lin, Shreya Shankar, and Aditya Parameswaran. 2024. It Took Longer than I was Expecting: Why is Dataset Search Still so Hard?. In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics* (Santiago, AA, Chile) *(HILDA 24)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3665939.3665959

[24] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 7969–7992. https://doi.org/10.18653/v1/2023.emnlp-main.495

[25] M. L. Kersten. 2015. Big Data Space Fungus. In *CIDR*. www.cidrdb.org. http://cidrdb.org/cidr2015/Papers/21_Abstract58KM.pdf

[26] Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. 2024. RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 1665–1680.

[27] Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. ARGS: Alignment as Reward-Guided Search. In *The Twelfth International Conference on Learning Representations*.

[28] Nikita Klyuchnikov, Davide Mottin, Georgia Koutrika, Emmanuel Müller, and Panagiotis Karras. 2019. Figuring out the user in a few steps: Bayesian multifidelity active search with cokriging. In *KDD*. 686–695.

[29] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=VD-AYtP0dve

[30] Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. 2022. Should I Run Offline Reinforcement Learning or Behavioral Cloning?. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=AP1MKT37rJ

[31] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[32] Conglong Li, Minjia Zhang, David G. Andersen, and Yuxiong He. 2020. Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo (Eds.). ACM, 2539–2554.

[33] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187* (2023).

[34] Yuanxing Liu, Weinan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li, and Wanxiang Che. 2023. Conversational Recommender System and Large Language Model Are Made for Each Other in E-commerce Pre-sales Dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 9587–9605. https://doi.org/10.18653/v1/2023.findings-emnlp.643

[35] Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. 2024. Distributional Preference Alignment of LLMs via Optimal Transport. *arXiv preprint arXiv:2406.05882* (2024).

[36] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proc. VLDB Endow.* 16, 4 (dec 2022), 738–746. https://doi.org/10.14778/3574245.3574258

[37] Apoorva Nitsure, Youssef Mroueh, Mattia Rigotti, Kristjan Greenewald, Brian Belgodere, Mikhail Yurochkin, Jiri Navratil, Igor Melnyk, and Jarret Ross. 2024. Risk Aware Benchmarking of Large Language Models. In *Forty-first International Conference on Machine Learning*.

[38] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: Lessons and challenges. *ACM Queue* 17, 2 (2019), 48–75.

[39] Bruno Ordozgoiti, Ananth Mahadevan, Antonis Matakos, and Aristides Gionis. 2022. Provable randomized rounding for minimum-similarity diversification. *Data Min. Knowl. Discov.* 36, 2 (2022), 709–738. https://doi.org/10.1007/S10618-021-00811-2

[40] Botao Peng, Panagiota Fatourou, and Themis Palpanas. 2021. SING: Sequence Indexing Using GPUs. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*. IEEE, 1883–1888.

[41] Abdul Quamar, Vasilis Efthymiou, Chuan Lei, Fatma Özcan, et al. 2022. Natural language interfaces to data. *Foundations and Trends in Databases* 11, 4 (2022), 319–414.

[42] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=HPuSIXJaa9

[43] Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. 2024. Parameter-Efficient Conversational Recommender System as a Language Processing Task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, 152–165. https://aclanthology.org/2024.eacl-long.9

[44] G. Rioux, A. Nitsure, M. Rigotti, K. Greenewald, and Y. Mroueh. 2024. Multivariate Stochastic Dominance via Optimal Transport and Applications to Models Benchmarking. *Advances in Neural Information Processing Systems (NeurIPS)* 38 (2024).

[45] Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. 2024. RAGE Against the Machine: Retrieval-Augmented LLM Explanations. In *International Conference on Data Engineering (ICDE)*. 5469–5472.

[46] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[48] Shreya Shankar, Haotian Li, Parth Asawa, Madelon Hulsebos, Yiming Lin, JD Zamfirescu-Pereira, Harrison Chase, Will Fu-Hinthorn, Aditya G Parameswaran, and Eugene Wu. 2024. Spade: Synthesizing assertions for large language model pipelines. *arXiv preprint arXiv:2401.03038* (2024).

[49] Harsha Vardhan Simhadri, George Williams, Martin Aumüller, Matthijs Douze, Artem Babenko, Dmitry Baranchuk, Qi Chen, Lucas Hosseini, Ravishankar Krishnaswamny, Gopal Srinivasa, Suhas Jayaram Subramanya, and Jingdong Wang. 2022. Results of the NeurIPS'21 Challenge on Billion-Scale Approximate Nearest Neighbor Search. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track (Proceedings of Machine Learning Research, Vol. 176)*, Douwe Kiela, Marco Ciccone, and Barbara Caputo (Eds.). PMLR, 177–189. https://proceedings.mlr.press/v176/simhadri22a.html

[50] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761* (2024).

[51] Maite Taboada. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2, 1 (2016), 325–347.

[52] Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[53] Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns* 4, 4 (2023), 100729. https://doi.org/10.1016/j.patter.2023.100729

[54] Immanuel Trummer. 2024. Generating Succinct Descriptions of Database Schemata for Cost-Efficient Prompting of Large Language Models. *Proceedings of the VLDB Endowment* 17, 11 (2024), 3511–3523.

[55] Michele Tufano, Anisha Agarwal, Jinu Jang, Roshanak Zilouchian Moghaddam, and Neel Sundaresan. 2024. AutoDev: Automated AI-Driven Development. *arXiv preprint arXiv:2403.08299* (2024).

[56] Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning. http://arxiv.org/abs/2406.14283

[57] Qitong Wang and Themis Palpanas. 2023. SEAnet: A Deep Learning Architecture for Data Series Similarity Search. *IEEE Trans. Knowl. Data Eng.* 35, 12 (2023), 12972–12986.

[58] Xi Wang, Hossein Rahmani, Jiqun Liu, and Emine Yilmaz. 2023. Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3609–3622. https://doi.org/10.18653/v1/2023.findings-emnlp.233

[59] Zeyu Wang, Peng Wang, Themis Palpanas, and Wei Wang. 2023. Graph- and Tree-based Indexes for High-dimensional Vector Similarity Search: Analyses, Comparisons, and Future Directions. *IEEE Data Eng. Bull.* 46, 3 (2023), 3–21. http://sites.computer.org/debull/A23sept/p3.pdf

[60] Zeyu Wang, Qitong Wang, Peng Wang, Themis Palpanas, and Wei Wang. 2023. Dumpy: A Compact and Adaptive Index for Large Data Series Collections. *Proc. ACM Manag. Data* 1, 1 (2023), 111:1–111:27.

[61] Jiuqi Wei, Botao Peng, Xiaodong Lee, and Themis Palpanas. 2024. DET-LSH: A Locality-Sensitive Hashing Scheme with Dynamic Encoding Tree for Approximate Nearest Neighbor Search. *Proc. VLDB Endow.* 17, 9 (2024), 2241–2254. https://www.vldb.org/pvldb/vol17/p2241-wei.pdf

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *International Conference on Neural Information Processing Systems*.

[63] Lukas Wertz, Jasmina Bogojeska, Katsiaryna Mirylenka, and Jonas Kuhn. 2023. Reinforced Active Learning for Low-Resource, Domain-Specific, Multi-Label Text Classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10959–10977. https://doi.org/10.18653/v1/2023.findings-acl.697

[64] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *International Conference on Neural Information Processing Systems*.

[65] Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024. The Shift from Models to Compound AI Systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.