Sell Data to AI Algorithms Without Revealing It: Secure Data Valuation and Sharing via Homomorphic Encryption

Michael Yang
University of Texas at Dallas

Ruijiang Gao
University of Texas at Dallas

Zhiqiang (Eric) Zheng
University of Texas at Dallas

Abstract

Traditional data-sharing practices require data owners to reveal the data to buyers to determine its value before they can negotiate a fair price, creating legal exposure, privacy risk, and asymmetric information that discourages exchange. We propose a Homomorphic Encryption (HE) framework that enables prospective buyers to quantitatively assess a dataset's utility for an AI algorithm while the data remains fully encrypted end-to-end. Our approach tackles the last-mile problem in building secure AI data marketplaces. We design a lightweight data utility evaluation method using HE protocols that allow buyers to score different data samples without actually having to obtain the raw data. The proposed method can work with popular gradient-based data valuation methods and can scale to Large Language Models (LLMs). By allowing organizations to determine the value of their data, without disclosing the data itself before the transaction, our work provides a practical path toward secure data monetization.

1 Introduction

Artificial intelligence is fundamentally driven by data, and advanced machine learning models require large, high-quality datasets to learn effectively (Zhang and Beltrán, 2020). For example, Llama 3 was pretrained on over 15 trillion tokens of publicly available text[1]. Recognizing the importance of domain-specific information, AI companies increasingly offer fine-tuning services that allow users to adapt foundation models to their proprietary datasets[2]. Consequently, firms holding valuable private data are exploring partnerships with AI developers to enhance products or inform internal decision making. For instance, Reddit signed formal licensing agreements with OpenAI to provide authorized access to Reddit content for model training (OpenAI and Reddit, 2024). Yet, in the absence of robust AI regulation, data exchange often remains ad hoc and opaque. Leading firms such as OpenAI initially relied on large-scale scraping of copyrighted online content without consent (Grynbaum and Mac, 2023), sparking lawsuits and public criticism. A prominent case is Getty Images' lawsuit against Stability AI, alleging unauthorized use of 12 million Getty photographs in training a generative art model (Brittain, 2023).

This chaotic environment persists because the industry has yet to solve a fundamental challenge: how to quantify the value of data in algorithmic predictions and decisions (Zhang et al., 2024).

[1] https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] https://platform.openai.com/docs/guides/fine-tuning/create-a-fine-tuned-model

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Lock-LLM Workshop at NeurIPS 2025.

Evaluating a dataset's impact typically requires its use in model training or testing. However, this process inherently exposes the data to the model owner, creating a significant risk for the data owner. If the model owner were to use the data without payment, holding them accountable is challenging, partly due to the lack of high-fidelity tools for detecting data misuse or theft (Grynbaum and Mac, 2023). From the buyer's perspective, there is also the risk of sellers exaggerating a dataset's quality – without testing it directly, the buyer might overpay based on inflated claims. One potential workaround is to introduce a trusted third party to facilitate the process. For instance, a neutral escrow service or a platform could hold the sensitive dataset and run the valuation on behalf of the buyer, then report the results. In practice, however, this approach still hinges on trust and carries significant drawbacks. Spiekermann (2019) have noted that a "lack of trust and security" in intermediaries or platforms is a major obstacle that keeps companies from participating in data sharing. The data owner must trust that the third-party evaluator has the ability to assess the value of her data and will not leak or misuse the data.

In light of these challenges, we propose an *encrypted data-valuation* framework that integrates privacy-preserving encrypted computation with modern data valuation methods, allowing buyers to compute utility scores on encrypted ciphertexts while sellers prove utility integrity through randomized, auditable challenge tests—all without exposing the underlying raw data possessed by the data owner. The proposed framework is shown in Figure 1, which we will discuss in detail in Section 3. The method employs the *gradient-based influence function* to quantify the marginal utility of a candidate record while keeping both the seller's data and the buyer's evaluation set encrypted throughout the protocol. As we shall see in Section 4, the proposed framework can recover the utility of data points with near-perfect accuracy with minor computing overhead, a property especially critical for large foundational models such as Large Language Models.

Our work makes the following contributions. First, we design a novel *encrypted data-valuation* framework that, for the first time, integrates modern data valuation techniques (specifically, gradient-based influence functions) with encrypted computation. Second, our framework resolves the dilemma in data commerce where buyers risk purchasing low-quality data while sellers risk data theft during evaluation. By enabling secure, direct testing on encrypted data, our work protects the intellectual property of the data owner while preventing the buyer from overpaying for a dataset of exaggerated value. We also theoretically demonstrate that the proposed method, under the correct hyper-parameter specifications, can recover the true data utility in a privacy-preserving manner. Finally, our empirical evaluation shows that the proposed method achieves near-perfect utility computation while preserving privacy across a diverse range of models.

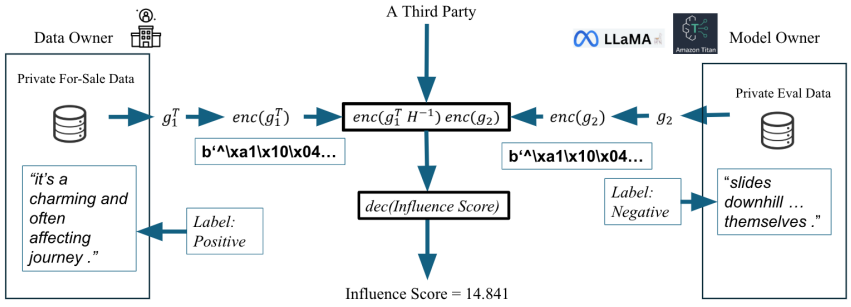

Figure 1: Proposed Secure Data Marketplace with Homomorphic Encryption.

2 Related Work

Data Markets The increasing demand for high-quality data has led to the emergence of data marketplaces(Zhang et al., 2024). Early research explored data market paradigms, including direct sales, query-based access, and trained-model exchanges (e.g., Balazinska et al., 2011; Koutris et al., 2015; Agarwal et al., 2019). However, a persistent challenge across all these designs is the "value-privacy dilemma": buyers cannot assess a dataset's true value without accessing it, yet sellers cannot provide access without risking their intellectual property (Fernandez et al., 2020; Spiekermann, 2019). Prior work has focused on market design and pricing mechanisms (Wang et al., 2021; Roughgarden, 2010), but has not provided a technical solution to this core trust problem. We address this

fundamental issue by providing a practical and secure protocol for pre-transaction utility evaluation to foster the trust required for these data markets to function effectively. As a result, our proposed framework attains near-perfect performance.

Data Attribution Methods Data attribution concerns how to evaluate the value of individual data points; therefore, it is different from other explanation methods, such as feature attribution methods like LIME or SHAP (Lundberg and Lee, 2017; Ribeiro et al., 2016). Foundational approaches like Data Shapley Value (Ghorbani and Zou, 2019) provide a theoretically principled framework, but are computationally prohibitive for large models as they require extensive retraining. Influence functions approximate a data point's value by measuring its impact on the model's loss or parameters using gradients, avoiding the need for retraining (Hampel, 1974; Koh and Liang, 2017). However, computing the influence function at scale can be prohibitively slow and memory-intensive (Hammoudeh and Lowd, 2024). Recent efforts have adapted influence function techniques to large models. TRAK (Park et al., 2023) forgoes computing a full Hessian and instead uses a simplified "influence-style" gradient similarity metric to trace influential data for vision and language models. DataInf (Kwon et al., 2023) leverages LoRA (Hu et al., 2021) to compute influence scores for LLM fine-tuning data. Our work is the first to bridge this gap by integrating scalable, gradient-based attribution methods with a formal cryptographic protocol, enabling secure valuation without revealing sensitive information.

Privacy-Aware Computing To solve the value-privacy dilemma, a secure computation protocol is required. Differential Privacy (DP) provides a mathematically rigorous definition of privacy for statistical data analysis (Dwork, 2006; Dwork and Roth, 2014). It ensures privacy by limiting the impact of any single participant's data, assuming honest participation. Liu et al. (2021) provided an end-to-end model marketplace that formalizes the interactions between data owners, a broker, and model buyers using Shapley value for fair data valuation and differential privacy to preserve privacy. Zheng et al. (2022) proposes a mechanism for federated learning that ensures privacy preservation through local differential privacy while motivating data owners to contribute high-quality model updates. We leverage Homomorphic Encryption to enable computation directly on encrypted data (Furukawa et al., 2017; Subramanyan et al., 2017; Gramaglia et al., 2017; Yao, 1986). Specifically, we use the CKKS scheme because it uniquely supports approximate arithmetic on the real-valued vectors inherent in gradients (Cheon et al., 2017a; Chillotti et al., 2020). While HE has been applied to model training and inference (Lee et al., 2022; Graepel et al., 2012), our key distinction is the novel application of CKKS to a specialized protocol for secure influence function calculation, creating the first practical bridge between scalable data valuation and strong cryptographic privacy.

3 Secure Data Market

3.1 Preliminary on Influence Function

Data valuation is used to quantify the contribution of individual data points to a model's behavior. A principled way to measure this contribution is to ask: "How would the model's performance on the test set change if an additional training example was added?". Influence functions offer a computationally efficient approximation and have become a popular way to quantify data utilities (Park et al., 2023; Kwon et al., 2023; Choe et al., 2024). Intuitively, it approximates the effect of upweighting a new data point z_i on the loss of a test point z_{test} without retraining the model. The first-order influence is given by the following formula:

$$I(z_i, z_{\text{test}}) \approx -\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z_i, \hat{\theta}) \tag{1}$$

where $\nabla_\theta L(z_i, \hat{\theta})$ and $\nabla_\theta L(z_{\text{test}}, \hat{\theta})$ are the gradients of the loss with respect to the model parameters $\hat{\theta}$ for the training and test points, respectively. $H_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ denotes the Hessian of the total training loss, which captures the curvature of the loss landscape at the converged model parameters $\hat{\theta}$. n indexes the number of training data samples. $H_{\hat{\theta}}^{-1}$ represents the inverse of the Hessian. Equation (1) can be derived by taking the derivative of the upweighted loss with respect to the weight (Xia and Henao, 2023).

The resulting scalar score, $I(z_i, z_{\text{test}})$, has a direct and intuitive interpretation for data valuation. A large negative influence score implies that the training point z_i was highly beneficial, as including it

in the training set helps reduce the loss on the test point z_{test}. Conversely, a large positive score suggests that z_i was harmful—perhaps an outlier or mislabeled example—as its presence increases the test loss. By ranking training data based on these influence scores, a model owner can quantitatively assess the value of each data point for a given task[3].

Applying this formula to large-scale models like LLMs presents two major scalability challenges. First, computing and inverting the Hessian matrix is intractable, as its size is quadratic in the number of model parameters, which can be in the billions. Second, computing and storing the gradient for every single example in a massive training dataset introduces prohibitive memory and computational costs. To overcome these issues, recent work has proposed random projection (Hu et al., 2021) or low-rank approximations (Choe et al., 2024) to overcome the computational hurdle. For example, in Hu et al. (2021), the influence function is approximated as $\text{Influence}(z_i, z_{\text{test}}; P) = (Pg_{\text{test}})^\top (PH_{\hat{\theta}}P^\top)^{-1}(Pg_i)$, where $P \in \mathbb{R}^{k\times d}$ is a projection matrix with the projected dimension k being far smaller than the original model dimension d. These influence function approximations enable data valuation for large-scale models with billions of parameters. In our experiments, we adopt this recipe by using LoRA to define the projection and Kronecker-factored Approximate Curvature(K-FAC) to implement the inverse preconditioner.

3.2 Secure Data Marketplace Design

We aim to establish a secure data marketplace in which buyers can assess the utility of sellers' data without disclosing their respective private assets. Data value is measured through influence functions that capture the marginal contribution of each sample to model predictions. As illustrated in Figure 1, the setting involves three parties under the standard semi-honest ("honest-but-curious") model: the *data owner* ($\mathcal{S}$), who holds a private dataset intended for sale; the *model owner* ($\mathcal{B}$), who possesses a proprietary model f_θ and a private evaluation set D_{eval} to assess the seller's data; and the *broker* ($\mathcal{T}$), an untrusted computation service that performs encrypted operations but may otherwise attempt to leak information.

Let f_θ be a model with parameters θ pre-trained by $\mathcal{B}$. Consider a candidate for-sale training sample $z_i = (x_i, y_i)$ held by a data seller, and a private evaluation example $z_{\text{eval}} = (x_{\text{eval}}, y_{\text{eval}})$ held by a potential buyer. The influence of z_i on the buyer's evaluation loss $\ell(z_{\text{eval}}; \theta)$ is defined as the first-order change in that loss if z_i were up-weighted as:

Theorem 3.1 (First-Order Influence on Loss (Koh and Liang, 2017))**.** *Suppose θ is a local minimiser of the empirical risk on D and let $H_\theta = \nabla^2_\theta L(\theta)$ be the Hessian of the total training loss at θ (assumed nonsingular and optionally regularized). Adding an infinitesimal weight ε on sample z_i changes the evaluation loss by:*

$$\text{Inf}\big(z_i \to z_{eval}\big) := \frac{d}{d\varepsilon}\,\ell\big(z_{eval};\theta_\varepsilon\big)\Big|_{\varepsilon=0} \;=\; -\,\nabla_\theta \ell\big(z_{eval};\theta\big)^\top\, H_\theta^{-1}\,\nabla_\theta \ell\big(z_i;\theta\big). \tag{2}$$

With the low-rank projection, the projected gradients are denoted by $\tilde{\mathbf{g}} = P^\top \mathbf{g}$. The influence formula can then be tractably approximated in this low-dimensional space with $\text{Inf}\big(z_i \to z_{\text{eval}}\big) \approx \tilde{\mathbf{g}}_i^\top \tilde{H}_\theta^{-1} \tilde{\mathbf{g}}_{\text{eval}}$. To compute this score, the data seller must provide their projected gradient $\tilde{\mathbf{g}}_i$, and the model owner (buyer) must provide their projected evaluation gradient $\tilde{\mathbf{g}}_{\text{eval}}$ and projected inverse Hessian $\tilde{H}_\theta^{-1}$. Sharing these components in plaintext would violate the privacy of both parties. Gradients often contain sensitive information about the underlying data (which is why companies like OpenAI restrict gradient access for safety reasons). Specifically: $\tilde{\mathbf{g}}_i$ reveals information about the seller's private data, while $\tilde{\mathbf{g}}_{\text{eval}}$ and $\tilde{H}_\theta^{-1}$ reveal information about the buyer's private evaluation query and proprietary model. Our goal is to design a cryptographic protocol that allows for the secure multi-party computation of the influence score. We achieve it through the following Lemma.

Lemma 3.2 (CKKS approximate homomorphism)**.** *Let $(\mathsf{pk}, \mathsf{sk})$ be keys for CKKS and $m_1, m_2 \in \mathbb{R}^d$. Then there exist relinearization and rescaling procedures such that*

$$D(\mathsf{sk},\, E_{\mathsf{pk}}(m_1) \oplus E_{\mathsf{pk}}(m_2)) = m_1 + m_2 + \varepsilon_{\text{add}}, \quad D(\mathsf{sk},\, E_{\mathsf{pk}}(m_1) \otimes E_{\mathsf{pk}}(m_2)) = m_1 \odot m_2 + \varepsilon_{\text{mul}},$$

[3]We note that while we only discuss the contributions of individual data points in this paper, partly due to space constraints. The proposed framework can be easily extended to the best dataset selection among datasets problem with subset-based data valuation methods (Hu et al., 2024).

where $\odot$ *is element-wise product and the errors satisfy* $\|\varepsilon_{\text{add}}\|, \|\varepsilon_{\text{mul}}\| \leq \delta(\textit{scale}, \textit{modulus chain}, \textit{depth})$.

Lemma 3.2 suggests that HE can perform the computation using an encrypted ciphertext and return the exact solutions of the computation using plaintext under additive and multiplicative operations, which, interestingly, are the only operations needed for influence function computation. Inspired by this observation, we introduce the secure data market design. We choose the CKKS (Cheon-Kim-Kim-Song) scheme (Cheon et al., 2017b) as the cryptographic foundation of our framework. The specific parameters chosen for our implementation, which control the trade-off between precision, security, and performance, are detailed in the Appendix and omitted due to space constraints. The protocol is illustrated in Figure 1 and Algorithm 1.

Algorithm 1 TIP: Trustworthy Influence Protocol

Require: From $\mathcal{S}$: Projected gradients $\{\tilde{\mathbf{g}}_i\}_{i=1}^N$
Require: From $\mathcal{B}$: Projected evaluation vector $\tilde{\mathbf{v}}_{\text{eval}} = \tilde{H}_\theta^{-1}\tilde{\mathbf{g}}_{\text{eval}}$
Require: Key ownership: $\mathcal{B}$ runs SETUP to obtain $(\mathsf{pk}_B, \mathsf{sk}_B)$; publishes pk_B; keeps sk_B private
Ensure: Buyer-only plaintext scores $\{s_i\}_{i=1}^N$; all intermediate values remain encrypted under pk_B

$(\mathsf{pk}_B, \mathsf{sk}_B) \leftarrow$ HE.SETUP ▷ *Action by Model Owner* $\mathcal{B}$, *once per session*
$\mathcal{B}$ publishes pk_B, shares with $\mathcal{S}$
procedure SECUREINFLUENCECOMPUTATION($\{\tilde{\mathbf{g}}_i\}_{i=1}^N, \tilde{\mathbf{v}}_{\text{eval}}, \mathsf{pk}_B$)
  $\mathsf{ct}_{\text{eval}} \leftarrow \mathsf{Encrypt}_{\mathsf{pk}_B}(\tilde{\mathbf{v}}_{\text{eval}})$ ▷ *Action by Model Owner* $\mathcal{B}$
  $\mathcal{B}$ sends $\mathsf{ct}_{\text{eval}}$ to Broker $\mathcal{T}$
  for each $i = 1, \ldots, N$ **do** ▷ *Action by Data Owner* $\mathcal{S}$ *and Broker* $\mathcal{T}$
    $\mathsf{ct}_i \leftarrow \mathsf{Encrypt}_{\mathsf{pk}_B}(\tilde{\mathbf{g}}_i)$ ▷ $\mathcal{S}$ *encrypts under* pk_B *(never learns* sk_B*)*
    $\mathcal{S}$ sends ct_i to Broker $\mathcal{T}$
    $\mathsf{ct}_{\text{prod}} \leftarrow \mathsf{ct}_i \odot \mathsf{ct}_{\text{eval}}$ ▷ Homomorphic element-wise product (under pk_B)
    $\mathsf{ct}_{\text{inf},i} \leftarrow \mathsf{RotateAndSum}(\mathsf{ct}_{\text{prod}})$ ▷ Homomorphic summation over slots
    $\mathcal{T}$ sends $\mathsf{ct}_{\text{inf},i}$ to $\mathcal{B}$ ▷ Encrypted result stays under pk_B
  end for
  $s_i \leftarrow \mathsf{Decrypt}_{\mathsf{sk}_B}(\mathsf{ct}_{\text{inf},i})$ for all $i = 1, \ldots, N$ ▷ *Only* $\mathcal{B}$ *can decrypt*
  return $\{s_i\}_{i=1}^N$
end procedure

Next, we verify the computational correctness of our protocol design. Theorem 3.3 implies that our framework enables the accurate computation of data utility, matching the results of traditional methods that assume unencrypted data, but within an encrypted domain. We omit the proof due to space constraints.

Theorem 3.3 (Encrypted Influence Approximation)**.** *Let* $s := \tilde{\mathbf{g}}_i^\top \tilde{H}_\theta^{-1}\tilde{\mathbf{g}}_{\text{eval}}$ *denote the true influence score in plaintext, and let* $\hat{s}$ *be the value obtained by decrypting the final ciphertext* $\mathsf{ct}_{\textit{inf}}$ *from the protocol. Under the condition of Lemma 3.2,* $\hat{s} = s + \Delta$*, where the error* $|\Delta|$ *is negligible and can be made arbitrarily small by adjusting the CKKS precision parameters.*

Zero-knowledge proofs (ZKPs). We note that the proposed Trustworthy Influence Protocol (TIP) can be further strengthened by Zero-knowledge proofs (ZKP) and an interactive proof system (IPS) to enhance trust in a secure AI data marketplace by letting buyers verify claims on whether the dataset has a utility above a certain threshold—without seeing the data—and letting sellers prove value before disclosure or payment. Our protocol extends naturally to an IPS: the seller acts as the prover, the buyer as the verifier, and a notary computes a public homomorphic inner-product oracle on encrypted inputs; the buyer then learns only the resulting utility score. Adding ZKP on top of this flow provides a crisp privacy guarantee: the verifier can be convinced the score is correct while learning nothing beyond that single number (no gradients, no evaluation set, no model details). In practice, one-shot proofs give a lightweight "proof-by-encryption", and optional multi-round challenges amplify confidence without additional disclosure. These additions slot into our current design without changing the core computation. We omit the details due to space constraints.

4 Experiment

We evaluate our encrypted data evaluation framework on three representative tasks of increasing scale and complexity: (1) image classification on MNIST (Deng, 2012), (2) sentiment analysis with BERT on SST-2 (Devlin et al., 2019), and (3) next-token prediction with GPT-2 on WikiText-2 (Radford et al., 2019). For each task, we report: **Model performance** (e.g., classification accuracy or perplexity) under normal, unencrypted training. **Fidelity of encrypted influence scores** is measured both by absolute error against the plaintext baseline and by Pearson correlation. **Runtime overhead** introduced by our homomorphic steps, specifically the time to encrypt and decrypt the logged gradients, and the time to compute the influence function under CKKS.

We demonstrated our FHE-based influence computation in Table 1. On the MNIST task (60000 training points), we recover a Pearson correlation of 1.00 and a mean absolute error of 2.16×10^{-5}, indicating that the encrypted and decrypted influence values are virtually indistinguishable from the plaintext reference. On the much larger BERT-SST2 task (66978 training points), we observe a slightly lower but still very high correlation of 0.9719 and an MAE of 1.84×10^{-5}. The total extra time climbs to about 10 016 s, but the per-sample cost (0.1495 s) remains essentially identical to MNIST. This confirms that our implementation scales linearly in the number of data points, with FHE batching and packing amortizing setup costs across all samples. Finally, on the GPT-2 MLP probe (21 saved gradients), we recover perfect correlation (1.0000) with an MAE of 1.12×10^{-5} in only 3.1s total ($\approx$ 0.1476 s per sample).

	Pearson Correlation	**Mean Average Error**	**Extra Running Time**	**Running Time Per Sample**
MLP(MNIST)	1.00	2.16e-05	890.19s	0.1483
BERT(SST-2)	0.97	1.84e-05	10016.07s	0.1495
GPT-2(Wikitext-2)	1.00	1.12e-05	3.1s	0.1476

Table 1: Experiment Results. The proposed method achieves near-perfect data utility computation with a computation time independent of model size.

High fidelity ensures FHE-based valuations match plaintext decisions. Also, since the additional FHE overhead is linear in the number of samples and stable across tasks, budgeting teams can accurately forecast the computing expense for any scale, from a small 21-sample probe to tens of thousands of records. This makes it straightforward to justify investments in secure computation when considering privacy compliance costs. We also conducted ablation studies on the CKKS scale parameter, which controls the precision of the encrypted computation. We omit the detailed discussion due to space constraint.

5 Conclusion

We introduced a novel homomorphic encryption framework designed to address the challenges in secure AI data marketplaces. Our framework allows potential data buyers to quantitatively assess a dataset's utility for AI algorithms while the data remains fully encrypted end-to-end. By enabling secure, direct testing on encrypted data, the proposed method protects the intellectual property of data owners and prevents buyers from overpaying for data with exaggerated value. The proposed approach is compatible with popular gradient-based data valuation methods and demonstrates scalability to Large Language Models. Experimental results show near-perfect accuracy in computing data utility with minor computational overhead.

Despite its advancements, the proposed framework has limitations. Although low-rank approximations and random projections are used to overcome some computational hurdles, the underlying complexity of homomorphic encryption, specifically the CKKS scheme, introduces trade-offs between precision, security, and performance. The choice of CKKS scale, for instance, directly impacts numerical fidelity, with smaller scales leading to severe quantization error. This necessitates careful parameter selection to ensure adequate precision without incurring unnecessary overhead.

References

Agarwal, A., Dahleh, M., and Sarkar, T. (2019). A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726.

Balazinska, M., Howe, B., and Suciu, D. (2011). Data markets in the cloud: An opportunity for the database community. *Proceedings of the VLDB Endowment*, 4(12):1482–1485.

Brittain, B. (2023). Getty images lawsuit says stability ai misused photos to train AI. *Reuters*.

Cheon, J. H., Kim, A., Kim, M., and Song, Y. (2017a). Homomorphic encryption for arithmetic of approximate numbers. In *Advances in cryptology–ASIACRYPT 2017: 23rd international conference on the theory and applications of cryptology and information security, Hong kong, China, December 3-7, 2017, proceedings, part i 23*, pages 409–437. Springer.

Cheon, J. H., Kim, A., Kim, M., and Song, Y. (2017b). Homomorphic encryption for arithmetic of approximate numbers. In Takagi, T. and Peyrin, T., editors, *Advances in Cryptology – ASIACRYPT 2017*, pages 409–437, Cham. Springer International Publishing.

Chillotti, I., Gama, N., Georgieva, M., and Izabachène, M. (2020). Tfhe: fast fully homomorphic encryption over the torus. *Journal of Cryptology*, 33(1):34–91.

Choe, S. K., Ahn, H., Bae, J., Zhao, K., Kang, M., Chung, Y., Pratapa, A., Neiswanger, W., Strubell, E., Mitamura, T., et al. (2024). What is your data worth to gpt? llm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Dwork, C. and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends® in Theoretical Computer Science.

Fernandez, R. C., Subramaniam, P., and Franklin, M. J. (2020). Data market platforms: Trading data assets to solve data problems. *arXiv preprint arXiv:2002.01047*.

Furukawa, J., Lindell, Y., Nof, A., and Weinstein, O. (2017). High-throughput secure three-party computation for malicious adversaries and an honest majority. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 225–255. Springer.

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

Graepel, T., Lauter, K., and Naehrig, M. (2012). Ml confidential: Machine learning on encrypted data. In *International conference on information security and cryptology*, pages 1–21. Springer.

Gramaglia, M., Fiore, M., Tarable, A., and Banchs, A. (2017). Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE.

Grynbaum, M. M. and Mac, R. (2023). The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27.

Hammoudeh, Z. and Lowd, D. (2024). Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *JASA*, 69(346):383–393.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Y., Hu, P., Zhao, H., and Ma, J. (2024). Most influential subset selection: Challenges, promises, and beyond. *Advances in Neural Information Processing Systems*, 37:119778–119810.

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

Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., and Suciu, D. (2015). Query-based data pricing. *Journal of the ACM (JACM)*, 62(5):1–44.

Kwon, Y., Wu, E., Wu, K., and Zou, J. (2023). Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*.

Lee, J.-W., Kang, H., Lee, Y., Choi, W., Eom, J., Deryabin, M., Lee, E., Lee, J., Yoo, D., Kim, Y.-S., et al. (2022). Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *iEEE Access*, 10:30039–30054.

Liu, J., Lou, J., Liu, J., Xiong, L., Pei, J., and Sun, J. (2021). Dealer: An end-to-end model marketplace with differential privacy. *Proceedings of the VLDB Endowment*, 14(6).

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

OpenAI and Reddit (2024). Openai and reddit partnership. *OpenAI Blog*.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. (2023). Trak: Attributing model behavior at scale. In *arXiv preprint arXiv:2303.14186*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144.

Roughgarden, T. (2010). Algorithmic game theory. *Communications of the ACM*, 53(7):78–86.

Spiekermann, M. (2019). Data marketplaces: Trends and monetisation of data goods. *Intereconomics*, 54(4):208–216.

Subramanyan, P., Sinha, R., Lebedev, I., Devadas, S., and Seshia, S. A. (2017). A formal foundation for secure remote execution of enclaves. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 2435–2450.

Wang, Z., Zheng, Z., Jiang, W., and Tang, S. (2021). Blockchain-enabled data sharing in supply chains: Model, operationalization, and tutorial. *Production and Operations Management*, 30(7):1965–1985.

Xia, M. and Henao, R. (2023). Reliable active learning via influence functions. *TMLR*.

Yao, A. C.-C. (1986). How to generate and exchange secrets. In *27th annual symposium on foundations of computer science (Sfcs 1986)*, pages 162–167. IEEE.

Zhang, J., Bi, Y., Cheng, M., Liu, J., Ren, K., Sun, Q., Wu, Y., Cao, Y., Fernandez, R. C., Xu, H., et al. (2024). A survey on data markets. *arXiv preprint arXiv:2411.07267*.

Zhang, M. and Beltrán, F. (2020). A survey of data pricing methods. *SSRN J*.

Zheng, S., Cao, Y., Yoshikawa, M., Li, H., and Yan, Q. (2022). Fl-market: Trading private models in federated learning. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1525–1534. IEEE.