Data Auctions for Retrieval Augmented Generation

Minbiao Han*
Google
minbiaohan@google.com

Seyed A. Esmaeili*
Princeton University
se5830@cs.princeton.edu

Michael Albert
Darden School of Business
University of Virginia
albertm@darden.virginia.edu

Haifeng Xu
University of Chicago
haifengxu@uchicago.edu

Abstract

We study the problem of data selling for Retrieval Augmented Generation (RAG) tasks in Generative AI applications. We model each buyer's valuation of a dataset with a natural coverage-based valuation function that increases with the inclusion of more relevant data points that would enhance responses to anticipated queries. Motivated by exclusive licensing for RAG, we focus on the scenario where each data point can be allocated to only one buyer. We show that the problem of welfare maximization in this setting is NP-hard even with two bidders, but design a polynomial-time $(1 - 1/e)$ approximation algorithm for any number of bidders. Unfortunately, however, this efficient allocation algorithm fails to be incentive compatible. The crux of our approach is a carefully tailored post-processing step called *data burning* which retains the $(1 - 1/e)$ approximation factor but achieves incentive compatibility. Our thorough experiments on synthetic and real-world image and text datasets demonstrate the practical effectiveness of our algorithm compared to popular baseline algorithms for combinatorial auctions.

1 Introduction

Generative AI (GenAI) has shown impressive performance in various domains such as text, code, image, and even video generation [1, 28, 29, 34, 57]. The impressive performance of GenAI has had profound impacts on numerous applications. Perhaps the most important and salient application of GenAI is its usage as a replacement for search engines and question-answer websites. In particular, it has become widespread to look up information using GenAI platforms such as ChatGPT, Perplexity, and Claude; traditional websites such as the Google and Bing have also developed similar "AI Overview" functionality. The answers provided by GenAI models are rich, detailed, and summarize information across various webpages and sources.

Although GenAI models have shown promise as a tool for acquiring knowledge, they are known to *hallucinate*, i.e., output invalid or even fully nonfactual answers [31]. Retrieval Augmented Generation (RAG) [5, 22, 40] has emerged as a very useful method that not only gives the user confidence in the GenAI answer but also enables the user to look up the exact source of a specific piece of information in the answer. In particular, by using RAG the output of the GenAI includes *references* to the sources of the information.

Another important usage of RAG is to incorporate information into a response that was not present during training. For example, common GenAI chatbots can converse intelligently about current

*These authors contributed equally to this work. Work done while both M. Han and S.A. Esmaeili were at the University of Chicago.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

events even though the events occurred after the training data cutoff date. This is due to news articles being incorporated into the context window through RAG. Many news organizations (The Associated Press, News Corp, and The Financial Times to give a few examples) have entered into licensing agreements with GenAI providers to allow for their news articles to be used in this way.

In this paper, we consider a setting where a collection of GenAI platforms are interested in buying information from a seller who possesses a large *database* of information for RAG based usage. These GenAI platforms anticipate queries about certain topics, are interested in augmenting their generated answers with the information from the database coupled with references. The seller looks to auctions off the database and can allocate a dataset from the database (i.e., a subset of data points) to each buyer and assign a price for each dataset. We focus on the more challenging setting where the seller does *exclusive* allocation (i.e., each data point can be allocated to at most one buyer). Our intended setting is licensed or controlled-access RAG, where exclusivity concerns contractual rights to ingest, index, retrieve from, and use a corpus or updated feed with documented provenance. These rights can remain valuable even when generated answers reveal facts or excerpts, but our model does not apply when underlying documents can be cheaply reconstructed and reused without effective restrictions. Exclusive licensing can also be motivated by cross-buyer externalities [3]; see Appendix A for examples and discussion.

Motivated by the above questions, we solve the mechanism design problem of allocating datasets for RAG based applications with private buyer preferences. Our contributions are summarized as follows. Due to space constraints, all proofs of theoretical results are provided in the appendix.

1. **Model and Valuation Function:** Section 2 introduces the formal model of our problem including the mechanism design objectives and a new RAG-based coverage function that naturally captures a buyer's value for any subset of data points and turns out to satisfy submodularity. Interestingly, we further show that in our setting, revenue maximization can be reduced to welfare maximization.

2. **Computational Hardness and Efficient Approximation:** Section 3 analyzes the computational complexity of our problem. We show that the demand oracle (see Definition 3.1), which is commonly used in combinatorial auctions, is NP-hard to implement in our setting. Moreover, welfare maximization is NP-hard even for special cases such as two bidders (though trivial for one bidder). On the positive side, we introduce a $(1 - 1/e)$-approximate algorithm for general welfare maximization. Unlike many previous submodular welfare maximization algorithms, our algorithm is practical, and involves a simple linear program and a rounding step.

3. **Incentive Compatible Welfare Maximizing Mechanism:** Section 4 accounts for the issue of incentives. We start by showing an interesting generalization of the well-known Myerson lemma [45] to our setting with combinatorial allocations. In particular, given any allocation satisfying a certain monotonic property, we can find a payment scheme that induces an incentive compatible mechanism, i.e., truthful bidding is optimal in expectation for each bidder. However, monotonicity turns out to be a quite delicate property – we find that almost all natural submodular maximization algorithms turn out to *not* be monotonic, and neither is our approximate algorithm from Section 3. However, we develop a novel post-processing technique for the output of our previous allocation algorithm that leads to monotonic allocation and meanwhile preserves the $1 - 1/e$ approximation. This helps us to achieve incentive compatibility at no additional performance cost to our original approximation algorithm. We also discuss potential further applications of our techniques to related problems.

4. **Experiments:** Section 5 tests our algorithm on real-world datasets and demonstrates its superior performance over popular baselines in terms of both revenue quality and incentive compatibility.

1.1 Additional Discussion on Related Work

Our work is closely related to research on AI and computational economics, as detailed below.

Mechanism design for generative AI. Conceptually, our work contributes to the recent literature on mechanism design for generative AI. While a significant body of work has explored the use of auctions to shape LLM outputs and integrate ads [16, 17, 25, 54, 55], our approach takes a different direction. We investigate methods for selling data to accomplish generative AI tasks using LLMs, offering a novel perspective on the intersection of economics and AI.

Data markets. Our work also intersects with the field of data market design, which encompasses data valuation and data selling. Previous work has introduced various ML markets [2, 9, 26, 41], primarily focusing on selling the instrumental value of data by creating ML models from aggregated market data and offering different versions to buyers based on their preferences. While Mehta et al. [43] explored a similar data selling problem as our paper, their focus was on pricing conventional data types, such as phone lists. Our exclusivity constraint couples buyers' feasible allocations, while their valuations depend only on their own allocated datasets. In a related but distinct area, there's a substantial body of research on data valuation using the concept of Data Shapley [24, 33, 52, 56, 60]. This method assesses the value of individual data points or groups by measuring their impact on model performance. Our approach diverges from this by modeling data value independently of the specific model used.

Combinatorial auctions. Our data auction problem also shares notable similarities with traditional combinatorial auction problems. A notable related work by Dobzinski et al. [14] presents an incentive compatible mechanism with a $1/\sqrt{n}$ approximation algorithm for general submodular value functions. Dughmi et al. [18] obtain a truthful-in-expectation $(1 - 1/e)$-approximation for matroid rank sum valuations through convex rounding, using a lottery-value oracle and expected polynomial time. Our construction replaces this expected-polynomial-time convex-rounding procedure with a compact LP and a polynomially bounded sequence of explicit rounding and monotonic post-processing steps. Using public coverage structure and scalar private values, we compute rounded expectations explicitly and use data burning to scale each bidder's expected coverage to $(1 - 1/e)$ times their monotone LP coverage. Unlike continuous greedy [59], this yields truthfulness in expectation while preserving the approximation guarantee. Other relevant studies in this area have limitations such as focusing on data auctions with single-minded buyers [12, 39, 44], requiring strong demand oracles for buyer values [6, 7, 15, 20, 37], or only studying the allocation problem under the assumption that all buyers' value information is public [13, 14, 38, 59].

2 Model and Problem Statement

Here we introduce our model, notation, and the formal statement of our problem. The seller owns a database of points $\mathcal{DB}$ that consists of a collection of n many points, i.e, $\mathcal{DB} = \{x_1, x_2, \ldots, x_n\}$. The points in the database could represent images, documents, or even articles that contain both textual and visual information that can be used for RAG. Our framework is general and agnostic to the nature of the points only requiring a measure of distance $d(.,.)$ or similarity $s(.,.)$ that can be calculated given any pair of points x_a and x_b. The choice for the distance/similarity measure is largely dependent on the setting, but there are many well-known standard measures that can be used. For example, if the database $\mathcal{DB}$ consists of a collection of images then a simple and meaningful measure would be the Euclidean distance between the embeddings of the images, i.e., $d(x_a, x_b) = \|f(x_a) - f(x_b)\|_2$ where $f(.)$ is some embedding function [11, 32]. If we had a collection of documents on the other hand, then we may use the cosine similarity between the term frequency vectors of the points x_a and x_b, i.e., $s(x_a, x_b) = \frac{\vec{t_a} \cdot \vec{t_b}}{\|\vec{t_a}\|\|\vec{t_b}\|}$ where $\vec{t_a}$ and $\vec{t_b}$ are the term frequency vectors of documents x_a and x_b, respectively [27]. For consistency, we assume that we have a distance $d(.,.)$ instead of a similarity measure $s(.,.)$, although our framework, theoretical results, and algorithms hold for both settings.

Coverage-Based Value Function: There are m buyers (bidders)[1] who are interested in (parts of) the database $\mathcal{DB}$ and want to use it in particular for applications such a RAG or standard information retrieval. It is critical to model a value function that accurately represents the value a bidder would extract from a subset of points in the database without requiring the training of a large machine learning model as the training process itself can be quite expensive.[2] Since we are motivated by RAG-based applications our value function is focused on measuring the approximate *coverage* a bidder obtains from a given subset of points (dataset) from the database. To fully characterize the coverage function, we need to introduce some notation. First, consider a bidder i and dataset of points $\mathcal{D} \subset \mathcal{DB}$ that has been given to him. The absence of a point j from $\mathcal{D}$ does not imply that it is not covered by the dataset $\mathcal{D}$ since there may exist a point $j' \in \mathcal{D}$ that is *sufficiently close* to j and can practically serve as a representative of j. However, this threshold of closeness can be dependent

[1] Since the buyers will submit bids, we will use the words buyers and bidders interchangeably.

[2] The cost of training large language models (LLMs) has surged in recent years, with expenses now reaching into the millions of dollars for state-of-the-art models [10].

on the point and the bidder. Therefore, we define the coverage variable C_{ij} and set $C_{ij} = 1$ only if $\exists j' \in \mathcal{D}$ such that $d(j, j') \leq r_{ij}$ where r_{ij} is the *radius* threshold for point j and bidder i, we set $C_{ij} = 0$ otherwise. Note that our model enables a fine-grained specification of the radius r_{ij}, i.e., at the level of a specific bidder and a specific point. r_{ij} determines the desired accuracy level buyer i wants for point j. Practically, the radius value r_{ij} can be set according to the topic (such as politics or sports) point j belongs to and the desired level of accuracy of bidder i for covering that topic.

Furthermore, since each point provides a different measure of value which can be bidder dependent, we associate a value $w_{ij} \geq 0$ for each point $j \in \mathcal{DB}$ and each bidder $i \in [m]$. Similar to the radius, our model enables the weights to be specified at the level of the point and the buyer. In a practical setting, each buyer might be interested in a specific topic and accordingly topics which the buyer finds more relevant will receive higher weights. Based on the above we define the total coverage bidder i gains from dataset $\mathcal{D}$ to be $\mathcal{C}_i$ where we have

$$\mathcal{C}_i(\mathcal{D}) = \textstyle\sum_{j \in \mathcal{DB}} w_{ij} \cdot C_{ij} \tag{1}$$

It is straightforward to verify that this coverage function is submodular – i.e., $\mathcal{C}_i(\mathcal{D}_1 \cup \{x\}) - \mathcal{C}_i(\mathcal{D}_1) \geq \mathcal{C}_i(\mathcal{D}_2 \cup \{x\}) - \mathcal{C}_i(\mathcal{D}_2)$ for all $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \mathcal{DB}$ and $x \notin \mathcal{D}_2$. Without loss of generality we assume that $\sum_{j \in \mathcal{DB}} w_{ij} = 1$,[3] note that this implies $0 \leq \mathcal{C}_i(\mathcal{D}) \leq 1$ for any dataset $\mathcal{D}$.

The full value a bidder i gains from dataset $\mathcal{D}$ is $v_i(\mathcal{D})$ which is

$$v_i(\mathcal{D}) = \theta_i \cdot \mathcal{C}_i(\mathcal{D}) \tag{2}$$

where $\theta_i \geq 0$ is a number that indicates the *importance* of the coverage to the bidder. Any possible value $\theta_i \in \Lambda = \{\theta_1, \cdots, \theta_{|\Lambda|}\}$ where Λ is a public set.

Public and Private Parameters: Based on the above, each bidder $i \in [m]$ is characterized by three sets of parameters $\{w_{ij}\}_{j \in \mathcal{DB}}, \{r_{ij}\}_{j \in \mathcal{DB}}$ and θ_i. We assume that the weight $\{w_{ij}\}_{j \in \mathcal{DB}}$ and radius parameters $\{r_{ij}\}_{j \in \mathcal{DB}}$ are known publicly. We make this assumption because it is likely possible to estimate $\{w_{ij}\}_{j \in \mathcal{DB}}$ and $\{r_{ij}\}_{j \in \mathcal{DB}}$. Specifically, the seller, through querying an existing Gen AI product, would likely be able to infer the distance $\{r_{ij}\}_{j \in \mathcal{DB}}$ for which an article is considered acceptable in a RAG context for a bidder. Similarly, by understanding the specialty of the bidder in the marketplace, the seller is likely to be able to estimate the relative value of different regions of the embedding space, i.e., the weights $\{w_{ij}\}_{j \in \mathcal{DB}}$. However, we generally expect that it will be impossible for the seller to estimate the bidder's overall value for RAG data θ_i and therefore assume that it is private. We acknowledge that these assumptions may not always hold, but it is well known that the multi-dimensional mechanism design problem is generally intractable. By assuming that only θ_i is private, we can formulate the problem as a single dimensional mechanism design problem.

Bidding Setting and Objectives: Each bidder $i \in [m]$ will report a value b_i privately. Once all bids are submitted, the seller will decide an *allocation* and a *payment*. The allocation is simply the dataset $\mathcal{D}_i \subset \mathcal{DB}$ that will be given to bidder i while the payment is the price $p(\mathcal{D}_i)$ of dataset $\mathcal{D}_i$.[4] As noted earlier a point may not be assigned to more than one bidder, thus $\forall i, i' \in [m]$ we have $\mathcal{D}_i \cap \mathcal{D}_{i'} = \emptyset$ if $i \neq i'$. The above implies that the utility a bidder i obtains is $u_i(\mathcal{D}_i) = v_i(\mathcal{D}_i) - p(\mathcal{D}_i)$.

As in standard mechanism design, the seller wants allocation and payment rules that lead to an *incentive compatible* mechanism, i.e., a mechanism where truthful bidding is a dominant strategy for each bidder. Formally, if b_i denotes bidder i's bid and $\boldsymbol{b}_{-i}$ denotes all other bids, then $u_i(\theta_i, \boldsymbol{b}_{-i}) \geq u_i(b'_i, \boldsymbol{b}_{-i})$ $\forall b'_i$ and $\forall \boldsymbol{b}_{-i}$. In other words, a bidder never gains higher utility by reporting a value other than θ_i regardless of all other bids. Note that we assume that the bid values b_i like the true values θ_i are restricted to the same set Λ.

Besides incentive compatibility, given a bidding profile $\boldsymbol{b}$ the seller wants to maximize the reported welfare defined as $\mathcal{W}(\boldsymbol{b}) = \sum_{i \in [m]} b_i \cdot \mathcal{C}_i(\mathcal{D}_i)$. True welfare uses θ_i in place of b_i, coinciding with reported welfare under truthful bidding. Note that the true welfare is the sum of the total utilities (including that of the seller) but it excludes the payment values since the seller's utility is the sum

[3] Given any collection of non-negative coefficients $\{w_{ij}\}_{j \in \mathcal{DB}}$ for a bidder i we can simply set $w_{ij} := \frac{w_{ij}}{\sum_{j \in \mathcal{DB}} w_{ij}}$. Note that we are implicitly assuming that $\sum_{j \in \mathcal{DB}} w_{ij} > 0$ which is reasonable as otherwise this would imply that the bidder has no interest in the whole database.

[4] When the price is specified at the level of each point $j \in \mathcal{DB}$ the price of a dataset $\mathcal{D}$ is simply the sum of the prices of the points it contains, i.e., $p(\mathcal{D}) = \sum_{j \in \mathcal{D}} p(j)$

of the payments across all bidders $\sum_{i\in[m]} p(\mathcal{D}_i)$ and therefore the payment cancels out. We will show in Section 3 that welfare maximization in our setting is NP-hard, therefore finding a welfare maximizing allocation is not possible to do in polynomial time unless P=NP. Accordingly, we aim at α-optimal incentive compatible mechanisms which (in addition to incentive compatibility) for a given bid profile $\boldsymbol{b}$ lead to a welfare value of at least $\alpha\mathcal{W}^*(\boldsymbol{b})$ where $\mathcal{W}^*(\boldsymbol{b})$ is the optimal welfare value for bid profile $\boldsymbol{b}$.

Additionally, we note that an important property that we target in designing our mechanism is *individual rationality* which states that any truthful bidder should obtain non-negative utility regardless of the other bids, i.e., $u_i(\theta_i, \boldsymbol{b}_{-i}) \geq 0\ , \forall \boldsymbol{b}_{-i}$. Moreover, if we use randomized algorithms, then we consider the expected value of the utilities and the welfare.

Finally, in Appendix E, we show that an α approximation welfare maximizing algorithm can be used to guarantee an α approximation for the seller's revenue $\sum_{i\in[m]} p(\mathcal{D}_i)$. In fact, if the welfare maximizing algorithm is incentive compatible and individually rational then so would the revenue maximizing algorithm. Therefore, for simplicity we focus on discussing only welfare maximization for the rest of the paper since revenue maximization is immediately implied.

3 Welfare Maximization: Computational Hardness and Approximation

Our objective is to maximize the social welfare $\mathcal{W}(\boldsymbol{b})$. Unfortunately, we show that the problem is actually NP-hard. Interestingly, all of our hardness results hold even when the weights are constant, i.e., $w_{ij} = w > 0,\ \forall j \in \mathcal{DB}, \forall i \in [m]$. However, before we show that welfare maximization is NP-Hard, we note that a common thread in combinatorial auctions has considered using the demand oracle (or query) [6, 14, 21, 36] which is defined as follows:

Definition 3.1. Given a price vector $\boldsymbol{p}$ over the points in the database and a bidder $i \in [m]$ the demand oracle returns the subset of points that maximize utility for bidder i, i.e., $\underset{\mathcal{D}\subset\mathcal{DB}}{\operatorname{argmax}} \big(v_i(\mathcal{D}) - p(\mathcal{D})\big)$.

The demand oracle can be used to design incentive compatible auctions with high welfare. In fact, the recent work of Assadi and Singla [6] for submodular bidders achieves an approximation ratio of $O(\frac{1}{(\log\log n)^3})$ having a small dependence on the number of items. However, in some settings the implementation of the demand oracle might require solving an NP-hard problem [13], which render demand oracle based algorithms of little use. The proposition below shows that this is the case in our setting as well.

Proposition 3.2. *The demand oracle cannot be implemented in polynomial time unless $P = NP$.*

We now study welfare maximization allocation. Notably, if there is only a single bidder, welfare maximization is trivial – just allocate all items to this bidder would maximize revenue. Surprisingly, this welfare maximization problem immediately becomes NP-hard even when there are 2 bidders.

Theorem 3.3. *Welfare maximization is NP-hard even in the following two special situations:*

- *There are only two bidders and the radius values in the value functions differ across bidders but do not differ across data points (i.e., $\forall i \in [m] : r_{ij} = r_i$).*
- *There are only three bidders and their value functions have the same radius (i.e., $r_{ij} = r\ \forall i, j$).*

The proofs of both hardness results are non-trivial, and are novel to the best of our knowledge. This is because previous hardness of submodular welfare maximization reduction does not apply to our setting since our valuation function is more special than general coverage functions, hence its intractability is more difficult to prove. The two claims in Theorem 3.3 are reduced from dominating set and domatic number problem [35] respectively; details are deferred to Appendix C.2 and C.3.

Despite the intractability of welfare maximization even with few bidders, next we introduce a polynomial time approximation algorithm that computes a $(1-1/e)$-optimal solution for *any* number of bidders. Before we show our algorithm, we introduce a convenient notation: for a point $j \in \mathcal{DB}$ and bidder $i \in [m]$ we denote by $N_i(j)$ the set of points in $\mathcal{DB}$ that are within a distance of r_{ij} from point j. Formally, $N_i(j) = \{j' \in \mathcal{DB} | d(j, j') \leq r_{ij}\}$.

The starting point of our algorithm is a linear program (LP) relaxation of the allocation problem as shown below. We use the symbols $\bar{X}_{ij}$ and $\bar{C}_{ij}$ with a bar for the LP decision variables to emphasize that they are fractional (i.e., not necessarily equal to 0 or 1, but in $[0, 1]$).

$$\max_{\bar{X}_{ij},\bar{C}_{ij}} \quad \sum_{i\in[m]} \theta_i \cdot \Big(\sum_{j\in\mathcal{DB}} w_{ij}\cdot \bar{C}_{ij}\Big) \tag{3a}$$

$$\forall i \in [m], \forall j \in \mathcal{DB} : 0 \le \bar{X}_{ij}, \bar{C}_{ij} \le 1;\ \bar{C}_{ij} \le \sum_{j'\in N_i(j)} \bar{X}_{ij'} \tag{3b}$$

$$\forall j \in \mathcal{DB} : \quad \sum_{i\in[m]} \bar{X}_{ij} \le 1 \tag{3c}$$

In the LP, $\bar{X}_{ij}$ represents the (relaxed) fractional allocation of point j to bidder i whereas $\bar{C}_{ij}$ indicates whether the point is covered for bidder i or not. Note that a point j may be covered for a bidder i even if it was not allocated to him since the bidder may have been allocated another point $j' \in N_i(j)$. Since the welfare-optimal allocation is a feasible solution to this LP, we have the following.

Lemma 3.4. *For any given bid profile $\boldsymbol{b}$ the optimal value of LP* (3) *satisfies $OPT_{LP}(\boldsymbol{b}) \ge \mathcal{W}^*(\boldsymbol{b})$.*

Our full algorithm **LPR** (Algorithm 1) is simple. We start by solving LP (3), then round the resulting values $\bar{X}_{ij}$ as follows, for a point $j \in \mathcal{DB}$ we allocate it to bidder i with probability $\bar{X}_{ij}$. Note that this rounding is possible since by constraint (3c) we have $\sum_{i\in[m]} \bar{X}_{ij} \le 1$, further by constraint (3b) we have $\bar{X}_{ij} \ge 0$. Note further that the sampling is also done in an *independent* manner from point to point. It follows that the resulting integer values X_{ij} satisfy $\mathbb{P}(X_{ij} = 1) = \bar{X}_{ij}$.

Algorithm 1 LP + Rounding (**LPR**)

input Database $\mathcal{DB}$, bidders $i \in [m]$, bidders' parameters $\{w_{ij}, r_{ij}\}_{j\in\mathcal{DB}, i\in[m]}$ and$\{\theta_i\}_{i\in[m]}$.

1: Run LP (3) and obtain the optimal solution $\bar{X}_{ij}$ and $\bar{C}_{ij}$.
2: For a given point $j \in \mathcal{DB}$ assign it to bidder i with probability $\bar{X}_{ij}$.
3: Form the dataset allocation $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m)$ using the values X_{ij}.

The resulting coverage values for the bidders (and the welfare as well) are dependent on C_{ij}. However, the values C_{ij} have a somewhat complicated dependence on the values of X_{ij}. We now introduce the following lemma which uses the optimal LP solution $\bar{X}_{ij}, \bar{C}_{ij}$ to give an explicit form for the expected value of C_{ij} as well as lower bounding the expected total coverage for each bidder.

Lemma 3.5. *Using Algorithm 1 with optimal solution $\bar{X}_{ij}$ and $\bar{C}_{ij}$ the final rounded solution satisfies*

$$\mathbb{E}[C_{ij}] = 1 - \prod_{j'\in N(j)} (1 - \bar{X}_{ij'}) \tag{4}$$

$$\mathbb{E}[\mathcal{C}_i] \ge (1 - 1/e)\cdot \bar{\mathcal{C}}_i \tag{5}$$

where $\bar{\mathcal{C}}_i = \sum_{j\in\mathcal{DB}} w_{ij}\bar{C}_{ij}$.

Since the above lemma provides a lower bound for the coverage each bidder receives, it follows that we can lower bound the welfare as shown in the theorem below.

Theorem 3.6. *Algorithm 1 has an approximation ratio of* $(1 - 1/e)$.

4 Incentive Compatible Welfare Maximization

In this section, we turn to the study of incentive compatibility of welfare maximization, i.e., designing a mechanism that incentivizes every buyer i to report their private type θ_i *truthfully* for the data auction. We first present a nice property of the allocation rule of an incentive compatible mechanism. Specifically, we generalize Myerson's results [45] to our setting showing that any incentive compatible mechanism's allocation rule has to necessarily be monotonic. In addition to the necessity of the monotonicity property, we show an efficient algorithm for calculating the payment in order to satisfy both incentive compatibility and individual rationality. First, we start with a formal definition of a monotonic allocation rule.

Definition 4.1. [Monotonic Allocation Rule] An allocation is monotonic if for every bidder $i \in [m]$ and bids $\boldsymbol{b}_{-i}$ then we have $\mathbb{E}[\mathcal{C}_i(\mathcal{D})] \ge \mathbb{E}[\mathcal{C}_i(\mathcal{D}')]$ if $b_i \ge b_i'$ where $\mathcal{D}$ and $\mathcal{D}'$ are the datasets allocated to bidder i under bid profiles $(b_i, \boldsymbol{b}_{-i})$ and $(b_i', \boldsymbol{b}_{-i})$, respectively.

A fundamental result of Myerson [45] reduces a mechanisms' truthfulness to monotonicity in allocation probability in single-item auctions. To handle our combinatorial allocation case, we start by showing a novel form of monotonicity for our combinatorial auction with coverage-based valuations.

Lemma 4.2. *For any incentive compatible mechanism, its allocation rule must be monotonic. Conversely, given a monotonic allocation rule, there always exists a corresponding efficiently computable payment such that truthful bidding is a dominant strategy and individually rational for each bidder.*

Lemma 4.2 offers a characterization of incentive compatible mechanisms for combinatorial allocation with our coverage-based value function. This differs from and in some sense strictly generalizes the classical monotonicity of allocating a single item, so we believe our lemma could be of independent interest. A relevant result is by Nisan and Ronen [49] who also show a generalization of Myerson's lemma to combinatorial auctions, but with single-minded value (i.e., the buyer either has value v if getting her demanded set, or value 0 otherwise). However, our valuation function here is fundamentally different.

At this point, one would hope that our **LPR** algorithm (Algorithm 1) from the previous section would be monotonic, hence truthfulness is directly achieved. Unfortunately, monotonicity turns out to be a very delicate property for allocation algorithms. Indeed, we find that almost all natural allocation approaches for submodular functions (e.g., greedy [8, 38] and continuous greedy [59]) are *not* monotone; neither is our **LPR** algorithm.

Fact 4.3. LPR *(Algorithm 1) is not monotonic in the sense of Definition 4.1.*

The above fact is shown by constructing a counter example. However, the core reason behind the failure is that the relation between the rounded coverage variables C_{ij} that decides the final coverage value and the LP coverage variables $\bar{C}_{ij}$ is complicated. More specifically, while the LP solution value in the constructed example is actually monotonic, due to rounding the final expected value for the coverage based on the rounded coverage variables C_{ij} is not monotonic.

The crux of our approach is that, although Algorithm 1 is not monotonic, we are able to make it monotonic through careful post-processing. Specifically, it can be shown that the optimal coverage values $\bar{\mathcal{C}}_1, \ldots, \bar{\mathcal{C}}_m$ each bidder receives from the optimal LP solution before rounding are, in fact, monotonic (see Lemma 4.5). As a result, we propose **LPRMONO** (Algorithm 2) which utilizes a post-processing of *data burning*. That is, instead of directly allocating the resulting datasets from Algorithm 1 to each bidder, the bidder is allocated the dataset with probability ρ_i and given nothing with probability $1 - \rho_i$. For each bidder $i \in [m]$ the probability is set to $\rho_i = (1 - 1/e) \cdot \frac{\bar{\mathcal{C}}_i}{\mathbb{E}[\mathcal{C}_i]}$, this ensures both that the probability value is valid (i.e., $\rho_i \in [0, 1]$) and that each bidder receives an expected coverage of exactly $(1 - 1/e) \cdot \bar{\mathcal{C}}_i$. Since for a bidder i the LP coverage value $\bar{\mathcal{C}}_i$ is monotonic it follows that if a bidder instead receives in expectation a positive constant of $(1 - 1/e)$ times the LP coverage value then the final coverage would also be monotonic. Further, while it is true that allocating nothing to a bidder with probability $1 - \rho_i$ would degrade the coverage for bidder i, the degradation is still within a factor of $(1 - 1/e)$ of the LP coverage value and we can thus show that this method leads to the same approximation factor of $(1 - 1/e)$. Therefore, we can establish the following theorem.

Algorithm 2 LP + Monotonic Rounding (**LPRMONO**)

input Database $\mathcal{DB}$, bidders $i \in [m]$, bidders' public parameters $\{w_{ij}, r_{ij}\}_{j \in \mathcal{DB}, i \in [m]}$, bidders' privately reported bids $\{\theta_i\}_{i \in [m]}$.

1: Run LP (3) and let the output be $\bar{X}_{ij}$ and $\bar{C}_{ij}$.
2: For a given point $j \in \mathcal{DB}$ assign it to bidder i with probability $\bar{X}_{ij}$.
3: Form the dataset allocation $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m)$ using the values X_{ij}.
4: For each bidder $i \in [m]$:
   -compute $\bar{\mathcal{C}}_i = \sum_{j \in \mathcal{DB}} w_{ij} \cdot \bar{C}_{ij}$.
   -compute $\mathbb{E}[\mathcal{C}_i] = \sum_{j \in \mathcal{DB}} w_{ij}\, \mathbb{E}[C_{ij}]$ using (4).
   -if $\bar{\mathcal{C}}_i > 0$ set $\rho_i = (1 - 1/e) \cdot \frac{\bar{\mathcal{C}}_i}{\mathbb{E}[\mathcal{C}_i]}$ else set $\rho_i = 0$.
5: Each bidder $i \in [m]$ with probability ρ_i gets dataset $\mathcal{D}_i$ and otherwise gets nothing.

Theorem 4.4. *Algorithm 2 is monotonic and has an approximation ratio of* $1 - 1/e$.

We defer the formal proof of this theorem to Appendix D.3. To illustrate its proof ideas, we present the two key lemmas below.

Lemma 4.5. *The optimal LP allocation $\bar{\mathcal{C}}_i$ resulting from LP* (3) *is monotonic for every bidder.*

Then we show that the monotonic allocation rule also guarantees a $(1-1/e)$ factor of the LP coverage for every bidder. Denoting the the final coverage resulting from Algorithm 2 with a prime $'$, i.e., $\mathcal{C}'_i$ we have the following lemma.

Lemma 4.6. *The expected coverage for any bidder $i \in [m]$ from Algorithm 2 is $\mathbb{E}[\mathcal{C}'_i] = (1-1/e)\cdot\bar{\mathcal{C}}_i$.*

We believe our $(1-1/e)$ algorithm, as well as our novel monotonic post-processing technique, are of independent interest. In fact, we show that our techniques can be modified to lead to a $(1-1/e)$ welfare maximization for another variant of coverage valuations as defined in Dobzinski and Schapira [13][5], we can in addition use a similar post-processing method to satisfy monotonicity for these coverage valuations, see Appendix F for details. One potential limitation of this approach is that whole-bundle burning can cause large losses in a bidder's realized coverage and high variance in utility. The granular alternative in Appendix D.4 randomizes between nested subsets of the rounded dataset and can reduce this variance while preserving expected coverage and welfare. It remains an intriguing open question whether one can design an algorithm that outperforms the $1-1/e$ approximation for our RAG-based coverage function. While this bound is known to be tight for general coverage functions [14, 19], the hope here lies in exploiting the distance-based membership relation in the RAG setting. However, tightly pining down the best ratio is beyond the scope of this paper.

5 Experiments

We evaluate the quality and performance of our algorithm for computing the expected truthful auction mechanism. Specifically, we consider three datasets for the selling process. While the first dataset is synthetic, the second and third are standard text and image datasets which we expect to have similar performance to RAG settings that use information such as text or images. The parameters for a given setting are m which is the number of buyers, n which is the size of the dataset $\mathcal{DB}$, Λ which is the set of possible types for each buyer, and d which is the dimension of the space. Further, we always use the Euclidean distance for all instances and datasets.

- **Synthetic** (**SYN**) dataset: We set $m = 2$, $n = 5$, $|\Lambda| = 10$, $d = 10$. And the database $\mathcal{DB}$ is generated uniformly within the range of $[0, 10]$ with a shape of $n \times d$. This is a relatively small toy dataset aimed to test all algorithms' performances, as we will show later, one benchmark algorithm is very computationally expensive and cannot solve large datasets efficiently.
- **20_newsgroups** (**NEWS**) dataset [53]: This is a popular collection of newsgroup documents used for text classification and machine learning tasks. The training dataset contains 11,314 newsgroup posts partitioned nearly evenly across 20 different classes. Then we generate the database $\mathcal{DB}$ by encoding the dataset with a powerful and efficient sentence embedding model all-MiniLM-L6-v2 [51], which produces an embedding dataset with a shape of $n = 11,314$ and $d = 384$. In addition, we set $m = 3$ and $|\Lambda| = 10$.
- **CIFAR-10** (**CIFAR**) dataset [58]: This is a widely used benchmark in machine learning and computer vision for image classification tasks. The training dataset contains 50,000 images in 10 classes, with 5,000 images per class. We downloaded the encoded embeddings from MatthewWilletts [42] with a shape of $n = 50,000$ and $d = 2048$. We also set $m = 3$ and $|\Lambda| = 10$.

For all three datasets, we set $\Lambda = \{0, 0.1, \cdots, 0.9\}$ and then each buyer's type θ_i where $i \in [m]$ is drawn independently and uniformly from Λ. In addition, we generate $r_{ij} = \alpha_i R_{jc}$ for all $i \in [m]$ and $j \in [n]$, where α_i's are uniformly randomly generated coefficients within range of $[0, 1]$ and R_{jc} is a class dependent average distance. Note that **NEWS** and **CIFAR** datasets come with different classes, and we assume that the **SYN** dataset only contains a single class. R_{jc} is computed by first sampling 100 random points within each class (all points for **SYN**) and then computing the average Euclidean distance among those 100 points. Finally, we assume w_{ij} is the same for all points in the

[5]The coverage valuation defined in Dobzinski and Schapira [13] is different from ours and does not generalize our coverage valuation. Nevertheless, we show that our techniques can be applied there as well.

same class, which is generated uniformly randomly within the range of $[0, 1]$. Then we normalize the randomly generated numbers such that $\sum_j w_{ij} = 1, \forall i \in [m]$.

We compare our Algorithm 1 (**LPR**) and Algorithm 2 (**LPRMONO**) with the following allocation algorithms: (1) The widely adopted greedy (**GREEDY**) allocation algorithm, as used in various recent works such as [4, 8, 38]; (2) The optimal approximation algorithm for submodular welfare maximization without incentive constraints, termed as continuous greedy algorithm (**CONGREEDY**) that is proposed by Vondrák [59]. These algorithms are widely recognized for their effectiveness in handling the complex valuations of item bundles, which directly aligns with our problem of valuing datasets in RAG tasks. We note that both **GREEDY** and **CONGREEDY** are designed for the allocation problem under the assumption that buyers' valuations are publicly known, achieving approximation ratios of $1/2$ and $1 - 1/e$, respectively [4, 8, 38, 59]. However, *neither* algorithm is guaranteed to satisfy *incentive compatibility*, since they rely on the assumption of public value information.

It's also worth mentioning that Dobzinski et al. [14] proposed an incentive-compatible allocation algorithm based on matching techniques, though it was developed for a different problem setting. Their algorithm focuses on the scenario where the number of items is smaller than the number of buyers. This algorithm achieves an approximation ratio of $1/\sqrt{n}$, is monotonic, and assigns one item to one buyer. However, it is not suitable for our scenario, where the number of items is much larger than the number of buyers, and is therefore excluded from the comparison.

We compare all algorithms' performance in terms of the welfare, measured as the total utility achieved by all buyers, as well as the runtime. First, we start by comparing the algorithms' performance on welfare. For **LPR** and **LPRMONO**, we run the rounding process (i.e., Lines 2-3 of Algorithm 1 and Lines 4-5 of Algorithm 2) for $150,000$ trials to get an estimation of the expected welfare. In addition, even though the optimal welfare of the allocation problem is NP-Hard to compute, we have the objective value from LP (3) as an upper bound on the optimal welfare. As a result, we compare each algorithm's approximation ratio with respect to the objective value of LP (3).

Algorithm	SYN	NEWS	CIFAR
LPR	0.9924 ±0.037	**0.9947** ±0.014	**0.9982** ±0.004
LPRMONO	0.6319 ±0.001	0.6325 ±0.002	0.6322 ±0.001
GREEDY	**0.9945** ±0.023	0.9908 ±0.014	0.9871 ±0.014
CONGREEDY	0.915 ±0.136	N/A	N/A

Table 1: Welfare approximation ratio relative to objective, averaged over 50 instances with standard deviation.

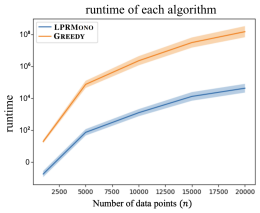


Figure 1: Runtime in seconds of each algorithm.

All welfare comparisons assume truthful bids ($b_i = \theta_i$), including for the non-IC algorithms **LPR**, **GREEDY**, and **CONGREEDY**. With the corresponding payments, **LPRMONO** makes truthful bidding a dominant strategy in expectation. From the experimental results in Table 1, we can see that **GREEDY** performs slightly better than our **LPR** on the small **SYN** dataset, but is outperformed by **LPR** when the size of dataset gets larger. In addition, we note that our **LPRMONO** indeed achieves a tight $1 - 1/e$ approximation ratio, as demonstrated by the experimental results. Finally, we would like to mention that **CONGREEDY** has a runtime complexity of $O\big((mn)^8\big)$ [59], making it impractical for larger datasets like **NEWS** and **CIFAR**. Consequently, we exclude **CONGREEDY** from welfare comparisons in those two datasets and the following runtime comparison.

The runtime results is shown by Figure 1. The difference between **LPR** and **LPRMONO** is negligible so only **LPRMONO** is presented. We vary the number of data points n and evaluate the runtime of each algorithm. All experiments are repeated on 50 random instances on **CIFAR**, and the plotted curves are averaged across the runs with error bars for 95% confidence level. We can see that the runtime of **GREEDY** is significantly larger compared to our algorithms. It is also worth noting that besides the runtime advantage of our algorithms, **LPRMONO** is incentive compatible.

In terms of monotinicity we show that **GREEDY** is not monotone through an example, see Appendix G for the full details and proof. Further, it is challenging to theoretically demonstrate that the **CONGREEDY** algorithm is not monotonic due to its complicated randomized nature which would require a lengthy manual simulation of the algorithm. Consequently, we assessed its monotonicity on a randomly generated instance from the **SYN** dataset. Given the stochastic nature of randomized

allocation algorithms, we estimated the expected allocation coverage $\mathbb{E}[\mathcal{C}_i(\mathcal{D})]$ for every buyer i at a given bid value by running the algorithm 150,000 times and averaging the outcomes.[6] Let $\hat{\mathcal{C}}_i(\mathcal{D})$ and $\hat{\mathcal{C}}_i(\mathcal{D}')$ denote our estimates of the expected coverage for buyer i at bids $b_i \geq b_i'$, with the other bids fixed. After selecting the instance and bid pair, we used fresh simulations for confirmation and observed $\hat{\mathcal{C}}_i(\mathcal{D}') - \hat{\mathcal{C}}_i(\mathcal{D}) > 0.02$. A one-sided Hoeffding test of $H_0 : \mathbb{E}[\mathcal{C}_i(\mathcal{D})] \geq \mathbb{E}[\mathcal{C}_i(\mathcal{D}')]$ gives the conservative p-value bound $p \leq 2e^{-30} < 1.9 \times 10^{-13}$; see Appendix H for the instance and test details.

6 Limitations

Our guarantees rely on a single-parameter model: coverage weights and radii, and hence neighborhoods, are public, while only each bidder's scalar value θ_i is private (Section 2). Allowing bidders to report weights or radii requires a multidimensional treatment beyond our analysis. The intended practical pathway is a seller licensing access to a corpus or continuously updated feed, with allocations specifying contractual usage rights. Deployment would require establishing the coverage parameters independently of bids and validating how well they capture buyers' retrieval needs. Our model provides a foundation for this further work.

Data burning preserves the $(1 - 1/e)$ approximation guarantee on expected welfare but can reduce achieved welfare relative to the rounded allocation before burning and create high variance in a bidder's realized coverage and utility. The granular alternative in Appendix D.4 weakly reduces this variance while preserving the expected coverage and welfare of whole-bundle burning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 2019.

[3] Anish Agarwal, Munther Dahleh, Thibaut Horel, and Maryann Rui. Towards data auctions with externalities. *Games and Economic Behavior*, 148:323–356, 2024.

[4] Arpit Agarwal and Eric Balkanski. Learning-augmented dynamic submodular maximization. *Advances in Neural Information Processing Systems*, 37:14148–14176, 2024.

[5] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, 2023.

[6] Sepehr Assadi and Sahil Singla. Improved truthful mechanisms for combinatorial auctions with submodular bidders. *ACM SIGecom Exchanges*, 18(1):19–27, 2020.

[7] Yair Bartal, Rica Gonen, and Noam Nisan. Incentive compatible multi unit combinatorial auctions. In *Proceedings of the 9th conference on Theoretical aspects of rationality and knowledge*, pages 72–87, 2003.

[8] Jeff Bilmes. Submodularity in machine learning and artificial intelligence. *arXiv preprint arXiv:2202.00132*, 2022.

[9] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the 2019 international conference on management of data*, pages 1535–1552, 2019.

[10] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*, 2024.

[6] The hypothesis test below accounts for sampling error in these estimates.

[11] Xinyan Dai, Xiao Yan, Kaiwen Zhou, Yuxuan Wang, Han Yang, and James Cheng. Convolutional embedding for edit distance. In *proceedings of the 43rd international ACM SIGIR conference on Research and Development in information retrieval*, pages 599–608, 2020.

[12] Shahar Dobzinski and Noam Nisan. Mechanisms for multi-unit auctions. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 346–351, 2007.

[13] Shahar Dobzinski and Michael Schapira. An improved approximation algorithm for combinatorial auctions with submodular bidders. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1064–1073, 2006.

[14] Shahar Dobzinski, Noam Nisan, and Michael Schapira. Approximation algorithms for combinatorial auctions with complement-free bidders. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 610–618, 2005.

[15] Shahar Dobzinski, Noam Nisan, and Michael Schapira. Truthful randomized mechanisms for combinatorial auctions. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 644–652, 2006.

[16] Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 713–722, 2024.

[17] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 144–155, 2024.

[18] Shaddin Dughmi, Tim Roughgarden, and Qiqi Yan. From convex optimization to randomized mechanisms: toward optimal combinatorial auctions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 149–158, 2011.

[19] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45 (4):634–652, 1998.

[20] Uriel Feige. On maximizing welfare when utility functions are subadditive. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 41–50, 2006.

[21] Uriel Feige and Jan Vondrák. Approximation algorithms for allocation problems: Improving the factor of 1-1/e. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 667–676. IEEE, 2006.

[22] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[23] Michael R Garey and David S Johnson. *Computers and intractability*, volume 174. freeman San Francisco, 1979.

[24] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[25] MohammadTaghi Hajiaghayi, Sebastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for llms via retrieval augmented generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=Ujo8V7iXmR`.

[26] Minbiao Han, Jonathan Light, Steven Xia, Sainyam Galhotra, Raul Castro Fernandez, and Haifeng Xu. A data-centric online market for machine learning: From discovery to pricing. *arXiv preprint arXiv:2310.17843*, 2023.

[27] Zellig S Harris. Distributional structure, 1954.

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[30] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[31] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[32] Woojay Jeon. Acoustic neighbor embeddings. *arXiv preprint arXiv:2007.10329*, 2020.

[33] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[34] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

[35] Haim Kaplan and Ron Shamir. The domatic number problem on some perfect graph families. *Information Processing Letters*, 49(1):51–56, 1994.

[36] Subhash Khot, Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. Inapproximability results for combinatorial auctions with submodular utility functions. In *Internet and Network Economics: First International Workshop, WINE 2005, Hong Kong, China, December 15-17, 2005. Proceedings 1*, pages 92–101. Springer, 2005.

[37] Ron Lavi and Chaitanya Swamy. Truthful and near-optimal mechanism design via linear programming. *Journal of the ACM (JACM)*, 58(6):1–24, 2011.

[38] Benny Lehmann, Daniel Lehmann, and Noam Nisan. Combinatorial auctions with decreasing marginal utilities. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 18–28, 2001.

[39] Daniel Lehmann, Liadan Ita Oćallaghan, and Yoav Shoham. Truth revelation in approximately efficient combinatorial auctions. *Journal of the ACM (JACM)*, 49(5):577–602, 2002.

[40] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[41] Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. Dealer: an end-to-end model marketplace with differential privacy. *Proceedings of the VLDB Endowment*, 14(6), 2021.

[42] MatthewWilletts. Resnet15 embeddings of cifar10. `https://github.com/MatthewWilletts/Embeddings`, 2025. Accessed: 2025-01-20.

[43] Sameer Mehta, Milind Dawande, Ganesh Janakiraman, and Vijay Mookerjee. How to sell a data set? pricing policies for data monetization. *Information Systems Research*, 32(4):1281–1297, 2021.

[44] Ahuva Mu'Alem and Noam Nisan. Truthful approximation mechanisms for restricted combinatorial auctions. *Games and Economic Behavior*, 64(2):612–631, 2008.

[45] Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

[46] National Football League and Genius Sports Group. National football league taps genius sports group as exclusive distributor of official league data. `https://investors.geniussports.com/news/`, April 2021. Press release; Genius Sports Group to distribute NFL's play-by-play, betting, and Next Gen Stats data feeds worldwide.

[47] National Football League and Genius Sports Group. Nfl extends and broadens partnership with genius sports through 2029. https://sccgmanagement.com/sccg-news/2025, June 2025. Press release on extended strategic partnership, extending NFL and Genius Sports collaboration through the 2029 season.

[48] NIQ Global Intelligence plc. 2025 annual report (Form 10-K). https://www.sec.gov/Archives/edgar/data/2054696/000162828026012572/niq-20251231.htm, February 2026.

[49] Noam Nisan and Amir Ronen. Computationally feasible vcg mechanisms. *Journal of Artificial Intelligence Research*, 29:19–47, 2007.

[50] Refinitiv and Tradeweb Markets. Refinitiv and tradeweb markets announce extension and expansion of exclusive redistribution agreement. https://www.tradeweb.com/newsroom/media-center/, November 2018. Press release on the expanded multi-year redistribution agreement providing additional Tradeweb market data across Refinitiv platforms.

[51] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

[52] Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. Cs-shapley: class-wise shapley values for data valuation in classification. *Advances in Neural Information Processing Systems*, 35: 34574–34585, 2022.

[53] SetFit. 20_newsgroups. https://huggingface.co/datasets/SetFit/20_newsgroups, 2025. Accessed: 2025-01-20.

[54] Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. *arXiv preprint arXiv:2405.05905*, 2024.

[55] Haoran Sun, Yurong Chen, Siwei Wang, Wei Chen, and Xiaotie Deng. Mechanism design for llm fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*, 2024.

[56] Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A Dunnmon, James Zou, and Daniel L Rubin. Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset. *Scientific reports*, 11(1):8366, 2021.

[57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[58] uoft_cs. cifar10. https://huggingface.co/datasets/uoft-cs/cifar10, 2025. Accessed: 2025-01-20.

[59] Jan Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 67–74, 2008.

[60] Jiachen T Wang, Prateek Mittal, and Ruoxi Jia. Efficient data shapley for weighted nearest neighbor algorithms. *arXiv preprint arXiv:2401.11103*, 2024.

A Motivation for Exclusive Data Selling

Exclusive licensing occurs at the access and distribution layers of data markets. For instance, the National Football League (NFL) designates Genius Sports (a sports data and technology company) as the exclusive distributor of official live game data as well as advanced statistics drawn from that NFL data (called Next Gen Stats) [46]. In fact, this agreement has been extended to 2029 [47]. Beyond sports, exclusivity is also employed at the distribution layer. For instance, in financial markets, the electronic trading venue Tradeweb has designated LSEG Data & Analytics (formerly Refinitiv), a major financial-information provider, as the exclusive redistributor of specified Tradeweb market-data products [50]. NIQ reports an exclusive data partnership with Fetch Rewards and has deployed Ask Arthur, a generative-AI tool for accessing its data and insights [48]. These examples motivate contractual restrictions on data access; our model imposes the specific constraint that each data point is licensed to at most one buyer. Cross-buyer externalities can make restricting rivals' access valuable [3], but are not included in our valuation model.

B Useful Facts and Lemmas

Fact B.1. $\forall x \in [0,1]$ *we have*

$$1 - e^{-x} \geq (1 - 1/e)x \tag{6}$$

Proof. Following a standard approach, we define the function $f(x) = 1 - e^{-x} - (1 - 1/e)x$. Therefore, we only need to show that $f(x) \geq 0\ \forall x \in [0,1]$. Note that $f(0) = f(1) = 0$. Further, the derivative is $f'(x) = e^{-x} - (1 - 1/e)$. Note that $f'(x) = 0$ at $x = \ln \frac{e}{e-1} \in [0,1]$. Further, by the form of the derivative $f'(x) = e^{-x} - (1 - 1/e)$, it is clear that in the interval $[0,1]$ the function f increases from 0 at $x = 0$ until $\ln \frac{e}{e-1}$ and then decreases until it has a value of 0 again at $x = 1$. Therefore, $f(x) \geq 0\ \forall x \in [0,1]$. □

Lemma B.2. *For any function f that can be written as*

$$f(Z) = \beta_1 f_1(Z) + \beta_2 f_2(Z) + \cdots + \beta_m f_m(Z) \quad , \quad Z \in S \tag{7}$$

where S is some set and $\beta_i > 0\ , \forall i \in [m]$, the value of $f_i(.)$ at any optimal solution is monotonic in β_i. I.e., let Z^* be the optimal solution for f given (β_i, β_{-i}) and let Z'^* be the optimal solution for for f given (β'_i, β_{-i}), then if $\beta_i > \beta'_i$ then $f_i(Z^*) \geq f_i(Z'^*)$.

Proof. The proof is by contradiction, so we will assume that $f_i(Z^*) < f_i(Z'^*)$ with $\beta_i > \beta'_i$. Note first the welfare under different bids can be written as

$$f(Z^*) = \beta_i \cdot f_i(Z^*) + \underbrace{\sum_{k \in [m], k \neq i} \beta_k \cdot f_k(Z^*)}_{V} \tag{8}$$

$$f(Z'^*) = \beta'_i \cdot f_i(Z'^*) + \underbrace{\sum_{k \in [m], k \neq i} \beta_k \cdot f_k(Z'^*)}_{V'} \tag{9}$$

We now have the following claim

Claim 1. $f_i(Z'^*) - f_i(Z^*) \leq \frac{V - V'}{\beta_i}$.

Proof. Since Z^* is an optimal solution then it follows that by using the solution Z'^* we would not get a higher welfare under β_i. Therefore, we have

$$\beta_i \cdot f_i(Z^*) + V \geq \beta_i \cdot f_i(Z'^*) + V' \tag{10}$$

By re-arranging the terms the inequality follows. □

Now we reach the following claim

Claim 2. $\beta'_i \cdot f_i(Z^*) + V > \beta'_i \cdot f_i(Z'^*) + V'$.

Proof. By taking the difference we get

$$\beta_i' \cdot f_i(Z^*) + V - \Big(\beta_i' \cdot f_i(Z'^*) + V'\Big) = \beta_i' \cdot \Big(f_i(Z^*) - f_i(Z'^*)\Big) + V - V' \tag{11}$$

$$\geq -\beta_i' \cdot \Big(\frac{V - V'}{\beta_i}\Big) + V - V' \qquad \text{(This follows by Claim 1)} \tag{12}$$

$$= \Big(V - V'\Big) \cdot \Big(1 - \frac{\beta_i'}{\beta_i}\Big) \tag{13}$$

$$> 0 \tag{14}$$

Where the last step follows since by Claim 1 we have $V - V' > 0$ since by assumption $f_i(Z'^*) - f_i(Z^*) > 0$ and $\beta_i > 0$. Further, since $\beta_i' < \beta_i$ then $1 - \frac{\beta_i'}{\beta_i} > 0$. Therefore, the multiplication of two positive number results in a positive. □

Looking at the last Claim 2 we see that using the solutuion Z^* for (β_i', β_{-i}) would result in a strictly higher value for $f(.)$ than using Z'^*, which contradicts the assumption that Z'^* is an optimal solution for (β_i', β_{-i}). Therefore, we must have $f_i(Z^*) \geq f_i(Z'^*)$. □

C Omitted Proofs of Section 3

C.1 Proof of Proposition 3.2

The reduction is from the dominating set problem. In the dominating set decision problem, we are given a graph $G = (V, E)$ and an integer k and we should decide if there exists a subset of vertices $V' \subset V$ such that $|V'| \leq k$ and every vertex in V is either in V' or has an edge connected to V'. The problem is known to be NP-complete [23].

The vertices will be the points in the database $\mathcal{DB}$ and the distance between any pair of points (vertices) will be the path distance. We let a bidder i have $r_{ij} = 1\ , \forall j \in \mathcal{DB}$ and set $b_i = 1$. We set all prices for all points to $p_j = p\ , \forall j \in \mathcal{DB}$ where $0 < p < \frac{1}{n}$. Note that for any non-trivial instance we have $n > k$ and therefore $\frac{1}{n} < \frac{1}{k}$. Further, the weights are the same for all points, i.e., we have $w_{ij} = \frac{1}{n}\ , \forall j \in \mathcal{DB}$. We will prove that the graph G has a dominating set of size at most k if and only if there exists a subset $V' \subset V$ such that $u_i(V') \geq 1 - pk$.

The first direction is easy to show. In particular, suppose that V' is a dominating set of size at most k then clearly $u_i(V') \geq 1 - pk$.

For the other direction, we will show that if G does not have a dominating set of size at most k then there exists no subset V' such that $u_i(V') \geq 1 - pk$. First, if $|V'| > k$ then $u_i(V') \leq 1 - p \cdot (k+1) < 1 - pk$. Therefore, we must have $|V'| = k - \ell$ where ℓ is a non-negative integer. We will argue that the size of the covered subset, denoted by $N(V')$, satisfies $N(V') \leq n - \ell - 1$. Otherwise, V' can be used to construct a dominating set of size at most k. To see that note, we can add at most ℓ uncovered vertices and we would cover the full graph using at most k many vertices which is a dominating set of size at most k. Thus, we must $|V'| = k - \ell$ and $N(V') \leq n - \ell - 1$. Now we upper bound the utility

$$\begin{aligned}
u_i(V') &\leq \frac{n - \ell - 1}{n} - p \cdot (k - \ell) \\
&= 1 - pk + p\ell - \frac{\ell + 1}{n} \\
&= 1 - pk - \frac{1}{n} + \ell(p - \frac{1}{n}) \\
&\leq 1 - pk - \frac{1}{n} \quad \text{(since } (p - \frac{1}{n}) < 0 \text{ as } p < \frac{1}{n}) \\
&< 1 - pk
\end{aligned}$$

C.2 Proof of the First Claim in Theorem 3.3

The reduction from dominating set to our problem is as follows. We will have two bidders ($m = 2$). For simpliclity, we set the bid values to $b_1 = b_2 = 1$. The weights for all points and bidders are set to the same value, it follows that $w_{ij} = \frac{1}{n}$, $\forall j \in \mathcal{DB}, \forall i \in [m]$. Each bidder has the same radius value across all points. In particular, we set $r_1 = 1$ and $r_2 = 0$. Given the graph $G = (V, E)$ from the dominating set problem, each vertex $v \in V$ will represent a point in the database $\mathcal{DB}$. For any two points (vertices) $v, v' \in \mathcal{DB}$ we set $d(v, v')$ to the path distance between $v, v' \in V$. Therefore, $d(v, v') = 0$ only if $v = v'$. Moreover, $d(v, v') = 1$ only if there exists an edge between v and v'. For all other cases, $d(v, v') > 1$.

Now for a subset of $V' \subset V$ define $N(V')$ to be the number the vertices of V' union the vertices that have an edge connected to a vertex in V'. It follows that if we give vertices V_1 and V_2 to the first and second bidder, respectively, then the welfare would be:

$$\begin{aligned} \mathcal{W} &= v_1(V_1) + v_2(V_2) \\ &= \frac{N(V_1) + |V - V_1|}{n} \end{aligned} \tag{15}$$

Therefore, to maximize utility we solve the following maximization:

$$\max_{V_1 \subset V} \ \frac{N(V_1) + |V - V_1|}{n} \tag{16}$$

The following claim proves the theorem.

Claim 3. G has a dominating set of size less than or equal to k if and only if the optimal welfare is at least $\mathcal{W}^* \geq 2 - \frac{k}{n}$.

Proof. Suppose G has a dominating set of size $\leq k$, then $\mathcal{W}^* \geq 2 - \frac{k}{n}$. To see that, set V_1 equal to the dominating set. Therefore, it follows that $N(V_1) = n$ and $|V - V_1| \geq n - k$ so by plugging these values in (15) we get a welfare of at least $2 - \frac{k}{n}$.

For the other direction. We first show that if a solution has a welfare $\mathcal{W} \geq 2 - \frac{k}{n}$ then $|V_1| \leq k$. This is the case since if $|V_1| > k$ then $|V - V_1| < n - k$ and therefore using (15) we get $\mathcal{W} < 2 - \frac{k}{n}$.

Now we argue that if $|V_1| = k - \ell \leq k$ where $\ell \geq 0$ is a non-negative integer then $N(V_1) \geq n - \ell$. Suppose not, i.e., $N(V_1) \leq n - \ell - 1$ then we have

$$\begin{aligned} \mathcal{W} &= \frac{N(V_1) + |V - V_1|}{n} \\ &\leq \frac{n - \ell - 1 + n - (k - \ell)}{n} \\ &= \frac{2n - k - 1}{n} \\ &< 2 - \frac{k}{n} \end{aligned}$$

Therefore, it must be that if $|V_1| = k - \ell$ then $N(V_1) \geq n - \ell$, but this implies that there exists a dominating set of size k. To see that, note that we can construct a dominating set from V_1 by adding (at most) the remaining ℓ many uncovered vertices, this leads to $N(V_1) + \ell = n - \ell + \ell = n$ using only $k - \ell + \ell = k$ many vertices which is a dominating set of size at most k. □

C.3 Proof of Second Claim in Theorem 3.3

Proof. The reduction is from the domatic number problem. First, given a graph $G = (V, E)$ a domatic partition of the graph is a set of disjoint vertices $V_1, \ldots, V_D$ such that for each $i \in \{1, \ldots, D\}$ the subset of vertices V_i is a dominating set of G. In the decision version of the domatic number problem we are given a graph $G = (V, E)$ and an integer k and we should decide if there is a domatic partition of G into at least k many subsets. In Kaplan and Shamir [35] it is shown that the domatic number problem is NP-complete for $k \geq 3$.

We will now reduce any instance of the domatic number problem for $k \geq 3$ to our welfare maximization problem. First, note for a given value of k if the graph has a domatic partition for $k' > k$ then it has a domatic partition of size k this follows since given a domatic partition $\{V_1, \ldots, V_{k'}\}$ we can reduce the number of subsets (possibly recursively) until we have a total of k many subsets by simply merging two subsets since the union of two dominating sets is also a dominating set.

The reduction is as follows. We will have k bidders and for simplicity, we will set $b_1 = b_2 = \cdots = b_k = b > 0$. Given the graph $G = (V, E)$ the vertices V will represent the database $\mathcal{DB}$. All weights $w_{ij} = \frac{1}{n}, \forall j \in \mathcal{DB}, \forall i \in [k]$. The distance between two points (vertices) will equal the path distance, hence vertices that have an edge between them are at a distance of 1. For our coverage function, we set the radius $r_{ij} = r = 1, \forall j \in \mathcal{DB}, \forall i \in [k]$. Therefore, if a bidder receives a subset of points $V' \subset V$ then his utility would be $v_i(V') = b\frac{N(V')}{n}$ where $N(V')$ is the total number of vertices that are either in V' or adjacent to a vertex in V'. We will now show that there is a domatic partition into k subsets if and only if the welfare is at least $b \cdot k$ in our problem.

First, suppose that there exists a domatic partition into k subsets $V_1, V_2, \ldots, V_k$ then it follows since each subset is dominating that the welfare is at least $\mathcal{W} = v_1(V_1) + v_2(V_2) + \cdots + v_k(V_k) = b \cdot (\frac{n}{n} + \frac{n}{n} + \cdots + \frac{n}{n}) = b \cdot k$.

For the other direction, suppose that we have found an allocation with $\mathcal{W} \geq bk$. By definition of the valuation function, it follows that for any bidder and any allocation $v_i(V') = b\frac{N(V')}{n} \leq b\frac{n}{n} = b$. But since $w \geq b \cdot k$ it follows that each bidder covers the entire set of points (vertices). Further, since we allocate each point (vertex) to only one bidder it follows that we have a partition of the graph into k disjoint subset of vertices each covering the entire set of points, i.e., a partition into k dominating sets. □

C.4 Proof of Lemma 3.4

Proof. To prove the lemma, we will show that an optimal solution is feasible and that objective (3a) gives its welfare value. First, note that an optimal solution which consists of a collection of datasets $\{\mathcal{D}_1^*, \ldots, \mathcal{D}_m^*\}$ corresponds to a set of integral assignments X_{ij}^* which for a point $j \in \mathcal{DB}$ and bidder $i \in [m]$ would equal 1 only if bidder i received point j and would be zero otherwise.

Since X_{ij}^* is an integral assignment ($X_{ij}^* \in \{0, 1\}$) it follows that $0 \leq X_{ij}^* \leq 1$ satisfying constraint (3b). Further, since no point can be assigned to more than one bidder we have $\sum_{i \in [m]} X_{ij}^* \leq 1$ satisfying constraint (3c). Moreover, the values X_{ij}^* would induce a collection of coverage values C_{ij}^*. Since $C_{ij}^* \in \{0, 1\}$ then we have $0 \leq C_{ij}^* \leq 1$ satisfying (3b). Finally, a point j is covered if and only if at least one point $j' \in N_i(j)$ is given to bidder i. If this is the case for a point j then $\sum_{j' \in N_i(j)} X_{ij'}^* \geq 1$ and $C_{ij}^* = 1$. If no point in the neighborhood of j is allocated to bidder i then we must have $\sum_{j' \in N_i(j)} X_{ij'}^* = 0$ and $C_{ij}^* = 0$. In both cases, we have $C_{ij}^* \leq \sum_{j' \in N_i(j)} X_{ij'}^*$ satisfying constraint (3b). From the above, it follows that $\sum_{i \in [m]} \theta_i \cdot \left(\sum_{j \in \mathcal{DB}} w_{ij} \cdot C_{ij}^* \right) = \mathcal{W}^*(\boldsymbol{b})$.

Since the LP also enables fractional assignments in $[0, 1]$ not just integral assignments in $\{0, 1\}$ for $\bar{X}_{ij}$ and $\bar{C}_{ij}$ it follows that $\text{OPT}_{\text{LP}}(\boldsymbol{b}) \geq \mathcal{W}^*(\boldsymbol{b})$. □

C.5 Proof of Lemma 3.5

Proof.

$$\begin{aligned}
\mathbb{E}(C_{ij}=1) &= 1-\mathbb{P}(C_{ij}=0) && (17)\\
&= 1-\mathbb{P}(\forall j'\in N_i(j), X_{ij'}=0) && (18)\\
&= 1-\prod_{j'\in N_i(j)}\mathbb{P}(X_{ij'}=0) \qquad \text{(Since the rounding is independent)}\\
&= 1-\prod_{j'\in N_i(j)}(1-\bar{X}_{ij'}) && (19)\\
&\geq 1-\prod_{j'\in N_i(j)}\exp(-\bar{X}_{ij'}) \qquad \text{(Since } 1-x\leq e^{-x}\text{)}\\
&= 1-\exp(-\sum_{j'\in N_i(j)}\bar{X}_{ij'}) && (20)\\
&\geq 1-\exp(-\bar{C}_{ij}) \qquad \text{(Since } \bar{C}_{ij}\leq \sum_{j'\in N_i(j)}\bar{X}_{ij'} \text{ by constraint (3b))}\\
&\geq (1-1/e)\bar{C}_{ij} \qquad \text{(By Fact B.1 and (3b) we have } 0\leq \bar{C}_{ij}\leq 1\text{)}
\end{aligned}$$

Line (19) proves Equation (4). The second point follows immediately by definition (1) and Line (21). □

C.6 Proof of Theorem 3.6

Proof. The resulting welfare is

$$\begin{aligned}
\mathbb{E}\Big[\sum_{i\in[m]}\theta_i\cdot\mathcal{C}_i\Big] &= \sum_{i\in[m]}\theta_i\cdot\mathbb{E}[\mathcal{C}_i] && (21)\\
&\geq (1-1/e)\sum_{i\in[m]}\theta_i\cdot\bar{\mathcal{C}}_i && (22)\\
&= (1-1/e)\text{OPT}_{\text{LP}}(\boldsymbol{b}) && (23)\\
&\geq (1-1/e)\mathcal{W}^*(\boldsymbol{b}) && (24)
\end{aligned}$$

□

D Omitted Proofs of Section 4

D.1 Proof of Lemma 4.2

Proof. We will focus on a single bidder $i\in[m]$ and hold the other buyer's bids $\boldsymbol{b}_{-i}$ fixed. Bidder i will receive datasets $\mathcal{D}$ and $\mathcal{D}'$ under bid profiles $(b,\boldsymbol{b}_{-i})$ and $(b',\boldsymbol{b}_{-i})$, respectively. Note that we have dropped the subscript i in $\mathcal{D}$,$\mathcal{D}'$,b, and b' for ease of notation. We also overload the notation and have $p(b)=p(\mathcal{D})$ and $p(b')=p(\mathcal{D}')$.

For the mechanism to be incentive compatible the following conditions must hold

$$\begin{aligned}
b\cdot\mathbb{E}[\mathcal{C}_i(\mathcal{D})]-\mathbb{E}[p(b)] &\geq b\cdot\mathbb{E}[\mathcal{C}_i(\mathcal{D}')]-\mathbb{E}[p(b')] && (25)\\
b'\cdot\mathbb{E}[\mathcal{C}_i(\mathcal{D}')]-\mathbb{E}[p(b')] &\geq b'\cdot\mathbb{E}[\mathcal{C}_i(\mathcal{D})]-\mathbb{E}[p(b)] && (26)
\end{aligned}$$

From (25) and (26) the following holds

$$(b-b')\cdot(\mathbb{E}[\mathcal{C}_i(\mathcal{D})]-\mathbb{E}[\mathcal{C}_i(\mathcal{D}')])\geq 0 \qquad (27)$$

The above inequality immediately implies that the allocation must be monotonic. I.e., $\mathbb{E}[\mathcal{C}_i(\mathcal{D})]\geq\mathbb{E}[\mathcal{C}_i(\mathcal{D}')$ if and only if $b\geq b'$. This proves the first part of the theorem.

To find the payment, first note that (25) and (26) imply the following inequalities:

$$b' \cdot \Big(\mathbb{E}[\mathcal{C}_i(\mathcal{D})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}')] \Big) \leq \mathbb{E}[p(b)] - \mathbb{E}[p(b')] \tag{28}$$

$$\leq b \cdot \Big(\mathbb{E}[\mathcal{C}_i(\mathcal{D})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}')] \Big) \tag{29}$$

We will now set the payment rule. Note that we will follow a deterministic payment rule, thus the payment value will always equal its expectation. The possible bid values belong to the set $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{|\Lambda|}\}$, where $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{|\Lambda|}$ and $\lambda_0 = 0$ is a dummy value. We denote by $\mathcal{D}_t$ the dataset which bidder i receives if he bids with value λ_t for some $t \in \{1, 2, \ldots, |\Lambda|\}$ with the other bidders $\boldsymbol{b}_{-i}$ held fixed. Further, we set $\mathcal{D}_0 = \emptyset$ hence $\mathbb{E}[\mathcal{C}_i(\mathcal{D}_0)] = 0$ and we also set $p(0) = 0$. At a bid value of b the payment for bidder i is the following:

$$p(b) = \sum_{\substack{t \in \{0,1,2,\ldots,|\Lambda|-1\}: \\ \lambda_t < b}} \lambda_t \cdot (\mathbb{E}[\mathcal{C}_i(\mathcal{D}_{t+1})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}_t)]) \tag{30}$$

Without loss of generality, we assume that for the two bid values b and b' we have $b > b'$. We will now prove that Inequalities (28) and (29) are satisfied by this payment.

$$b' \cdot \Big(\mathbb{E}[\mathcal{C}_i(\mathcal{D})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}')] \Big) \tag{31}$$

$$= b' \Big(\sum_{\substack{t \in \{1,2,\ldots,|\Lambda|-1\}: \\ b' \leq \lambda_t < b}} (\mathbb{E}[\mathcal{C}_i(\mathcal{D}_{t+1})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}_t)]) \Big) \qquad \text{(since it is a telescoping sum)}$$

$$\leq \sum_{\substack{t \in \{1,2,\ldots,|\Lambda|-1\}: \\ b' \leq \lambda_t < b}} \lambda_t \cdot (\mathbb{E}[\mathcal{C}_i(\mathcal{D}_{t+1})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}_t)]) \tag{32}$$

$$= p(b) - p(b') \quad \text{(by Equation (30))} \tag{33}$$

This proves inequality (28). Further, from Lines (32) and (33) we have

$$\begin{aligned} p(b) - p(b') &= \sum_{\substack{t \in \{1,2,\ldots,|\Lambda|-1\}: \\ b' \leq \lambda_t < b}} \lambda_t \cdot (\mathbb{E}[\mathcal{C}_i(\mathcal{D}_{t+1})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}_t)]) \\ &\leq b \sum_{\substack{t \in \{1,2,\ldots,|\Lambda|-1\}: \\ b' \leq \lambda_t < b}} (\mathbb{E}[\mathcal{C}_i(\mathcal{D}_{t+1})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}_t)]) \\ &= b \cdot \Big(\mathbb{E}[\mathcal{C}_i(\mathcal{D})] - \mathbb{E}[\mathcal{C}_i(\mathcal{D}')] \Big) \end{aligned}$$

This proves Inequality (29). This shows a payment rule that can be efficiently computed and that would lead to an incentive compatible mechanism. Note that we assume that the expected value of the allocation can be found for any bid value. Our main algorithm (Algorithm 2) in Section 4 gives such an allocation.

We now prove that individual rationality holds. Let D_{θ_i} be the dataset allocated when i bids with value θ_i then we have

$$\begin{aligned} \theta_i \, \mathbb{E}[\mathcal{C}_i(D_{\theta_i})] - p(\theta_i) &\geq \theta_i \, \mathbb{E}[\mathcal{C}_i(\mathcal{D}_1)] - p(\lambda_1) \quad \text{(since the mechanism is incentive compatible)} \\ &= \theta_i \, \mathbb{E}[\mathcal{C}_i(\mathcal{D}_1)] - \lambda_0 \, \mathbb{E}[\mathcal{C}_i(\mathcal{D}_1)] \qquad \text{(using (30))} \\ &= (\theta_i - \lambda_0) \, \mathbb{E}[\mathcal{C}_i(\mathcal{D}_1)] \\ &\geq 0 \end{aligned}$$

□

D.2 Proof of Fact 4.3

Proof. This can be shown by an example. Specifically, suppose that we only have three points and two bidders ($n = 3$ and $m = 2$). Let the radius values be the same for all bidders and points, i.e.,

$r_{ij} = r,\ \forall j \in \mathcal{DB}, \forall i \in [m]$. Further, suppose that all of the points are in each others neighborhoods, i.e., $\forall j, j' \in \mathcal{DB} : d(j, j') \leq r$. Moreover, we set $w_{ij} = \frac{1}{3}\ \forall j \in \mathcal{DB}, \forall i \in [m]$.

Suppose the second bidder always has $b_2 = 1$ but the first bidder has $b_1 = 2$ and $b'_1 = 1$. Consider the following LP optimal solution under b'_1:

$$\begin{array}{lll} \bar{X}'_{11} = 1 & \bar{X}'_{12} = 0 & \bar{X}'_{13} = 0 \\ \bar{X}'_{21} = 0 & \bar{X}'_{22} = 1 & \bar{X}'_{23} = 1 \end{array}$$

Note that this leads to an optimal coverage of 1 for both bidders in the LP solution and when rounded, i.e., $\bar{\mathcal{C}}_1 = \bar{\mathcal{C}}_2 = \mathbb{E}[\mathcal{C}_1] = \mathbb{E}[\mathcal{C}_2] = 1$.

Now consider the following LP optimal solution under b_1:

$$\begin{array}{lll} \bar{X}_{11} = \frac{1}{2} & \bar{X}'_{12} = \frac{1}{2} & \bar{X}_{13} = \frac{1}{2} \\ \bar{X}_{21} = \frac{1}{2} & \bar{X}'_{22} = \frac{1}{2} & \bar{X}_{23} = \frac{1}{2} \end{array}$$

Although the LP solution is optimal for both bidders, the rounded allocation is smaller in expectation. This follows since by Equation (4) we have $\mathbb{E}[\mathcal{C}_1] = \mathbb{E}[\mathcal{C}_2] = 1 - \frac{1}{2^3} < 1$. □

D.3 Proof of Theorem 4.4

Proof. First we start with the following claim that shows that the LP coverage values are monotonic.

Proof of Lemma 4.5. This follows immediately as a special case of Lemma B.2. □

Now we establish the validity of the probability values ρ_i for each bidder $i \in [m]$.

Claim 4. For each bidder $i \in [m]$ we have $\rho_i \in [0, 1]$.

Proof. First, if $\bar{\mathcal{C}}_i = 0$ then it follows that $\rho_i = 0$ and is therefore in $[0, 1]$. So we focus now on the case where $\bar{\mathcal{C}}_i > 0$, we have the following:

$$\begin{aligned} \rho_i &= (1 - 1/e) \cdot \frac{\bar{\mathcal{C}}_i}{\mathbb{E}[\mathcal{C}_i]} && (34) \\ &\leq (1 - 1/e) \cdot \frac{1}{(1 - 1/e)} \qquad \text{(By Inequality (5))} && (35) \\ &= 1 && (36) \end{aligned}$$

Further, it is clear that we would have $\rho_i > 0$. □

We now introduce the following lemma which lower bounds the final expected coverage $\mathbb{E}[\mathcal{C}'_i]$ a bidder i receives from Algorithm 2.

Proof of Lemma 4.6. Note that when a non-empty dataset is allocated to the bidder the expected coverage will be the same as the one from Algorithm 1 which is $\mathbb{E}[\mathcal{C}_i]$. Therefore, we have

$$\begin{aligned} \mathbb{E}[\mathcal{C}'_i] &= (1 - \rho_i) \cdot 0 + \rho_i \cdot \mathbb{E}[\mathcal{C}_i] && (37) \\ &= (1 - 1/e) \cdot \frac{\bar{\mathcal{C}}_i}{\mathbb{E}[\mathcal{C}_i]} \cdot \mathbb{E}[\mathcal{C}_i] && (38) \\ &= (1 - 1/e) \cdot \bar{\mathcal{C}}_i && (39) \end{aligned}$$

□

Now we can prove the main claims of the theorem. We start by proving the approximation ratio. By the above lemma we immediately have $\sum_{i\in[m]} \theta_i \, \mathbb{E}[\mathcal{C}'_i] = (1 - 1/e) \cdot \sum_{i\in[m]} \theta_i \bar{\mathcal{C}}_i = (1 - 1/e) \cdot \text{OPT}_{\text{LP}}(\boldsymbol{b}) \geq (1 - 1/e) \cdot \mathcal{W}^*(\boldsymbol{b})$.

Now to prove monotonicity, note that by the Lemma each bidder receives exactly coverage of $(1-1/e)\cdot\bar{\mathcal{C}}_i$. Therefore, if a bidder deviates from b_i to b'_i then we have $(1-1/e)\cdot\bar{\mathcal{C}}_i \geq (1-1/e)\cdot\bar{\mathcal{C}}'_i$ if $b_i \geq b'_i$ by Lemma 4.5. □

D.4 Granular Data Burning

The final step of Algorithm 2 either allocates the entire rounded dataset or discards it. We give a granular implementation that weakly reduces the resulting coverage variance while preserving expected coverage. Fix a bid profile and a bidder i, and let $\mathcal{D}_i$ be the rounded dataset before burning. Write $\mu_i = \mathbb{E}[\mathcal{C}_i(\mathcal{D}_i)]$ and use the same probability $\rho_i = (1 - 1/e)\bar{\mathcal{C}}_i/\mu_i$ as in Algorithm 2, with $\rho_i = 0$ when $\bar{\mathcal{C}}_i = 0$. Conditional on a realized $\mathcal{D}_i$, set $x = \mathcal{C}_i(\mathcal{D}_i)$ and $y = \rho_i x$.

Fix a deterministic ordering of the database, independent of bids, and restrict it to $\mathcal{D}_i$. Let P_i^j contain its first j points, for $j = 0, \ldots, k$, where $k = |\mathcal{D}_i|$. Thus $P_i^0 = \emptyset$, $P_i^k = \mathcal{D}_i$, and monotonicity of coverage gives

$$0 = c_0 \le c_1 \le \cdots \le c_k = x, \qquad c_j = \mathcal{C}_i(P_i^j).$$

Choose the nearest attained coverage levels bracketing the target:

$$a = \max\{c_j : c_j \le y\}, \qquad b = \min\{c_j : c_j \ge y\}.$$

These levels exist since $0 \le y \le x$. If $a = b = y$, allocate a prefix attaining y deterministically, choosing the shortest such prefix. In particular, allocate nothing when $y = 0$. Otherwise, allocate a prefix attaining a with probability $(b-y)/(b-a)$ and one attaining b with probability $(y-a)/(b-a)$. Let $\widetilde{\mathcal{D}}_i$ be the allocated prefix and $\widetilde{\mathcal{C}}_i = \mathcal{C}_i(\widetilde{\mathcal{D}}_i)$. All discarded points remain unallocated. The construction requires at most $|\mathcal{D}_i| + 1$ coverage evaluations and is polynomial time under our explicit coverage model.

Proposition D.1. *Replacing the final step of Algorithm 2 by granular burning preserves feasibility and each bidder's expected coverage, and weakly reduces each bidder's coverage variance. With the payment rule* (30)*, the mechanism remains incentive compatible and individually rational in expectation, with the same expected welfare and* $(1 - 1/e)$ *approximation guarantee.*

Proof. For $a < b$, the conditional mean is

$$\mathbb{E}[\widetilde{\mathcal{C}}_i \mid \mathcal{D}_i] = \frac{a(b-y) + b(y-a)}{b-a} = y.$$

The same equality holds when $a = b = y$. Let $\mathcal{C}_i'$ denote coverage after the original whole-bundle burning step. Its conditional mean is y and its conditional variance is $y(x-y)$. The granular rule instead has

$$\begin{aligned}\mathrm{Var}(\widetilde{\mathcal{C}}_i \mid \mathcal{D}_i) &= (y-a)(b-y) \\ &\le y(x-y) = \mathrm{Var}(\mathcal{C}_i' \mid \mathcal{D}_i),\end{aligned}$$

The inequality follows from $a \ge 0$ and $b \le x$, and also holds in the deterministic case. Since the conditional means agree for every rounded dataset, the law of total variance gives $\mathrm{Var}(\widetilde{\mathcal{C}}_i) \le \mathrm{Var}(\mathcal{C}_i')$. Moreover,

$$\mathbb{E}[\widetilde{\mathcal{C}}_i] = \mathbb{E}[\rho_i \mathcal{C}_i(\mathcal{D}_i)] = \rho_i \mu_i = (1 - 1/e)\bar{\mathcal{C}}_i.$$

This equality holds for every bid profile, so expected coverage as a function of bids and every term in payment rule (30) are unchanged. Consequently, each report gives the same expected utility as before, preserving incentive compatibility and individual rationality in expectation. Since the payment is deterministic for a fixed bid profile, utility variance is θ_i^2 times coverage variance and therefore also weakly decreases. Expected welfare and its approximation guarantee are unchanged as well. Finally, each allocated prefix is a subset of its original rounded dataset, so allocations remain disjoint and feasible. □

E Revenue Maximizing Auction using Virtual Welfare

We can convert our welfare maximizing results to a $(1 - 1/e)$-approximate revenue maximizing auction using the standard Myerson virtual valuation approach. The revenue maximizing mechanism requires knowledge of the distribution of buyer types, and in this setting, this corresponds to a knowledge of the distribution of θ. We will assume that each buyer's valuation θ_i is independent of all other buyers, and that $\theta_i \sim F_i(\theta)$ where $F_i(\theta_i)$ is the cumulative distribution function, denote by $f_i(\theta_i)$ the probability mass function. Note that throughout this section, we differ from the notation

in the main body of the paper in that we refer to the set of bidder types as Θ_i for each type where $\boldsymbol{\Theta} = \times_i \Theta_i$ and $\boldsymbol{\Theta}_{-i} = \times_{j \neq i} \Theta_j$. Additionally, we refer to the ordered discrete set of types for bidder i as $\{\theta_{i,1}, \theta_{i,2}, ..., \theta_{i,|\Theta|}\}$. Finally, we denote the coverage that bidder i received given bidder profile $(\theta_i, \boldsymbol{\theta}_{-i})$ as $\mathcal{C}_i(\theta_i, \boldsymbol{\theta}_{-i}))])$. We can then write the expected revenue from a truthful mechanism as follows:

$$\mathbb{E}_{F(\boldsymbol{\theta})}\left[\sum_{i\in\{m\}} p_i(\theta_i, \boldsymbol{\theta}_{-i})\right] = \sum_{i\in\{m\}} \mathbb{E}_{F(\boldsymbol{\theta})}[p_i(\theta_i, \boldsymbol{\theta}_{-i})] \tag{40}$$

By equation 30,

$$p(\theta_{i,t}, \boldsymbol{\theta}_{-i}) = \sum_{\substack{t\in\{0,1,2,...,|\Theta_i|\}:\\ \theta_{i,t}<\theta_i}} \theta_{i,t} \cdot (\mathbb{E}[\mathcal{C}_i(\theta_{i,t+1}, \boldsymbol{\theta}_{-i})] - \mathbb{E}[\mathcal{C}_i(\theta_{i,t}, \boldsymbol{\theta}_{-i})])$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{F(\boldsymbol{\theta})}[p_i(\theta_i, \boldsymbol{\theta}_{-i})] = \mathbb{E}_{F(\boldsymbol{\theta})}\left[\sum_{\substack{t\in\{0,1,2,...,|\Theta_i|\}:\\ \theta_{i,t}<\theta_i}} \theta_{i,t} \cdot (\mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i})] - \mathbb{E}[\mathcal{C}_i(\theta_{i,t}, \boldsymbol{\theta}_{-i}))])\right] \\
&= \sum_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f_{-i}(\boldsymbol{\theta}) \sum_{\substack{t\in\{0,1,2,...,|\Theta_i|\}:\\ \theta_{i,t}<\theta_i}} \theta_{i,t} \cdot (\mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i}))] - \mathbb{E}[\mathcal{C}_i(\theta_{i,t}, \boldsymbol{\theta}_{-i})]) \\
&= \sum_{\boldsymbol{\theta}_{-i}\in\boldsymbol{\Theta}_{-i}} f_{-i}(\boldsymbol{\theta}_{-i}) \left(\sum_{\theta_i\in\Theta_i} f(\theta_i) \sum_{\substack{t\in\{0,1,2,...,|\Theta_i|\}:\\ \theta_{i,t}<\theta_i}} \theta_{i,t} \cdot (\mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i}))] - \mathbb{E}[\mathcal{C}_i(\theta_{i,t}, \boldsymbol{\theta}_{-i})])\right) \\
&= \sum_{\boldsymbol{\theta}_{-i}\in\boldsymbol{\Theta}_{-i}} f_{-i}(\boldsymbol{\theta}_{-i}) \sum_{t\in\{0,1,2,...,|\Theta_i|\}} \left(\sum_{\substack{\theta_i\in\Theta_i:\\ \theta_i\geq\theta_{i,t}}} f(\theta_i)\theta_{i,t} \cdot (\mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i}))] - \mathbb{E}[\mathcal{C}_i(\theta_{i,t}, \boldsymbol{\theta}_{-i})])\right) \\
&= \sum_{\boldsymbol{\theta}_{-i}\in\boldsymbol{\Theta}_{-i}} f_{-i}(\boldsymbol{\theta}_{-i}) \sum_{t\in\{0,1,2,...,|\Theta_i|-1\}} ((1-F(\theta_{i,t}))\theta_{i,t} \cdot (\mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i}))] - \mathbb{E}[\mathcal{C}_i(\theta_{i,t}, \boldsymbol{\theta}_{-i})])) \\
&= \sum_{\boldsymbol{\theta}_{-i}\in\boldsymbol{\Theta}_{-i}} f_{-i}(\boldsymbol{\theta}_{-i}) \left(-(1-F_i(\theta_{i,0})) \cdot \theta_{i,0} \cdot \mathbb{E}[\mathcal{C}_i(\theta_{i,0}, \boldsymbol{\theta}_{-i})]\right) \\
&\quad - \sum_{t\in\{0,1,2,...,|\Theta_i|-1\}} \mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i})]((1-F(\theta_{t+1}))\theta_{t+1} - (1-F(\theta_{i,t}))\theta_{i,t})) \\
&= \sum_{\boldsymbol{\theta}_{-i}\in\boldsymbol{\Theta}_{-i}} f_{-i}(\boldsymbol{\theta}_{-i}) \left(-(1-F_i(\theta_{i,0})) \cdot \theta_{i,0} \cdot \mathbb{E}[\mathcal{C}_i(\theta_{i,0}, \boldsymbol{\theta}_{-i})]\right) \\
&\quad + \sum_{t\in\{0,1,2,...,|\Theta_i|-1\}} f(\theta_{i,t})\, \mathbb{E}[\mathcal{C}_i(\theta_{t+1}, \boldsymbol{\theta}_{-i})](\theta_{t+1} - (\theta_{t+1} - \theta_{i,t})\frac{1-F(\theta_{i,t})}{f(\theta_{i,t})}))
\end{aligned}$$

Note that the first term $-(1-F_i(\theta_{i,0})) \cdot \theta_{i,0} \cdot \mathbb{E}[\mathcal{C}_i(\theta_{i,0}, \boldsymbol{\theta}_{-i})]$ is always negative, so we can maximize revenue by ensuring that we do not allocate to any type with the lowest possible θ_i. This will ensure that the first term is zero, and it leaves the revenue maximizing equation as:

$$\begin{aligned}
&\sum_{i\in\{m\}} \mathbb{E}_{F(\boldsymbol{\theta})}[p_i(\theta_i, \boldsymbol{\theta}_{-i})] = \\
&\sum_{i\in\{m\}} \sum_{\boldsymbol{\theta}_{-i}\in\boldsymbol{\Theta}_{-i}} f_{-i}(\boldsymbol{\theta}_{-i}) \sum_{t\in\{0,1,2,...,|\Theta_i|\}} f(\theta_{i,t})\, \mathbb{E}[\mathcal{C}_i(\theta_{i,t+1}, \boldsymbol{\theta}_{-i})](\theta_{i,t+1} - (\theta_{i,t+1} - \theta_{i,t})\frac{1-F(\theta_{i,t})}{f(\theta_{i,t})})) \\
&= \mathbb{E}_{\boldsymbol{\theta}}[\sum_{i\in\{m\}} (\theta_{i,t+1} - (\theta_{i,t+1} - \theta_{i,t})\frac{1-F(\theta_{i,t})}{f(\theta_{i,t})})\, \mathbb{E}[\mathcal{C}_i(\theta_{i,t+1}, \boldsymbol{\theta}_{-i})]]
\end{aligned}$$

Note that we can treat the first term, $(\theta_{i,t+1} - (\theta_{i,t+1} - \theta_{i,t})\frac{1-F(\theta_{i,t})}{f(\theta_{i,t})})$, as a scaled type, a virtual type, denoted $\phi(\theta_{i,t+1})$. Then if we maximize $\mathbb{E}_{\boldsymbol{\theta}}[\phi(\theta_{i,t+1})\, \mathbb{E}[\mathcal{C}_i(\theta_{i,t+1}, \boldsymbol{\theta}_{-i})]]$ we will maximize

revenue. However, this is just a welfare maximization problem with virtual types replacing the true type. Therefore, a monotone allocation algorithm that maximizes welfare for the virtual types, will maximize revenue. Also, if an allocation algorithm approximately maximizes virtual welfare, then it will also approximately maximize revenue with the same approximation factor. Note that, as is standard in revenue maximization through virtual values, if the virtual value is not monotone in the true type, we can use an ironing procedure to ensure monotonicity. Then the monotonicity of the virtual value in the true type ensures incentive compatibility and individual rationality as in Lemma 4.2.

F $(1 - 1/e)$ Welfare Maximizing Algorithm for Standard Coverage Functions

First, we start by introducing the coverage functions as defined in Dobzinski and Schapira [13]. Note that for the setting, Dobzinski and Schapira [13] algorithm requires using the demand oracle whereas Dughmi et al. [18] runs in expected polynomial time. We have a ground set U and a collection of L many subsets of U $R_1, \ldots, R_L$. We have m many bidders (buyers). We can allocate each set in $\{R_1, \ldots, R_L\}$ to only one bidder. If a buyer $i \in [m]$ receives a collection of sets $R_{i,1}, \ldots, R_{i,L_i}$ then his valuation would be $v_i(\{R_{i,1}, \ldots, R_{i,L_i}\}) = \theta_i \cdot |\cup_{\ell \in [L_i]} R_{i,\ell}|$ where $\theta_i \geq 0$ is the private type (importance) parameter. Note this generalizes the function from Dobzinski and Schapira [13] since we associate a parameter θ_i for each bidder i whereas in Dobzinski and Schapira [13] all bidders have $\theta_i = 1$.

For an element $j \in U$ we define $N(j) = \{R_\ell | j \in R_\ell\}$. Now we introduce the below LP which is essentially modified from 3.

$$\max_{\bar{X}_{i\ell}, \bar{C}_{ij}} \quad \sum_{i \in [m]} \theta_i \Big(\sum_{j \in U} \bar{C}_{ij} \Big) \tag{41a}$$

$$\forall i \in [m], \forall j \in U, \forall \ell \in [L]: \quad 0 \leq \bar{X}_{i\ell}, \bar{C}_{ij} \leq 1 \tag{41b}$$

$$\forall \ell \in [L]: \quad \sum_{i \in [m]} \bar{X}_{i\ell} \leq 1 \tag{41c}$$

$$\forall i \in [m], \forall j \in U: \quad \bar{C}_{ij} \leq \sum_{\ell \in N(j)} \bar{X}_{i\ell} \tag{41d}$$

The same methods of Section 3 can be followed straightforwardly to show a $1 - 1/e$ approximation. Further, a similar data burning method from 4 can be used to guarantee monotonicity.

G Counterexample Demonstrating the Non-monotonicity of GREEDY

Consider a $\mathcal{DB}$ with three data points on a line $A - B - C$, where the distances between (A, B), (A, C), and (B, C) are $1, 2, 1$, respectively. We have three bidders all having the same radius of $r_{ij} = 1$ and weight of $w_{ij} = \frac{1}{3}$ for all points. Further, the bidders have $\theta_1 = 1$, $\theta_2 = \frac{3}{4}$, and $\theta_3 = \frac{13}{25}$. The greedy algorithm will loop over the data points starting from point A then B and finally C. At each iteration t the marginal gain $v_i(S_{i,t} \cup \{j\}) - v_i(S_{i,t})$ is calculated for each bidder i where $S_{i,t}$ is the set of points given to bidder i before the t^{th} iteration is calculated and the point j is the one considered at iteration t. Point j is then allocated to the bidder with the maximum gain.

Now, consider the truthful bid profile of $\boldsymbol{b} = (b_1 = 1, b_2 = \frac{3}{4}, b_3 = \frac{13}{25})$. By running the greedy algorithm we can see that the solution would result in bidder 1 obtaining only point A and hence a coverage of $\mathcal{C}_1 = \frac{2}{3}$. On the other hand, if bidder 1 deviates to $b'_1 = \frac{7}{10}$ (with the other bids fixed), then he would instead obtain point B leading to $\mathcal{C}'_1 = 1 > \mathcal{C}_1 = \frac{2}{3}$. All comparisons are strict under both bid profiles. This clearly violates monotonicity.

H Empirical Counterexample Demonstrating the Non-Monotonicity of the CONGREEDY Algorithm

As introduced in Section 5, our **SYN** dataset has $m = 2$, $n = 5$, $|\Lambda| = 10$, $d = 10$. And the dataset $\mathcal{DB}$ is generated uniformly within the range of $[0, 10]$ with a shape of $n \times d$. Specifically, the randomly generated buyer types are $\theta_1 = 0.4$ and $\theta_2 = 0.7$. The weights $w_{ij} = 0.2$ for all i and j.

The radius $r_{ij} = 11.50893871$ for all j when $i = 1$ is the first buyer; and $r_{ij} = 10.48592194$ for all j when $i = 2$. Finally, the randomly generated dataset $\mathcal{DB}$ is as follows.

$$\mathcal{DB} = \begin{bmatrix} 2.60299691 & 8.70395688 & 1.85039927 & 0.19661425 & 9.53252032 & 6.80450805 & 4.86588127 & 9.6502682 & 3.93398739 & 0.79557571 \\ 3.51407424 & 1.63635163 & 9.83166821 & 8.80628184 & 4.94063468 & 4.00959241 & 4.51291463 & 7.20876849 & 2.47768284 & 6.22779952 \\ 1.42448816 & 2.01176282 & 0.81217729 & 9.53472295 & 0.5573827 & 5.99536483 & 7.2299763 & 9.70289719 & 8.21569457 & 5.27551067 \\ 3.3147673 & 3.53982196 & 0.79030301 & 5.5591438 & 1.65794463 & 2.95237831 & 8.40608485 & 3.62562009 & 3.40813978 & 7.10401744 \\ 0.31707752 & 4.72387866 & 7.31576623 & 7.73215774 & 2.41244792 & 3.06197046 & 1.36649638 & 5.9573572 & 1.62568866 & 8.02327214 \end{bmatrix}$$

For this instance, we fixed the bidder, the bid pair $b_i \geq b_i'$, and the other bids before running $N = 150{,}000$ fresh independent simulations at each bid. Let $\mu = \mathbb{E}[\mathcal{C}_i(\mathcal{D})]$ and $\mu' = \mathbb{E}[\mathcal{C}_i(\mathcal{D}')]$. Under $H_0 : \mu \geq \mu'$, a gap $\hat{\mathcal{C}}_i(\mathcal{D}') - \hat{\mathcal{C}}_i(\mathcal{D}) \geq t > 0$ requires either $\hat{\mathcal{C}}_i(\mathcal{D}') - \mu' \geq t/2$ or $\mu - \hat{\mathcal{C}}_i(\mathcal{D}) \geq t/2$. Since coverage lies in $[0, 1]$, Hoeffding's inequality [30] and a union bound give

$$\Pr_{H_0}\Big(\hat{\mathcal{C}}_i(\mathcal{D}') - \hat{\mathcal{C}}_i(\mathcal{D}) \geq t\Big) \leq 2e^{-Nt^2/2}.$$

The reported gap exceeds $t = 0.02$, yielding the conservative p-value bound $2e^{-30} < 1.9 \times 10^{-13}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: Answer: [Yes]

   Justification: In this paper, we considered the mechanism design problem of data selling to a set of buyers for RAG-based usage. We showed that the problem has inherent computational barriers that require using approximation algorithms. We introduced a generalization of Myerson's lemma which proves that any incentive compatible mechanism must have a monotonic allocation. We proved that simple algorithms such as the greedy approximation are not actually monotonic. The examples of non-monotonic algorithms included our novel LP-rounding algorithm which achieves a $1 - 1/e$ approximation. Interestingly, we showed however that it can be post-processed with an additional data burning step to guarantee monotonicity without any degradation to the $1 - 1/e$ approximation factor. Finally, we conducted experiments over real world datasets and demonstrated the superiority of our algorithm.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: Section 6 discusses the public coverage assumptions, the practical licensing setting, and the welfare and variance effects of data burning.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All proofs of theoretical results are provided in the appendix

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed is disclosed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and data is submitted as supplemental materials and will be open-sourced.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setting details can be found in Section 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Statistical significance are reported for Table 1 and Figure 1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide sufficient information in Section 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper focused on data selling for RAG-based usage. Our main focus was on the mechanism design issue. Through mathematical analysis we proved that our algorithms enforces strategy proofness and makes truthful bidding a dominant strategy for the buyers. We do not see immediate negative effects from our paper. However, in terms of societal impacts, there are some issues which our analysis does not consider. For example, what are the long-term societal effects when such data selling occurs? Since the data points are likely to be generated by individuals how do we handle issues such as privacy and copy rights? Such considerations require engaging multiple stakeholders and as such are beyond the scope of this paper.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Data and models used in the paper are properly credited and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.