Do Data Agents Need Semantic Metadata? A Comparative Study in Agentic Data Retrieval

Shiyu Chen, Tarfah Alrashed, Alon Halevy, and Natasha Noy

Google, Mountain View, CA, USA
{shiyuc,tarfah,halevy,noy}@google.com

Abstract. In the era of autonomous agents, machine-actionable data is critical for data-driven workflows. For more than a decade, semantic metadata like `schema.org` has anchored the FAIR principles (Findable, Accessible, Interoperable, and Reusable) for machine-actionable data and enabled discovery tools like Google Dataset Search. However, the rise of Large Language Models (LLMs) capable of navigating the unstructured web raises a fundamental question: Is semantic metadata still necessary for agentic data discovery, or can agents reliably retrieve actionable data directly from the web? We present a comparative analysis of agentic data retrieval across two distinct environments: a Baseline Agent searching billions of open-web documents, and a Semantic Agent leveraging a corpus of 90 million datasets using `schema.org`. We deploy an "LLM-as-a-judge" evaluation pipeline, mapped directly to the FAIR principles, to assess the semantic relevance, data accessibility, and computational utility of the retrieved data. Our results reveal a clear divergence. The Semantic Agent excels at retrieving actionable data, achieving a 44.9% higher precision for metadata-rich registries and a 46.6% higher precision for pages with machine-readable downloads among its returned results. Conversely, the Baseline Agent frequently suffers "Last-Mile Utility" failures, retrieving prose-heavy pages (20.1% of results) and portal landing pages (8.5%) rather than actual data pages. While the Baseline Agent achieves higher coverage by answering 40% more questions, the Semantic Agent delivers greater accuracy, achieving 65.7% higher overall precision in retrieving FAIR-compliant datasets. We conclude that while unstructured retrieval supports broad exploratory tasks, structured ecosystems remain the indispensable foundation for reliable, execution-oriented autonomous workflows.

Keywords: Semantic Metadata · Data Discovery · Large Language Models · Autonomous Agents · Retrieval-Augmented Generation

1 Introduction

As the world becomes increasingly data-driven, researchers and practitioners rely heavily on open data to answer scientific questions and to understand complex phenomena [38]. This deep reliance on data has catalyzed a shift in scientific communication, making publication of primary data the norm across many scientific

disciplines [31]. Users do not merely need to read about data; they frequently need to find it and to act upon it—for example, to conduct independent analyses of primary sources or to replicate experimental results.

Today, there are tens of millions of datasets scattered across the Web. Finding specific, usable data in this vast, unstructured space is a massive challenge [4]. We built Google Dataset Search [22] specifically for this use case. Recognizing that traditional Web search engines are optimized for narrative prose rather than structured data, we relied on semantic metadata—specifically `schema.org` [11] and W3C DCAT [2][3]—to construct a dedicated vertical search engine. By incentivizing publishers to add explicit semantic metadata for the datasets, this structured ecosystem organized the chaotic subset of the Web describing datasets into a reliable, deterministic corpus for data discovery [5].

Historically, semantic metadata facilitated the FAIR principles—ensuring data is Findable, Accessible, Interoperable, and Reusable [34]. These principles emphasize machine-actionability: the capacity for computational systems to utilize this data with minimal human intervention. Today, the emergence of autonomous agents [35, 25, 28] transforms FAIR from a human-centric best practice into a technical requirement. For an agent, *findability* is merely the starting point. Bridging the gap to data execution requires machine-actionable *accessibility*—such as direct APIs and machine-readable formats—so agents can process the data seamlessly. Furthermore, *interoperability* via semantic attributes ensures agents understand dataset schemas, streamlining the integration of extracted data into downstream workflows. Finally, *reusability* provides the provenance and licensing context needed to verify whether or not it is appropriate for the agent to use the asset, preventing the blind extraction of unverified data. Ultimately, the advent of agents does not diminish the need for FAIR data; rather, it makes FAIR the essential foundation for reliable autonomous workflows.

Simultaneously, modern LLMs have demonstrated a remarkable capability to navigate the unstructured web directly. A dataset-discovery agent built entirely on traditional web search [21, 37] can extract context straight from HTML prose, effectively bypassing traditional metadata structures. This mechanism stands in contrast to the structured ecosystem of the Semantic Web, where semantic metadata constitutes the ground truth—unambiguous, machine-readable statements guaranteeing schema and provenance. The capability with which LLMs navigate the unstructured pages raises a critical question: if an agent locates data without these semantic guarantees, is that asset truly FAIR and actionable, or simply findable?

This friction drives our core research question: In an agent-led system, do we still need semantic metadata to empower data discovery, or can LLMs natively bridge the gap through unstructured retrieval?

To answer this question, we present a comparative analysis of agentic data retrieval across two environments: Using a similar agent architecture, we evaluate a **Baseline Agent** (billions of open-web documents accessed via general web search) against a **Semantic Agent**. The Semantic Agent operates over a corpus

of 90 million dataset metadata records from Google Dataset Search, which relies on the presence of `schema.org` semantic markup (Section 3).

To assess retrieval quality in a scalable way, we deploy a multi-dimensional "LLM-as-a-judge" pipeline [36] mapped directly to the FAIR principles. Using keyword-based queries from the NTCIR-16 Data Search benchmark [15], our autoraters evaluate the retrieved assets across semantic relevance, data accessibility, and computational utility (Section 4).

The following **research questions** drive our comparative framework:

- **Data Actionability**: Does semantic metadata provide an advantage in surfacing actionable datasets compared to unstructured web search systems? *Hypothesis:* Semantic Agent will yield a significantly higher success rate for locating actionable, downloadable datasets.
- **The "Last Mile" Utility**: When unstructured web search fails in identifying actionable data, what is the reason? Is it contextual divergence—where the model loses focus on the data objective in favor of the surrounding prose—or explicit hallucination? *Hypothesis:* Baseline Agent often fails the "last mile" of data retrieval, landing on web pages requiring further navigation or complex extraction rather than immediate data payloads.
- **Exploratory Breadth vs. Precision**: How do the two agents balance exploratory recall against retrieval precision? *Hypothesis:* The Baseline Agent will achieve higher overall query recall by acting as a broader discovery superset for niche topics lacking semantic markup, whereas the Semantic Agent will deliver significantly higher precision and more focused results.

In summary, this paper makes the following contributions:

- We provide a comparative study evaluating the end-to-end utility of a semantic metadata ecosystem for agentic data retrieval.
- We introduce a multi-dimensional, LLM-as-a-judge evaluation pipeline that translates the human-centric FAIR data principles into metrics for autonomous systems, extending evaluation beyond traditional semantic relevance.
- We publish the exact prompts used by our autoraters to ensure full transparency and foster reproducibility of our LLM-as-a-judge methodologies.

2 Related Work

The challenge of agentic data retrieval sits at the intersection of RAG, dataset discovery, and Semantic Web technologies. Despite the reasoning power of LLMs, integrating them with external search remains a research challenge, specifically regarding data reliability and utility.

Autonomous Agents and Web Retrieval Autonomous agents built on RAG [16] leverage frameworks like ReAct [35], Toolformer [28], and ToolLLM [27] for multi-step reasoning and tool execution. For web tasks, systems such as WebGPT [21] and WebArena [37] demonstrate that agents can navigate raw HTML

to fulfill complex intents, but DOM traversal introduces high computational overhead and boilerplate noise that degrades reasoning [12]. Although distilling pages into Markdown or accessibility trees [37, 7] and pruning HTML [30] improve context management, agents with multi-step navigation still suffer from misgrounding and poor state tracking [1]. To bypass these bottlenecks, research has pivoted toward explicit API-calling for hallucination-free execution [26]. Building on these findings, rather than optimizing multi-step scraping for data discovery, we restrict our agents to the search tools. By comparing agents querying different web ecosystem endpoints, we demonstrate that the metadata endpoint natively solves "last mile" failure of data retrieval. This approach sidesteps the computational overhead of multi-page scraping entirely while landing on a page with actionable data.

Evaluating Dataset-Discovery Engines Unlike general web search, dataset retrieval requires meeting specific spatial, temporal, and format constraints [22, 6]. The NTCIR tasks [15] formalized ad-hoc retrieval evaluation, yet focused primarily on matching natural language queries to metadata for human consumption. As consumers shift from humans to autonomous agents, *findability* becomes a dead end without machine-actionable *accessibility*. Our work extends the FAIR principles [34] into the agentic domain, evaluating the Google Dataset Search ecosystem [22, 4, 29] not merely as a discovery tool, but as a critical infrastructure layer that solves the operational bottlenecks of autonomous execution.

Automated Evaluation and LLM-as-a-Judge Agentic tasks are inherently open-ended. Thus, traditional exact-match metrics and rigid lexical comparisons fall short of capturing true task success [18]. Consequently, the field has increasingly adopted the "LLM-as-a-judge" paradigm [36], leveraging the zero-shot reasoning capabilities of strong foundational models to evaluate complex outputs against specialized rubrics. Frameworks such as MT-Bench [36] and G-Eval [19] have demonstrated that LLM evaluators achieve high correlation with human annotators, particularly for nuanced semantic tasks. In the context of data retrieval, human evaluation of FAIR compliance is prohibitively slow and subjective. By explicitly mapping our multi-dimensional LLM-as-a-judge pipeline to the FAIR principles, our methodology extends the automated evaluation paradigm beyond conversational fluency into the rigorous assessment of operational data utility.

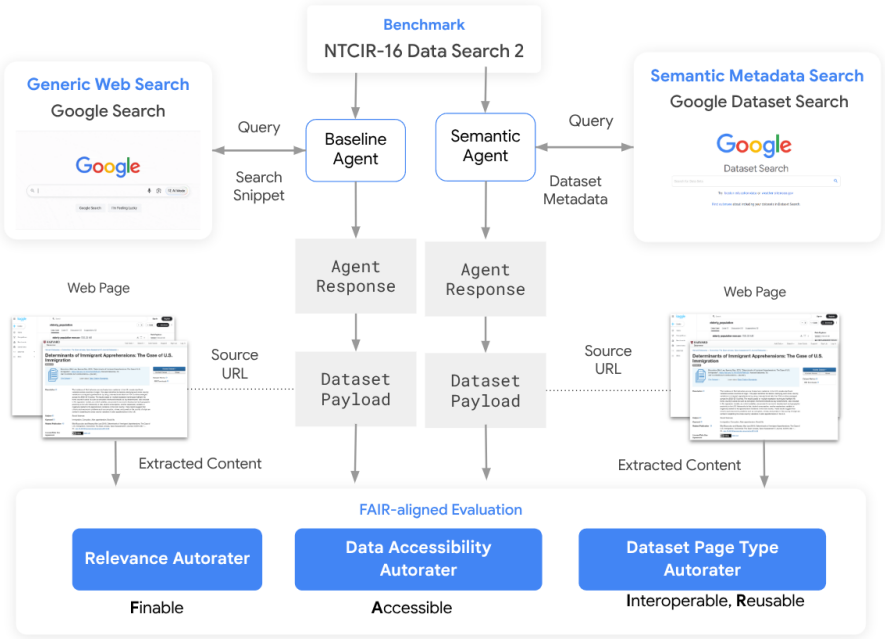
Structured Metadata and the Semantic Web in LLMs. While pairing LLMs with Knowledge Graphs (KGs) reduces hallucinations [24], existing evaluations often rely on closed, domain-specific KGs [14, 9] or static RAG frameworks where non-textual webpage metadata is injected primarily as a pre-filtered prompt feature [8]. These designs reduce the LLM to a static synthesizer of provided evidence rather than an autonomous agent tasked with active information discovery. Our work evaluates agent behavior across two architectures. In the Baseline Agent, the agent actively navigates noisy search results to uncover real

datasets. In the Semantic Agent, `schema.org` metadata acts as a high-precision corpus filter for datasets, streamlining the high-entropy task of autonomous discovery. Comparing these modes isolates how semantic standards mitigate contextual divergence during autonomous data retrieval.

3 System Architecture & Experimental Setup

To isolate the impact of semantic metadata on dataset discovery, we designed two nearly identical workflows. Rather than analyzing granular variables like index size, we treat `schema.org` metadata as a filter that fundamentally shapes corpus composition, search scale, and tool-execution logic.

Fig. 1. Comparative System Architecture. Similar agent logic is evaluated across unstructured Baseline Agent and Semantic Agent dataset search environments. Both feed a unified, FAIR-aligned evaluation of relevance, accessibility, and utility.



3.1 Agentic Framework and Setup

To ensure experimental parity, we contrast a Semantic Agent against a Baseline Agent using identical underlying architectures (Figure 1)—specifically, the

Agent Development Kit (ADK) [10] powered by Gemini 2.5 Pro. We isolate the target corpus and its respective search ecosystem as the primary independent variables. The agent prompts are highly symmetrical and tailored solely to optimize retrieval for their respective search environments, reflecting how users naturally interact with these platforms [1].

- **Baseline Agent (Unstructured Corpus):** The agent queries the Google Search index via a standard search tool. Representing the unstructured web ecosystem, this baseline corpus is a superset of tens of billions of web documents, where inclusion relies purely on general crawlability. To optimize data retrieval, the agent explicitly appends data-seeking keywords (e.g., "dataset") to the user's intent. This query expansion simulates real-world search behavior, preventing a trivially weak baseline by ensuring the agent successfully targets data rather than general web content. The retrieved payload consists of URLs and text snippets synthesized from page content.
- **Semantic Agent (Structured Corpus):** The agent queries the Google Dataset Search index via a metadata search tool. Representing the structured ecosystem, this curated corpus aggregates all pages with `schema.org/Dataset` markup. To maintain quality, we use a classification model to filter out roughly 80% of pages with invalid or misused annotations [4]. The remaining corpus contains 90 million dataset metadata records. Because the ecosystem natively restricts results to dataset entities, the query expansion is redundant. Therefore, the agent extracts the core keyword to query the index directly without the need for query expansion.

To eliminate confounding factors, both agents are governed by a set of controlled variables: (1) both utilize the same agentic framework (ADK) and underlying Large Language Model (Gemini 2.5 Pro); (2) prompts restrict agents to returning only datasets present in their tool-execution results, paired with a model temperature of 0 to maximize factual adherence and reproducibility; (3) agents are limited to returning a maximum of three highly relevant datasets per query to standardize evaluation volume; (4) findings must be output as a JSON array with dataset name, and source URL to facilitate automated evaluation; and (5) agents must output a deterministic fallback ("No relevant datasets found.") if no datasets are discovered.

3.2 Real-World Indices vs. Curated Corpora

We deliberately chose live, web-scale indices (Google Search and Google Dataset Search) to evaluate agentic retrieval under authentic conditions. While we adopt a query set from an established benchmark—specifically NTCIR-16—to ensure a reproducible evaluation of genuine user intent, we explicitly bypass the benchmark's underlying static document collections. Rather than operating in artificially sanitized environments—such as the NTCIR-16 data collections themselves, fixed web crawls (e.g., ClueWeb22 [23]), or centralized, well-formatted

[1] prompts: https://doi.org/10.6084/m9.figshare.32141311

Table 1. Agent Comparison

Component	Baseline Agent	Semantic Agent
Agent / Model	ADK Agent / Gemini 2.5 Pro	ADK Agent / Gemini 2.5 Pro
Target Index	Google Search Index	Google Dataset Search Index
Inclusion Filter	General web crawlability	Semantic metadata markup
Corpus Scale	Billions of web pages	∼90 Million dataset metadata
Retrieved Payload	URL and synthesized text from web pages	URL and semantic metadata

dataset repositories (e.g., Zenodo, Hugging Face)—our approach forces agents to navigate the messy, decentralized open web. Deploying these queries against billions of live documents proves the system's scalable, "in-use" viability. Furthermore, live indices capture newly published data and real-world `schema.org` adoption, avoiding the staleness of static corpora.

4 Evaluation Methodology

To compare the two agents, we use a public benchmark for dataset retrieval, develop evaluation metrics that explicitly address the FAIRness of the datasets, and design the autoraters to ensure scalable, low-variance evaluations.

4.1 Benchmark Dataset

We evaluate our systems on the NTCIR-16 Data Search 2 dataset [15], an established ad-hoc dataset retrieval benchmark spanning multiple domains like economics, demographics, and public health. Specifically, we use queries from its Information Retrieval (IR) subtask.

From this subtask, we isolated the English keyword-based questions (N=58) from the IR subtask to simulate the authentic data discovery intent. Because these queries were crowdsourced from human workers tasked with expressing real-world information needs derived from actual web pages, they simulate authentic data discovery intent (e.g., "air quality baltimore"). Unlike synthetic or verbose prompts, these human queries are frequently underspecified and context-dependent. Utilizing this query set forces the agents to navigate the genuine complexities of matching user intent to dataset metadata, providing a realistic assessment of how these systems will perform in production environments.

4.2 Evaluation Metrics

We evaluate the agents for autonomous workflows across three FAIR-inspired dimensions: discovering relevant web pages (Findable), ensuring programmatic access (Accessible), and verifying the computational utility based on page types (Interoperable/Reusable).

- **Relevance (Findable):** Evaluates semantic alignment between the NTCIR-16 query and dataset scope, assigning a score from -1 to 2 (summarized in Table 2). We deliberately score "Exploratory Matches" as Highly Relevant. In real-world data discovery, users frequently issue broad queries that lack explicit geographic, temporal, or demographic constraints. Returning a highly specific dataset in response to a broad, underspecified query proves the agent successfully navigated the correct semantic domain without violating any explicit user constraints.
- **Data Accessibility (Accessible):** Measures data accessibility via a multi-level rubric (Table 3) to assess complexity of data extraction and the presence of technical barriers.
- **Dataset Page Type (Interoperable & Reusable):** Measures computational utility by categorizing the dataset pages into a page type rubric (summarized in Table 4).

Table 2. LLM Autorater Rubric for Scoring Query-to-Dataset Relevance.

Score	Classification	Description
2	Highly Relevant	Contains requested data (exact match or superset), or provides concrete data for broad, unconstrained exploratory queries.
1	Partially Relevant	On-topic but incomplete or mismatched (e.g., wrong year/location, or a narrow subset of an explicitly constrained query).
0	Irrelevant	Entirely off-topic or unrelated to the query.
-1	Not Applicable	The page is unreachable. The relevance cannot be evaluated.

4.3 Autorater Execution

Autorater Evaluation Framework We use Gemini 2.5 Pro as an LLM-judge to evaluate our three metrics.[2] Rather than relying on agent-generated snippets, the autoraters evaluate live content extracted directly from the agent-provided URLs and serialized into Markdown. To ensure reproducibility and prevent evaluation drift from link rot, we freeze this extracted content at query time and provide it to the autoraters alongside the dataset payload from agent response.

Chain-of-Thought and Evidence Extraction. To prevent evaluation hallucinations, autoraters follow a strict Chain-of-Thought (CoT) [33] protocol. Before outputting a final rating, they must extract explicit evidence (e.g., quotes) from

[2] prompts: https://doi.org/10.6084/m9.figshare.32141311

Table 3. LLM Autorater Rubric for Scoring Data Accessibility Levels.

Level	Classification	Description
6	Machine-Readable	Contains direct links or export options to structured data files (e.g., .rdf, .ttl, .csv, .json) or API endpoints.
5	Structured	Data in structured formats (e.g., tables) is presented in a page requiring parsing but not NLP.
4	Unstructured	Data points (e.g., statistics, coordinates) are embedded in prose, requiring NLP for extraction.
3	Presentation-Bound	Data is confined to static images, non-machine-readable documents, or interactive visual interfaces
2	Non-Data	Page is reachable but lacks data.
1	Unreachable	Page is unreachable (e.g., 403/404 errors).

Table 4. LLM Autorater Classification Rubric for Retrieved Dataset Page Types.

Classification	Description
DATA_REGISTRY	Dataset landing pages with interoperable metadata records (e.g., DOIs, data dictionaries, provenance).
RAW_DATA	Machine-readable files (e.g., .rdf, .csv) or APIs.
DATA_EXPLORER	Interactive interfaces (e.g., dashboards, dynamic maps) or standalone charts without metadata.
DATA_NARRATIVE	Prose-heavy content using data to support storytelling (e.g., papers, news articles, reports).
DISCOVERY_PORTAL	Gateways or search catalogs for querying across multiple distinct datasets.
NO_DATA	General web content lacking a specific dataset focus (e.g., organizational homepages, marketing).
UNREACHABLE	Page is unreachable (e.g., 403/404 errors).

the serialized snapshot and map it to the rubric. This ensures evaluations are grounded in the scraped content rather than the model's parametric memory.

Relevance Prompt Design. Our relevance evaluation adapts the UMBRELA prompt [32], leveraging its few-shot examples and structured reasoning to distinguish between superficial mentions and primary data sources. To align with the NTCIR-16 Data Search benchmark, we converted UMBRELA's original 1–3 scale to a 0–2 ranking. This maps our outputs directly to established human-labeled ground truths while preserving the framework's zero-shot reasoning strengths.

Autorater Validation and Manual Review

Autorater Validation. To validate our autoraters, two authors independently annotated a diverse 11% sample of query-retrieval pairs to establish a "gold set." Prior to analysis, we excluded non-ordinal 'Unreachable' categories. We then mapped the autorater's classifications to progressive ordinal scales for relevance and accessibility and and derived the scale for computational utility directly from the dataset page type.We evaluated agreement using Linear-Weighted Cohen's Kappa (κ) to penalize minor ordinal mismatches less severely than major ones. Human annotators achieved high reliability, with Kappa scores of 0.70 (Page Type), 0.93 (Accessibility), and 0.95 (Relevance), resolving discrepancies via consensus. Comparing the autorater to this gold set yielded strong Kappa scores of 0.73, 0.78, and 0.74, respectively, demonstrating close alignment with human judgment.

Manual Evaluation for Edge Cases. Approximately 31% of retrievals were routed to a human evaluation pipeline (29% from Baseline Agent and 33% from Semantic Agent). As discussed in Section 7, our internal scraper sometimes encounters complex layouts or security policies, classifying these scraping errors as "Undetermined" and flagging them for review. To avoid selection bias from discarding these edge cases, two authors independently annotated this subset. We evaluated inter-rater reliability using Linear-Weighted Cohen's Kappa, achieving scores of 0.72 (Dataset Type), 0.88 (Accessibility), and 0.70 (Relevance). Following consensus resolution, these human-annotated labels were integrated into the final dataset, ensuring a complete and statistically sound evaluation of both actual data availability and accessibility.

5 Results

We now present the experimental outcomes of our comparative analysis across the 58 English keyword-based queries from the NTCIR-16 dataset. The Baseline Agent answered 56 queries, retrieving 164 datasets. The Semantic Agent answered 40 queries, retrieving 112 datasets.

5.1 Relevance

Both systems exhibited similar performance in retrieving datasets (Figure 2) relevant to the questions. For the "Highly Relevant" (Score 2), the Baseline Agent retrieved 99 datasets (60.4%), while the Semantic Agent retrieved 68 datasets (60.7%).

5.2 Data Accessibility

The Semantic Agent demonstrated a significant accessibility advantage, returning pages with machine-readable data in 71.4% of retrievals compared to 48.7% for the Baseline Agent. Accounting for the differences in total retrieval volumes, this represents a 46.6% relative increase in precision (Figure 3). By leveraging

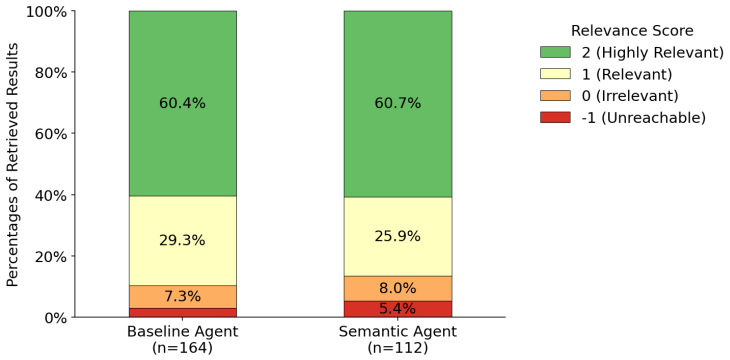


Fig. 2. Comparison of the agent results by relevance scores.

semantic metadata, the agent avoided non-computational roadblocks, achieving relative reductions of 46.6% in Narrative/Unstructured Data (data embedded in narrative prose), 62.9% in Presentation-Bounded Data (pages with only charts or interactive dashboards, without metadata), and 76.3% in Non-Data (false positive pages lacking data entirely).

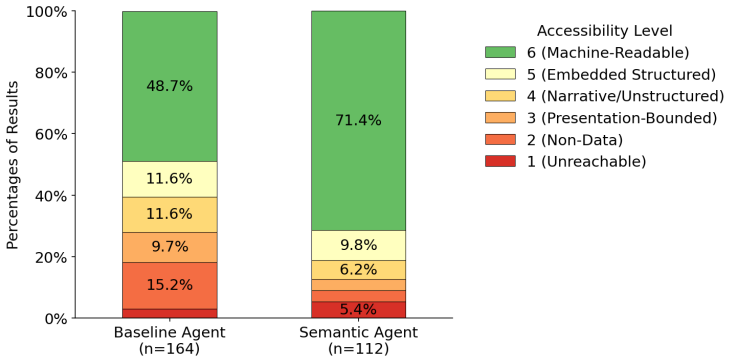


Fig. 3. Comparison of the agent results by data accessibility levels

5.3 Dataset Page Type

The distribution of retrieved dataset type pages varied between the two systems (Figure 4). The Semantic Agent's results consisted primarily of DATA_REGISTRY entries. These accounted for 88.4% of its total retrievals compared to 61.0% for the Baseline Agent, yielding a 44.9% relative increase in precision. While the Baseline Agent retrieved a wider variety of alternative page types, the Semantic

Agent streamlined this distribution by achieving relative reductions of 86.6% in DATA_NARRATIVE pages (prose-heavy pages, such as news articles or academic papers discussing data), 55.7% in DATA_EXPLORER pages (data exploration pages without metadata), and 100% in DISCOVERY_PORTAL pages (search engines or gateways for multiple datasets).

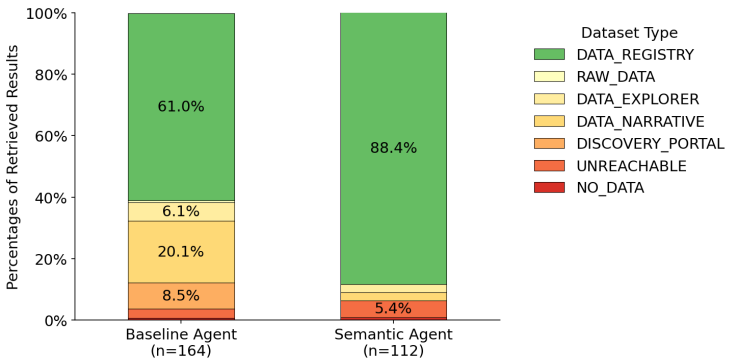


Fig. 4. Comparison of the agent results by dataset page type

5.4 Agentic FAIRness

To evaluate an agent's capacity to identify machine-actionable registry entries that ensure metadata compliance and direct data accessibility. We define a dataset as fully "FAIR-compliant" if it achieves a perfect composite score across three criteria: a relevance score of 2 (Highly Relevant), Dataset Accessibility at Level 6 (Machine-Readable), and Dataset Page Type of DATA_REGISTRY.

Dataset-Level Precision To assess the overall precision of the retrieved data, we measured the dataset-level success rate—the proportion of FAIR-compliant dataset URLs out of the total number of URLs retrieved by each system across all queries. The Baseline Agent achieved a 28.0% precision rate (46 of 164 URLs). In contrast, the Semantic Agent achieved a precision of 46.4% (52 of 112 URLs, $p < 0.01$). This represents a 65.7% relative improvement, demonstrating that nearly half of the Semantic Agent's retrievals met the most strict FAIR criteria.

Query-Level Result Density To evaluate the signal-to-noise ratio, we measured the concentration of FAIR-compliant datasets per answered query, capped at a maximum of 3 results per question. The Semantic Agent achieved a result density of 1.30 (52 FAIR-compliant dataset URLs / 40 answered queries), utilizing 43.3% of its top-3 capacity. The Baseline Agent managed a density of 0.82 (46 FAIR-compliant URLs / 56 answered queries), achieving a 27.4% utilization rate. Due to the query size, this difference lacks statistical significance ($p > 0.05$).

6 Discussion and Future Work

Based on our findings, we revisit our core hypotheses, discuss practical strategies for autonomous agents navigating gated and constrained data, and propose a hybrid path balancing high-precision actionability with broad exploratory reach.

6.1 Actionability, Utility, and Breadth

We now return to our three key hypotheses in Section 1: data actionability, last-mile utility, and exploratory breadth. The Semantic Agent achieves superior precision by retrieving machine-actionable data for autonomous execution. Conversely, while the Baseline Agent prioritizes exploratory breadth, it frequently fails at the "last mile" of data retrieval, where unstructured noise hinders autonomous action.

Data Actionability Our results validate the data-actionability hypothesis: the Semantic Agent demonstrated an advantage in navigating to these metadata-rich registries. More importantly, the Semantic Agent was more successful at identifying pages with machine-readable download links.

The "Last Mile" Utility The Baseline Agent often retrieves prose-heavy web pages and navigational portals. Navigating and scraping these unstructured sources makes it harder for agents to isolate the actionable data. Specifically, web DOM syntactic noise causes severe token bloat, while dense prose degrades the context window, triggering retrieval failures [17]. Additionally, portals trap agents in redundant search loops instead of delivering data payloads directly[1]. Our results support this "Last Mile" utility hypothesis: while open-web agents can discover the relevant pages, they fail to locate actionable data.

Exploratory Breadth The Baseline Agent answered more questions than the Semantic Agent due to the limited adoption of `schema.org` markup. This broader coverage is most pronounced in the web's "long tail"— research domains and repositories where dataset pages often lack `schema.org` annotations.

Precision and Agentic Actionability In contrast to recall, the Semantic Agent method showed a clear precision advantage over the Baseline Agent in retrieving FAIR-compliant datasets. For an autonomous agent, this advantage translates to a "fail-fast" mechanism that favors an empty state over a probabilistic guess, preventing downstream execution failures.

6.2 Autonomy Strategy for Gated Data

Real-world data discovery frequently encounters paywalls and authentication barriers. FAIR principles recognize these controls as standard research logistics

rather than metadata failures—mandating that data remain "as open as possible, as closed as necessary" [20]. Rather than treating these barriers as absolute roadblocks, an autonomy strategy can employ Human-in-the-Loop (HITL) handoffs [13]. An agent uses structured metadata to verify a dataset's utility, identifies the access blocker, and pauses to request human intervention for credentials or payment. Once the human clears the blocker, the agent resumes autonomously downloading and parsing the machine-readable payload.

6.3 Scaling to Multi-Faceted Queries

While our evaluation utilized keyword-based queries, it presents a compelling opportunity to explore complex, multi-faceted natural language queries. Evaluating the systems under metadata constraints exposes a contrast in their underlying mechanics. To satisfy specific constraints in an unstructured ecosystem, the Baseline Agent must rely on probabilistic retrieval to identify candidate URLs, followed by high-friction scraping and parsing to verify attributes. Conversely, because structured corpora natively index explicit semantic properties (e.g., `license`, `dataType`), the Semantic Agent translates multi-faceted constraints directly into a native tool call with deterministic filter and reduces post-hoc verification overhead. As illustrated below, a multi-faceted natural language query can be converted into a metadata-constrained tool call:

Multi-faceted Query:
"Find **tabular** datasets on **household consumption** data updated **within the last 3 months** with **non-commercial** licensing"

Metadata Tool call:

```
search_datasets(
    keywords="household consumption",
    last_updated="3 months",
    license="noncommercial",
    dataType="tabular")
```

6.4 Hybrid Path

The choice to use semantic metadata should be driven by the agent's primary objective: breadth of knowledge versus reliability of action. Discovery-oriented tasks (e.g., exploratory research) require unstructured systems to achieve broad open-web recall, though this approach introduces higher noise that necessitates greater computational overhead or human intervention. Conversely, autonomous workflows (e.g., real-time analytics, code generation) require a semantic-metadata approach to guarantee machine-actionability, as false positives are far costlier than empty states.

To bridge the gap between reliability and breadth, we propose a hybrid architecture. Under this model, agents first query the high-precision semantic-

metadata layer. If this initial search yields an empty state, the system falls back to the unstructured approach to cast a wider net.

7 Limitations

Our work has several limitations, specifically corpus coverage, ranking mechanisms, web scraping constraints, and queryset scale.

Coverage of the Structured Corpus. Our evaluation of the structured ecosystem is limited to the Google Dataset Search index (90 million records), which represents only a fraction of global data. Because inclusion strictly requires `schema.org/Dataset` or DCAT markup, our study excludes dataset pages using alternative markup vocabularies or lacking semantic annotation. Ultimately, this strict filtering ensures operational reliability but trades away absolute recall and exploratory breadth of the open web.

Proprietary Ranking Mechanisms. Google Search and Google Dataset Search share the core infrastructure, but their internal ranking mechanisms remain "black boxes." Rather than isolating this algorithmic delta, our evaluation measures end-to-end utility—assessing the empirical advantage of participating in the structured data ecosystem versus relying on the unstructured web.

Automated Scraping. Our internal scraper has limitations that prevent it from processing complex web layouts and penetrating enterprise-level security policies. As a result, the autoraters could not evaluate 31% of the pages. We subsequently directed this subset to a human evaluation pipeline (4.3) to mitigate bias, in order to prevent the scraping constraints from skewing evaluation results. Future work could mitigate this limitation by using advanced scraping tools.

Queryset Scale. Although our queryset yields statistically significant dataset-level metrics, its size may not fully capture the extreme diversity of the web's "long tail." Furthermore, the query limits generalizability to the multi-faceted queries discussed in 6.3. Future work should employ larger, cross-disciplinary datasets to comprehensively evaluate agentic retrieval.

8 Conclusion

Autonomous workflows shift the primary focus of system utility from discovery to computational actionability. While unstructured retrieval supports broad exploratory tasks, structured ecosystems remain the indispensable foundation for reliable, execution-oriented autonomous workflows. In the agentic paradigm, the FAIR principles have transitioned from human-centric best practices into technical requirements. The structured data ecosystem fuels the autonomous data-driven workflow, significantly increasing the probability that retrieved data assets are not just findable, but instantly machine-actionable.

9 Use of Generative AI

We used Gemini to generate initial drafts of some sections based on our discussion notes, as well as a copy-editing tool to improve grammar and overall readability throughout the manuscript. In addition, we used the model to critically review early drafts, incorporating its feedback to refine our arguments and restructure the narrative flow.

References

1. Aghzal, M., Stein, G.J., Yao, Z.: Why Do LLM-based Web Agents Fail? A Hierarchical Planning Perspective (2026)
2. Albertoni, R., Browning, D., Cox, S.J.D., Gonzalez Beltran, A., Perego, A., Winstanley, P.: Data catalog vocabulary (DCAT) - version 3. W3C recommendation, World Wide Web Consortium (W3C) (2024), https://www.w3.org/TR/vocab-dcat-3/
3. Albertoni, R., Browning, D., Cox, S.J.D., Gonzalez-Beltran, A.N., Perego, A., Winstanley, P.: The W3C data catalog vocabulary, version 2: Rationale, design principles, and uptake. Data Intelligence **6**(2), 457–487 (2024). https://doi.org/10.1162/dint_a_00241
4. Alrashed, T., Paparas, D., Benjelloun, O., Sheng, Y., Noy, N.: Dataset or Not? A Study on the Veracity of Semantic Markup for Dataset Pages. In: The Semantic Web – ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings. p. 338–356. Springer-Verlag, Berlin, Heidelberg (2021), https://doi.org/10.1007/978-3-030-88361-4_20
5. Benjelloun, O., Chen, S., Noy, N.: Google Dataset Search by the numbers. In: International Semantic Web Conference (ISWC-2020), In-Use Track (2020), https://arxiv.org/abs/2006.06894
6. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset Search: A Survey. The VLDB Journal **29**(1), 251–272 (Aug 2019), https://doi.org/10.1007/s00778-019-00564-x
7. Chezelles, T.L.S.D., Gasse, M., Drouin, A., Caccia, M., Boisvert, L., Thakkar, M., Marty, T., Assouel, R., Shayegan, S.O., Jang, L.K., Lù, X.H., Yoran, O., Kong, D., Xu, F.F., Reddy, S., Cappart, Q., Neubig, G., Salakhutdinov, R., Chapados, N., Lacoste, A.: The BrowserGym Ecosystem for Web Agent Research (2025), https://arxiv.org/abs/2412.05467
8. Chiang, C.H., Lee, H.y.: Do Metadata and Appearance of the Retrieved Webpages Affect LLM's Reasoning in Retrieval-Augmented Generation? In: Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A., Chen, H. (eds.) Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. pp. 389–406. Association for Computational Linguistics, Miami, Florida, US (Nov 2024). https://doi.org/10.18653/v1/2024.blackboxnlp-1.24, https://aclanthology.org/2024.blackboxnlp-1.24/
9. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R.O., Larson, J.: From Local to Global: A Graph RAG Approach to Query-Focused Summarization (2025), https://arxiv.org/abs/2404.16130
10. Google: Agent Development Kit (ADK) (2026), https://adk.dev/, accessed: 2026-04-22

11. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: evolution of structured data on the web. Commun. ACM **59**(2), 44–51 (Jan 2016), https://doi.org/10.1145/2844544
12. Gur, I., Nachum, O., Miao, Y., Safdari, M., Huang, A., Chowdhery, A., Narang, S., Fiedel, N., Faust, A.: Understanding HTML with Large Language Models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 2803–2821. Association for Computational Linguistics, Singapore (Dec 2023), https://aclanthology.org/2023.findings-emnlp.185/
13. Humphreys, P.C., Raposo, D., Pohlen, T., Thornton, G., Chhaparia, R., Muldal, A., Abramson, J., Georgiev, P., Santoro, A., Lillicrap, T.: A data-driven approach for learning to control computers. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 9466–9482. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/humphreys22a.html
14. Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, X., Wen, J.R.: StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 9237–9251. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.574, https://aclanthology.org/2023.emnlp-main.574/
15. Kato, M.P., Ohshima, H., Liu, Y.H., Chen, H.L.: NTCIR-16 Data Search 2 (2022)
16. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NeurIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
17. Li, X., Lyu, T., Yang, Y., Shan, L., Yang, S., Zhang, L., Huang, Z., Liu, Q., Li, Y.: Escaping the Context Bottleneck: Active Context Curation for LLM Agents via Reinforcement Learning (2026)
18. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J.: AgentBench: Evaluating LLMs as Agents. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. ICLR '24, vol. 2024, pp. 52989–53046. https://openreview.net/forum?id=zAdUB0aCTQ
19. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: NLG Evaluation using GPT-4 with Better Human Alignment. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522. Association for Computational Linguistics, Singapore (Dec 2023), https://aclanthology.org/2023.emnlp-main.153/
20. Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., Wilkinson, M.D.: Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services and Use **37**(1), 49–56 (2017)
21. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., But-

ton, K., Knight, M., Chess, B., Schulman, J.: WebGPT: Browser-assisted question-answering with human feedback (2022), https://arxiv.org/abs/2112.09332

22. Noy, N., Burgess, M., Brickley, D.: Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In: The World Wide Web Conference. p. 1365–1375. WWW '19, Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3308558.3313685
23. Overwijk, A., Xiong, C., Callan, J.: ClueWeb22: 10 Billion Web Documents with Rich Information. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3360–3362. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3536321, https://doi.org/10.1145/3477495.3536321
24. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering **36**(7), 3580–3599 (2024). https://doi.org/10.1109/TKDE.2024.3352100
25. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. UIST '23, Association for Computing Machinery, New York, NY, USA (2023), https://doi.org/10.1145/3586183.3606763
26. Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E.: Gorilla: Large Language Model Connected with Massive APIs. Advances in Neural Information Processing Systems **37**, 126544–126565 (2024)
27. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al.: ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In: The Twelfth International Conference on Learning Representations. ICLR '24, https://openreview.net/forum?id=dHng2O0Jjr
28. Schick, T., Dwivedi-Yu, J., Dessí, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language Models Can Teach Themselves to Use Tools. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NeurIPS '23, Curran Associates Inc., Red Hook, NY, USA
29. Sostek, K., Russell, D.M., Goyal, N., Alrashed, T., Dugall, S., Noy, N.: Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search. Harvard Data Science Review (Special Issue 4) (apr 2 2024), https://hdsr.mitpress.mit.edu/pub/psnc8zsr
30. Tan, J., Dou, Z., Wang, W., Wang, M., Chen, W., Wen, J.R.: HtmlRAG: HTML is Better Than Plain Text for Modeling Retrieved Knowledge in RAG Systems. In: Proceedings of the ACM on Web Conference 2025. p. 1733–1746. WWW '25, Association for Computing Machinery, New York, NY, USA (2025), https://doi.org/10.1145/3696410.3714546
31. Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., Sepp, T.: Data sharing practices and data availability upon request differ across scientific disciplines. Scientific Data **8**(1), 192 (2021), https://doi.org/10.1038/s41597-021-00981-0
32. Upadhyay, S., Pradeep, R., Thakur, N., Craswell, N., Lin, J.: UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. arXiv:2406.06519 (2024)

33. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 24824–24837 (2022)
34. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data **3**(1), 1–9 (2016), https://doi.org/10.1038/sdata.2016.18
35. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: ReAct: Synergizing Reasoning and Acting in Language Models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=WE_vluYUL-X
36. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NeurIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
37. Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., Neubig, G.: WebArena: A Realistic Web Environment for Building Autonomous Agents. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. ICLR '24, vol. 2024, pp. 15585–15606
38. Zuiderwijk, A., Shinde, R., Jeng, W.: What Drives and Inhibits Researchers to Share and Use Open Research Data? A Systematic Literature Review to Analyze Factors Influencing Open Research Data Adoption. PLoS one **15**(9), e0239283 (2020), https://doi.org/10.1371/journal.pone.0239283