



GUIXU: Valuation-Driven Data Discovery for Autonomous AI Agents with On-Chain Attestation

Yifan Wu*
Zhejiang University
yifan.wu@zju.edu.cn

Yuchen Peng*
Zhejiang University
zjupengyc@zju.edu.cn

Jiaqi Chai*
Zhejiang University
chaijiaqi@zju.edu.cn

Yufei Qian*
Zhejiang University
qianyufei@zju.edu.cn

Xilin Li
Zhejiang University
lixilin@zju.edu.cn

Ke Chen
Zhejiang University
chenk@zju.edu.cn

Lidan Shou†
Zhejiang University
should@zju.edu.cn

ABSTRACT

Autonomous agents increasingly rely on external data to complete downstream tasks such as model training and decision support. However, existing data discovery systems remain largely retrieval-oriented: they surface candidate datasets from heterogeneous sources, but provide limited support for estimating task-specific utility, selecting cost-effective datasets under budget constraints, or incorporating trustworthy feedback from prior usage. This paper presents GUIXU, a valuation-driven data discovery system for autonomous agents. GUIXU employs a three-phase valuation pipeline with proxy-label propagation and multi-round knapsack optimization for task-aware data valuation. GUIXU integrates agentic payment protocol to enable budget-constrained data procurement workflows. GUIXU leverages on-chain data market and attestation signals for verifiable data discovery. Our demonstration highlights how GUIXU enables an agent to move beyond keyword-based dataset retrieval toward task- and budget-aware, trustworthy data discovery and procurement. Attendees can interactively explore the full workflow, from NL task specification and multi-source search to data valuation and verifiable transaction feedback.

PVLDB Reference Format:

Yifan Wu, Yuchen Peng, Jiaqi Chai, Yufei Qian, Xilin Li, Ke Chen, Lidan Shou. Guixu: Valuation-Driven Data Discovery for Autonomous AI Agents with On-Chain Attestation. PVLDB, 19(12): 4562 - 4565, 2026.
doi:10.14778/3827998.3828066

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/guixu-project/guixu>.

1 INTRODUCTION

Data discovery, defined as the process of identifying datasets that satisfy the information needs of data-driven applications [3], has become a foundational capability in modern AI and data systems.

*These authors contributed equally to this work.

†Lidan Shou is the corresponding author.

‡Also affiliated with: The State Key Laboratory of Blockchain and Data Security; Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security. This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3828066

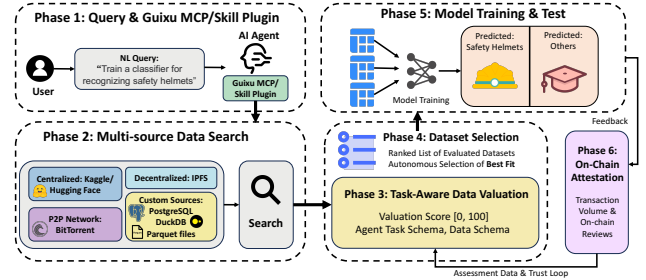


Figure 1: Overview of the GUIXU workflow. A user's natural-language query is processed through six phases.

Its importance is amplified in agentic settings, where autonomous agents such as OpenClaw and OpenAI Codex must independently locate, evaluate, and procure external data to complete complex tasks, including model training, prediction, and decision support. In such workflows, data discovery is not merely a retrieval problem: agents must determine which datasets are genuinely useful for a given objective, which combination is worth acquiring under a budget constraint, and which sources are trustworthy enough to use without human oversight.

State of the Art and Limitations. Existing dataset search platforms, including Google Dataset Search [1], Kaggle, and Hugging Face, have exposed agent-accessible interfaces via Model Context Protocol (MCP) or skill plugins, enabling keyword-based search across heterogeneous sources. However, these systems operate in a fundamentally *retrieval-first* manner: they surface candidate datasets but provide no support for task-aware data valuation, budget-constrained selection, or trust-aware procurement. Autonomous agents, therefore, still lack an integrated mechanism for translating a high-level task request into a cost-effective data acquisition decision.

To make this gap concrete, consider the following running example used throughout this paper: “Train an image classifier that determines whether workers in construction site images are wearing safety helmets correctly, with a total data procurement budget of \$2.00.” A practical system must do far more than retrieve datasets containing the keyword *helmet*: it must quantify each dataset’s utility relative to the task objective, account for acquisition cost, and incorporate trust signals when deciding whether a source is worth purchasing. Existing tools leave these decisions to manual inspection, which is fundamentally misaligned with autonomous agent execution.

Specifically, we identify three limitations of existing data discovery systems. **(1) Insufficient Task-Aware Data Valuation:** Existing systems lack a standardized schema of agent tasks and their associated data requirements, and thus cannot reason about which datasets are genuinely valuable for a given task beyond superficial keyword relevance. This causes agents to surface functionally incompatible datasets, such as massive image sets that trigger processing timeouts due to task-irrelevant samples, or datasets whose volume fails to meet the minimum execution thresholds of agent-generated code. **(2) Incompatibility with Agentic Payment:** Existing platforms do not support decentralized agentic payment protocols such as x402 [2] and machine payment protocol (MPP) [5]. Datasets relevant to niche scenarios are rarely free, yet how to optimally procure a portfolio of datasets under a fixed budget remains largely unaddressed. **(3) Absence of On-Chain Attestation Signals:** Blockchain ledgers expose valuable trust signals (e.g., transaction volume, on-chain reviews, and seller reputation) that can substantially improve data valuation, yet existing discovery systems rarely incorporate them.

Our Approach. This paper presents GUIXU, a valuation-driven data discovery system for autonomous agents. GUIXU reframes dataset discovery as a *task-aware valuation and procurement* problem rather than a pure retrieval problem. As illustrated in Figure 1, a user’s natural-language query flows through six phases: NL query parsing, multi-source data search, task-aware valuation, autonomous dataset selection, model training and testing, and on-chain attestation that feeds trust signals back into future valuations. This end-to-end workflow is realized by three tightly integrated layers.

First, the *frontend and protocol layer* provides the GUIXU Hub marketplace and an MCP server interface through which AI agents access all discovery capabilities, together with a multi-protocol agentic payment framework that dynamically routes transactions via the x402 agentic payment protocol or an on-chain escrow contract depending on the transaction context. Second, the *valuation-driven discovery layer* parses an NL query into a structured task schema and data schema, searches heterogeneous data sources, and scores each candidate dataset through a multi-signal coarse ranking, a proxy-label propagation phase, and a multi-round knapsack optimization, thereby enabling agents to autonomously select the most valuable datasets for a given task. Third, the *on-chain data market layer* handles trustworthy data transactions via escrow-based smart contracts and a decentralized key management network, while its on-chain attestation module records buyer reviews and seller reputation and feeds these trust signals back into the valuation model to continuously improve discovery quality. In our running example, GUIXU enables an agent to autonomously discover, evaluate, purchase, and assemble helmet-detection training data, completing the loop from natural-language intent to verified model training. We view GUIXU as an initial system step toward more autonomous and trustworthy data discovery and procurement workflows.

Audience Interaction. GUIXU is designed as a live, interactive system demonstration. Users can submit natural-language tasks, inspect candidate datasets collected from multiple sources, observe how valuation signals affect ranking and dataset portfolio selection, and complete a verifiable on-blockchain procurement workflow. The system therefore demonstrates not only how autonomous

agents can search for data, but also how they can reason about data utility, cost, and trust in a principled and operational manner.

2 OVERVIEW OF GUIXU

GUIXU comprises three layers: a *frontend and protocol layer* (Section 2.1) exposing the Hub marketplace, MCP interface, and agentic payment framework; a *valuation-driven discovery layer* (Section 2.2) transforming natural-language queries into optimal dataset portfolios; and an *on-chain data market layer* (Section 2.3) handling trustworthy transactions via escrow contracts, decentralized key management, and on-chain attestation.

2.1 GUIXU Frontend and Protocols

Next, we delineate the key components of the GUIXU ecosystem: the GUIXU Hub frontend module, the agent-native MCP plugin, and a specialized protocol tailored for autonomous agentic payment.

GUIXU Hub Platform. GUIXU Hub serves as the marketplace frontend, providing cataloging, data discovery, and transaction management for data assets. Sellers publish datasets with structured metadata, including schema definitions, semantic tags, pricing, and access policies. Consumers interact via full-text search, on-chain transaction synchronization, and on-chain review aggregation.

Model Context Protocol. GUIXU encapsulates all data acquisition capabilities as a Model Context Protocol (MCP) server, exposing them to AI agents via JSON-RPC over stdio and HTTP transports. The server advertises a typed tool catalog spanning intent parsing, dataset search, quality evaluation, purchase, and feedback, enabling an agent to orchestrate the complete data procurement pipeline within a single MCP session.

Agentic Payment Protocol. GUIXU supports a multi-protocol agentic payment framework (e.g., x402 and MPP) for autonomous agent transactions. A payment router inspects the transaction context to dynamically select the settlement mechanism: micropayments default to the x402 protocol with atomic on-chain settlement, whereas high-value transactions are routed to an escrow contract that supports fund locking.

2.2 Valuation-Driven Data Discovery

Once the agent has connected to GUIXU via the MCP interface and configured its payment credentials through the protocol layer above, the system proceeds to the core discovery pipeline.

NL Query Parser. A natural-language intent parser transforms the user’s free-form query into a structured profile guiding all downstream stages. It issues a zero-temperature, JSON-constrained LLM call to extract a task schema (e.g., task type, target entity, keywords) and a data schema (e.g., budget ceiling, dataset size bounds, column names). The profile is further enriched with local context, such as hardware profiles and agent memory, ensuring it reflects both explicit intent and implicit computational constraints. **Task-aware Data Valuation.** Given the parsed query profile and the candidate datasets returned by the GUIXU multi-source search module, the valuation module estimates the task-specific utility of each dataset through a three-phase pipeline, as shown in Figure 2.

Phase 1: Multi-Signal Coarse Ranking. Each candidate dataset is scored by a weighted combination of five task-aware signals: (i) *relevance*, cosine similarity between the task description embedding

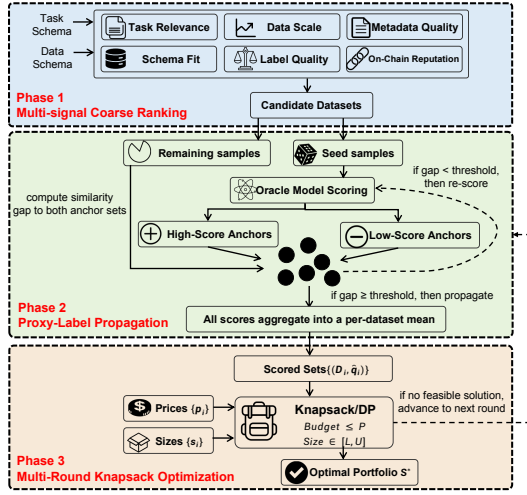


Figure 2: Three-phase pipeline of task-aware data valuation.

and dataset representation; (ii) *schema fit*, the degree to which the dataset schema satisfies task-required columns; (iii) *data scale*, a saturating logarithmic function of sample count and byte size; (iv) *label quality*, a composite indicator of class balance, missing-value rate, and target-column annotations; and (v) *metadata completeness*, measuring the availability of descriptions, summaries, and provenance metadata. (vi) *on-chain reputation*, an additive adjustment blending trade activity frequency with aggregated review sentiment from the attestation layer (Section 2.3). The top- k datasets are selected to proceed as a batch. GUIXU computes the heuristic weight of each signal using the Shapley value algorithm, based on analysis of historical dataset performance.

Phase 2: Proxy-Label Propagation. To estimate per-sample utility without invoking an expensive LLM oracle on every sample, the system employs a proxy-label propagation strategy. For each candidate dataset, a small seed subset of samples is randomly selected and scored by a strong generative model (the *LLM oracle*), producing per-sample utility judgments. The scored seeds are then partitioned into *high-score anchors* and *low-score anchors* based on configurable thresholds. For each remaining unscored sample, the system computes its average textual similarity to both anchor sets. If the similarity gap Δ between the two anchor affinities exceeds a user-configurable threshold θ , the sample inherits the mean score of its nearest anchor set. Larger θ means higher accuracy but more oracle cost, and vice versa. Samples for which $\Delta < \theta$ are deemed *uncertain* and are sent back to the LLM oracle for re-scoring. The final per-dataset score $\hat{q}_i$ is the mean of all sample-level scores (seed, propagated, and re-judged). This design ensures that the majority of samples are scored via lightweight similarity propagation, while LLM invocations are reserved for the small fraction of ambiguous cases, yielding substantial cost savings without sacrificing scoring fidelity.

Phase 3: Multi-Round Knapsack Optimization. Given the scored candidates $\{(D_i, \hat{q}_i)\}$ together with their prices $\{p_i\}$ and sizes $\{s_i\}$, the system selects an optimal dataset portfolio S^* that maximizes aggregate utility subject to a budget constraint $\sum_{i \in S} p_i \leq P$ and

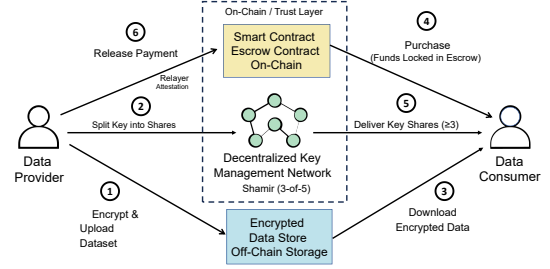


Figure 3: End-to-end transaction lifecycle of the on-chain data market.

a size constraint $L \leq \sum_{i \in S} s_i \leq U$. This is formulated as a two-dimensional knapsack problem and solved via dynamic programming over discretized price units, with a size-indexed frontier at each budget level to prune dominated states. If no feasible solution exists within the current candidate batch, the system expands the pool by advancing to the next batch of coarse-ranked datasets, triggering a new round of sample-efficient scoring before re-running the optimization; hence the term *multi-round*. The process terminates when a feasible portfolio is found or all batches are exhausted.

2.3 On-Chain Data Market

Overview. As illustrated in Figure 3, the end-to-end transaction lifecycle proceeds in six stages: ① the provider encrypts the dataset and uploads the encrypted datasets to off-chain storage; ② the decryption key is split into five shares via Shamir’s Secret Sharing and distributed to independent Keyguard nodes in the GUIXU Decentralized Key Management Network; ③ the data consumer locates the encrypted file by its content identifier and downloads it from off-chain storage; ④ the consumer initiates an on-chain purchase, locking funds in the listing’s escrow contract; ⑤ a designated relay-layer verifies successful key-share delivery by collecting delivery receipts and submits an attestation transaction that triggers escrow release; ⑥ upon payment confirmation, at least three Keyguard nodes deliver their shares to the consumer, who reconstructs the decryption key locally.

On-Chain Attestation. Each dataset is governed by a smart contract deployed on the *Base* chain that encodes the listing price, payment token, and cryptographic commitments. Upon purchase, funds are locked in escrow and released only upon confirmed delivery, with a refund option after a protocol-defined timeout. GUIXU derives trust signals directly from on-chain evidence: a chain indexing client extracts consumer reviews from purchase transaction metadata and aggregates them into per-listing summaries that feed into the valuation module’s on-chain signal dimension. In parallel, a seller reputation module computes a profile based on sales volume, consumer count, and activity span, mapped to a tiered trust classification. Each dataset is further accompanied by a W3C Verifiable Credential attesting to its content-addressed identifier, providing a cryptographically verifiable provenance chain.

Decentralized Key Management Network (DKMN). GUIXU Hub is the data market to apply decentralized key management techniques to dataset key custody, instantiating a 3-of-5 Shamir Secret Sharing scheme across independent Keyguard nodes via the Lit Protocol [4] to guarantee that no single party ever holds the complete decryption key.

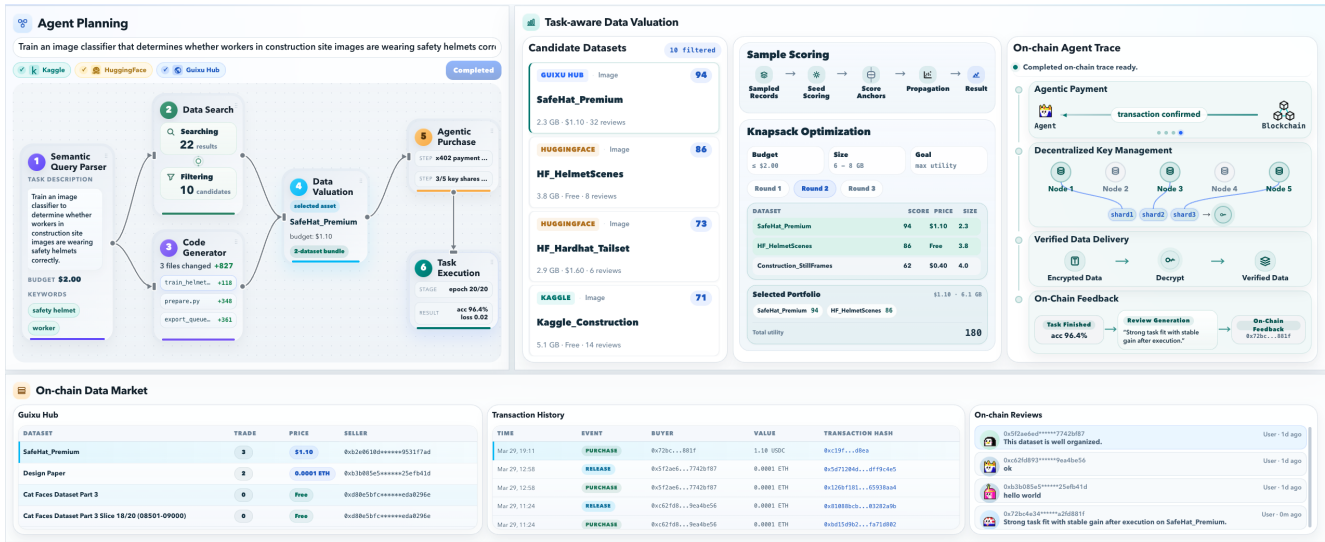


Figure 4: The Graphical User Interface of GUIXU.

3 DEMONSTRATION

Demonstration Setup and Interaction. GUIXU showcases an agent-native interface for valuation-driven data discovery and data acquisition for autonomous AI agents. The interface is connected to the GUIXU MCP server, the GUIXU Hub backend, heterogeneous data sources, and smart contracts. Attendees will be able to observe how GUIXU turns a natural-language task into agentic planning, task-aware data valuation, and on-chain interaction under budget constraints. As shown in Figure 4, we provide two demonstration scenarios by varying the agent’s task and budget inputs.

Scenario 1: Data Discovery with No Budget. *Our first demonstration evaluates how agents rely only on free datasets to build classification models.* The Agent first issues a task to check whether the cat is captured by the monitor with no budget. In response, the *Agent Planning* panel shows how GUIXU parses the request into a structured task profile and automatically restricts discovery to free sources such as Kaggle and Hugging Face. The *Task-aware Data Valuation* panel then visualizes the returned candidate datasets, the sample-scoring process, and the knapsack-based selector under the zero-budget constraint, allowing attendees to inspect how data is chosen to maximize task utility. Throughout this scenario, attendees can click any operator nodes and dataset panels to inspect intermediate outputs and observe that the valuation-driven selection yields a measurable accuracy gain even at zero budget.

Scenario 2: Data Acquisition under Budget Constraints. *Our second demonstration evaluates how agents autonomously select paid datasets when higher task utility justifies the cost.* The agent starts the safety-helmet classification task, but with a budget of \$2.00 and access to the GUIXU Hub. The *Agent Planning* and *Task-aware Data Valuation* panels indicate that paid datasets are involved, and the optimization process also incorporates budget constraints to maximize utility. In the *On-chain Agent Trace* panel, attendees can follow the full transaction life cycle: agentic payment, decentralized key management, and verified data delivery. In parallel, the

On-Chain Data Market panel displays the dataset listing, its full transaction history, and accumulated on-chain reviews.

Key Benefits for the Audience. GUIXU offers attendees three distinctive perspectives on agent-native data discovery:

- **Agent-Native Data Discovery.** Unlike conventional dataset search that matches keywords against metadata, GUIXU lets attendees experience a discovery paradigm driven by an autonomous agent that reasons about task requirements and market feedback. This gives the audience deep insights into how task-defined utility reshapes dataset selection.
- **Transparent and Interpretable Valuation.** GUIXU visualizes each stage of its valuation pipeline, from coarse ranking signals and sample-level scoring to knapsack optimization. This allows attendees to inspect not only *which* datasets are selected, but also *why* and how the selection is made.
- **Interactive On-Chain Attestation.** GUIXU further enables attendees to inspect how on-chain attestation transforms each transaction into a verifiable, traceable event. By examining transaction history and on-chain reviews, the audience can assess the market’s trustworthiness and how these attestation signals are incorporated into future valuation decisions.

ACKNOWLEDGMENTS

This work was supported by the Pioneer R&D Program of Zhejiang (No. 2024C01021) and the Leading Talent of Technological Innovation Program (No. 2023R5214) of Zhejiang Province.

REFERENCES

- [1] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference* (San Francisco, CA, USA) (WWW ’19). Association for Computing Machinery, New York, NY, USA, 1365–1375.
- [2] Coinbase. 2026. x402. <https://www.x402.org/>. Accessed: 2026-03-29.
- [3] Norman W. Paton, Jiaoyan Chen, and Zhenyu Wu. 2023. Dataset Discovery and Exploration: A Survey. *ACM Comput. Surv.* 56, 4, Article 102 (Nov. 2023), 37 pages.
- [4] Lit Protocol. 2026. Lit Protocol. <https://litprotocol.com>. Accessed: 2026-03-29.
- [5] Stripe. 2026. Machine Payment Protocol. <https://mpp.dev/>. Accessed: 2026-03-29.