Exploring Importance Sampling for Aggregation on Natural Language Predicates

Andrew Wang
Faculty Guide: Dr. Amol Deshpande

Abstract

Large Language Models (LLMs) enable powerful semantic queries over unstructured data, but executing queries at scale is computationally prohibitive. Approximate Query Processing (AQP) techniques such as stratified sampling (e.g., ABAE) and clustering-based sampling (UQE) offer efficiency gains, but the role of importance sampling remains underexplored. In this report, we investigate importance sampling for aggregate queries with expensive natural language predicates. We compare importance sampling, uniform sampling, and ABAE on a simulated dataset of embedding distances and truth labels for predicate queries. Our findings suggest that while importance sampling can theoretically minimize variance, its practical effectiveness hinges on accurately modeling the relationship between returning true for the predicate and the embedding distance. By contrast, ABAE demonstrates robust and stable performance.

Introduction

Large Language Models (LLMs) have introduced new general-purpose capabilities for querying unstructured data such as text, images, and audio. Systems like the Universal Query Engine (UQE) propose a structured query language for unstructured datasets by extending SQL with natural-language predicates. However, executing such queries requires running inference on large models, which is costly in both latency and compute, motivating the use of approximate query processing (AQP).

Although AQP as a field is fairly mature, its application to machine-learning predicates is still novel. In its most basic form, databases can simply perform uniform sampling. However, this leads to relatively high variance, which can be reduced by incorporating prior information via machine learning methods. To clarify, the problem is to estimate queries of the following form:

> *SELECT COUNT(*)*
> *FROM table*
> *WHERE query(“The agent successfully canceled the customer’s flight.”) = TRUE*

For these queries, UQE proposes using embedding models to convert text passages into semantically meaningful vectors, clustering these vectors, and performing stratified sampling on these clusters. Stratified sampling achieves lower variance than uniform sampling when the

strata are correlated with the measured variable. In UQE's case, the goal is that elements with similar embedding vectors respond overwhelmingly with either TRUE or FALSE. For simple datasets, designed for text classification tasks before the advancement of LLMs, this is often a significant improvement over uniform sampling. However, we can further improve this through more advanced sampling methods and the inclusion of additional information.

Methods

UQE relies on a strong assumption that, if false, can actually *increase* variance. It assumes that the answer to the query is correlated with the raw embedding vectors of the elements. But we can easily imagine queries where this is not true.

For example, consider the AirDialogue dataset (2018), one of the datasets UQE used for evaluation. It contains chat logs between airline agents and customers, and each chat is labeled with one of five classes: **booked**, **changed**, **no flight found** (very rare, so UQE ignored it), **canceled**, and **no reservation**.

- **booked**: the agent successfully books a flight requested by the customer
- **changed**: the agent successfully changes an existing reservation
- **canceled**: the agent cancels an existing reservation
- **no reservation**: the agent cannot find the customer's reservation to change or cancel

The example query above asks for the number of chats in the **"canceled"** class. For this query, conversations where the agent cancels a flight and those where they do not are semantically quite different, so clustering by embedding works well. However, consider a different query like: **"The agent says thanks twice or more."** This property is not strongly tied to the overall semantic meaning of the conversation. In that case, clustering by embedding is unlikely to reduce variance and may not improve over uniform sampling.

ABAE proposes a more complex solution that still utilizes stratified sampling. Although it was originally designed for computer vision use cases on videos using proxy models, we replace key components with textual analogues. Whereas UQE only embeds and clusters the elements once, textual ABAE uses the query's embedding to create more meaningful strata to sample from. It computes the embedding distances between the query and the population and stratifies the population based on quantiles of this distance. Furthermore, it follows a two-stage sampling procedure. In the first stage, it samples an equal number of elements from each stratum to estimate within-stratum means and variances. Then, it optimally assigns sampling weights to each stratum for the second stage. The incorporation of the relationship between the query and the population, as well as weighted stratified sampling, allows ABAE to perform very well.

ABAE's assumption is more likely to be met than UQE's, as it assumes that the query is more "similar" to its positive answers than to its negative ones. It is important to note that more advanced embedding models can be used for question answering. Some accept instructions,

resulting in vector embeddings that capture more specific features than the general meaning of the input text. We can more safely assume that the embedding distance is correlated with the query answers, as long as a sufficiently capable embedding model is used. Interestingly, since ABAE only performs stratified sampling, it does not require that the distance over its entire range be correlated with the answer. Instead, it only requires that samples with similar distances answer similarly, so even if distances were reversed, it would still perform the same.

This last fact suggests that improvements can be made by assuming an inverse relationship between the probability of a positive answer and distances. Formally, let D_i denote the distance between the query and element i, and let Y_i be 1 if the element satisfies the predicate. We can define a non-increasing function f: [0, ∞) -> [0, 1] such that $P(Y_i = 1 \mid D_i = d) = f(d)$. For now, we will also forgo the use of strata and jump directly to importance sampling. In essence, importance sampling reduces variance in our case if elements with higher P(Y) are sampled more often. To illustrate, if our distances range from 0.0 to 1.0, but we happen to know that only elements with distances <0.1 have P(Y) > 0, then we should only sample from them (based on their $P(Y_i)$). This is similar to how ABAE assigns weights to quantiles, but in theory, we can perform better by using certain assumptions about the function f.

Algorithm (in pseudocode)

```
N is the size of the population
n is the sampling budget
Y in reality would be the expensive LLM/oracle call

ImportanceSamplingProportion(D, Y, N, n):
     D, Y ← sort elements based on D
     f ← fit the function f according to the above definition, or
already know it
     proposal_weights ← sqrt(f(D))
     weight_distribution ← proposal_weights / sum(proposal_weights)
     sample_indices ← sample n indices from weight_distribution
     importance_weights ← (1 / N) /
weight_distribution[sample_indices]
     proportion_estimate ← mean(importance_weights *
Y[sample_indices])
     return proportion_estimate
```

Derivation For the Proposal Weight

To minimize the variance of the estimator $\hat{\mu} = \frac{1}{n}\sum \frac{Y_k}{Nq_k}$, we minimize the second moment. For binary variables where $E[Y_i^2] = E[Y_i] = f_i$:

$$E_q\left[\left(\frac{Y}{Nq}\right)^2\right] = \sum_{i=1}^{N} q_i \frac{E[Y_i^2]}{(Nq_i)^2}$$
$$= \frac{1}{N^2}\sum_{i=1}^{N} \frac{f_i}{q_i}$$

We minimize this sum subject to $\sum q_i = 1$ using the Lagrangian:

$$\mathcal{L} = \sum_{i=1}^{N} \frac{f_i}{q_i} + \lambda\left(\sum_{i=1}^{N} q_i - 1\right)$$
$$\frac{\partial \mathcal{L}}{\partial q_i} = -\frac{f_i}{q_i^2} + \lambda = 0$$

Solving for q_i yields the optimal distribution:

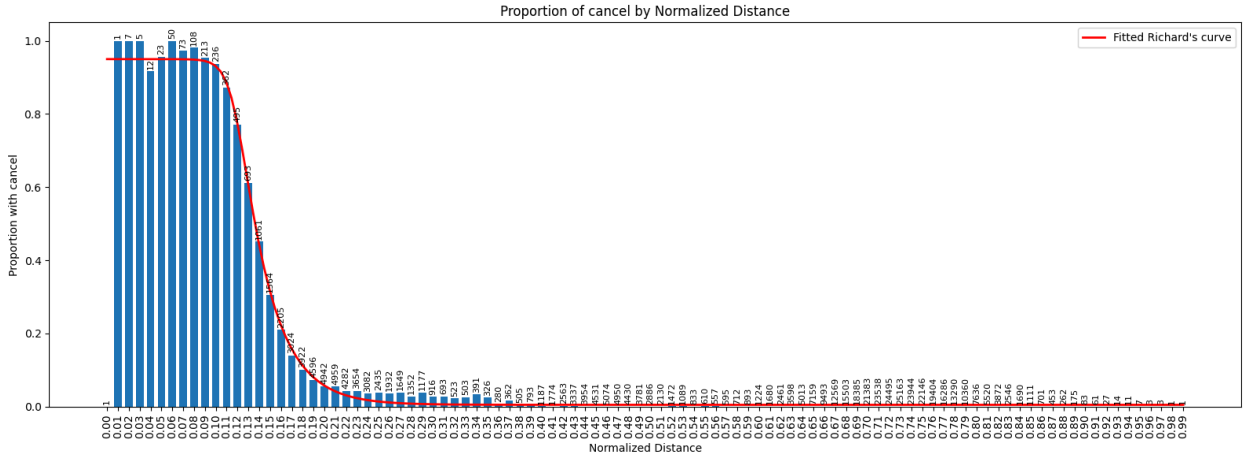
$$q_i^2 = \frac{f_i}{\lambda}$$
$$q_i \propto \sqrt{f_i}$$

Experiments

We generated a simulated dataset of (distance, Y) tuples to bypass the overhead of live LLM inference and to control the data distribution for our importance sampling tests. Simulation parameters were calibrated using the 'AirDialogue' dataset. We embedded the dataset using intfloat/multilingual-e5-large-instruct, labeling dialogues where the agent canceled a reservation as positive samples. We then measured the distances between these embeddings and the prompt: *'Instruct: Given a chat history, retrieve passages that satisfy the criteria\n Query: The*

agent successfully canceled the customer's reservation.' These measurements allowed us to quantify the correlation between embedding distance and predicate satisfaction.

To model the embedding distances, we normalized the range to [0, 1]. Since the observed histogram exhibited a slightly skewed shape within these bounds, we generated the simulated distances using a Beta distribution. Modeling the predicate satisfaction Y was more complex, as it required identifying the underlying probability function. While the data roughly aligned with a standard logistic curve, we found that the Generalized Logistic Function (Richards' curve) provided a significantly more accurate fit. Therefore, we adopted the generalized form to ensure a faithful simulation.

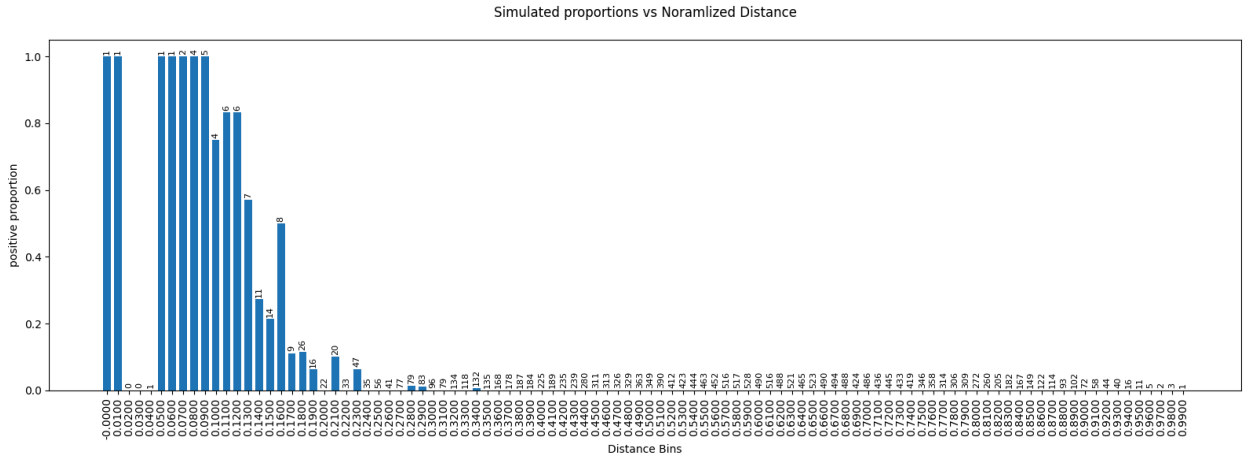


We benchmarked our importance sampling method against ABAE and Uniform Sampling. We excluded UQE because our simulation operates on scalar distances rather than full vector representations. The experimental setup consisted of a population size of 20,000 and a total sampling budget of 5,000, with results averaged over 5,000 independent runs. For the ABAE baseline, we utilized five strata and split the budget evenly (50/50) between the pilot and refinement stages.

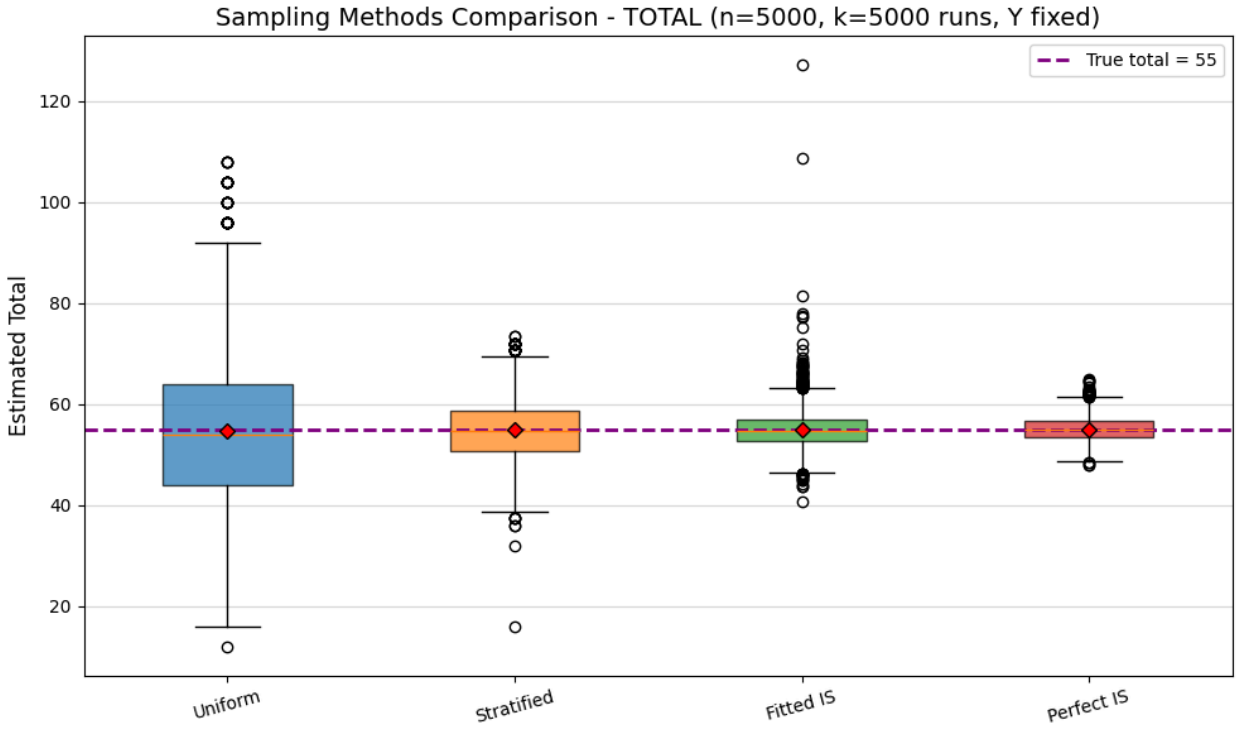
We evaluated two variants of importance sampling. The first, 'Fitted IS,' is a two-stage approach. It allocates 1/3 of the budget to estimate the propensity function via stratified sampling across 100 uniform distance bins. Unlike ABAE, samples from this pilot stage are not included in the final estimator. We employed standard logistic regression for this estimation; although the data were generated using the generalized logistic (Richards') curve, we selected the simpler model to ensure successful fitting given the limited pilot data. The second variant, 'Oracle IS' (or 'Perfect'), utilizes the true generating function f(). To ensure a fair comparison with the inference phase of the Fitted method, this variant uses only the remaining 2/3 of the budget. Finally, we conducted two versions of the experiment: one with fixed label assignments (Y) and another where labels were regenerated for each run.

Results

First, we consider the case where we do not regenerate the Y labels. Here is an example of a simulated dataset:



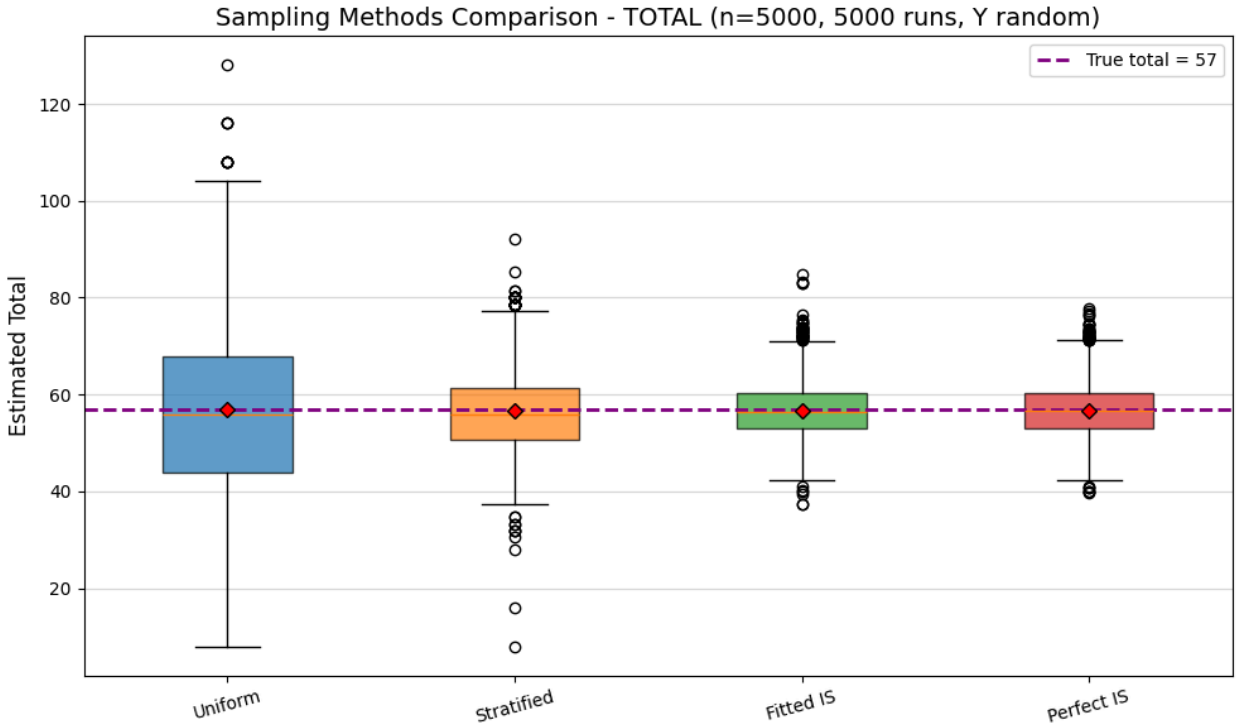
And the results (using counts and standard deviation for readability):



Method	Standard Deviation
Uniform Sampling	14.9278
ABAE	5.7250
Fitted Importance Sampling	3.5669

Perfect Importance Sampling	2.4035

As shown in the table above, both importance sampling methods produced lower standard deviations than ABAE or Uniform Sampling. However, Fitted IS was prone to rare but severe outliers, a problem not observed in Perfect IS. This instability appears to be data-dependent: specific fixed realizations of Y caused significant spikes in variance, while others did not. To address this dependency, the following section analyzes performance under the 'regenerated labels' condition, where Y is randomized for every run.



Method	Standard Deviation
Uniform Sampling	15.77
ABAE	7.63
Fitted Importance Sampling	5.44
Perfect Importance Sampling	5.42

Regenerating the Y labels for each run introduced additional variance, resulting in higher standard deviations across the board. However, the general advantage of importance sampling persists. Notably, in this setting, the gap between Fitted IS and Perfect IS has mostly closed, suggesting that the 'Fitted' approach is robust on average, even if it suffers from rare outliers in specific fixed scenarios.

Discussion

Our results show that by assuming a more precise and accurate relationship between predicate satisfaction and embedding distance, we are able to achieve lower variance at the cost of worse outliers. Additionally, the outliers we observe indicate that when our assumptions are incorrect and we also get unlucky, our estimates become highly inaccurate. In other words, a stratified sampling approach like ABAE is far more robust to a variety of data distributions, especially since it does not assume any relationship between strata, only within strata. Our approach assumes that the data follows a specific type of function, though exact matches are not necessary, as indicated by the performance of Fitted IS, which only fits the basic logistic curve.

Future Work

One area for future investigation is testing the method's robustness on a wider variety of data distributions. We also propose a hybrid strategy that combines stratified and importance sampling. Instead of assigning a weight to every single element, we could group K elements together and assign a single weight to that group. This would act as a form of regularization, preventing any single sample from having an exploding weight and balancing the trade-off between outliers and average variance. Additionally, we could explore how to best allocate the sampling budget between the pilot stage and the final estimation to further reduce variance.

Finally, we could replace the simple embedding distance with a lightweight proxy model to estimate the score f() directly. Our current technique is limited because vector distance often misses subtle semantic nuances, such as negation or specific counting of details, that a smarter proxy model could capture.

Dai, H. et al. *UQE: A Query Engine for Unstructured Databases*. NeurIPS 2024.

Kang, D. et al. *Accelerating Approximate Aggregation Queries with Expensive Predicates (ABAE)*. VLDB 2021.