Information Processing and Management 63 (2026) 104855

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Information-aware valuation and dynamic pricing for textual data in mobile environments

Wenze Xiong [a], Yetong Wang [b], Wanxin Li [b,*], Hao Guo [c,*], Jie Zhang [b,*], Haoyu Wang [a]

[a] Business School, University of Auckland, 12 Grafton Road, Auckland 1010, New Zealand
[b] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou 215123, Jiangsu, China
[c] School of Software, Northwestern Polytechnical University, Taicang Campus, Suzhou 215400, Jiangsu, China

ARTICLE INFO

Keywords:
Mobile textual data
Data valuation
Dynamic pricing

ABSTRACT

Data has become a fundamental asset in the information age. However, the absence of generalizable valuation and pricing frameworks poses significant challenges for data trading. This study proposes an information-aware valuation method for textual data based on weighted perplexity entropy (WPE), along with a dynamic pricing mechanism grounded in reinforcement learning (RL). For data valuation, we first evaluate the proposed WPE method across three heterogeneous textual datasets, followed by robustness assessments under noise injection scenarios. For data pricing, we formulate it as a Markov Decision Process (MDP) with the objective of profit maximization, incorporating an early submission incentive to encourage timely data contributions. Two different pricing strategies are developed, considering both immediate and long-term profits. Experimental results demonstrate that our WPE valuation method and RL-based pricing strategies consistently outperform existing valuation and pricing baselines. Specifically, our pricing strategies can increase an additional 16%, 80%, and 230% profit compared to three selected baselines. These findings further underscore the theoretical relevance of linking information theory with data economics, and the practical value of enabling more efficient and informed data acquisition.

1. Introduction

Data refers to digitalized information, encompassing a wide range of forms, from poetry and social media posts to Non-fungible Tokens and patents (Veldkamp, 2023). In the era of big data, the potential value of data is continuously being tapped (Lu et al., 2024). Comparable value creation has been reported in retail, finance, and public administration (Jiang et al., 2024; Kromidha, 2023; Zhou et al., 2025). With the rapid increase in mobile devices (Douch et al., 2022), data generation has become increasingly ubiquitous across diverse platforms and environments, expanding the scope and complexity of data valuation and pricing challenges.

Despite the demonstrated economic potential of data, determining how much a dataset is worth (valuation) and how it should be traded in the market (pricing) remain two interrelated but distinct challenges. Valuation aims to quantify the intrinsic worth of data based on its informational content, quality, and potential utility. Pricing, on the other hand, converts this valuation into a transaction price by incorporating market demand, competition, and strategic considerations. A robust pricing mechanism cannot be constructed

* Corresponding authors.
E-mail addresses: wenze.xiong@auckland.ac.nz (W. Xiong), yetong.wang19@alumni.xjtlu.edu.cn (Y. Wang), wanxin.li@xjtlu.edu.cn (W. Li), haoguo@nwpu.edu.cn (H. Guo), jie.zhang01@xjtlu.edu.cn (J. Zhang), haoyu.wang@auckland.ac.nz (H. Wang).

https://doi.org/10.1016/j.ipm.2026.104855
Received 14 November 2025; Received in revised form 9 April 2026; Accepted 21 April 2026
Available online 30 April 2026

without a sound and justifiable valuation method. This paper addresses both challenges by proposing an information-aware valuation method for textual data, which is then integrated into a dynamic pricing framework.

The urgency of developing systematic data valuation and pricing mechanisms has become particularly evident as data generation is now an integral part of daily life. Every online interaction, from search histories to user-generated content, produces valuable data that is often monetized without compensating the individuals who create it (Fleckenstein et al., 2023). This mismatch between value creation and value distribution highlights the need for rigorous valuation approaches to determine the economic value of data, as well as pricing mechanisms that facilitate profitable and efficient transactions. This challenge is particularly pronounced in mobile environments, where distributed devices generate data (Ghafari & Mansouri, 2025) while resource constraints create additional limitations for traditional valuation approaches. While we use the term "mobile environment" to describe mobile-enabled data marketplaces in which data are submitted through mobile devices, the current dataset in our research does not explicitly encode mobility traces (e.g., location trajectories or network handoffs). Therefore, we position "mobile" primarily as a deployment context rather than a modeling assumption. The methodological contributions of this work, WPE-based valuation and reinforcement-learning pricing, are general and applicable beyond mobile-only scenarios.

Within this broader context, textual data occupies a distinctive position. As a fundamental form of digital information, textual data underpins advances in Natural Language Processing (NLP), sentiment analysis, and numerous intelligent applications (Liu et al., 2026). However, as outlined by Malieckal et al. (2024), the data exhibits intangibility, complexity, heterogeneity, and the knowledge obtained from a specific issue can determine the value of data. These properties challenge conventional valuation methods, which aim to estimate the intrinsic worth of data before it is traded. Existing algorithmic approaches (Cong et al., 2022; Miao et al., 2020; Sim et al., 2022) and quality-based models (Xing & Wang, 2024; Yang et al., 2019) often over-rely on specific dataset characteristics and lack generalizability across domains. Moreover, many of these valuation methods focus on structural or statistical attributes without fully capturing the semantic and contextual richness of textual data, leading to incomplete or biased value estimates.

Beyond valuation, pricing mechanisms determine the actual transaction price by incorporating market demand, competition, and strategic considerations. Economic approaches such as auctions and Stackelberg games (Agarwal et al., 2024; Duan et al., 2023; Xiao et al., 2020) provide flexible market structures but often overlook long-term profitability of the data pricing. More recent computational models, such as multi-armed bandits (MAB) (Xu et al., 2024), frame pricing as the discrete selection of prices from fixed options based on estimated value and user acceptance. While effective in short-term settings, these models are limited by their discrete price space, myopic optimization, and poor adaptability to dynamic conditions. This motivates the adoption of reinforcement learning (RL), which models pricing as a sequential decision process with a continuous action space, enabling adaptive and long-term strategies in realistic data trading environments.

1.1. Research objectives

To address the dual challenges of data valuation and pricing, this paper proposes an information-aware framework. We aim to overcome the limitations of existing work by first creating a reliable method to value textual data based on its information content. We then use this valuation to build a dynamic pricing mechanism for real-world data marketplaces.

This research is guided by two primary objectives:

(1) To develop an efficient and unsupervised valuation method for textual data. The goal is to create a metric that captures the data's intrinsic informational worth without needing expensive labels or computationally heavy retraining.

(2) To design an adaptive and profitable dynamic pricing system. The goal is to build a mechanism that learns to make smart pricing decisions to maximize profit, while respecting budget limits and handling uncertain user behavior.

1.2. Contributions

The key contributions of our work are summarized as follows:

- We introduce an **information-aware** valuation framework for textual data based on weighted perplexity entropy (**WPE**). By combining linguistic features (POS, dependency roles) with statistical measures (TF-IDF), WPE enables fine-grained assessment of sample informativeness. We validate its effectiveness across multiple datasets and under noise perturbations.
- We design a **dynamic pricing** mechanism that integrates WPE into a reinforcement learning framework. By modeling the trading process as a Markov Decision Process (MDP), our method adapts prices under budget and acceptance uncertainty using two Proximal Policy Optimization strategies: **PPO-A** (immediate profit) and **PPO-B** (long-term profit).
- We benchmark our valuation and pricing methods against existing approaches. Results show that WPE yields more consistent and robust rankings, while our dynamic pricing strategies deliver higher cumulative profits under realistic trading conditions.

2. Related work

Data trading involves two closely related but distinct tasks: **data valuation**, which estimates the intrinsic worth of data based on its characteristics and potential utility, and **data pricing**, which determines the transaction price in a market setting by incorporating valuation results with market demand, competition, and strategic considerations. Existing work in both areas has drawn from both economic theory and computational techniques (Hao et al., 2023).

2.1. Data valuation methods

Valuation approaches aim to quantify the intrinsic value of data before it enters the market. Economic-oriented approaches adapt classical appraisal theories, originally developed for physical assets, to the context of data. Quality-driven approaches link multidimensional data quality metrics to estimated value (Xing & Wang, 2024; Yang et al., 2019), while cost-based, revenue-based, and market-based models assess value from the perspective of production cost, expected returns, or comparable transactions (Mehta et al., 2021, 2022).

Technological developments have enabled algorithmic valuation frameworks that estimate the marginal contribution of data to analytical or predictive tasks. Cong et al. (2022) provides an overview of how ML models are priced for end users during deployment. Sim et al. (2022) delivers a comprehensive technical overview, formally analyzing ML data valuation through its constituent components and corresponding properties. These valuation methods provide an essential foundation for subsequent pricing mechanisms.

Another influential line of work in data valuation leverages the **shapley value**, a cooperative game-theoretic concept originally used for fairly distributing payoffs among players. In the context of machine learning, the Shapley value quantifies the marginal contribution of each data point to model performance. Jia et al. (2019) propose approximation methods for efficient computation and later develop an exact algorithm for nearest-neighbor classifiers (Jia et al., 2019). Ghorbani and Zou (2019) introduce a truncated Monte Carlo sampling scheme (TMC-Shapley), demonstrating empirical effectiveness across multiple ML tasks, and subsequently extend the framework to distributional Shapley (G-Shapley), where the value of a data point is defined relative to an underlying data distribution (Ghorbani et al., 2020).

Other researchers have investigated the role of **perplexity** as a proxy for data quality and downstream performance. Prior work has shown systematic relationships between web corpus validation loss and benchmark performance, suggesting that correlations between perplexity and benchmark scores can serve as the basis for data selection policies. Validation losses on text corpora are widely used as indicators of downstream utility when comparing large language models trained on the same data distribution (Hoffmann et al., 2022; Kaplan et al., 2020). These correlations hold even across models with different architectures. Such findings support the view that perplexity captures meaningful information about the usefulness of data. Building on this insight, our work extends perplexity into a weighted perplexity entropy (WPE) framework, which integrates linguistic and statistical weights to provide a more fine-grained and theoretically grounded measure for valuing textual data.

2.2. Data pricing mechanisms

Pricing mechanisms convert valuation outcomes into actual transaction prices, considering market conditions and strategic interactions. Economic-based pricing often employs auction designs to allocate data to buyers with the highest valuations, such as one monopolistic data seller and n potential buyers (Agarwal et al., 2024) and many-to-many trading frameworks (Duan et al., 2023). Smart-contract-enabled auctions prevent collusion without the need for trusted intermediaries (Xiong & Xiong, 2021). Game-theoretic models, including non-cooperative games (Inegbedion et al., 2023), Stackelberg games (Xiao et al., 2020) and two period game model (Chen et al., 2026), capture strategic dynamics between sellers and buyers, and some integrate privacy loss into pricing to ensure fair profit distribution (Yang et al., 2024).

From a computational perspective, query-based pricing assigns prices to database queries using predefined view prices and has been extended to handle incomplete datasets (Miao et al., 2020). Privacy-preserving pricing adjusts compensation according to quantified privacy exposure (Shen et al., 2022), thereby balancing utility and privacy in sensitive domains, such as healthcare. Blockchain-enabled mechanisms facilitate secure, transparent, and automated pricing and payment processes (Heideman et al., 2024).

Online learning approaches have been adopted to model sequential data acquisition and adaptive pricing. The multi-armed bandit (MAB) framework offers a natural paradigm for balancing exploration and exploitation as platforms iteratively interact with data owners. Classical strategies, such as ϵ-greedy, upper confidence bound (UCB), and Thompson sampling, have been adapted for use in dynamic pricing problems. Contextual bandits extend this paradigm to scenarios where pricing depends on data features or user profiles. The multi-armed bandit framework has been widely applied across diverse domains (Li & Duan, 2025; Wang et al., 2022; Xu et al., 2024).

Reinforcement Learning (RL) has emerged as a promising approach for dynamic pricing due to its adaptability and continuous learning capability. As described by Sutton et al. (1998), RL algorithms iteratively interact with the environment through trial and error, discovering optimal pricing strategies without assuming a fixed market model. Applications in domains such as electric vehicle charging, retailers' pricing decisions, and perishable products demonstrate the effectiveness of RL-based approaches in handling time-varying conditions and diverse real-world environments (Hao et al., 2022; Kavoosi et al., 2025; Liu et al., 2024; Wu et al., 2023).

2.3. Comparison with existing works

As shown in Table 1, existing data valuation methods exhibit clear limitations across four critical dimensions. No Label Needed evaluates whether a method can function without supervised labels, which is essential when labeled data are scarce or expensive. No Retraining Needed reflects the extent to which a method avoids repeated model retraining, a property that strongly influences computational scalability. Online Deployment measures whether a method can adapt in dynamic or streaming environments,

Table 1
Comparative analysis of valuation mechanisms with existing studies

References	Valuation approach	No label needed	No retraining needed	Online deployment	Semantic salience
Ghorbani and Zou (2019)	Shapley Value	×	×	×	×
Xing and Wang (2024)	Quality	✓	✓	×	×
Li et al. (2017)	Information Entropy	✓	✓	✓	×
Xu et al. (2024)	Bayesian Posterior	✓	✓	✓	×
Proposed WPE	Weighted Perplexity	✓	✓	✓	✓

Table 2
Comparative analysis of pricing mechanisms with existing studies.

Reference	Pricing approach	Online deployment	Budget awareness	Long-term profitability	Contextual adaptation
Agarwal et al. (2024)	Auctions	✓	✓	×	×
Xiao et al. (2020)	Game Theory	×	×	×	×
Miao et al. (2020)	Query	×	×	×	×
Xu et al. (2024)	Contextual Bandits	✓	×	×	✓
Xu et al. (2024)	Bandits with Knapsack	✓	✓	×	✓
Proposed PPO-A	Reinforcement Learning	✓	✓	×	✓
Proposed PPO-B	Reinforcement Learning	✓	✓	✓	✓

providing valuation in real time rather than relying solely on offline computation. Finally, semantic salience encompasses the ability to incorporate linguistic and semantic features beyond surface-level statistics. This capability is crucial for textual data tasks. Within this framework, Shapley-based approaches fail across all dimensions, as they rely on supervision, repeated retraining, and remain offline without any token importance modeling. Quality-based methods eliminate the need for labels and retraining. However, their static design prevents online deployment, and they remain agnostic to semantic signals. Information-entropy-based approaches, as well as the recent VAP method, further relax these constraints by supporting unsupervised and online valuation without retraining, yet they still operate purely at the statistical level, blind to semantic structure. In contrast, our proposed WPE uniquely satisfies all four criteria, combining unsupervised and computation-efficient valuation with online adaptability and explicit integration of linguistic semantics.

For pricing mechanisms (Table 2), existing approaches face different trade-offs. Budget Awareness indicates whether the mechanism accounts for the platform's limited budget to prevent inefficient allocation or premature exhaustion. The Long-term Profitability captures whether cumulative rewards across multiple interactions are optimized, rather than focusing only on immediate profit. Contextual Adaptation refers to decisions that condition on the characteristics of individual data samples (e.g., estimated value) and platform states (e.g., budget ratio, time step), as opposed to non-contextual methods that apply uniform or heuristic offers regardless of sample-specific features. Traditional auctions and game-theoretic methods lack Long-term Profitability. Query-based pricing enables interactive discovery, but it is limited to discrete, non-contextual decisions. Contextual bandits introduce online learning with contextual information but overlook budget considerations, whereas bandits with knapsack (BwK) address budget constraints but still focus on short-term, discrete objectives without long-term profit optimization. Our RL-based strategies overcome these limitations: PPO-A enables continuous and budget-aware pricing tailored to contextual features for immediate profit, while PPO-B further extends this to long-term profit optimization, achieving dynamic profit maximization under budget constraints.

Unlike existing works that address valuation or pricing separately, our framework unifies both dimensions. The WPE method provides online, semantics-aware sample assessment, while RL-based pricing leverages these valuations for budget-constrained acquisition decisions, enabling dynamic pricing that balances immediate profit with long-term profit maximization.

3. Mechanism overview

In mobile and distributed environments, user-generated textual data arrives continuously, often with diverse quality and varying relevance to downstream tasks. Platforms operating under limited resources, such as communication bandwidth, attention span, or monetary budget, must decide which data samples to acquire and how much to pay for them, to build useful models for applications like health monitoring or sentiment classification. This challenge is further complicated by the sequential nature of data arrival, where the platform acts as a data acquirer, continuously interacting with individual users who each contribute unique data samples with potentially varying value.

This paper addresses the challenge of information-aware and dynamic data acquisition in such environments. As illustrated in Fig. 1, we consider a setting where the platform sequentially observes data samples and interacts with data owners by offering monetary compensation in exchange for their data. At each time step, the platform must determine a price to offer based on its internal valuation and remaining budget, while users either accept or reject the offer depending on their private reservation prices. The objective is to maximize the cumulative profit of the acquired dataset within a limited budget, which naturally gives rise to a sequential decision-making problem under budget and information constraints.

To tackle this problem, we propose a two-stage framework that decouples valuation and pricing. Stage 1 (Information-aware Valuation) computes fine-grained informativeness scores for each text using a masked language model augmented with linguistic and statistical weights, including part-of-speech tags, dependency roles, and TF-IDF features. This weighted perplexity entropy (WPE)

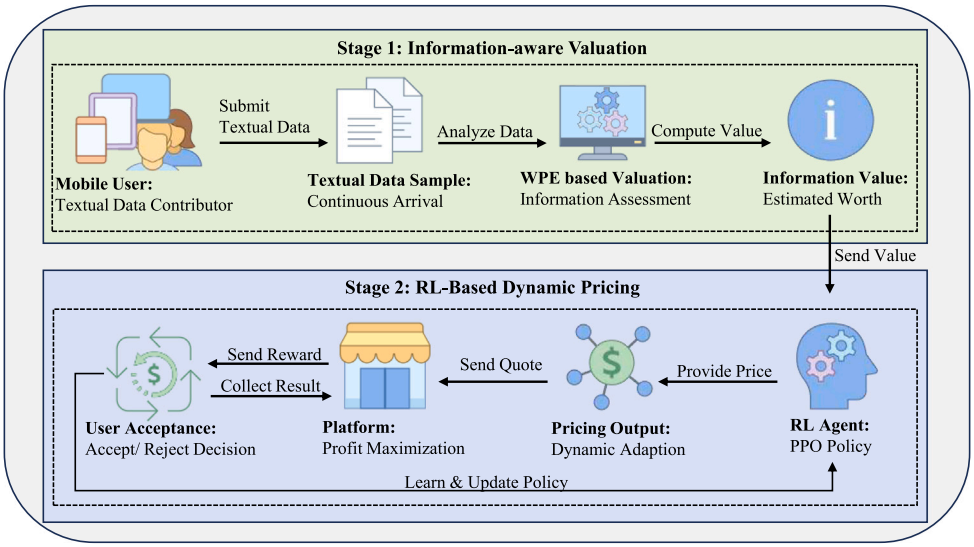


Fig. 1. Information-aware valuation and RL-based dynamic pricing in mobile environments.

approach enables semantic-aware assessment without requiring labels or model retraining. Stage 2 (RL-based Dynamic Pricing) formulates data acquisition as a Markov Decision Process and trains a Proximal Policy Optimization (PPO) agent to post prices under budget constraints and stochastic user acceptance behavior, balancing early exploitation of high-value samples with long-term budget efficiency.

We adopt PPO as it combines the stability of trust-region updates with the scalability of policy-gradient methods, making it particularly suitable for continuous action spaces and long-horizon episodes. In addition, we adopt a model-free formulation because accurately specifying a transition model for user acceptance behavior is challenging in data acquisition markets, where reservation prices, submission timing, and user heterogeneity may vary over time. Model-based RL therefore requires strong assumptions about behavioral dynamics, which may introduce model misspecification bias in practice. Nevertheless, model-based approaches constitute a promising extension when reliable behavioral models or interaction logs are available. Inverse reinforcement learning (IRL) typically relies on expert demonstrations to infer reward functions, whereas in our setting, optimal pricing policies are not directly observable.

In the current study, we intentionally adopt a two-stage design in which the valuation module remains fixed during pricing. This separation is motivated by both methodological and practical considerations. First, allowing valuation and pricing to be jointly updated would introduce a strongly non-stationary learning signal: changes in valuation would alter the reward landscape faced by the pricing agent, making policy optimization less stable and complicating the attribution of performance improvements. Second, maintaining a separable valuation stage preserves interpretability. The WPE metric is designed to provide an explainable estimate of informational value, enabling analysis of linguistic and statistical contributions, which would be more difficult in an end-to-end formulation.

Third, in many practical data acquisition systems, data quality assessment and pricing decisions are operationally separated, making a modular design both realistic and tractable. Recent literature explicitly supports this separation as a foundational design choice. Research on data market architectures demonstrates that decoupling quality evaluation from pricing functions enables more robust and scalable marketplace operations (Bauer-Hänsel et al., 2024). The modular approach allows quality assessment components to focus exclusively on technical metrics, such as accuracy, completeness, consistency, and timeliness, while pricing mechanisms can independently optimize for market dynamics, consumer utility, and profit maximization (Lin et al., 2025). This separation aligns with established data governance principles where quality control operates as an independent validation layer before economic valuation processes commence (Bernardo et al., 2024).

Our formulation of valuation and pricing captures the key challenges of mobile data markets: asynchronous data arrival, valuation under information asymmetry, strategic pricing under uncertainty, and decision-making under dynamic resource constraints. By integrating semantic-aware valuation with budget-conscious reinforcement learning, our framework establishes a principled foundation for learning-based data procurement in information-centric ecosystems, moving beyond static or myopic schemes toward dynamic pricing that optimizes both immediate and long-term profit maximization.

4. Weighted perplexity entropy valuation

In this section, we propose a new data valuation method based on predictive uncertainty under masked language models (MLMs), namely Weighted Perplexity Entropy (WPE). Unlike traditional valuation metrics that rely on data label information or model parameter space, WPE captures the informativeness of a text sample by aggregating the prediction uncertainty of token-level masking, modulated by linguistic and statistical importance weights.

4.1. Predictive entropy under MLMs

Let a text sample be $T = [t_1, \ldots, t_n]$. We uniformly sample a subset of token indices $I \subseteq \{1, \ldots, n\}$ with size $m = |I|$ (mask ratio ρ). For each $i \in I$, the masked language model (MLM) produces a predictive distribution $P_i(\cdot)$ over the vocabulary by replacing t_i with a `[MASK]` token.

The token-level negative log-likelihood is defined as

$$\ell_i = -\log P_i(t_i^{\text{true}}), \tag{1}$$

where t_i^{true} is the ground-truth token at position i. The sample-level predictive entropy is then

$$H(T) = \frac{1}{m} \sum_{i \in I} \ell_i. \tag{2}$$

Equivalently, the corresponding perplexity is $\text{PP}(T) = \exp(H(T))$.

4.2. Incorporating linguistic and statistical weights

While the unweighted perplexity entropy $H(T)$ provides a baseline measure of sample-level uncertainty, it assumes uniform contribution across tokens. To capture token-level variation in informativeness, we introduce a weight w_i that modulates each token's entropy term.

The composite weight is defined as

$$w_i = w_i^{(\text{POS})} \cdot w_i^{(\text{DEP})} \cdot w_i^{(\text{TF-IDF})}, \tag{3}$$

where each factor highlights a distinct property:

4.2.1. Part-of-speech weight

$$w_i^{(\text{POS})} = \omega_{\text{pos}(t_i)}, \tag{4}$$

with $\omega_{\cdot}$ a fixed mapping from POS tags to scores (content words such as nouns and verbs receive larger values).

4.2.2. Dependency role weight

$$w_i^{(\text{DEP})} = \omega_{\text{dep}(t_i)}, \tag{5}$$

where central syntactic roles (e.g., subjects, objects, roots) are emphasized through higher weights.

4.2.3. Statistical rarity weight

$$w_i^{(\text{TF-IDF})} = \text{TF}(t_i, T) \cdot \left(\log\left(\frac{1+N}{1+\text{DF}(t_i)} \right) + 1 \right), \tag{6}$$

where $\text{TF}(t_i, T)$ is the term frequency in sample T, N is the corpus size, and $\text{DF}(t_i)$ is the document frequency. The +1 terms are standard smoothing constants.

These factors assign higher weights to tokens that are syntactically central, semantically informative, or statistically distinctive. A detailed summary of specific parameter setting of three composite weight can be found in Appendix Tables A.1, A.2 and A.3. In the next subsection, they are combined with token-level perplexity entropy to define the weighted valuation function.

4.3. Weighted perplexity entropy valuation

Combining token-level perplexity entropy with the importance weights, we define the final valuation function for a text sample T:

$$G(T) = \frac{1}{|I|} \sum_{i \in I} w_i \, \ell_i \tag{7}$$

This function integrates model uncertainty with token-level salience, assigning a higher value to samples that are both more difficult to predict and linguistically informative. The resulting score, $G(T)$, can be utilized in downstream applications, such as sample ranking, noise filtering, and dynamic pricing.

The information value $G(T)$ is designed to quantify the expected contribution of a textual data item to downstream learning and decision-making processes. From an economic perspective, acquiring data generates value not directly through the data itself, but through its impact on the performance of models or systems that rely on the data. Higher information value implies a greater potential to reduce predictive uncertainty, improve generalization, or accelerate learning efficiency when the data is incorporated into a downstream task. Under this view, the economic value of a data item can be interpreted as the expected

marginal benefit arising from such performance improvements, rather than as a deterministic monetary gain. Accordingly, we treat $G(T)$ as a principled proxy for the expected economic utility of acquiring a data item, abstracting away from task-specific payoff structures while preserving the relative ordering of data usefulness. This abstraction enables a unified valuation formulation in which heterogeneous textual samples can be compared and valued based on their anticipated contribution to downstream outcomes.

The weighting mechanism is not introduced as arbitrary parameter tuning, but as an interpretable decomposition of token salience along complementary linguistic dimensions (frequency, syntactic role, and dependency structure). While these weights are heuristic, they provide structured, explainable priors for modulating token-level uncertainty contributions.

4.4. Theoretical justification

To clarify the conceptual positioning of WPE, we emphasize that its contribution lies in the valuation mapping rather than in entropy theory or language model design. Traditional perplexity aggregates token-level surprisal under a uniform expectation operator. WPE generalizes this aggregation into a weighted expectation, thereby relaxing the implicit uniform-token assumption when perplexity is used as a proxy for data value. This modification operates at the functional level of valuation, not at the level of language modeling architecture.

The proposed WPE valuation is grounded in information theory and data utility analysis. The unweighted entropy $H(T)$, defined as the expected negative log-likelihood of true tokens under a language model, measures predictive uncertainty. From this perspective, we assume that samples with higher $H(T)$ carry greater informational potential and are therefore more likely to reduce perplexity when used in training, consistent with the value-of-information principle.

The weighting scheme further connects to linguistic informativeness: $w_i^{(\text{POS})}$, $w_i^{(\text{DEP})}$, $w_i^{(\text{TF-IDF})}$ emphasize syntactically central or statistically rare tokens, thereby amplifying their entropy contribution. Consequently, the weighted valuation $G(T)$ integrates predictive uncertainty with token-level salience, serving as a proxy for the marginal utility of a sample, which is consistent with the classical view of data value as a marginal contribution.

The predictive uncertainty is not a universal measure of data utility. From an information-theoretic perspective, entropy quantifies uncertainty (Shannon information), whereas utility is inherently task-dependent and relates to the marginal reduction in predictive loss or posterior uncertainty. Our framework does not claim that high uncertainty universally implies high utility. Rather, under standard likelihood-based training dynamics, predictive entropy can be interpreted as an approximation to expected information gain with respect to the model's current posterior. In this sense, WPE serves as a proxy for expected marginal impact on model updates, rather than intrinsic semantic value. We acknowledge that highly predictable samples may still carry substantial utility, for example when they represent prototypical or structurally critical patterns. Therefore, WPE should be viewed as a task-conditioned proxy metric, whose validity depends on the learning objective and model assumptions. To partially bridge the gap between information quantity and information utility, we incorporate linguistic importance weights and empirically validate WPE through downstream task performance improvements. However, we do not claim that WPE universally captures all forms of data utility.

4.5. Implementation and computational considerations

Algorithm 1 Weighted Perplexity Entropy (WPE) Computation

Require: Text sample $T = [t_1, t_2, \ldots, t_n]$, pre-trained MLM $\mathcal{M}$, masking ratio ρ
Ensure: WPE(T)

```
Compute POS tags and dependency labels for all tokens in T
Compute TF-IDF scores for all tokens in T
I ← sample({1, ..., n}, ⌊ρ · n⌋)
total_entropy ← 0
for each i ∈ I do
    T_\i ← mask(T, i)
    P_i ← M(T_\i)
    ℓ_i ← − log P_i(t_i^true)
    w_i ← w_i^(POS) · w_i^(DEP) · w_i^(TF-IDF)
    total_entropy ← total_entropy + w_i · ℓ_i
end for
return total_entropy/|I|
```

Algorithm 1 presents the complete WPE computation procedure for practical implementation. The computational complexity of WPE is mainly determined by MLM inference, which requires $O(|\mathcal{I}| \cdot n)$ operations, where $|\mathcal{I}|$ is the number of masked positions and n is the sequence length. POS tagging, dependency parsing, and TF–IDF scoring each add $O(n)$ per sample, while corpus-level TF–IDF statistics incur a one-time cost of $O(|\mathcal{D}| \cdot \bar{n})$. Thus, the overall per-sample complexity is $O(|\mathcal{I}| \cdot n + n)$, making WPE computationally feasible for large-scale text valuation.

In mobile textual data pricing scenarios, this computational efficiency is particularly important given the resource constraints and real-time processing requirements. The method can be deployed on mobile devices or edge servers to provide on-demand data valuation for dynamic pricing mechanisms. The label-free nature of WPE makes it particularly suitable for mobile environments

where ground-truth labels are often unavailable or expensive to obtain. Furthermore, the modular design allows for domain-specific customization of importance weights, enabling adaptation to different mobile application contexts such as social media posts, messaging data, or location-based text content. The weighting scheme can also be tuned or learned in a task-specific manner, allowing for further customization to meet diverse valuation objectives.

4.6. Properties of data valuation metric

4.6.1. Weight separability

An important property of our valuation function $G(T)$ is *weight separability*, which refers to the ability to independently analyze and interpret the effect of each weight component in the aggregated score.

$$G(T) = \frac{1}{m} \sum_{i \in \mathcal{I}} w_i \cdot \ell_i \quad \text{with} \quad w_i = w_i^{(\text{POS})} \cdot w_i^{(\text{DEP})} \cdot w_i^{(\text{TF-IDF})} \tag{8}$$

Due to the multiplicative structure of w_i, the contribution of each token t_i to the final valuation $G(T)$ can be decomposed into three interpretable dimensions:

- **Syntactic Importance:** $w_i^{(\text{POS})}$ reflects the token's grammatical category and highlights the relative value of content vs. function words.
- **Structural Centrality:** $w_i^{(\text{DEP})}$ captures the token's role in dependency parsing, emphasizing core components such as subjects and roots.
- **Statistical Rarity:** $w_i^{(\text{TF-IDF})}$ quantifies the distinctiveness of the token within the corpus.

Formally, this separability allows us to express the overall score as a weighted sum over log-likelihoods modulated by interpretable dimensions. Thus, $G(T)$ is not a black-box metric, but a modular and explainable score that facilitates empirical insights into what makes a token or a sample valuable.

4.6.2. Sampling unbiasedness under masked token selection

A desirable property of the WPE valuation is its robustness to the stochasticity introduced by masking. Since the valuation is computed based on a randomly selected subset of tokens $\mathcal{I} \subseteq \{1, \ldots, n\}$, it is important to ensure that such sampling does not bias the final score. This motivates the following property:

Let $G^{(t)}(T)$ denote the WPE valuation computed on a text sample T with a random subset $\mathcal{I}^{(t)}$ of masked positions in the tth trial. If each subset $\mathcal{I}^{(t)}$ is sampled uniformly without replacement from the token positions, then:

$$\mathbb{E}_{\mathcal{I}}[G(T)] = \mathbb{E}_{\mathcal{I}} \left[-\frac{1}{m} \sum_{i \in \mathcal{I}} w_i \cdot \log P_i(t_i^{\text{true}}) \right] \approx \frac{1}{M} \sum_{t=1}^{M} G^{(t)}(T) \tag{9}$$

Here M denotes the number of independent masking trials, and $G^{(t)}(T)$ is the valuation obtained under the tth random mask subset $\mathcal{I}^{(t)}$.

In other words, the expected value of WPE over multiple random maskings approximates the full valuation computed over all token positions. This holds as long as the masking strategy is uniform, regardless of the specific `mask_ratio` used.

Thus, WPE provides an unbiased estimate of a sample's informational value even when computed on a subset of tokens. This makes it computationally efficient while maintaining estimation reliability across different runs.

4.6.3. Label-free applicability

A notable advantage of the WPE valuation is its independence from ground-truth labels. Unlike many data valuation methods that rely on supervised training signals or task-specific loss functions, WPE operates solely based on the model's predictive uncertainty over masked tokens and token-level linguistic/statistical weights.

As the valuation relies solely on model-internal uncertainty, it requires no supervision signal and remains applicable to both labeled and unlabeled regimes. This makes it especially valuable in:

- Pretraining corpus filtering: selecting high-value samples before any fine-tuning.
- Domain adaptation scenarios: where target-domain labels are unavailable.
- Privacy-constrained tasks: where using ground-truth labels may expose sensitive information.

In essence, WPE provides a principled, task-agnostic, and model-aware measure of sample quality that remains valid even before task-specific annotations are applied.

5. Reinforcement learning for dynamic data pricing

5.1. MDP design

We formulate the dynamic data pricing problem as a Markov Decision Process (MDP), where the platform sequentially interacts with data owners and decides a price to offer for each incoming data sample. The goal is to develop a pricing policy that maximizes cumulative profit while adhering to budget constraints and considering the stochastic nature of user responses.

5.1.1. Environment

At each time step $t = 1, 2, \ldots, T$, the platform observes a new data sample with estimated value G_t (e.g., based on WPE valuation) and must decide a price p_t to offer. The data owner then either accepts or rejects the offer, based on a latent acceptance function. If accepted, the sample is purchased, and the platform's budget is reduced by p_t. Otherwise, the transaction is skipped, and the budget remains unchanged.

At the beginning of each episode, the platform is assigned a budget B_0 sampled from a Gaussian distribution centered around a predefined ratio (e.g., 50%) of the total episode value. This stochastic budget initialization captures uncertainty in available resources across episodes and encourages the learning of more robust and adaptive pricing strategies. The platform aims to optimally allocate its budget over the T-step horizon.

5.1.2. State space

The agent's state at time t is represented as:

$$s_t = \left(G_t, B_t, t\right) \tag{10}$$

where:

- G_t is the estimated value of the current sample,
- B_t is the normalized remaining budget (i.e., remaining budget divided by initial budget B_0),
- t is the normalized step index, representing the current progress in the episode,

5.1.3. Action space

At each time step t, the agent selects a continuous action $a_t \in [0, 1]$, representing the proportion of the sample's estimated value G_t (as obtained from the valuation model) that the platform offers as the price:

$$p_t = a_t \cdot G_t \tag{11}$$

Unlike conventional discrete-arm pricing strategies, this formulation enables fine-grained and adaptive pricing tailored specifically to the estimated value of each individual data sample.

5.1.4. Transition dynamics

After determining the action a_t, the environment calculates the actual offered price $p_t = a_t \cdot G_t$ and includes an additional *early submission incentive* E_t, designed to encourage earlier data acquisitions within the episode:

$$E_t = \kappa \cdot G_t \cdot \left(1 - \frac{t}{T}\right) \tag{12}$$

where $\kappa \in (0, 1)$ denotes the early incentive coefficient, controlling the magnitude of the time-decaying bonus. Thus, the effective total payment offered by the platform at time t is:

$$P_t = p_t + E_t \tag{13}$$

Next, the environment samples a binary outcome $x_t \in \{0, 1\}$ indicating acceptance ($x_t = 1$) or rejection ($x_t = 0$) of the offered price, governed by a probabilistic acceptance function $\phi(G_t, p_t)$. We adopt a *soft budget constraint*: purchases may temporarily exceed the remaining budget B_t, in which case a penalty Penalty_t is applied. The budget state is updated as Eq. (14), where B_t denotes the actual remaining budget in monetary units.

$$B_{t+1} = \begin{cases} B_t - P_t, & \text{if } x_t = 1, \\ B_t, & \text{otherwise.} \end{cases} \tag{14}$$

5.1.5. User acceptance model

We assume that each data owner has a private reservation price $r_t^{(\text{user})}$, which is independently sampled from a uniform distribution:

$$r_t^{(\text{user})} \sim \mathcal{U}(r_{\min}, r_{\max}) \tag{15}$$

Given an offered price p_t, the probability that the user accepts the offer is governed by a sigmoid function of the pricing surplus:

$$\phi(p_t, r_t^{(\text{user})}) = \sigma\left(\gamma_{acc} \cdot (p_t - r_t^{(\text{user})})\right) = \frac{1}{1 + e^{-\gamma_{acc}(p_t - r_t^{(\text{user})})}} \tag{16}$$

where $\gamma_{acc} > 0$ controls the steepness of the response curve. This probabilistic model captures heterogeneous user behavior and introduces realistic stochasticity into the environment. Larger values of γ_{acc} result in more deterministic responses, while smaller values allow for noisier acceptance behavior.

5.1.6. Modeling assumptions

The uniform reservation-price distribution and the sigmoid acceptance function are adopted as a baseline behavioral model, primarily for analytical clarity and stable simulation. These assumptions allow us to derive tractable acceptance probabilities and to focus on the interaction between data valuation and dynamic pricing. We acknowledge that real-world user price sensitivity may deviate from this simplified setting. In practice, reservation prices can exhibit skewness, heavy tails, or multimodal patterns due to user heterogeneity, budget constraints, and contextual factors, while acceptance behavior may vary across user groups and application domains. Our framework does not rely on these assumptions being exact. Rather, they serve as a reference model for theoretical derivation and initial policy learning, and the robustness of the proposed pricing strategy under alternative distributions and acceptance functions is systematically examined in the experimental sensitivity analysis.

We further note that user acceptance in real-world mobile environments can depend on contextual factors (e.g., time, location, cognitive load), privacy/data sensitivity, and repeated-interaction effects. Modeling such context-dependent and strategic behaviors would require additional user-side signals or longitudinal transaction logs, which are not available in our current experimental setting. Therefore, we adopt a parsimonious baseline and evaluate robustness under alternative distributions and acceptance functions.

5.2. Formal MDP definition

Formally, our dynamic pricing problem is defined as a Markov Decision Process (MDP) tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma_{disc})$, where:

- State space $\mathcal{S}$: $s_t = (G_t, B_t, t)$, where G_t is the current sample valuation, B_t is the normalized remaining budget ($B_t := \frac{\text{remaining budget}}{B_0}$), and t is the normalized time index.
- Action Space $\mathcal{A}$: A continuous action $a_t \in [0, 1]$, representing the pricing ratio $p_t = a_t \cdot G_t$.
- Transition Dynamics $\mathcal{P}$: Acceptance outcome $x_t \sim \text{Bernoulli}\left(\phi(p_t, r_t^{(\text{user})})\right)$, where $r_t^{(\text{user})} \sim \mathcal{U}(r_{\min}, r_{\max})$, and budget evolution $B_{t+1} = B_t - P_t$ if accepted and sufficient budget remains.
- Reward Function $\mathcal{R}$: Defined as immediate or total profit minus soft budget penalty. Its detailed design (including PPO-A and PPO-B) is discussed in the next section.
- Discount Factor γ_{disc}: Typically set as $\gamma_{disc} = 1$ for finite-horizon optimization.

It is important to clarify how the proposed pricing mechanism differs from conventional reinforcement learning-based dynamic pricing. In most existing reinforcement learning pricing studies, the objective is to optimize revenue or demand under inventory or demand-curve uncertainty (Kavoosi et al., 2025; Qiao et al., 2024). In contrast, our setting focuses on budget-constrained data acquisition, where each transaction corresponds to acquiring a data sample with heterogeneous and exogenously estimated value G_t. The agent therefore learns a pricing policy that allocates a limited budget across heterogeneous data items to maximize cumulative net utility, which differs fundamentally from traditional price-demand optimization problems.

5.3. Reward function

We propose two reward formulations, each capturing distinct optimization goals:

5.3.1. Budget penalty mechanism

To ensure that the policy respects budget constraints while allowing controlled flexibility during training, we employ a tiered soft-budget penalty function. Let P_t be the total payment at time t, and B_t the remaining budget. Define the budget excess as:

$$\Delta_t = \max(P_t - B_t, 0) \tag{17}$$

The penalty function employs three thresholds relative to the remaining budget B_t:

- Light penalty region: $0 \le \Delta_t \le \alpha_{\text{light}} B_t$
- Medium penalty region: $\alpha_{\text{light}} B_t < \Delta_t \le \alpha_{\text{heavy}} B_t$
- Heavy penalty region: $\Delta_t > \alpha_{\text{heavy}} B_t$

Specifically, the penalty at time t is computed as follows:

$$\text{Penalty}_t = w_{\text{light}} \Delta_t^{\text{light}} + w_{\text{medium}} \Delta_t^{\text{medium}} + w_{\text{heavy}} \Delta_t^{\text{heavy}} \tag{18}$$

where each term is defined as:

$$\Delta_t^{\text{light}} = \min(\Delta_t, \alpha_{\text{light}} B_t) \tag{19}$$

$$\Delta_t^{\text{medium}} = \min\left(\max(\Delta_t - \alpha_{\text{light}} B_t, 0), (\alpha_{\text{heavy}} - \alpha_{\text{light}}) B_t\right) \tag{20}$$

$$\Delta_t^{\text{heavy}} = \max(\Delta_t - \alpha_{\text{heavy}} B_t, 0) \tag{21}$$

The parameters α_{light}, α_{heavy}, and weights w_{light}, w_{medium}, w_{heavy} control the penalty strength, guiding the policy toward careful but flexible budget management.

The penalty scaling parameters (e.g., α_{light}, α_{heavy}) should be interpreted as context-dependent factors rather than universal constants. In practical deployments, these parameters can be calibrated to reflect the economic trade-off between short-term budget violations and the long-term value of acquired data. Conceptually, the calibration can be guided by estimating the marginal economic benefit of additional data in the downstream task, for example through historical A/B testing, model performance sensitivity analysis, or cost-benefit evaluation. Applications with high marginal value of information may justify a more relaxed penalty (larger α), whereas cost-sensitive environments may require stricter budget discipline (smaller α).

5.3.2. Reward function for PPO-A

The immediate reward at each step t aims to maximize instantaneous profit, explicitly accounting for the early submission incentive and the soft budget penalty. It is defined as:

$$r_t^{(A)} = \begin{cases} (G_t - P_t) - \text{Penalty}_t, & \text{if accepted} \\ -\text{Penalty}_t, & \text{otherwise} \end{cases} \tag{22}$$

where the effective payment P_t includes an early submission incentive:

$$P_t = p_t + \kappa \cdot G_t \cdot \left(1 - \frac{t}{T}\right), \tag{23}$$

and Penalty_t represents the tiered budget penalty defined in Eq. (18). Thus, PPO-A directly optimizes short-term transaction efficiency under controlled budget management.

5.3.3. Reward function for PPO-B

Similar to PPO-A, the total payment P_t includes an early submission bonus. This augmented payment is used consistently across all components of the reward function defined below. In PPO-B, the reward function is designed to optimize the total cumulative profit over the episode, integrating several strategic considerations to balance short-term and long-term objectives:

$$r_t^{(B)} = r_t^{(\text{immediate})} + r_t^{(\text{value-density})} + r_t^{(\text{total-performance})} - \text{Penalty}_t \tag{24}$$

The penalty term Penalty_t follows the same tiered penalty mechanism described in Eq. (18). Specifically, each component is defined as:

- **Immediate profit**:

$$r_t^{(\text{immediate})} = \begin{cases} G_t - P_t, & \text{if accepted} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

- **Value density bonus**:

$$r_t^{(\text{value-density})} = \begin{cases} \delta \cdot r_t^{(\text{immediate})}, & \text{if } G_t > \text{AvgValue}(t), \\ & \text{and accepted} \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

  where δ is the coefficient, and $\text{AvgValue}(t)$ is the average valuation of accepted samples until step t. This component prevents the policy from over-investing in low-value but easily accepted samples, thereby improving the overall quality of the acquired dataset under budget constraints.

- **Total performance bonus**:

$$r_t^{(\text{total-performance})} = \begin{cases} \lambda \cdot r_t^{(\text{immediate})} \cdot \frac{\sum_{i=1}^{T} r_i^{(\text{immediate})}}{B_0}, \\ \text{if final step and total profit ratio} > \xi \\ 0, \text{otherwise} \end{cases} \tag{27}$$

  where λ is the scaling coefficient, and ξ is the final profit ratio threshold. By linking the final reward to the overall profit-to-budget ratio, this component enforces long-term efficiency: the agent is incentivized to sustain profitable strategies across the entire horizon rather than maximizing short-term gains.

PPO-B thus systematically encourages optimal long-term decision-making, guiding the agent toward sustained profitability while maintaining effective budget management.

5.3.4. Objective

The platform aims to learn an optimal pricing policy $\pi(s)$ that maximizes the expected cumulative reward over T rounds. Unlike a hard-constraint formulation with $B_t \geq 0$, our environment adopts a *soft budget constraint*: the budget state B_t is allowed to temporarily fall below zero, but such overspending incurs a penalty Penalty_t as defined in Eq. (18). Accordingly, the optimization problem is expressed as:

$$\max_{\pi} \ \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} \left(r_t - \text{Penalty}_t\right)\right], \tag{28}$$

where r_t includes immediate profit components and incentive terms, while Penalty_t discourages excessive budget overruns. This soft-constraint design strikes a balance between exploring uncertain user acceptance behaviors and sustainable long-term budget management. In reality, this relaxation can be interpreted as a short-term financing or credit mechanism, where temporary budget overruns incur explicit penalty costs, analogous to interest or overdraft fees in real-world financial systems. Such mechanisms are common in platform operations where strict real-time budget matching is not required, but financial discipline is enforced through cost penalties.

5.4. Policy optimization

To solve the dynamic pricing MDP introduced above, we employ *Proximal Policy Optimization* (PPO) (Schulman et al., 2017), a widely used policy-gradient method that balances sample efficiency and training stability. PPO is particularly suitable for continuous control tasks, such as adaptive pricing, where the action space spans a real-valued interval and fine-grained policy updates are crucial.

In this framework, the platform maintains two neural networks: a policy network $\pi_\theta(p_t \mid s_t)$ that maps the current state to a distribution over price offers, and a value network $V_\psi(s_t)$ estimating the expected return from a given state.

5.4.1. Advantage estimation

We compute the advantage estimate by subtracting the predicted state value from the immediate reward:

$$A_t = r_t - V_\psi(s_t) \tag{29}$$

These advantages are normalized per training batch to stabilize updates:

$$\hat{A}_t = \frac{A_t - \text{mean}(A_t)}{\text{std}(A_t) + \epsilon}, \quad \epsilon = 10^{-8} \tag{30}$$

5.4.2. Value function update

The value network $V_\psi(s_t)$ is updated by minimizing the mean squared error between its predictions and the empirical returns:

$$\mathcal{L}_{\text{value}}(\psi) = \mathbb{E}_t\left[\left(V_\psi(s_t) - r_t\right)^2\right], \tag{31}$$

where r_t denotes the observed return at timestep t (we approximate returns by the immediate rewards). This provides a learned baseline to reduce variance in policy gradient estimation.

5.4.3. Entropy regularization

To encourage adequate exploration and prevent premature convergence to deterministic policies, we introduce an entropy regularization term into the PPO policy loss. The final policy optimization objective becomes:

$$\begin{aligned}\mathcal{L}_{\text{policy}}(\theta) = -\mathbb{E}_t\left[\min\left(\rho_t(\theta)A_t, \text{clip}\left(\rho_t(\theta), 1-\epsilon, 1+\epsilon\right)A_t\right)\right] \\ -\beta \cdot \mathcal{H}\left(\pi_\theta(\cdot|s_t)\right)\end{aligned} \tag{32}$$

where:

- $\rho_t(\theta) = \frac{\pi_\theta(p_t|s_t)}{\pi_{\theta_{\text{old}}}(p_t|s_t)}$ is the policy probability ratio.
- A_t is the estimated advantage at timestep t.
- ϵ is the PPO clipping parameter.
- $\mathcal{H}(\pi_\theta(\cdot|s_t))$ represents the entropy of the policy distribution at state s_t, with coefficient β controlling the strength of entropy regularization.

This regularization ensures sustained exploration throughout training, enabling the agent to more thoroughly search for optimal pricing strategies.

5.4.4. Training procedure

In practice, each batch of collected trajectories is reused for multiple PPO update epochs. This iterative update scheme improves sample efficiency while maintaining training stability.

6. Experiments

In this section, we present experiments from two complementary directions: data valuation and dynamic data pricing. We first assess the proposed WPE metric in terms of effectiveness, robustness, and comparison with established valuation methods. We then evaluate reinforcement learning–based pricing strategies through profit benchmarking, ablation studies, and scenario-based adaptability tests. All experiments are conducted on real-world datasets to ensure practical generalization.

6.1. Experimental setup

6.1.1. Datasets

We evaluate our data valuation methods using three real-world textual datasets that represent typical mobile-generated content across diverse domains:

Medical Abstract Classification Dataset (MACD) (Schopf et al., 2023) contains medical abstracts for disease category classification, where each abstract describes patient conditions to assist medical professionals in diagnosis. This dataset comprises 14438 medical abstracts across 5 disease categories: neoplasms (3163 samples), digestive system diseases (1494 samples), nervous system diseases (1925 samples), cardiovascular diseases (3051 samples), and general pathological conditions (4805 samples).

Drug Reviews Dataset (DRD) (Gräßer et al., 2018) provides patient reviews on specific drugs, reflecting real user experiences and opinions. The dataset contains 4143 patient reviews across multiple medications and medical conditions. Each entry includes the drug name, patient condition, three aspect-based text reviews (benefits, side effects, and overall comments), a 10-star overall rating, and categorical ratings for side effects and effectiveness.

Douban Movie Short Comments Dataset(Douban)[1] contains 2.14 million Chinese user comments on 28 movies from the Douban platform. Each entry includes comment ID, movie name (in both English and Chinese), username, posting date, star rating (1–5 scale), comment text, and like counts.

For data valuation, we use these three datasets in WPE performance on data valuation, while only use DRD in the robustness test and sota comparison. For dynamic pricing experiments, we partition the DRD dataset with approximately 51% allocated for training the learning environment and 49% reserved as a holdout set for testing, maintaining similar distributions across both sets.

6.1.2. Evaluation metrics

The evaluation process involves two distinct but related perspectives: data valuation using WPE and dynamic pricing using PPO.

For data valuation (WPE), the primary metric is incremental accuracy, which measures the classification accuracy as samples ranked by their WPE scores are gradually added into the training set. This metric directly reflects whether WPE can identify and prioritize informative samples that accelerate model learning. The second is Kendall's Tau correlation coefficient (τ), which measures the robustness of valuation rankings under token-level perturbations such as keyword deletion and word shuffling. Formally, given two rankings R_1 and R_2 of n text samples, Kendall's Tau is defined as:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{33}$$

where C represents the number of concordant pairs and D represents the number of discordant pairs between the two rankings. Higher τ values indicate stronger robustness, as the relative valuation ranks are preserved despite perturbations.

For dynamic pricing (PPO), performance is primarily measured by the cumulative profit per episode, which represents the total reward accumulated by the pricing agent across a complete transaction sequence. Additionally, the convergence rate of the training process indicates learning stability and efficiency, while the average offered price during transactions reflects the agent's pricing behavior and market adaptiveness. These indicators provide a comprehensive view of both profitability and behavioral dynamics in the pricing policy.

6.1.3. Baseline models

To effectively validate our method, this study selected six baseline methods and re-implemented them on our research dataset to facilitate performance comparisons. Baselines are executed on the same artefacts and cohort, ensuring apples-to-apples evaluation.

We divide baselines into two groups according to the utility of them: (A) Valuation Methods: TMC-Shapley, G-Shapley and Random (for valuation); (B) Pricing methods: Half-Fix, LinUCB and Random(for pricing).

- **TMC-Shapley** (Ghorbani & Zou, 2019): The method estimates the contribution of each training sample by measuring its marginal impact across multiple retraining permutations. This method provides an accurate but computationally intensive approximation of true Shapley values.
- **G-Shapley** (Ghorbani et al., 2020): The method offers a more scalable alternative by approximating influence scores using gradient information from a trained model. Despite being efficient, G-Shapley is still model-dependent and sensitive to training stability.
- **Random (Valuation):** We remove training samples in a uniformly random order as a signal-free control. This baseline approximates the expected performance trajectory under no valuation information.
- **Influence Functions** (Koh & Liang, 2017) This method estimates the importance of each training sample by approximating the effect of removing the sample on the validation loss using a second-order influence approximation. It provides a model-based sensitivity analysis of data utility, but the approximation accuracy may degrade in high-dimensional settings or when model assumptions are violated.
- **Beta-Shapley** (Kwon & Zou, 2021): This method extends the classical Shapley value by weighting marginal contributions across different coalition sizes using a Beta distribution. It provides a flexible game-theoretic valuation framework, but still requires Monte Carlo approximation and can be computationally expensive for large datasets.

[1] https://www.kaggle.com/datasets/utmhikari/doubanmovieshortcomments/data

Table 3
Experimental setup and key configurations.

Component	Setting
WPE valuation	
Language model	BERT-base-uncased
Mask ratio	20% tokens per text
Token length	truncated to 64 tokens
Weighting factors	POS, dependency, and IDF
Valuation baselines	
Classifier	ridge classifier (consistent across methods)
Training split	85% training, 15% testing
Random seed	42
Compared methods	TMC-S, G-S, Influence, Beta-S, Random
PPO pricing framework	
State variables	value estimate, budget ratio, time index
Action definition	price ratio $a_t \in [0, 1]$
Policy network	MLP (3–128–64–64)
Training episodes	3000
Learning rate	3×10^{-4}
Clip ratio	0.2
Mini-batch size	256
Acceptance model	reservation price $r \sim \mathrm{Uniform}(0, 10)$
Reward components	profit + early bonus + budget penalty
Penalty thresholds	$\alpha_{\mathrm{light}} = 0.05,\ \alpha_{\mathrm{heavy}} = 0.15$
Penalty weights	$w_{\mathrm{light}} = 0.1,\ w_{\mathrm{medium}} = 1.0,\ w_{\mathrm{heavy}} = 10.0$
Value-density coefficient	$\delta = 0.15$
Total-performance coefficient	$\lambda = 0.5$
Pricing Baselines	
Random Seed	42
Training split	51% training, 49% testing
Random pricing	price sampled uniformly from $[0, g_t]$
Fixed pricing	fixed price ratio $p_t = 0.5 g_t$
LinUCB	action grid $\{0.1 g_t, \ldots, 1.0 g_t\}$

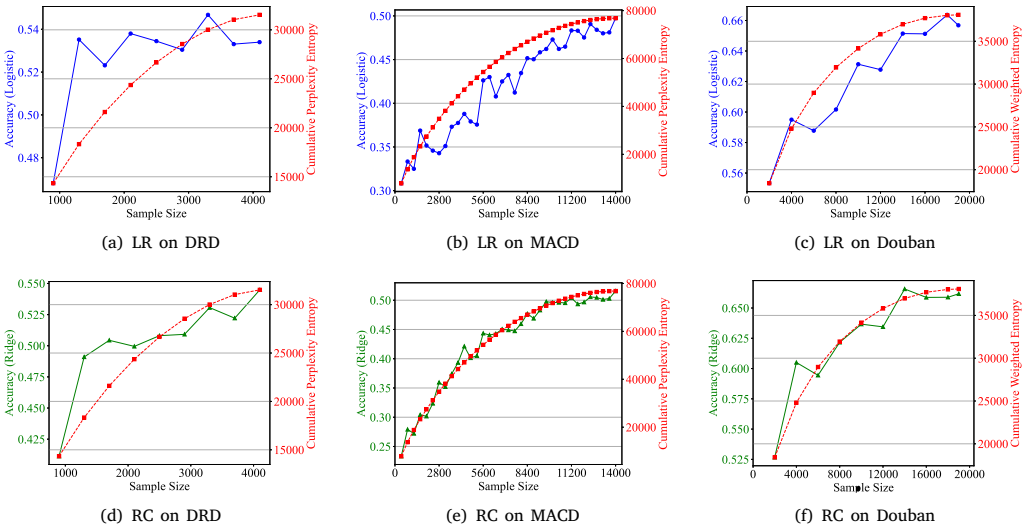
- **Half-Fix:** A simple non-learning heuristic that posts a fixed offer rate (e.g., 50% of the value proxy) for every sample until the budget is exhausted. This baseline reflects a stable "one-size-fits-all" policy without adaptation to sample heterogeneity or remaining budget, providing a conservative lower bound for adaptive methods.
- **LinUCB (Contextual Bandit)** (Xu et al., 2024): A linear upper-confidence-bound bandit that treats pricing as contextual action selection over a discretized set of offer rates. At each step it picks the rate with the highest UCB score given the current context (e.g., sample value proxy and remaining-budget features), then updates its linear model online. LinUCB balances exploration and exploitation without modeling long-term returns, thus benchmarking myopic yet adaptive pricing.
- **Random (Pricing):** For each incoming sample, we draw an offer rate uniformly at random from a fixed range and compute the price accordingly, subject to the same budget constraint and acceptance rule as the RL policies.

6.1.4. Implementation details

Table 3 is the Experimental Setup table. For data valuation, the Weighted Perplexity Entropy (WPE) method integrates three linguistic features into its entropy estimation, which is based on perplexity. These features are TF-IDF weighting, part-of-speech tagging, and dependency relations. This design allows for a detailed, unsupervised evaluation of each sample without needing to retrain the model.

For dynamic pricing, the environment simulates a sequential data acquisition market in which the platform interacts with numerous data providers. Each data sample is assigned a latent value determined by WPE, and a seller's acceptance decision follows a sigmoid function comparing the offered price to their private reservation price. Each episode comprises 2100 transactions, and the initial budget is drawn from a Gaussian distribution centered at 50% of the total dataset value.

Two Proximal Policy Optimization (PPO) variants are implemented to model the dynamic pricing process. The first, PPO-A, focuses on maximizing immediate profit under a soft budget constraint, aiming for short-term efficiency. In contrast, PPO-B extends the reward function by introducing both a value-density bonus and a total-performance bonus, encouraging more strategic, long-term optimization. In both settings, the policy and value networks are lightweight three-layer MLPs (128–64–64, ReLU) trained over 3000 episodes using the Adam optimizer (learning rate is 3×10^{-4}), a batch size of 256, and a PPO clipping ratio of 0.2. This configuration ensures a balance between training stability and computational efficiency, making the approach suitable for deployment in resource-constrained mobile or edge environments.



(a) LR on DRD (b) LR on MACD (c) LR on Douban

(d) RC on DRD (e) RC on MACD (f) RC on Douban

Fig. 2. Effectiveness of WPE for data valuation across multiple datasets.

6.2. WPE performance on data valuation

To assess WPE's capability as a valuation metric, samples are ranked in descending order of their WPE scores and incrementally added to the model's training set. At each stage, cumulative entropy and corresponding classification accuracy are recorded.

In Figs. 2(a) and 2(d), we observe that the cumulative perplexity entropy rises steadily and then flattens. As shown in Fig. 2(a), when applying logistic regression on the DRD dataset, model accuracy improves rapidly as higher-valued data points are added and then converges at about 0.53, indicating that the marginal contribution of each additional sample diminishes over time. A similar trend is observed in the ridge classification setting on the same dataset, as shown in Fig. 2(d), where accuracy increases from approximately 0.41 to over 0.525, and the slow growth in accuracy can be witnessed after 1200 samples. These results suggest that WPE effectively prioritizes data samples that contribute more substantially to model performance, especially in the initial stages of learning.

We further extend the analysis to the MACD dataset, whose sample scale and distribution differ from DRD. Figs. 2(b) and 2(e) illustrate the results for logistic regression and ridge classification, respectively. Compared to DRD, the accuracy curve starts from a lower baseline but demonstrates a similar, consistent improvement as more samples are added, and again exhibits a diminishing slope, supporting the submodular property of the valuation metric in this context as well.

Additionally, we evaluate our method on the Douban Dataset using a randomly selected subset of 20000 samples. The Douban dataset presents unique characteristics with its Chinese text content and diverse user-generated movie reviews, providing an opportunity to test our approach across different languages and cultural contexts. As demonstrated in Figs. 2(c) and 2(f), accuracy improves from 0.55 to 0.65 for logistic regression and from 0.525 to 0.665 for ridge classification, and both showed a clear diminishing marginal effect in the later stages. The results validate the robustness of our data valuation approach, confirming that our method successfully identifies informative samples across diverse textual domains.

6.3. Stability under textual perturbations

To examine the stability of our valuation method against realistic textual variations that commonly occur in real-world scenarios, we design a comprehensive robustness evaluation framework. We introduce two types of input perturbations that simulate different forms of text corruption: (1) **Keyword Deletion**, where selected informative tokens are randomly removed to simulate information loss scenarios that may occur due to data preprocessing errors or incomplete text transmission; and (2) **Word Order Shuffling**, where token sequences are randomly permuted to simulate syntactic distortion that might result from automatic translation systems or text generation processes.

To systematically evaluate the contribution of different weighting components to overall robustness, we conduct ablation experiments across five configurations: (1) All components enabled, (2) Without TF-IDF weighting, (3) Without POS weighting, (4) Without dependency weighting, and (5) No weighting mechanisms. This design allows us to isolate the impact of each linguistic component on the stability of our valuation method.

The robustness evaluation results, as presented in Table 4, demonstrate the effectiveness of our comprehensive weighting strategy. The configuration using all three weight types consistently yields higher correlations across two types of perturbations, with τ values

Table 4
Kendall's Tau correlation under different perturbations and ablation settings.

Weight configuration	Keyword deletion	Word shuffling
All Components	0.6123	0.4427
Without TF-IDF	0.5432	0.3284
Without POS	0.6132	0.4338
Without Dependency	0.6111	0.4502
No Weighting (None)	0.5071	0.3438

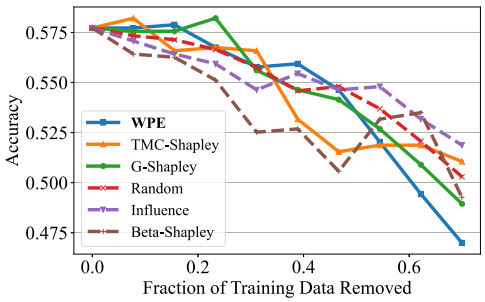


Fig. 3. The impact of sequentially removing high-value data on model prediction effectiveness.

of 0.6123 for keyword deletion and 0.4427 for word shuffling, suggesting that combined linguistic and statistical signals contribute to more stable valuations under noise.

The ablation analysis reveals important insights about component contributions. Removing TF-IDF components results in the most significant drop in Kendall's Tau, particularly under word shuffling (from 0.4427 to 0.3284), indicating that term frequency information plays a central role in preserving ranking consistency when syntactic structure is disrupted. In contrast, removing POS or dependency weights individually leads to a smaller reduction in Kendall's Tau under a single noise scenario, but results in lower stability across different noise conditions, suggesting partial redundancy among linguistic features as well as their complementary roles in maintaining robustness. The configuration with no weighting mechanisms achieves the lowest robustness across both perturbation types, confirming the effectiveness of our proposed weighting strategy in enhancing valuation stability under syntactic and semantic distortions.

6.4. Comparison with valuation baselines

To validate the effectiveness of the proposed WPE as a data valuation method, we compare it with three other baselines in two experiments. The first experiment removes training samples in descending order of their estimated value, and measures the resulting test accuracy. As shown in Fig. 3, removing high-value samples identified by WPE consistently degrades test accuracy, confirming that WPE captures critical training instances. Importantly, WPE exhibits a smoother and more monotonic decline over time, reflecting a stable and trustworthy ranking signal. By contrast, the baseline methods (TMC-Shapley, G-Shapley, Random, Influence Functions and Beta-Shapley) display larger fluctuations and occasional stagnation, indicating their valuation rankings are noisier and less reliable than WPE.

The second experiment removes samples from low to high estimated value (Fig. 4). An effective valuation should tolerate the removal of low-contributing samples with minimal accuracy loss. WPE demonstrates strong robustness in the early stage, whereas G-Shapley and the random baseline deteriorate more quickly. TMC-Shapley initially fluctuates notably, indicating sensitivity to noise. Influence Functions and Beta-Shapley display some stagnation. Once high-value regions are reached, WPE's accuracy declines more sharply, confirming that it effectively places valuable samples in the later stages of the deletion order.

These results demonstrate that WPE offers a reliable and efficient alternative for ranking data by utility. While Shapley-based methods provide useful baselines, their reliance on model-specific dynamics and retraining constraints limits scalability. WPE avoids these pitfalls by leveraging linguistic structure and masked prediction uncertainty, making it a promising approach for fine-grained sample valuation.

6.5. Learning optimal pricing policies

We begin by examining the training dynamics of our two PPO-based pricing strategies: **PPO-A** (maximizes immediate profit) and **PPO-B** (incorporates long-term profit). Fig. 5 depicts the cumulative profit (in thousands) achieved by each strategy over 3000 training episodes. Both strategies exhibit an initial exploration phase, during which the agent learns the dynamics of the action space.

PPO-A shows slow initial convergence, reaching stable performance around episode 500, while PPO-B requires fewer episodes to fully converge. Once converged, PPO-B consistently outperforms PPO-A throughout the training process. PPO-A stabilizes at

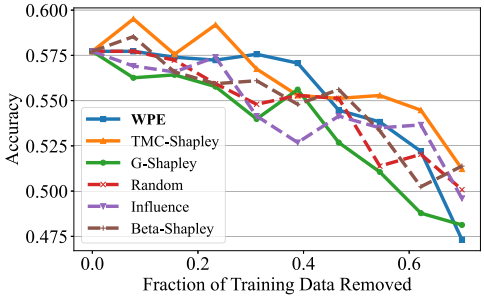


Fig. 4. The impact of sequentially removing low-value data on model prediction effectiveness.

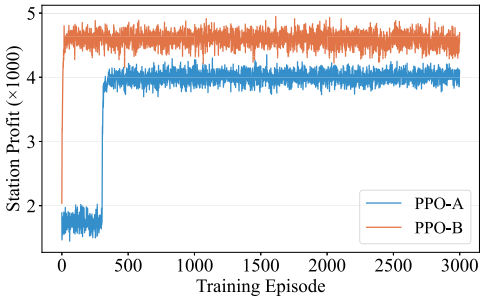


Fig. 5. Training progress comparison between PPO-A and PPO-B.

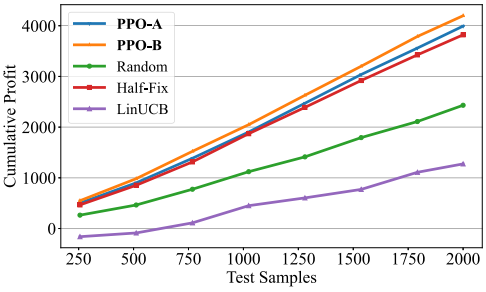


Fig. 6. Cumulative profit of five pricing mechanisms.

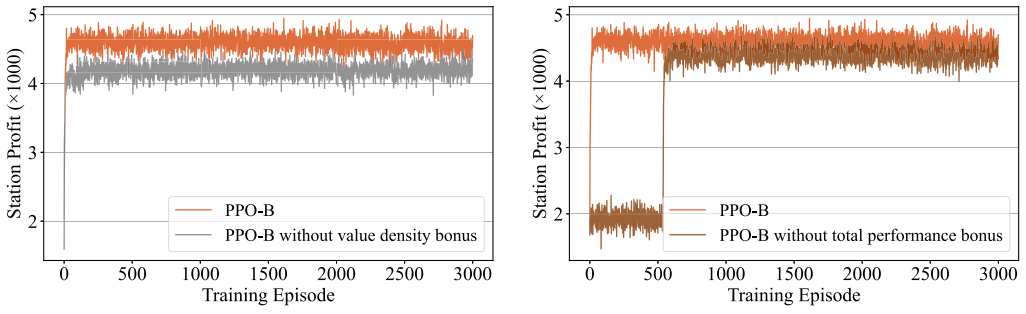
approximately $4.0k per episode, while PPO-B achieves superior performance at around $4.6k per episode. Both strategies exhibit stable learning curves after convergence, reflecting the robustness of PPO's clipped surrogate objective.

The convergence characteristics observed in Fig. 5 are particularly relevant for mobile deployment scenarios. PPO-B's faster convergence makes it suitable for environments requiring rapid policy adaptation, such as when mobile market conditions change frequently. Additionally, PPO-B's superior long-term performance makes it ideal for stable mobile marketplaces that prioritize sustained profitability.

6.6. *Comparing pricing strategies*

We compare our learned PPO policies against three baseline strategies. Fig. 6 compares the cumulative profit trajectories of all five approaches across 2000 test samples.

The Random strategy exhibits the medium overall return due to its lack of informed decision-making, achieving approximately 2400 in final cumulative profit. The Half-Fix baseline achieves modest gains around 3700 but fails to adapt to budget depletion or temporal dynamics. LinUCB improves upon random pricing by leveraging contextual features, including sample value, remaining budget ratio, and normalized time, but performs poorly with only 1300 final profit due to its linear limitations. Both PPO-A and PPO-B substantially outperform all baselines, with PPO-A reaching around 4000 and PPO-B attaining the highest final profit of approximately 4300 by successfully optimizing long-term reward under budget constraints. These results clearly demonstrate the superiority of our reinforcement learning policies in dynamic, budget-limited environments.



(a) Without ValueDensity (b) Without TotalPerformance

Fig. 7. Impact of value density and total performance on PPO-B.

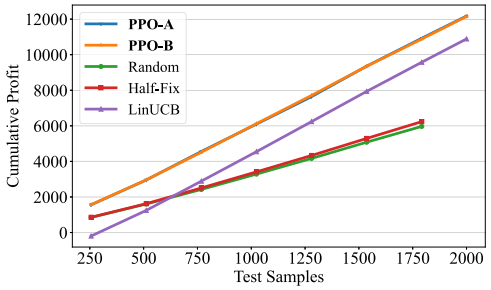


Fig. 8. Cumulative profit of five pricing mechanisms under the zero reservation price ($r_{\text{user}} = 0$) scenario. The Random and Half-fix strategy uses up their budget before reaching 2000 samples.

6.7. Pricing ablation and model adaptability

To explain the performance gap between PPO-A and PPO-B, we conducted an ablation study. As shown in Fig. 7(a), removing the value density bonus reduces PPO-B's converged profit by 0.5 (on the × 1000 scale). In Fig. 7(b), by contrast, removing the total performance bonus not only reduces converged profit, but also slows learning, requiring 500 additional episodes to reach convergence.

The result of Fig. 7 demonstrates the value of incorporating value density bonus and total performance bonus in PPO-B's reward function, which guides the agent toward more strategic budget allocation and higher-utility data acquisitions. PPO-B reflects the complex, long-term optimization strategies that yield high profits.

To test the adaptability of our policies, we simulate an extreme market condition where every user's private reservation price is set to zero ($r_{\text{user}} = 0$), making them highly likely to accept any offer. Figs. 8 and 9 summarize the resulting cumulative profit and average offered price trajectories for all five strategies.

As shown in Fig. 8, both PPO-A and PPO-B rapidly accumulate profit, reaching approximately 12000 cumulative profit. Their nearly identical curves indicate that when reservation prices are zero, optimizing for immediate reward (PPO-A) and long-term reward (PPO-B) yield equivalent policies. This further demonstrates that both PPO strategies remain robust even under extreme market conditions. LinUCB demonstrates adaptive learning capabilities, achieving a final profit of around 11,000 through online adjustments, whereas Random and Half-Fix remain at around 6000 due to their lack of budget awareness and dynamic adjustment capabilities.

Fig. 9 reveals the underlying pricing behaviors, where PPO agents strategically lower their average offers to approximately 0.3–0.4, exploiting the zero reservation price environment. This contrasts sharply with the Random and Half-Fix strategies, which maintain static pricing at around 3.8–4.0 throughout the episode. LinUCB exhibits initial exploration with high prices around 8.0 before rapidly converging to minimal pricing around 0.7 after 500 samples. These results confirm that our PPO-based approach can fully exploit market conditions by dynamically balancing acceptance probability and remaining budget to maximize cumulative profit.

6.8. Computational efficiency and scalability

We report the wall-clock computational cost of (i) WPE valuation per text sample, (ii) PPO training, and (iii) policy inference latency. All measurements are obtained on [AMD Ryzen 9 8945HX with Radeon Graphics] and [NVIDIA GeForce RTX 5070 Ti Laptop GPU], using PyTorch [2.8] with CUDA [12.8]. For GPU timing, we use `torch.cuda.synchronize()` to avoid asynchronous bias.

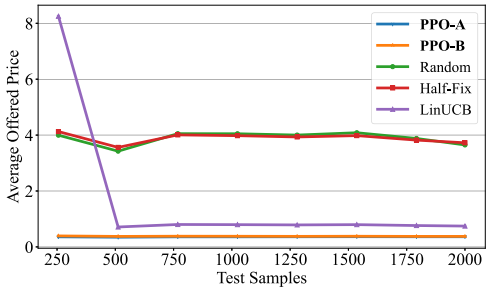


Fig. 9. Average offered price of five pricing mechanisms under the zero reservation price ($r_{\text{user}} = 0$) scenario.

Table 5
Sensitivity analysis of pricing performance under alternative user-behavior assumptions. U/N/E denote uniform/truncated-normal/truncated-exponential reservation-price distributions; Sig/Lin denote sigmoid/linear-threshold acceptance functions; γ_{acc} controls acceptance steepness.

Scenario	PPO-A (Immediate Profit)			PPO-B (Total Profit)		
	Final	Mean(last10)	Std(last10)	Final	Mean(last10)	Std(last10)
U+Sig(γ_{acc}=5)	3992.87	2219.29	1180.77	4197.44	2364.76	1233.41
N+Sig(γ_{acc}=5)	4110.51	2308.09	1213.94	3705.29	2061.36	1084.62
E+Sig(γ_{acc}=5)	7203.79	4054.43	2106.14	7489.12	4204.67	2196.66
U+Lin(γ_{acc}=5)	4025.39	2234.15	1194.14	4251.08	2364.85	1264.15
U+Sig(γ_{acc}=2)	4024.33	2223.91	1181.96	3952.39	2222.80	1147.90
U+Sig(γ_{acc}=10)	3956.26	2225.56	1158.80	3949.89	2259.63	1167.59

6.8.1. WPE valuation time and scalability

Table A.7 reports the per-sample runtime of WPE under the default configuration. The average valuation time is 129.7 ms per text, with MLM inference accounting for over 85% of the total cost, indicating that the main computational bottleneck lies in masked language model forward passes rather than linguistic preprocessing. Table A.8 evaluates scalability under longer sequences and larger mask ratios. As expected, runtime increases approximately linearly with the number of masked tokens, while throughput decreases smoothly, suggesting stable scaling behavior without abrupt performance degradation. In Table A.9 Embedding extraction benefits from batching, where throughput increases more than fivefold when batch size grows from 1 to 64, demonstrating good GPU utilization.

6.8.2. PPO training duration and policy inference time

Table A.10 reports the PPO training time over 3000 episodes. Training completes within approximately 8–9 min, with an average episode runtime of about 0.16–0.17 s. The relatively low runtime is due to the lightweight policy network and vectorized acceptance simulation, which avoids expensive environment rollouts. Table A.11 shows that policy inference requires approximately 0.2–0.3 ms per decision, remaining stable across batch sizes, indicating that the proposed pricing mechanism is suitable for real-time deployment.

6.9. Sensitivity analysis

6.9.1. User-behavior assumptions

To examine the robustness of our pricing conclusions, we conduct a systematic sensitivity analysis by perturbing (a) the reservation-price distribution, (b) the acceptance functional form, and (c) the acceptance steepness parameter. This sensitivity analysis is designed to test whether our conclusions depend on the baseline uniform and sigmoid user model, by explicitly perturbing both the reservation-price distribution and acceptance functional form. We consider six scenarios that span common deviations from the baseline: (i) reservation-price distribution: Uniform (U), truncated Normal (N), and truncated Exponential (E); (ii) acceptance function: Sigmoid (Sig) and Linear-threshold (Lin); (iii) sensitivity parameter: $\gamma_{acc} \in \{2, 5, 10\}$ controlling the steepness of acceptance. We summarize the results using the final cumulative profit and the mean/standard deviation over the last 10 evaluation batches.

Table 5 shows that our core conclusions remain qualitatively stable across alternative assumptions. First, replacing the sigmoid acceptance model with a linear-threshold alternative (U+Lin, γ_{acc}=5) yields very similar profitability levels to the baseline (U+Sig, γ_{acc}=5), suggesting that the effectiveness of learned pricing does not hinge on a specific functional form. Second, perturbing the steepness parameter γ_{acc} from 5 to 2 or 10 produces only moderate changes, indicating that the learned strategies are not overly sensitive to the assumed degree of price responsiveness. Third, changing the reservation-price distribution (N/E) alters the absolute profit level, especially under the truncated-exponential case where acceptance becomes easier on average, but the learned policies remain effective and do not collapse under distribution shifts. Overall, the sensitivity results support the robustness of the proposed pricing framework under plausible deviations from the baseline user model.

6.9.2. Sensitivity to valuation parameters

For evaluating the robustness of the proposed WPE valuation, we conduct several sensitivity analyses with respect to the masking ratio, component ablation, and manually specified linguistic weights. First, we vary the masking ratio used in the entropy estimation. The results, reported in Appendix Table A.4, show that moderate changes in the masking ratio lead to limited variation in ranking consistency and downstream performance remains comparatively stable. Second, we perform component ablation by removing POS, dependency, and TF-IDF weights individually. As shown in Appendix Table A.5, removing any component slightly reduces ranking stability, indicating that each component contributes to the valuation signal. Finally, we conduct a sensitivity analysis by perturbing a subset of representative linguistic categories to assess sensitivity to manually defined parameters. We focus on representative linguistic categories to ensure that the results can be clearly presented and interpreted. The results in Appendix Table A.6 suggest that WPE remains stable under reasonable variations of these weights, supporting the generalizability of the proposed method.

Although the current implementation uses heuristic initial weights, the WPE framework is not restricted to fixed parameters. The weighting scheme can be extended to a learnable formulation in which linguistic weights are treated as tunable hyperparameters or optimized jointly with downstream objectives. This opens a pathway toward adaptive data valuation models that automatically adjust to domain-specific characteristics.

6.9.3. Sensitivity to pricing parameters

We conducted sensitivity analyses on key PPO hyperparameters, including the entropy coefficient β, value-density weight δ, and long-term reward weight λ. As reported in Appendix Tables A.12, A.13, A.14, the proposed pricing policy exhibits stable performance across a wide range of hyperparameter values. In particular, the final profit and budget utilization vary only moderately across different settings, indicating that the effectiveness of the proposed mechanism does not depend on carefully tuned parameters. We also observe that moderate values of δ and λ tend to achieve slightly better profit and utilization, suggesting that both short-term and long-term reward components contribute to stable learning dynamics.

7. Discussion

7.1. User-behavior modeling

We acknowledge that the baseline assumption of independently and uniformly distributed reservation prices abstracts away from more complex user behaviors that may arise in real-world data markets. In practice, users may exhibit correlated valuations, strategic responses to pricing policies, or adaptive behavior through learning over repeated interactions. Explicitly modeling such behaviors would require richer market feedback, longitudinal user data, or game-theoretic formulations, which are beyond the scope of the current study. Instead, we focus on establishing a tractable baseline model and evaluate its robustness by testing alternative reservation-price distributions and acceptance functions in the experimental sensitivity analysis. These results indicate that the proposed pricing framework remains effective under plausible deviations from the uniform assumption. An important avenue for future research lies in incorporating correlated or strategic patterns of user behavior. Building on this, potential extensions could explore context-aware and privacy-aware user models. For instance, adapting reservation prices or acceptance decisions based on temporal or contextual factors, or on data-sensitivity indicators, whenever such signals are available.

From a market-structure perspective, the present framework models the platform as the primary decision maker that posts acquisition prices, while data providers respond according to their reservation prices. This formulation can be interpreted as a simplified Stackelberg-type setting and is commonly adopted in empirical studies of data procurement, where the focus is on pricing dynamics under budget and information constraints rather than on full market equilibrium. We acknowledge that real-world data markets are inherently bilateral and that sellers may exhibit strategic behavior over repeated interactions. In this study, however, our objective is to improve platform-side pricing decisions under uncertainty and streaming arrivals of data. Incorporating richer game-theoretic or mechanism-design models would shift the focus toward equilibrium analysis and market design.

Beyond supporting robustness under alternative user-behavior assumptions, the sensitivity analysis also provides an indirect lens for interpreting policy behavior. Comparing PPO-A and PPO-B across alternative reservation-price distributions suggests that the learned strategies do not merely preserve profitability under distribution shifts, they also adjust their pricing style. In easier-acceptance environments (e.g., truncated-exponential reservation prices), the long-term policy is more likely to exploit lower-price, higher-volume acquisitions with smoother budget usage, while its reward design also helps prevent over-allocation to low-value samples. In less permissive environments, where lower offers are less likely to be accepted, PPO-B may respond by reducing marginal transactions and becoming more selective and conservative. We acknowledge that these behavioral interpretations are inferred from aggregate policy outcomes and reward design, rather than from direct structural identification of user-platform interactions. A promising direction for future work is therefore to augment sensitivity experiments with policy diagnostics, such as average offered price, acceptance rate, accepted-sample value, and budget-consumption trajectories, to make the adaptation mechanism directly observable.

7.2. Robustness to budget enforcement rules

We conduct alternative hard-constraint experiments for PPO-B. Fig. 10 illustrates the training convergence of the three pricing strategies. We observe that PPO-B (Soft) exhibits a relatively smooth and stable improvement in cumulative profit over training episodes. In contrast, PPO-B (Hard) converges to a lower reward level, reflecting the stricter budget feasibility constraint in the hard-constraint mechanism. Fig. 11 shows that the lower convergence plateau of the hard variant is consistent with its more conservative purchasing behavior, as transactions are strictly bounded by real-time budget availability without soft penalty-based flexibility.

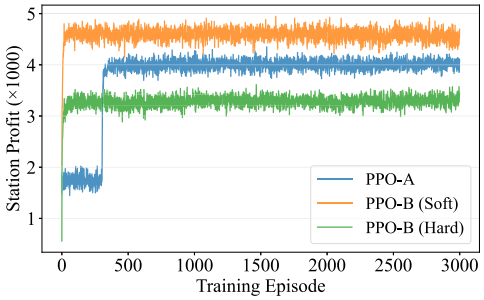


Fig. 10. Training progress comparison between PPO-A and PPO-B with soft constraint and hard constraint.

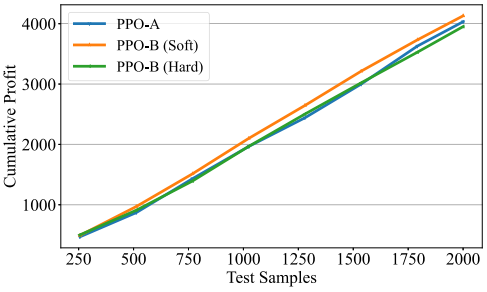


Fig. 11. Cumulative profit of PPO-A and PPO-B with soft constraint and hard constraint.

7.3. Operational example

To illustrate how the proposed framework operates in practice, consider a mobile health review platform that collects user-generated textual feedback to improve a symptom-detection model. As new reviews arrive, the platform first evaluates each text using the proposed information valuation module to estimate its potential contribution to the downstream model. Based on this estimated value and the remaining acquisition budget, the pricing module determines a dynamic offer for the data provider. Users then decide whether to share their data in response to the offered price. The acquired reviews are incorporated into the learning process, improving model performance over time and influencing subsequent valuation and pricing decisions. This closed-loop interaction between valuation, pricing, and learning enables the platform to allocate its data acquisition budget more efficiently while adapting to user responses in a dynamic environment.

7.4. Deployment considerations in mobile environments

The proposed WPE valuation is lightweight compared with model retraining-based valuation approaches. The RL pricing policy is executed only at the platform side, and the inference step requires only a forward pass of a small neural network, making real-time pricing feasible in mobile-driven data collection systems.

In practical deployments, textual data may be evaluated locally or at trusted edge servers, and only valuation scores or transaction outcomes need to be transmitted to the platform. This architecture reduces the need to transmit raw data in certain scenarios and can help mitigate privacy concerns.

Intermittent connectivity is another common characteristic of mobile environments. The proposed framework processes submissions in an asynchronous manner, allowing delayed or temporarily unavailable submissions to be incorporated without modifying the pricing mechanism. This design makes the system naturally compatible with mobile or edge-based data collection scenarios.

7.5. Theoretical and practical implications

This research provides both theoretical and practical implications. From a theoretical perspective, it bridges two previously disconnected domains, which are information theory and data economics. By introducing Weighted Perplexity Entropy, the study provides a formal measure of textual information value grounded in predictive uncertainty, enriching the theoretical foundation of data valuation. Moreover, the reinforcement learning–based pricing mechanism extends economic decision theory into the context of digital assets, modeling data trading as a sequential decision problem under uncertainty.

From a practical perspective, the proposed framework offers an implementable solution for data marketplaces and AI model training pipelines. It allows platforms to determine which data is worth acquiring and how much to pay, optimizing budget use and improving data quality. This is particularly relevant for federated learning, decentralized data sharing, and AI model fine-tuning, where pricing fairness and efficiency are crucial.

8. Conclusion

This paper presents a comprehensive framework addressing the dual challenges of data valuation and pricing in mobile environments through an information-aware approach. Our Weighted Perplexity Entropy (WPE) method effectively captures the semantic richness and informational content of textual data, demonstrating superior robustness and ranking consistency across diverse datasets compared to existing Shapley-based methods. The reinforcement learning-based dynamic pricing mechanism successfully adapts to budget constraints and market conditions, achieving higher cumulative profits than baseline strategies in mobile data trading scenarios. An important extension of the present framework is to enable online adaptation of the valuation module. In such a setting, the valuation function could be updated using realized transaction outcomes, downstream performance improvements, or reinforcement learning signals, forming a joint optimization loop between valuation and pricing. More generally, this could be formulated as a bilevel or end-to-end learning problem, in which valuation serves as a learnable component that adapts to task-specific utility. Exploring such formulations is an important direction for future research.

CRediT authorship contribution statement

Wenze Xiong: Writing – original draft, Writing – review & editing, Data curation, Visualization, Validation, Methodology, Formal analysis. **Yetong Wang:** Writing – review & editing, Software, Visualization. **Wanxin Li:** Writing – review & editing, Supervision, Conceptualization, Methodology, Funding acquisition. **Hao Guo:** Writing – review & editing, Supervision, Funding acquisition. **Jie Zhang:** Writing – review & editing, Supervision. **Haoyu Wang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially supported by the Jiangsu Province Science and Technology Youth Talent Promotion Program (Grant No. JSTJ-2025-144), the Natural Science Basic Research Program of Shaanxi (Grant No. 2025JC-YBMS-688), the Key R&D Programs of Taicang (Grant No. TC2024SF10), and the XJTLU Research Development Fund (Grant No. RDF-22-02-106).

Appendix

See Tables A.1–A.14.

Data availability

Data will be made available on request.

Table A.1
Part-of-Speech (POS) weights used in WPE.

POS Tag	Weight	Description
NOUN	1.2	Nouns representing entities or concepts
VERB	1.2	Verbs describing actions or states
ADJ	1.1	Adjectives expressing attributes or opinions
ADV	1.0	Adverbs modifying verbs or adjectives
PRON	0.8	Pronouns referring to entities
DET	0.5	Determiners such as articles and quantifiers
ADP	0.6	Prepositions and postpositions
CCONJ	0.6	Coordinating conjunctions
PART	0.5	Particles (e.g., “to”, negations)
INTJ	0.3	Interjections or discourse markers
X	0.3	Other or unknown tokens

Table A.2
Dependency role (DP) weights used in WPE.

Dependency	Weight	Description
nsubj	1.2	Nominal subject of a clause
dobj	1.2	Direct object of a verb
ROOT	1.1	Root of the dependency tree (main predicate)
amod	1.0	Adjectival modifier
advmod	1.0	Adverbial modifier
det	0.6	Determiner modifying a noun
mark	0.5	Subordinating conjunction or marker
cc	0.5	Coordinating conjunction
punct	0.0	Punctuation
discourse	0.3	Discourse elements (e.g., fillers)

Table A.3
TF-IDF weighting scheme.

Component	Description
TF-IDF weight	IDF score from corpus
Scaling	Multiplied directly in WPE
Range	Data-driven (not manually fixed)

Table A.4
Sensitivity to the masking ratio in WPE. Weight scale is fixed at 1.0.

Setting	Mask ratio	Weight scale	Kendall τ	Final Acc
Base	0.2	1.0	1.0000	0.5626
R1	0.1	1.0	0.7167	0.5748
R2	0.3	1.0	0.8507	0.5650

Table A.5
Component ablation of WPE under a fixed masking ratio (0.2). "w/o" removes a component by setting its factor to 1.0 (i.e., no contribution).

Setting	POS	DEP	TF-IDF	τ	Final Acc
Full WPE	True	True	True	1.0000	0.5626
w/o POS	False	True	True	0.9510	0.5585
w/o DEP	True	False	True	0.9517	0.5691
w/o TF-IDF	True	True	False	0.6295	0.5585
No Weight	False	False	False	0.5365	0.56

Table A.6
Fine-grained perturbation of manually specified POS/DEP/TF-IDF weights in WPE (mask ratio fixed at 0.2). Scales multiply the corresponding base weights for the targeted category.

Setting	POS scale	DEP scale	TF-IDF scale	Kendall τ	Final Acc
Base	1.0	1.0	1.0	1.0000	0.5626
NOUN↓	0.8	1.0	1.0	0.9436	0.5683
NOUN↑	1.2	1.0	1.0	0.9534	0.5772
VERB↓	0.8	1.0	1.0	0.9534	0.5545
VERB↑	1.2	1.0	1.0	0.9592	0.5772
ADJ↓	0.8	1.0	1.0	0.9630	0.5756
ADJ↑	1.2	1.0	1.0	0.9670	0.5545
nsubj↓	1.0	0.8	1.0	0.9833	0.5569
nsubj↑	1.0	1.2	1.0	0.9849	0.5293
dobj↓	1.0	0.8	1.0	0.9771	0.5634
dobj↑	1.0	1.2	1.0	0.9799	0.5878
ROOT↓	1.0	0.8	1.0	0.9735	0.5650
ROOT↑	1.0	1.2	1.0	0.9760	0.5797
TFIDF↓	1.0	1.0	0.8	0.9821	0.5650
TFIDF↑	1.0	1.0	1.2	0.9878	0.5642

Table A.7
Per-sample WPE valuation time (default setting).

Component	Mean (ms)	Median (ms)	p90 (ms)	Share
Tokenization	1.96	1.72	3.29	1.5%
spaCy tagging	15.39	12.75	26.55	11.9%
MLM inference	112.15	72.03	131.78	86.6%
Total	129.68	94.10	148.04	100%

Table A.8
Scalability of WPE valuation under longer sequences and different mask ratios.

Setting	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.4$
$L = 64$	57.9 ms/17.25 sps	129.7 ms/7.71 sps	185.6 ms/5.39 sps
$L = 128$	98.6 ms/10.14 sps	192.3 ms/5.20 sps	389.1 ms/2.57 sps
$L = 256$	164.6 ms/6.07 sps	328.1 ms/3.05 sps	596.5 ms/1.68 sps

Table A.9
BERT embedding extraction throughput under different batch sizes.

Batch size	Throughput (samples/s)	Batch latency (ms)
1	132.2	7.56
16	581.0	27.45
64	690.7	91.31
256	712.0	338.68

Table A.10
PPO training time (3000 episodes).

Model	Total time (s)	Mean time/episode (s)	p90 (s)
PPO-A	479.23	0.160	0.201
PPO-B	518.40	0.173	0.205

Table A.11
Policy inference latency (per forward pass, sim_T=2048).

Model	bs=1	bs=16	bs=64	bs=256
PPO-A	0.267 ms	0.258 ms	0.232 ms	0.233 ms
PPO-B	0.217 ms	0.207 ms	0.231 ms	0.235 ms

Table A.12
Sensitivity analysis of entropy coefficient β. Values are reported as mean $\pm$ standard deviation across random seeds.

β	Final profit	Budget utilization	Acceptance rate	Mean price ratio
0.000	4005.39 ± 59.40	0.518 ± 0.046	0.353 ± 0.022	0.477 ± 0.043
0.005	4008.67 ± 21.67	0.494 ± 0.014	0.339 ± 0.009	0.444 ± 0.012
0.010	3963.66 ± 52.03	0.528 ± 0.009	0.358 ± 0.010	0.464 ± 0.019
0.050	3942.06 ± 59.48	0.508 ± 0.017	0.334 ± 0.007	0.401 ± 0.011

Table A.13
Sensitivity analysis of value-density weight δ. Values are reported as mean $\pm$ standard deviation across random seeds.

δ	Final profit	Budget utilization	Acceptance rate	Mean price ratio
0.000	3910.80 ± 66.89	0.486 ± 0.066	0.332 ± 0.043	0.414 ± 0.075
0.050	3844.57 ± 176.76	0.409 ± 0.109	0.294 ± 0.056	0.352 ± 0.099
0.150	3963.66 ± 52.03	0.528 ± 0.009	0.358 ± 0.010	0.464 ± 0.019
0.300	3898.57 ± 118.42	0.467 ± 0.104	0.325 ± 0.052	0.400 ± 0.098

Table A.14
Sensitivity analysis of long-term reward weight λ. Values are reported as mean ± standard deviation across random seeds.

λ	Final profit	Budget utilization	Acceptance rate	Mean price ratio
0.000	3970.25 ± 76.21	0.545 ± 0.012	0.364 ± 0.004	0.468 ± 0.004
0.100	3981.04 ± 61.93	0.520 ± 0.049	0.350 ± 0.025	0.446 ± 0.027
0.300	3963.66 ± 52.03	0.528 ± 0.009	0.358 ± 0.010	0.464 ± 0.019
0.500	3969.36 ± 59.32	0.534 ± 0.022	0.355 ± 0.008	0.449 ± 0.017

References

Agarwal, A., Dahleh, M., Horel, T., & Rui, M. (2024). Towards data auctions with externalities. *Games and Economic Behavior*, *148*, 323–356.

Bauer-Hänsel, I., Liu, Q., Tessone, C. J., & Schwabe, G. (2024). Designing a blockchain-based data market and pricing data to optimize data trading and welfare. *International Journal of Electronic Commerce*, *28*(1), 3–30.

Bernardo, B. M. V., São Mamede, H., Barroso, J. M. P., & dos Santos, V. M. P. D. (2024). Data governance & quality management—Innovation and breakthroughs across different fields. *Journal of Innovation & Knowledge*, *9*(4), Article 100598.

Chen, W., Hu, Y., Sui, R., Guan, Z., & Liu, Y. (2026). Competitive e-commerce platforms' data provision and pricing strategies with different attribution behaviors of users. *Expert Systems with Applications*, *298*, Article 129466.

Cong, Z., Luo, X., Pei, J., Zhu, F., & Zhang, Y. (2022). Data pricing in machine learning pipelines. *Knowledge and Information Systems*, *64*(6), 1417–1455.

Douch, S., Abid, M. R., Zine-Dine, K., Bouzidi, D., & Benhaddou, D. (2022). Edge computing technology enablers: A systematic lecture study. *IEEE Access*, *10*, 69264–69302.

Duan, J., Tian, L., Mao, J., & Li, J. (2023). Optimal social welfare: A many-to-many data transaction mechanism based on double auctions. *Digital Communications and Networks*, *9*(5), 1230–1241.

Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A review of data valuation approaches and building and scoring a data valuation model. *Harvard Data Science Review*, *5*(1).

Ghafari, R., & Mansouri, N. (2025). Reinforcement learning-based solution for resource management in fog computing: A comprehensive survey. *Expert Systems with Applications*, *276*, Article 127214.

Ghorbani, A., Kim, M., & Zou, J. (2020). A distributional framework for data valuation. In *International conference on machine learning* (pp. 3535–3544). PMLR.

Ghorbani, A., & Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning* (pp. 2242–2251). PMLR.

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 international conference on digital health* (pp. 121–125).

Hao, J., Deng, Z., & Li, J. (2023). The evolution of data pricing: From economics to computational intelligence. *Heliyon*, *9*(9).

Hao, L., Jin, J., & Xu, Y. (2022). Laxity differentiated pricing and deadline differentiated threshold scheduling for a public electric vehicle charging station. *IEEE Transactions on Industrial Informatics*, *18*(9), 6192–6202.

Heideman, T., Kumara, I., Van Den Heuvel, W. J., & Tamburri, D. A. (2024). Smart contracts as data quality consensus enforcers in data markets. In *International symposium on business modeling and software design* (pp. 130–148). Springer.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Inegbedion, H., Asaleye, A., & Obadiaru, E. (2023). Competitive behaviour of major GSM firms' internet data pricing in Nigeria: A game theoretic model approach. *Heliyon*, *9*(1).

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., Spanos, C. J., & Song, D. (2019). Efficient task-specific data valuation for nearest neighbor algorithms. arXiv preprint arXiv:1908.08619.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. J. (2019). Towards efficient data valuation based on the shapley value. In *The 22nd international conference on artificial intelligence and statistics* (pp. 1167–1176). PMLR.

Jiang, X., Lin, J., Wang, C., & Zhou, L. (2024). How far is reality from vision: An online data-driven method for brand image assessment and maintenance. *Information Processing & Management*, *61*(5), Article 103769.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kavoosi, A., Tavakkoli-Moghaddam, R., Sajedi, H., Tajik, N., & Tafakkori, K. (2025). Dynamic pricing and inventory control of perishable products by a deep reinforcement learning algorithm. *Expert Systems with Applications*, *291*, Article 128570.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885–1894). PMLR.

Kromidha, E. (2023). Identity mediation strategies for digital inclusion in entrepreneurial finance. *International Journal of Information Management*, *72*, Article 102658.

Kwon, Y., & Zou, J. (2021). Beta shapley: a unified and noise-reduced data valuation framework for machine learning. arXiv preprint arXiv:2110.14049.

Li, H., & Duan, L. (2025). Competitive multi-armed bandit games for resource sharing. *IEEE Transactions on Mobile Computing*, *24*(9), 8393–8404.

Li, X., Yao, J., Liu, X., & Guan, H. (2017). A first look at information entropy-based data pricing. In *2017 IEEE 37th international conference on distributed computing systems* (pp. 2053–2060). IEEE.

Lin, J., Huang, Z., & Tang, Y. (2025). Pricing for data assets based on data quality, quantity and utility on the perspective of consumer heterogeneity. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, S., Wang, J., Wang, R., Zhang, Y., Song, Y., & Xing, L. (2024). Data-driven dynamic pricing and inventory management of an omni-channel retailer in an uncertain demand environment. *Expert Systems with Applications*, *244*, Article 122948.

Liu, J., Yi, B., Zhang, H., Shen, X., Song, L., Lei, Y., & Zheng, H. (2026). Modeling semantic representation with LLM-enhanced for knowledge-aware recommendation. *Information Processing & Management*, *63*(2, Part A), Article 104387.

Lu, Y., Wang, J., Liu, L., & Yang, H. (2024). Get by how much you pay: A novel data pricing scheme for data trading. *Information Processing & Management*, *61*(6), Article 103849.

Malieckal, M., Gurtoo, A., & Majumdar, R. (2024). Data pricing for data exchange: Technology and AI. In *Proceedings of the 2024 7th artificial intelligence and cloud computing conference* (pp. 392–401).

Mehta, S., Dawande, M., Janakiraman, G., & Mookerjee, V. (2021). How to sell a data set? Pricing policies for data monetization. *Information Systems Research*, *32*(4), 1281–1297.

Mehta, S., Dawande, M., Janakiraman, G., & Mookerjee, V. (2022). An approximation scheme for data monetization. *Production and Operations Management*, *31*(6), 2412–2428.

Miao, X., Gao, Y., Chen, L., Peng, H., Yin, J., & Li, Q. (2020). Towards query pricing on incomplete data. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 4024–4036.

Qiao, W., Huang, M., Gao, Z., & Wang, X. (2024). Distributed dynamic pricing of multiple perishable products using multi-agent reinforcement learning. *Expert Systems with Applications*, *237*, Article 121252.

Schopf, T., Braun, D., & Matthes, F. (2023). Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th international conference on natural language processing and information retrieval* (pp. 6–15). New York, NY, USA: Association for Computing Machinery.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Shen, Y., Guo, B., Shen, Y., Duan, X., Dong, X., Zhang, H., Zhang, C., & Jiang, Y. (2022). Personal big data pricing method based on differential privacy. *Computers & Security*, *113*, Article 102529.

Sim, R. H. L., Xu, X., & Low, B. K. H. (2022). Data valuation in machine learning:“ ingredients", strategies, and open challenges. In *IJCAI* (pp. 5607–5614).

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*: *Vol. 1, No. 1*, MIT press Cambridge.

Veldkamp, L. (2023). Valuing data as an asset. *Review of Finance*, *27*(5), 1545–1562.

Wang, Y., Zhang, B., Ma, J., & Jin, Q. (2022). Earning while learning: An adversarial multi-armed bandit based real-time bidding scheme in deregulated electricity market. *IEEE Transactions on Network Science and Engineering*, *9*(6), 3991–4000.

Wu, C., Bi, W., & Liu, H. (2023). Proximal policy optimization algorithm for dynamic pricing with online reviews. *Expert Systems with Applications*, *213*, Article 119191.

Xiao, Z., He, D., & Du, J. (2020). A Stackelberg game pricing through balancing trilateral profits in big data market. *IEEE Internet of Things Journal*, *8*(16), 12658–12668.

Xing, A., & Wang, H. (2024). Pricing and sample set strategies of data providers under quality information asymmetry. *Journal of the Operational Research Society*, *75*(2), 278–296.

Xiong, W., & Xiong, L. (2021). Anti-collusion data auction mechanism based on smart contract. *Information Sciences*, *555*, 386–409.

Xu, A., Zheng, Z., Li, Q., Wu, F., & Chen, G. (2024). VAP: Online data valuation and pricing for machine learning models in mobile health. *IEEE Transactions on Mobile Computing*, *23*(5), 5966–5983.

Yang, M., Feng, H., Wang, X., Wu, X., Wang, Y., & Ren, C. (2024). Data pricing with privacy loss compensation for cyber-physical systems: A Stackelberg game based approach. *Internet Technology Letters*, *7*(3), Article e443.

Yang, J., Zhao, C., & Xing, C. (2019). Big data market optimization pricing model based on data quality. *Complexity*, *2019*(1), Article 5964068.

Zhou, W., An, L., Han, R., & Li, G. (2025). Classification and severity assessment of disaster losses based on multi-modal information in social media. *Information Processing & Management*, *62*(5), Article 104179.