Introducing the Data Access Agent Benchmark

Your Name

November 13, 2024

Abstract

An open problem in enterprise AI deployment is building systems that can effectively access, process, and reason over private organizational data. While language models have shown impressive capabilities with public knowledge, their ability to work with private enterprise data remains limited. Most enterprise AI assistants today can only handle basic predefined workflows and struggle with novel requests or complex data operations. To address this challenge and measure the effectiveness of AI systems in enterprise settings, we are releasing the Data Access Agent Benchmark (DAAB), a comprehensive evaluation framework for assessing AI systems' ability to work with private enterprise data.

1 Introduction

An open problem in enterprise AI deployment is building systems that can effectively access, process, and reason over private organizational data. While language models have shown impressive capabilities with public knowledge, their ability to work with private enterprise data remains limited. Most enterprise AI assistants today can only handle basic predefined workflows and struggle with novel requests or complex data operations. To address this challenge and measure the effectiveness of AI systems in enterprise settings, we are releasing the Data Access Agent Benchmark (DAAB), a comprehensive evaluation framework for assessing AI systems' ability to work with private enterprise data.

The ability of an AI system to autonomously access and operate on enterprise data - what we call "agentic data access" - is crucial for building truly useful enterprise AI solutions. DAAB is designed to evaluate complete AI systems, which may include multiple models, tools, and retrieval pipelines, rather than focusing on individual models or components in isolation.

Current benchmarks face several limitations when applied to enterprise settings:

Data Privacy Context: Existing benchmarks predominantly focus on public knowledge. While they can incorporate retrieval and tool use, they fail to capture the unique challenges of private organizational data. Enterprise environments typically involve sensitive, proprietary information based on dynamic security rules.

Data Source Complexity: Traditional benchmarks evaluate performance on a single source of data or database. Enterprise data typically spans multiple systems and formats. Integration challenges are often overlooked in current evaluations

Task Authenticity: Most benchmarks rely on synthetic or academic tasks (factual question-answering on wikipedia like data being most common). Even advanced datasets like GAIA29 fail to capture typical business queries. Real business questions often involve a combination of information retrieval, exploration and decision making.

Data Dynamism: Existing benchmarks typically use static datasets. Enterprise systems must handle dynamic, continuously updating data sources.

User-goal Evaluation: Traditional benchmarks focus on isolated capabilities (like retrieval or reasoning) or specific tasks (factual question-answering or fact verification) in a restricted way e.g. single attempt answer, rather than assessing how a typical user would use an AI system to achieve their goal.

With DAAB, our goal is to create a set of real-world questions/user-goals with characteristics such as i) Real-world relevance: Questions are derived from actual enterprise use cases across different domains like Customer Support, Sales, HR and Engineering Management. ii) Comprehensive evaluation: The benchmark tests various complexity levels of data access and computation, from simple retrieval to complex multi-step operations. iii) System-agnostic: The benchmark is meant to evaluate any AI system architecture, whether it uses a single large model or multiple specialized components.

2 Prior Art

2.1 Evolution of LLM Benchmarks

The landscape of AI benchmarks has evolved significantly over the last few years. Early benchmarks like SQuAD[1], TriviaQA[2], and NaturalQuestions[3] focused primarily on evaluating question-answering capabilities over public information (mainly using Wikipedia as source of truth). For database question-answering, specialized datasets like Spider[4], WikiSQL[5], and BIRD[6] have emerged to evaluate a language model's ability to translate natural language queries into SQL statements (text-to-SQL). With the emergence of Large Language Models (LLMs), these traditional benchmarks have been largely surpassed. Recent evaluations like MMLU[7], AGIEval[8], Big-Bench (Hard)[9, 10] have ramped up efforts to assess broader reasoning and language understanding capabilities more appropriate for modern LLMs. There is still significant energy in both academia and industry in crafting representative datasets for benchmarking different LLM capabilities as seen most recently by OpenAI's SimpleQA[11] benchmark.

2.2 Benchmarks for LLMs with External Systems

While traditional benchmarks evaluate LLM capabilities relying on the implicit knowledge embedded within model parameters, a new set of benchmarks has emerged to assess LLM performance when augmented with external systems. Such augmentation is necessary when responses need to be grounded[12] in facts that are present in these external systems. Though fine-tuning LLMs[13] with additional data represents one augmentation approach, practical considerations like training costs, keeping models up-to-date, and security considerations have limited its industry adoption. Instead, current industry practices primarily fall into two categories:

1. Retrieval-augmented approaches[14, 15, 16]: This approach incorporates explicit retrieval steps of relevant data from a large database and provided alongside the user input. Most classic benchmarks like TriviaQA[2] and NaturalQuestions[3] have been used for evaluating retrieval-augmented techniques by using a Wikipedia database. There are various benchmarks like MTEB[17] which measure the capabilities of text embedding models (like S-BERT[18]) typically used in the retrieval phase. More complex benchmarks like HotPotQA[19] (for multi-hop questions) and FEVER[20] (for fact verification) enhance the task complexity for LLMs and illustrate the need to use retrieval-augmented approaches. More recently, benchmarks like FRAMES[21] and CRAG[22] have emerged to make the questions in the dataset more real-world.

2. Tool use[23, 24]: Tool use enables LLMs to dynamically request and interact with external tools - from simple calculators to data APIs and web search capabilities. To measure the ef-

ficiency of an LLM to use tools correctly, benchmarks like ToolQA[25], API-bank[26], API-bench[27] and ToolBench[28] are used for evaluating tool use capability of the LLM. More recently, tool use for general assistants has paved the way for benchmarks like GAIA[29].

3 About the Data Access Agent Benchmark

3.1 Dataset

DAAB contains approximately 150 questions across different enterprise domains. Each question is categorized along multiple dimensions:

Domain: The domain the question belongs to. As of writing, this could be one of the 5 domains: Customer support, Email+calendar, Sales, HR and Engineering Management. More about these domains is described in Appendix A. Although these are specific domains, we hope that the use-cases in these domains are generalizable to any domain and hence provide valuable insights into different kinds of questions that a human user might ask an AI system.

Agentic Complexity Level: Complexity level is one of 3 values: Low, Medium, or High.

- Low: Questions which use a single planning component e.g. Get user with id=1. This is a straightforward lookup on the user table.
- Medium Questions which require two planning components composed together e.g. Show me travel confirmation from American Airlines. This question requires looking up emails and classifying them as travel confirmation from American Airlines
- High: Questions which require more than two planning components composed together e.g. Summarize any interesting lead activities that I should look into for prospecting. This question requires retrieval of lead activities followed by classifying them as relevant for prospecting and finally summarizing all the results

Agentic Complexity Aspects: These are one of: Smart search strategy, Following connections, NLP compute and logic compute.

- Smart search strategy: This involves being able to choose a reasonable strategy for retrieving relevant data. This is especially important when it's not obvious what or how much data to retrieve
- Following connections: This involves getting related data by following connections to it. Usually insights and decision making is done by looking at a set of related data
- NLP compute: This involves extracting valuable information from unstructured or textual data
- Logic compute: This involves numeric or logic based computation

Use-case Category: A fundamental use-case pattern. These are described in detail in the next section

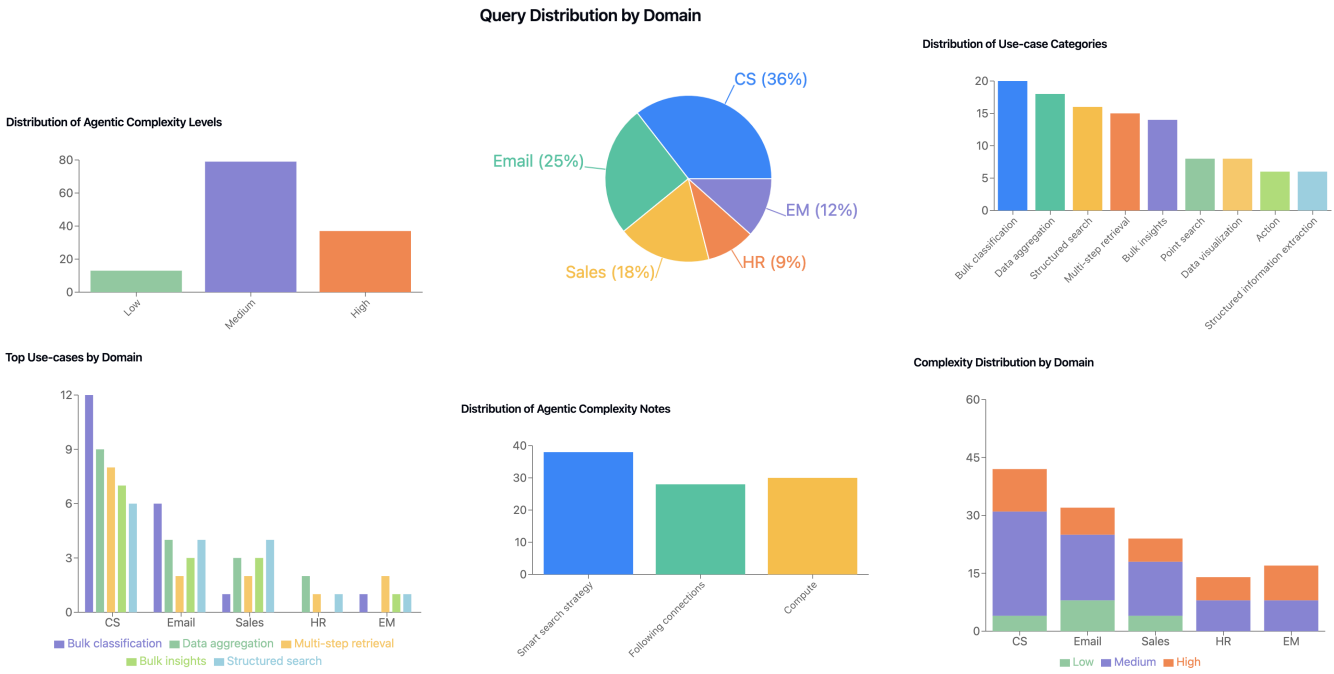
3.2 Use Case Categories

The benchmark identifies several fundamental use-case categories.:

1. *Attribute search*: Attribute search is retrieving data by filtering or ordering on an attribute e.g. get user with id=1 or get latest ticket

2. *Multi-step data retrieval*: Tasks requiring data from multiple source to be fetched e.g. get the projects, invoices and usage for tina.jackson@gray-smith.com
3. *Data aggregation*: Tasks requiring computing statistics over raw data e.g. what's the average length of our sales cycle this year
4. *Summarization/General analysis*: Looking at sample data and aggregates to give an overview e.g. How are project creations looking this month
5. *Bulk insights*: Analyzing multiple data items and extracting key insights from each e.g. Give summary of all tickets created today
6. *Bulk classification*: Classify multiple data items into preset categories e.g. categorize all unlabled tickets as one of operational, feature request, bug report, how-to or others.
7. *Decision making*: Involves analyzing different data and making informed choices as to next course of action e.g. help me resolve ticket 1234
8. *Clustering*: Identifying similar properties in bulk data e.g. Extract any common identified pain points for my opportunities
9. *Information search*: Searching for all relevant documents relevant for a query string e.g. product roadmap for q3
10. *Point search*: Finding specific items which satisfy complex descriptions e.g. Get the email where Tina mentioned performance issues
11. *Structured information extraction*: Extracting key-value pairs from unstructured text e.g. extract the invoice numbers, amounts, and due dates from uber last week
12. *Data visualization*: Human-friendly data transformation e.g. what is the breakup of lead sources this week?
13. *Action*: Tasks that have a state changing effect e.g. create reminder 10 mins before my next meeting

3.3 Question Distribution



4 Evaluation Methodology

The Data Access Agent Benchmark employs a goal-oriented evaluation metric[**42**], recognizing that while many dataset questions could be approached in a zero-shot (or question-answer) setting, users typically interact with the agent through an assistant. This interaction often evolves into a dialogue where users provide helpful hints to guide task completion based on previous responses.

Most evaluation studies on dialogue systems follow the PARADISE[**43**] framework. This framework evaluates user satisfaction through two key metrics: dialogue cost and task success. Dialogue cost measures interaction costs, such as the number of conversation turns, while task success assesses whether the system successfully achieves the user's goal. We can use a weighted combination of both metrics to come up with an overall score, although we believe that the task success alone may be reasonably considered as the main metric.

Given the challenges of evaluating dialogue responses on enterprise settings[**19**], we will mainly rely on human assessment of the dialogues to measure the metrics.

5 Discussion: Technical Approaches and their Limitations

We will briefly overview few common approaches to build systems on enterprise data and discuss their limitations.

RAG: RAG is a type of retrieval-augmented approach[**10**],[**11**],[**36**] where text-embedding based search is used to enhance the grounding of the responses generated by the LLM. More recently, techniques like GraphRAG[**41**] aim to augment text-embedding search with other techniques (like building knowledge graphs) to improve task performance. Although few use-cases like information search (described later) cater well to RAG approaches, it faces key limitations: lack of temporal awareness and attribute filtering, challenging to maintain user authorization with embeddings based search, and perhaps most importantly: limited to textual data.

text-to-SQL: Database question-answering leverages text-to-SQL[**12**] systems for data retrieval, but is limited in it's scope. The reality of enterprise data is that it's usually spread across multiple databases and is stored in different column formats (complex types like JSON/STRUCT also becoming quite common). Moreover, SQL systems can only ever answer filter-based search or data aggregation kind of questions. SQL's inherent limitations make it unsuitable for tasks like summarization, classification and business-logic based computations.

Tool use: Tool use[**25**], [**30**] in LLMs provides the most flexible approach for retrieving data and executing tasks, but faces practical challenges: missing APIs, poor documentation for APIs, and increased hallucination risk when handling large data inputs for these tools. These limitations significantly impact the reliability of tool use based approaches in real-world environments.

6 Extending DAAB and Future Work

7 Conclusion

The Data Access Agent Benchmark represents a significant step forward in evaluating enterprise AI systems' ability to work with private data. By providing a standardized way to assess these capabilities, we hope to drive progress in building more effective enterprise AI solutions that can truly understand and work with organizational data.

The complete benchmark dataset is available at [TODO: Add link to dataset].

We invite the AI community to use this benchmark in evaluating their enterprise AI systems and welcome feedback on making it even more useful for measuring progress in this crucial area.

References

[1] Pranav Rajpurkar et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *arXiv preprint arXiv:1606.05250* (2016).

[2] Mandar Joshi et al. "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension". In: *arXiv preprint arXiv:1705.03551* (2017).

[3] Tom Kwiatkowski et al. "Natural Questions: A Benchmark for Question Answering Research". In: *Transactions of the Association for Computational Linguistics* (2019).

[4] Tao Yu et al. "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

[5] Victor Zhong, Caiming Xiong, and Richard Socher. "Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning". In: *CoRR* abs/1709.00103 (2017).

[6] Jinyang Li et al. "Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs". In: *arXiv preprint arXiv:2305.03111* (2023).

[7] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". In: *arXiv preprint arXiv:2009.03300* (2020).

[8] Unknown. "AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models". In: (2023).

[9] Aarohi Srivastava et al. "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". In: *arXiv preprint arXiv:2206.04615* (2022).

[10] Mirac Suzgun et al. "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them". In: *arXiv preprint arXiv:2305.06617* (2023).

[11] Jason Wei et al. "SimpleQA: A Factuality Benchmark for Language Models". In: *arXiv preprint* (2024).

[12] Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. "Grounding and Evaluation for Large Language Models: Practical Challenges and Lessons Learned (Survey)". In: ACM, Aug. 2024, 6523–6533.

[13] Guanghui Chen et al. "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation". In: *arXiv preprint arXiv:2305.16938* (2023).

[14] Kelvin Guu et al. "REALM: Retrieval-Augmented Language Model Pre-Training". In: *arXiv preprint arXiv:2002.08909* (2020).

[15] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *arXiv preprint arXiv:2005.11401* (2020).

[16] Sebastian Borgeaud et al. "Improving language models by retrieving from trillions of tokens". In: *arXiv preprint arXiv:2112.04426* (2022).

[17] Niklas Muennighoff et al. "MTEB: Massive Text Embedding Benchmark". In: *arXiv preprint arXiv:2210.07316* (2023).

[18] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019.

[19] Zhilin Yang et al. “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering”. In: *arXiv preprint arXiv:1809.09600* (2018).

[20] James Thorne et al. “FEVER: a large-scale dataset for Fact Extraction and VERification”. In: *arXiv preprint arXiv:1803.05355* (2018).

[21] Bill Yuchen Lin et al. “FRAMES: Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation”. In: *arXiv preprint arXiv:2402.03783* (2024).

[22] Xiao Yang et al. *CRAG – Comprehensive RAG Benchmark*. 2024.

[23] Timo Schick et al. “Toolformer: Language Models Can Teach Themselves to Use Tools”. In: *arXiv preprint arXiv:2302.04761* (2023).

[24] Shunyu Yao et al. “ReAct: Synergizing Reasoning and Acting in Language Models”. In: *arXiv preprint arXiv:2210.03629* (2023).

[25] Unknown. “ToolQA: A Dataset for LLM Question Answering with External Tools”. In: (2023).

[26] Unknown. “API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs”. In: (2023).

[27] Shishir Patil Sun et al. “Gorilla: Large Language Model Connected with Massive APIs”. In: *arXiv preprint arXiv:2305.15334* (2023).

[28] Yujia Qin et al. “On the Tool Manipulation Capability of Open-source Large Language Models”. In: *arXiv preprint arXiv:2305.16504* (2023).

[29] Grégoire Mialon et al. “GAIA: A benchmark for General AI Assistants”. In: (2023).