Beyond Labels: Information-Efficient Human-in-the-Loop Learning using Ranking and Selection Queries

Belén Martín-Urcelay, Yoonsang Lee, Matthieu R. Bloch, Christopher J. Rozell,

***Abstract*—Integrating human expertise into machine learning systems often reduces the role of experts to labeling oracles, a paradigm that limits the amount of information exchanged and fails to capture the nuances of human judgment. We address this challenge by developing a human-in-the-loop framework to learn binary classifiers with rich query types, consisting of item ranking and exemplar selection. We first introduce probabilistic human response models for these rich queries motivated by the relationship experimentally observed between the perceived implicit score of an item and its distance to the unknown classifier. Using these models, we then design active learning algorithms that leverage the rich queries to increase the information gained per interaction. We provide theoretical bounds on sample complexity and develop a tractable and computationally efficient variational approximation. Through experiments with simulated annotators derived from crowdsourced word-sentiment and image-aesthetic datasets, we demonstrate significant reductions on sample complexity. We further extend active learning strategies to select queries that maximize information rate, explicitly balancing informational value against annotation cost. This algorithm in the word sentiment classification task reduces learning time by more than 57% compared to traditional label-only active learning.**

***Index Terms*—Human oracle, sample complexity, active learning**

I. Introduction

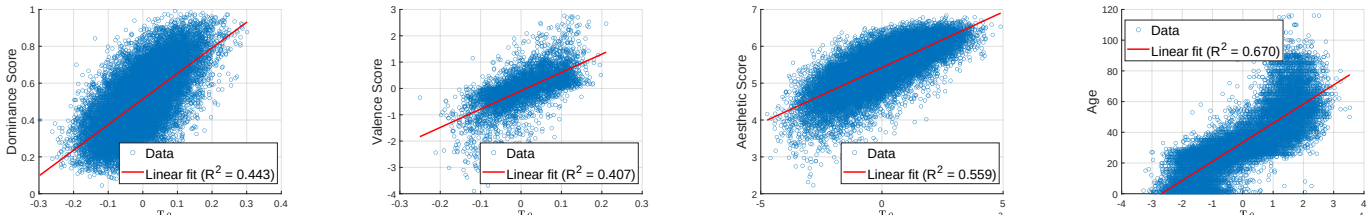
Integrating machine learning systems with the nuanced judgments of human experts is a critical goal in domains ranging from image [1] and text [2] generation to word sense disambiguation [3] and medical diagnosis [4]. However, transferring human expertise to computational systems remains challenging. Although humans excel at making complex judgments, they often struggle to articulate the precise features or explicit logic that guide their decisions [5]. This gap between intuitive expertise and machine-interpretable explanations largely reduces the role of humans to that of labeling oracles [6], [7]. As oracles, experts provide labels that enable supervised learning algorithms to approximate the implicit decision functions.

A drawback of this oracle-based paradigm is the substantial human effort and cost required to acquire large labeled datasets. Active learning [8] has emerged as an efficient approach to mitigate this cost. Active learning strategies select informative examples to query, thereby significantly reducing the number of required labels. Despite these advances, the information obtained from such queries is constrained by the simplistic nature of labels [9]. In practice, queries are often limited to binary classifications, for which the information gained is inherently capped at just one bit [10], [11]. This information bottleneck raises the question: can we move beyond simple labeling to *richer queries that provide more information per human interaction, while remaining intuitive for humans?*

Even when presented with binary labeling tasks (e.g., "Is the weather today good or bad?"), humans may engage in richer cognitive processes. Humans often categorize items by recalling exemplars and comparing to reference points (e.g. "This is the worst weather we have had all week"). Despite this capacity for comparative reasoning, most human-in-the-loop strategies artificially constrain experts to binary labeling roles. To harness the underlying context-rich evaluation, we consider alternative query types: ranking items by attribute strength or selecting the most representative exemplar from a list. Capitalizing on these richer query types requires quantitative models of human responses that render such queries machine-interpretable and optimizable, as well as an algorithm that selects informative, cost-aware queries and learns from the corresponding human answers. We address these requirements through the following contributions:

- **Human response models for rich queries on off-the-shelf embeddings.** We introduce probabilistic response models to ranking and exemplar-selection queries. These models are based on the observed relationship between the perceived score of an item and its distance to the decision boundary in the embedding space. This enables the use of rich queries that capture more information per interaction than traditional labeling.
- **Theoretical guarantees and empirical improvements in sample complexity.** We derive theoretical bounds for the expected stopping time. Empirical evaluations with simulated annotators based on human data demonstrate up to 85% reduction in human interactions compared to traditional active labeling approaches.
- **Cost-aware information-rate optimization with human timing models.** We formulate query selection as maximizing expected bits of information per second rather than bits per interaction. Using response time models derived from a crowdsourced study, our cost-aware approach reduces annotation time by more than half compared to label-only active learning on word sentiment

B. Martín-Urcelay (e-mail: burcelay3@gatech.edu), Y. Lee, M. R. Bloch and C. J. Rozell are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.



a) Word2Vec embedding for NRC dataset

b) Word2Vec embedding for SocialSent dataset

c) ViT-L/14 embedding for AVA dataset

d) InceptionResnetV1 for UTKFace dataset

Fig. 1: Scores for words (a, b) and image (c, d) attributes as a function of the inner product between their pre-defined embedding and the ground truth classifier. We observe there exists an approximately affine relationship. Pre-trained embeddings naturally encode score information as distance from decision boundary, enabling information-rich queries beyond binary labels.

classification tasks, demonstrating practical time savings beyond sample-complexity gains.

Parts of this work have been previously presented at the Conference of Decision and Control [12]. This version substantially extends it in four directions: We introduce ranking queries that gather full orderings with labels in a single interaction; we provide sample-complexity guarantees that explicitly quantify how query type and embedding dimension affect stopping time; we demonstrate generalization beyond word sentiment by adding experiments on image aesthetic classification; and we incorporate an information-rate objective together with empirically fitted human timing models to optimize time costs. Together, these additions turn the original proof-of-concept into a more theoretically grounded and cost-aware framework for rich-query human-in-the-loop learning. The paper is organized as follows. In Section II we introduce the components of our method: human response models, an information-theoretic question selection strategy, and a tractable active learning algorithm for query item selection. In Sections III-A and III-B we present theoretical and empirical results on sample complexity. Finally, in Section III-C we describe crowdsourced experiments, derive human time cost models, and show expected time savings from our query-type selection based on information rate for the word sentiment classification task.

II. Proposed Method

To leverage rich query types, we develop a quantitative framework comprising three components. First, we construct a human response model that predicts the likelihood of annotator answers to ranking and exemplar selection queries. Second, recognizing that different queries impose varying cognitive loads, we present an active learning algorithm that balances informational value against human effort. Third, we derive variational approximations and greedy heuristics that make Bayesian inference tractable in high-dimensional embedding spaces. Together, these components enable principled optimization over expressive query types while accounting for realistic cost constraints.

A. Human Response Model

Our goal is to learn a decision boundary $\boldsymbol{\theta}$ in an embedding space that accurately predicts the implicit classification rules of human annotators. We focus on unitary norm binary linear classifiers $\boldsymbol{\theta} \in \mathbb{R}^d, \|\boldsymbol{\theta}\|_2 = 1$ that label every embedded item $\mathbf{x} \in \mathbb{R}^d$ in a d-dimensional embedding space as $y = \text{sign}(\boldsymbol{\theta}^T\mathbf{x})$. These linear classifiers generalize to non-linear problems if there exists a mapping from the original non-linear space to a higher dimensional embedding in which the data is linearly separable. Given a set of items labeled by the annotator, a common method to learn a linear classifier is *logistic regression*, in which the label probability is given by

$$\mathbb{P}\left[y = 1|\mathbf{x}\right] = \left(1 + \exp\left(w(\boldsymbol{\theta}^T\mathbf{x})\right)\right)^{-1}, \tag{1}$$

with $w \in \mathbb{R}$ representing the inverse of the scale parameter.

The main drawback of reducing annotators to labelers is that each response conveys at most one bit of information. Consequently, accruing enough information to learn an accurate classifier often requires an impractically large number of queries in data-scarce settings. To mitigate this information bottleneck, we design queries that solicit information beyond binary labels. Not only must these queries be intuitive for human annotators to respond to, but the queries also require a response model that relates item embeddings to the classifier in a manner quantifiable for computational analysis. Our key observation is that off-the-shelf embeddings naturally exhibit geometry that aligns with perceived scores: We expect human annotators to be uncertain when classifying items whose embeddings lie close to the boundary between classes. Conversely, we expect annotators to become more confident as the items lie further away from the boundary, implicitly assigning higher and lower scores as we move in opposite directions from the classification boundary. Figure 1 demonstrates that this relationship holds across a variety of embeddings, tasks and datasets. Specifically, we examine popular word and image embeddings [13]–[15] for tasks ranging from word sentiment [16] and dominance [17] analysis, to image aesthetic perception [18] and age categorization from a face [19]. In all cases, *we consistently observe a linear relationship between an item score and the inner product between its embedding and the Minimum Mean-Square Error (MMSE) classifier*. We provide additional details in Appendix A. Unlike previous work [6], [20] that learns task-specific embedding spaces, we exploit this naturally occurring linear relationship. This empirical

observation motivates us to formalize the relationship between embeddings and scores as follows;

Assumption II.1. Annotators associate a score to an item with embedding $\mathbf{x}_i$ as

$$\text{score}(\mathbf{x}_i) = a\mathbf{x}_i^T\boldsymbol{\theta} + b + \delta_i, \tag{2}$$

where δ_i represents the noise associated with $\mathbf{x}_i$, with a query dependent extreme value distribution. The scalars a and b, which describe the affine relationship, are dataset and attribute dependent.

The score model enables us to derive probability distributions of human responses for a large variety of queries. We model the noise in Equation (2) so that it leads to the Boltzmann choice model used in behavioral economics [21]. We validate the distribution choices empirically in Appendix A-B. Consider the question q_{high} = "Select and label the item with highest score;" and the noise distribution $\delta_i \sim \text{Gumbel-Max}(\mu, \sigma)$. In that case, the probability that the i-th item is selected from a set $\mathcal{S} = \{\mathbf{x}_k\}_{k=1}^{|\mathcal{S}|}$ is

$$\begin{aligned}\mathbb{P}[i|\mathcal{S}, \boldsymbol{\theta}, q_{\text{high}}] &= \mathbb{P}[\text{score}(\mathbf{x}_i) > \text{score}(\mathbf{x}_j), \forall j \neq i] \\ &= \mathbb{P}[\delta_j - \delta_i < a(\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\theta}, \forall j \neq i] \\ &= \frac{\exp\left(\frac{a}{\sigma}\mathbf{x}_i^T\boldsymbol{\theta}\right)}{\sum_{\mathbf{x}\in\mathcal{S}}\exp\left(\frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}\right)},\end{aligned} \tag{3}$$

because this is a logit choice probability [22]. Analogously, for the question q_{low} ="Select and label the item with lowest score" and noise distribution $\delta_i \sim \text{Gumbel-Min}(\mu, \sigma)$, the probability that the annotator selects the i-th item is

$$\mathbb{P}[i|\mathcal{S}, \boldsymbol{\theta}, q_{\text{low}}] = \frac{\exp\left(-\frac{a}{\sigma}\mathbf{x}_i^T\boldsymbol{\theta}\right)}{\sum_{\mathbf{x}\in\mathcal{S}}\exp\left(-\frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}\right)}. \tag{4}$$

We extend the model to ranking queries q_{rank} = "Rank the items from highest to lowest score and indicate which is the last positive example in the ranked list." Given a set of items $\mathcal{S} = \{\mathbf{x}_k\}_{k=1}^{|\mathcal{S}|}$, let $\mathbf{r} = [r_1, ..., r_{|\mathcal{S}|}]$ represent the permutation that orders these items, where $\mathbf{x}_{r_j}$ is the item ranked at position j. We model the annotator's ranking as a sequential selection process:

$$\begin{aligned}\mathbb{P}\left[\mathbf{r}|\mathcal{S}, \boldsymbol{\theta}, q_{\text{rank}}\right] &= \prod_{j=1}^{|\mathcal{S}|-1}\mathbb{P}\left[r_j|\mathcal{R}_j, \boldsymbol{\theta}, q_{\text{high}}\right] \\ &= \prod_{j=1}^{|\mathcal{S}|-1}\frac{\exp\left(\frac{a}{\sigma}\mathbf{x}_{r_j}^T\boldsymbol{\theta}\right)}{\sum_{\mathbf{x}\in\mathcal{R}_j}\exp\left(\frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}\right)},\end{aligned} \tag{5}$$

where $\delta_i \sim \text{Gumbel-Max}(\mu, \sigma)$ and $\mathcal{R}_j = \mathcal{S} \setminus \{\mathbf{x}_{r_1}, ..., \mathbf{x}_{r_{j-1}}\}$. This response model is known as the Plackett-Luce model [23], [24]. By asking the annotators to mark the "last positive example" in the ordered list, we obtain a threshold $\ell \in \{0, 1, ..., |\mathcal{S}|\}$ that implicitly captures the labels for all items. Namely, if $\ell = 0$ all items will receive a negative label. Otherwise, all items in positions $i \leq \ell$ receive a positive label, while the remaining items receive a negative label.

The queries q_{high} and q_{low} request a listwise choice, gathering up to $\log_2|\mathcal{S}|$ more bits of information per query than a traditional binary label. The ranking query q_{rank} receives a full ordering over $\mathcal{S}$ and complete labeling further alleviating the information gain bottleneck. Table I summarizes the differences between the proposed rich queries and the traditional labeling query.

B. Question Selection

To minimize sample complexity, we wish to select the question and set size such that the information gained from the query $I(\boldsymbol{\theta}; o|q, |\mathcal{S}|)$ is maximized. In practice, feasible queries are often constrained by cognitive and interface limitations [25]. Under our response model, ranking queries subsume exemplar selection queries; thus, they provide strictly more information. Therefore, to minimize sample complexity, when the domain allows it, we should select q_{rank}. We also empirically observe that actively selecting between the questions q_{high} and q_{low} does not have a significant impact on performance, so when asking selection queries, we alternate between both uniformly at random. Moreover, we find that the information gain is increasing with $|\mathcal{S}|$; thus, to minimize sample complexity, we select the largest feasible $|\mathcal{S}|$.

In many applications, however, the primary objective is to minimize a cost different from sample complexity, such as cognitive effort or total response time, and not all queries incur the same cost [26]–[28]. To balance informativeness against response burden, we propose selecting queries that maximize the *information rate*, i.e., the expected bits of information gained per unit cost:

$$R = \frac{I(\boldsymbol{\theta}; o|q, |\mathcal{S}|)}{\mathbb{E}[\text{cost}(q, |\mathcal{S}|)]}. \tag{6}$$

Note that to maximize this rate, we need a model of the cost for each feasible question and item size combination, which we estimate in Section III-C1 for a word classification task.

C. Algorithm

We propose an online machine learning algorithm to learn a binary classifier from human feedback. Figure 2 illustrates the principle of our approach. At each iteration t, the *item selector* chooses the item set $\mathcal{S}$ such that the expected annotator response to the preselected question q is as informative as possible, i.e., the set $\mathcal{S}$ that maximizes the mutual information between the underlying classifier and the annotator response. Next, the annotator answers the query with o_t. In the case of q_{high} or q_{low}, the answer is an item and its corresponding label $o_t = (i_t, y_t)$; in the case of a ranking question q_{rank}, the answer involves an item ordering and threshold separating positive from negative items $o_t = (\mathbf{r}, \ell)$. The *classifier estimator* collects the response from the annotator and leverages this information to update the estimator of the classifier $\boldsymbol{\theta}$. The posterior $\mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t]$, where $\mathcal{F}_t = \{o_k, q_k, \mathcal{S}_k\}_{k=1}^t$ denotes the history, is updated using Bayes rule to leverage the likelihood functions in Equations (1), (3), (4) and (5). The algorithm continues querying the oracle until the uncertainty is sufficiently reduced. We measure uncertainty as the determinant of the posterior covariance matrix $|\boldsymbol{\Sigma}_t|$, which quantifies the volume of the uncertainty region, and we terminate when $|\boldsymbol{\Sigma}_t| \leq \epsilon^d$,

TABLE I: Comparison of Query Types for Human-in-the-Loop Learning

Query Type	**Outcome Space** (N)	**Information per Query (bits)**	**Expected Response Time (s)**	**Interactions to 75% Accuracy**
Label only	2	≤ 1	4.37	1310
Label + Selection (q_{high}, q_{low})	$2\|\mathcal{S}\|$	$\leq 1 + \log_2 \|\mathcal{S}\|$ ($\|\mathcal{S}\| = 4$: $\sim$3 bits)	$4.01 + 0.63\|\mathcal{S}\|$ ($\|\mathcal{S}\| = 4$: 6.5s)	468 ($\|\mathcal{S}\| = 4$)
Label Threshold + Ranking (q_{rank})	$(\|\mathcal{S}\| + 1)!$	$\leq \log_2(\|\mathcal{S} + 1\|!)$ ($\|\mathcal{S}\| = 4$: $\sim$4.3 bits)	$-0.32 + 4.41\|\mathcal{S}\|$ ($\|\mathcal{S}\| = 4$: 17.3s)	191 ($\|\mathcal{S}\| = 4$)

Information per query represents the theoretical maximum mutual information $I(\theta; \mathcal{O}|q, S)$. Expected response times are modeled from crowdsourced experiments with human participants on word sentiment classification tasks (Section III-C). Performance metric (75% accuracy) is based on word sentiment classification experiments with **active** word selection (Figure 3d).

Algorithm 1 Ideal Human-in-the-Loop Learning (HiLL)

1: **Input:** $\mathcal{X}, \mathbf{q}, |\mathcal{S}|, \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_0], \epsilon^d$
2: $t = 0$
3: **while** $|\boldsymbol{\Sigma}_t| > \epsilon^d$ **do**
4: $\quad \mathcal{S}_t \leftarrow \underset{\mathcal{S} \in \binom{|\mathcal{X}|}{|\mathcal{S}|}}{\text{argmax}} \mathbb{E}\left[I(\boldsymbol{\theta}; o|q_t, \mathcal{S}, \mathcal{F}_{t-1})\right]$
5: $\quad o_t \leftarrow$ human's response to the query
6: $\quad \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t] = \frac{\mathbb{P}[o_t|\boldsymbol{\theta}, q_t, \mathcal{S}_t]\mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_{t-1}]}{\mathbb{P}[o_t|q_t, \mathcal{S}_t]}$
7: $\quad t = t + 1$
8: **end while**

Algorithm 2 Approximate HiLL for Selection

1: **Input:** $\mathcal{X}, |\mathcal{S}|, \boldsymbol{\mu}_0, \Sigma_0, \epsilon^d, N$
2: $t = 1$
3: **while** $|\boldsymbol{\Sigma}_t| > \epsilon^d$ **do**
4: $\quad q_t \leftarrow$ sample uniformly from $\{q_{\text{high}}, q_{\text{low}}\}$
5: $\quad S_t \leftarrow$item_set_selection$(q_t, \mathcal{X}, |\mathcal{S}|, \boldsymbol{\mu}_{t-1}, \Sigma_{t-1}, N)$
6: $\quad i_t, y_t \leftarrow$ human response to the query
7: $\quad \boldsymbol{\mu}_t, \Sigma_t \leftarrow$ belief_update$(\mathcal{S}, i_t, y_t, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$
8: $\quad t = t + 1$
9: **end while**

where $\epsilon \in \mathbb{R}$ is the per dimension threshold. The approach is summarized in Algorithm 1.

Unfortunately, Algorithm 1 is intractable in high dimensional settings, because the Bayesian update lacks a closed-form solution and the set of possible items grows combinatorially. The next subsections provide approximations to make the computations feasible. The corresponding tractable implementations are described in Algorithm 2 and Algorithm 3.

1) Approximation of Belief Update: As line 6 of Algorithm 1 indicates, we use a Bayesian approach to update the belief of the classifier given the latest observation. When the question is $q_t \in \{q_{\text{high}}, q_{\text{low}}\}$ and the answer is $o_t = (i_t, y_t)$, the posterior is given by

$$\mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t] = \frac{\mathbb{P}[i_t|\boldsymbol{\theta}, \mathcal{S}_t, q_t]\mathbb{P}[y|\mathbf{x}_i, \boldsymbol{\theta}]}{\mathbb{P}[i_t, y_t|\mathcal{S}_t, q_t, \mathcal{F}_{t-1}]}\mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_{t-1}],$$

where $\mathcal{F}_0$ is the empty set $\emptyset$.

The likelihood functions are not conjugates of the prior, so no analytical closed form expression exists to compute the posterior. Although, Black Box Variational Inference (BBVI) [29] is commonly used to approximate the posterior, our closed-form derivation of the variational updates for this

Algorithm 3 Approximate HiLL for Ranking

1: **Input:** $\mathcal{X}, |\mathcal{S}|, \boldsymbol{\mu}_0, \Sigma_0, \epsilon^d, N$
2: $t = 1$
3: **while** $|\boldsymbol{\Sigma}_t| > \epsilon^d$ **do**
4: $\quad S_t \leftarrow$item_set_selection$(q_{\text{rank}}, \mathcal{X}, |\mathcal{S}|, \boldsymbol{\mu}_{t-1}, \Sigma_{t-1}, N)$
5: $\quad \mathbf{r}_t, l_t \leftarrow$ human response to the query
6: $\quad$ **for** $i = 1, 2, ..., |\mathcal{S}|$ **do**
7: $\quad\quad y \leftarrow 1$ if $i \leq l$, else -1
8: $\quad\quad \boldsymbol{\mu}_t, \Sigma_t \leftarrow$ belief_update$(\mathcal{S}, r_i, y, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$
9: $\quad$ **end for**
10: $\quad t = t + 1$
11: **end while**

Algorithm 4 belief_update

1: **Input:** $\mathcal{S}, i_t, y_t, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}, w, K$
2: $\boldsymbol{\mu}_t, \Sigma_t \leftarrow \boldsymbol{\mu}_{t-1}, \Sigma_{t-1}$
3: **while** $\boldsymbol{\mu}_t, \Sigma_t$ not converged **do**
4: $\quad$ **while** $\boldsymbol{\mu}_t, \Sigma_t$ not converged **do**
5: $\quad\quad \xi^2 = w^2\mathbf{x}^T\boldsymbol{\Sigma}_t\mathbf{x} + w^2(\mathbf{x}^T\boldsymbol{\mu}_t)^2$
6: $\quad\quad \boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t-1}^{-1} + 2\frac{\tanh(\xi/2)}{4\xi}w^2\mathbf{x}\mathbf{x}^T$
7: $\quad\quad \boldsymbol{\mu}_t = \boldsymbol{\Sigma}_t\left[\boldsymbol{\Sigma}_{t-1}^{-1}\boldsymbol{\mu}_{t-1} + \left(y - \frac{1}{2}\right)w\mathbf{x}\right]$
8: $\quad$ **end while**
9: $\quad \boldsymbol{\mu}_t, \Sigma_t \leftarrow \underset{\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q}{\text{argmin}} \text{KL}\left(\mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)||\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\right)$
$\quad\quad\quad\quad -K\mathbf{x}_i^T\boldsymbol{\mu}_q$
$\quad\quad\quad\quad + \log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\mu}_q + \frac{1}{2}\mathbf{x}_j^T\boldsymbol{\Sigma}_q\mathbf{x}_j\right)$
10: **end while**
11: **Output:** $\boldsymbol{\mu}_t, \Sigma_t$

Algorithm 5 item_set_selection

1: **Input:** $q, \mathcal{X}, |\mathcal{S}|, \boldsymbol{\mu}, \Sigma, N, w, K$
2: $\mathcal{S} \leftarrow \{\}$
3: **for** $i = 1, 2, ..., |\mathcal{S}|$ **do**
4: $\quad \{\widehat{\boldsymbol{\theta}}_n\}_{n=1}^N \leftarrow$ sample i.i.d. from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$
5: $\quad \mathbf{s} \leftarrow$ select item from dataset with Eq. (9)
6: $\quad \mathcal{S} \leftarrow \{\mathcal{S}, \mathbf{s}\}$
7: **end for**
8: **Output:** $\mathcal{S}$

setting, which we describe next, provides a lower variance and computationally cheaper approximation.

We approximate the classifier's density function as a Gaussian distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We may then compute the posterior mean and variance given an item label by an iterative process [30] described in lines 4 to 7 of Algorithm 4. In a similar fashion, we approximate the classifier posterior given the item selected with a variational approach. We look for the variational distribution $q(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ closest, in terms of the Kullback-Leibler (KL) distance, to the true posterior. This is equivalent to finding the distribution that maximizes the log Evidence Lower BOund (ELBO) [31]

$$\begin{aligned}\text{ELBO}(q) = &- \text{KL}(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})) + \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[K\mathbf{x}_i^T\boldsymbol{\theta}\right] \\ &- \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log \sum_{j=1}^{|\mathcal{S}|} \exp\left(K\mathbf{x}_j^T\boldsymbol{\theta}\right)\right], \quad (7)\end{aligned}$$

where $\mathbf{x}_i$ is the embedding of the item selected by the human, $K = \frac{a}{\sigma}$ for q_{pos} queries or $K = -\frac{a}{\sigma}$ for q_{neg} queries, and the prior is $p(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$. The first term in Equation (7) is the KL divergence between two Gaussian distributions,

$$\begin{aligned}\text{KL}(q\|p) = &\frac{1}{2}\left[\log\frac{|\Sigma_p|}{|\Sigma_q|} - d + \boldsymbol{\mu_q}^T\Sigma_p^{-1}\boldsymbol{\mu_q} + \boldsymbol{\mu_p}^T\Sigma_p^{-1}\boldsymbol{\mu_p}\right] \\ &- \boldsymbol{\mu_q}^T\Sigma_p^{-1}\boldsymbol{\mu_p} + \frac{1}{2}tr\left\{\Sigma_p^{-1}\Sigma_q\right\}.\end{aligned}$$

Because of the linearity property of the expectation, we compute the second term in Equation (7) as

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}\left[K\mathbf{x}_i^T\boldsymbol{\theta}\right] = K\mathbf{x}_i^T\boldsymbol{\mu}_q.$$

The third term in Equation (7) has no closed-form solution, but following [32], we apply Jensen's inequality to obtain an upper bound

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim q}&\left[\log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\theta}\right)\right] \\ &\leq \log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\mu}_q + \frac{1}{2}K^2\mathbf{x}_j^T\boldsymbol{\Sigma}_q\mathbf{x}_j\right).\end{aligned}$$

We approximate the posterior distribution given the item selection as a Gaussian distribution whose mean and covariance are obtained by maximizing the ELBO lower bound

$$\begin{aligned}\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} = &\operatorname*{argmin}_{\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q} \text{KL}(q||p) - K\mathbf{x}_i^T\boldsymbol{\mu}_q \\ &+ \log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\mu}_q + \frac{1}{2}K^2\mathbf{x}_j^T\boldsymbol{\Sigma}_q\mathbf{x}_j\right). \quad (8)\end{aligned}$$

Putting these together, Algorithm 4 approximates the posterior by accounting for both the label and item selection. We first update the posterior according to the label received. Then we update the posterior according to the selected item with Equation (8). We repeat both updates until convergence.

When $q_t = q_{\text{rank}}$, we compute the posterior as a recursion of q_{high} queries such that

$$\begin{aligned}\mathbb{P}[\boldsymbol{\theta} \mid \mathcal{F}_t] = \prod_{j=1}^{|\mathcal{S}|}&\frac{\mathbb{P}[r_j \mid \boldsymbol{\theta}, \mathcal{R}_j, q_{\text{high}}]\,\mathbb{P}[y_t \mid \mathbf{x}_{r_j}, \boldsymbol{\theta}]}{\mathbb{P}\left[r_j, y_j \mid \mathcal{R}_j, q_{\text{high}}, \mathcal{F}_{t-1}, \{(r_k, y_k)\}_{k=1}^{j-1}\right]} \\ &\times \mathbb{P}[\boldsymbol{\theta} \mid \mathcal{F}_{t-1}],\end{aligned}$$

where $y_j = \mathbb{1}[j \leq l_t]$. Said differently, Algorithm 4 is applied recursively, starting from the top item in the ranked list.

2) Active Learning Heuristic for Item Set Selection: Active learning [7], [33] looks for the most informative items for human annotation, such that the sample complexity is minimized. This implies querying about the items that provide the most information about the ground truth on expectation. However, this maximization, as defined in line 4 of Algorithm 1, requires computing the posterior over every possible item set and annotator response. There are combinatorially many options to compare, so even using the belief approximation, this approach is often computationally intractable.

To select the items in the query, we approximate the information gain based on query by committee [34]. At each iteration t, we sample N particles $\widehat{\boldsymbol{\theta}}_n \overset{\text{i.i.d}}{\sim} \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t]$. We maximize the disagreement between the prediction of each particle and the mean prediction among all particles as,

$$\begin{aligned}\mathcal{S}_t = \operatorname*{argmax}_{\mathcal{S}\in\binom{|\mathcal{X}|}{|\mathcal{S}|}} &\; H\left[\frac{1}{N}\sum_{n=1}^{N} p_n(o|\mathcal{S})\right] \\ &- \frac{1}{N}\sum_{n=1}^{N} H\left[p_n(o|\mathcal{S})\right],\end{aligned}$$

where $p_n(o|\mathcal{S}) := \mathbb{P}\left[o_t = o \mid \widehat{\boldsymbol{\theta}}_n, q_t, \mathcal{S}\right]$ represents the probability mass function over answers $o \in \mathcal{O}$ for query q_t conditioned on the drawn classifier $\widehat{\boldsymbol{\theta}}_n$, and $H[p] := -\sum_{o\in\mathcal{O}} p(o)\log_2 p(o)$ is the Shannon entropy. The first term

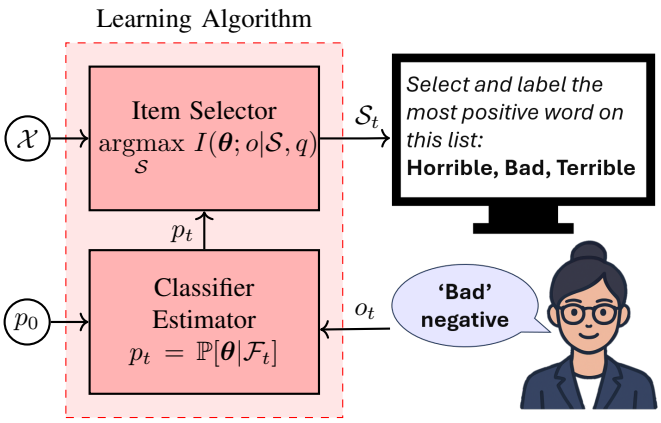


Fig. 2: Block diagram for human-in-the-loop learning for sentiment word classification. At each interaction, the human annotator receives the query with items that maximize the information gain about the ground truth classifier $\boldsymbol{\theta}$. In the example, we ask the annotator to select a word from a list, and provide its label. The answer to the query is used to update the estimator of the classifier and select the next query items.

in the objective function promotes queries with a high uncertainty of the expected output, which avoids queries for which the answer is predictable and thus not very informative. The second term attempts to minimize uncertainty due to intrinsic noise, for example, discouraging asking labels of neutral words such as "table" for which the uncertainty mostly comes from the labeling noise from humans, and not from a lack of exploration.

Actively selecting the item set significantly improves the performance, but there exist combinatorially many sets $\binom{|\mathcal{X}|}{|\mathcal{S}|}$ over which to maximize. To avoid this computational burden, we greedily aggregate a single item

$$\mathbf{s} = \underset{\mathbf{s}\in\mathcal{X}}{\operatorname{argmax}} \quad H\left[\frac{1}{N}\sum_{n=1}^{N} p_n\left(o|\{\mathcal{S},\mathbf{s}\}\right)\right] - \frac{1}{N}\sum_{n=1}^{N} H\left[p_n\left(o|\{\mathcal{S},\mathbf{s}\}\right)\right] \tag{9}$$

to the set until we reach size $|\mathcal{S}|$. Algorithm 5 summarizes the implementation of the active learning heuristic.

III. Results

A. Theoretical Sample Complexity Bounds

We estimate the classifier at step t using the unbiased estimator $\widetilde{\boldsymbol{\theta}}_t = \mathbb{E}[\boldsymbol{\theta}|\mathcal{F}_t]$. Using the arithmetic-geometric mean inequality, we bound this estimator Mean-Square Error (MSE) as

$$\mathrm{MSE}_t = \operatorname{trace}(\Sigma_{\boldsymbol{\theta}|\mathcal{F}_t}) \geq d|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_t}|^{1/d}, \tag{10}$$

where $\Sigma_{\boldsymbol{\theta}|\mathcal{F}_t} = \mathbb{E}\left[(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}_t)(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}_t)^T|\mathcal{F}_t\right]$. Analogously to [6], we observe that a necessary condition to obtain a low MSE is for the determinant of the posterior covariance $|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_t}|$ to be low. Next, we bound the expected number of iterations necessary to achieve a low enough posterior covariance. To facilitate our analysis, we first introduce a set of assumptions. These assumptions are not only useful for analytical tractability but also reflect common conditions or simplifications that align with real-world scenarios.

Assumption III.1. The annotator answer is independent of the history given the classifier, i.e., $p(o_t|\boldsymbol{\theta}, q_t, \mathcal{S}_t, \mathcal{F}_{t-1}) = p(o_t|\boldsymbol{\theta}, q_t, \mathcal{S}_t)$.

Assumption III.2. The label of an item is conditionally independent of the question and the rest of items in the query given the item, i.e., $p(y_t|\mathbf{x}_t, q_t, \mathcal{S}_t) = p(y_t|\mathbf{x}_t)$.

While Assumption III.2 may not be perfectly accurate in every instance, this simplification is justified because the intrinsic score perceived by a human towards an item is primarily determined by the item itself.

Assumption III.3. The information gain of the classifier given a human answer is lower bounded by a positive constant L. For ranking queries $I(\boldsymbol{\theta}; \mathbf{r}_t, l_t|\mathcal{F}_{t-1}) \geq L_r > 0$, for exemplar selection queries $I(\boldsymbol{\theta}; i_t, y_t|\mathcal{F}_{t-1}) \geq L_s > 0, \forall t$.

This assumption implies that the query pool is sufficiently diverse and that the human feedback is strictly more informative than random noise.

Assumption III.4. The prior distribution of the classifier p_0 is uniform over a hypercube $[-M, M]^d$, for some $M > 0.5$.

We assume a bounded parameter space to ensure realizability. Given this constraint, the uniform distribution is the least informative choice, as it is the unique maximum-entropy prior [35].

Our main result is the following:

Theorem III.5. *Let $T_\epsilon = \min\{t : \left|\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\mathcal{F}_t}\right|^{1/d} < \epsilon\}$ be the stopping time of Algorithm 1. Under Assumptions II.1 through III.4, $\mathbb{E}[T_\epsilon]$ is bounded as*

$$\frac{d}{2}\frac{\log_2 \frac{2M^2}{\pi e \epsilon}}{\log_2 N} \leq \mathbb{E}[T_\epsilon] \leq \frac{d}{2L}\log_2 \frac{e^4 d^2 M^2}{2\sqrt{2}(d+2)\epsilon} - 1,$$

with $N = (|\mathcal{S}|+1)!$ and $L = L_r$ for ranking queries $q = q_{rank}$, and with $N = 2|\mathcal{S}|$ and $L = L_s$ for exemplar selection queries $q \in \{q_{high}, q_{low}\}$.

Proof: Let $h(\boldsymbol{\theta}; \mathcal{F}_t) := \mathbb{E}_{\boldsymbol{\theta}|\mathcal{F}_t}[-\log \mathbb{P}(\boldsymbol{\theta}|\mathcal{F}_t)]$ be the entropy of the posterior distribution after observing t interactions. For the lower bound, we note that Algorithm 1 selects the query deterministically as a function of its latest classifier estimator, i.e., $I(\boldsymbol{\theta}; q_t, \mathcal{S}_t|\mathcal{F}_{t-1}) = 0$, so that

$$\begin{aligned}\mathbb{E}_{\mathcal{F}_t}\left[h(\boldsymbol{\theta}; \mathcal{F}_t)\right] &= h(\boldsymbol{\theta}; \mathcal{F}_0) - \sum_{j=1}^{t} I(\boldsymbol{\theta}; o_j, q_j, \mathcal{S}_j|\mathcal{F}_{j-1}) \\ &= d\log_2 2M - \sum_{j=1}^{t} I(\boldsymbol{\theta}; q_j, \mathcal{S}_j|\mathcal{F}_{j-1}) \\ &\quad - \sum_{j=1}^{t} I(\boldsymbol{\theta}; o_j|q_j, \mathcal{S}_j, \mathcal{F}_{j-1}) \\ &\geq d\log_2 2M - t\log_2 N. \end{aligned} \tag{11}$$

The first equality follows from the chain rule of mutual information, the second equality follows from Assumption III.4, and the last inequality holds because $I(\boldsymbol{\theta}; o_t|q_t, \mathcal{S}_t, \mathcal{F}_{t-1})$ is maximized for uniform outcomes. Equation (11) implies

$$\mathbb{E}[T_\epsilon] \geq \frac{d\log_2 2M - \mathbb{E}\left[h(\boldsymbol{\theta}; \mathcal{F}_{T_\epsilon})\right]}{\log_2 N}. \tag{12}$$

As Gaussian distributions maximize entropy for a given covariance [35, Theorem 8.6.5], so that

$$h(\boldsymbol{\theta}; \mathcal{F}_{T_\epsilon}) \leq \frac{d}{2}\log 2\pi e|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_{T_\epsilon}}|^{1/d} \leq \frac{d}{2}\log 2\pi e\epsilon. \tag{13}$$

Combining Equation (12) and (13), we obtain

$$\mathbb{E}[T_\epsilon] \geq \frac{d}{2}\frac{\log_2 \frac{2M^2}{\pi e\epsilon}}{\log_2 N}.$$

To obtain the upperbound, we define the random variable $U_t := \frac{Z_t}{L} - t$, where $Z_t := -h(\boldsymbol{\theta}; \mathcal{F}_t)$. From Lemma B.5 we know U_t is a submartingale that fulfills the conditions of the optional stopping theorem [36], so that

$$\begin{aligned}\frac{\mathbb{E}[Z_{T_\epsilon}]}{L} - \mathbb{E}[T_\epsilon] \geq \frac{\mathbb{E}[Z_0]}{L} - \mathbb{E}[0] \\ \iff \frac{\mathbb{E}[Z_{T_\epsilon}] - \mathbb{E}[Z_0]}{L} \geq \mathbb{E}[T_\epsilon].\end{aligned} \tag{14}$$

By Assumption III.4 the initial entropy is

$$Z_0 = -h(\boldsymbol{\theta}; \mathcal{F}_0) = -d \log_2 2M. \tag{15}$$

From Assumption III.3 and Lemma B.1,

$$\mathbb{E}[Z_{T_\epsilon}] \leq \mathbb{E}[Z_{T_\epsilon - 1}] - L. \tag{16}$$

Since $\mathbb{P}(\boldsymbol{\theta}|\mathcal{F}_t)$ is log-concave, we invoke [6, Lemma 3.1.] to obtain

$$\begin{aligned} Z_{T_\epsilon - 1} = -h(\boldsymbol{\theta}; \mathcal{F}_{T_\epsilon - 1}) &\leq -\frac{d}{2} \log_2 \frac{2|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_{T_\epsilon - 1}}|^{1/d}}{e^4 d^2/(4\sqrt{2}(d+2))} \\ &\leq -\frac{d}{2} \log_2 \frac{8\sqrt{2}(d+2)\epsilon}{e^4 d^2} \\ &= \frac{d}{2} \log_2 \frac{e^4 d^2}{8\sqrt{2}(d+2)\epsilon}. \end{aligned} \tag{17}$$

Substituting Equations (15), (16) and (17) into Equation (14), we obtain the desired upper bound of the expectation

$$\begin{aligned} \mathbb{E}[T_\epsilon] &\leq \frac{1}{L} \left(\mathbb{E}[Z_{T_\epsilon - 1}] - L + d \log_2 2M \right) \\ &\leq \frac{d}{2L} \left(\log_2 \frac{e^4 d^2}{8\sqrt{2}(d+2)\epsilon} + \log_2 4M^2 \right) - 1 \\ &\leq \frac{d}{2L} \log_2 \frac{e^4 d^2 M^2}{2\sqrt{2}(d+2)\epsilon} - 1. \end{aligned}$$

■

Theorem III.5 shows that the estimator uncertainty ϵ decays exponentially with the number of queries $\mathbb{E}[T_\epsilon]$. This result extends the upper bound in [6] to non-equiprobable queries. The parameter M appears because we allow for more freedom in the prior. The more constrained the prior is, i.e., the lower M, the fewer interactions we need. In contrast to the lower bound of [6], we generalize beyond binary queries; thus, the denominator $\log_2 N$ appears in the lower bound, suggesting a lower stopping time may be possible as the query complexity increases. We empirically corroborate the dependence of the stopping time on the question type and item set size in the next section. The lower and upper bounds on the stopping time are monotonically increasing with d, which is consistent with the intuition that more interactions are required to learn classifiers in higher-dimensional embedding spaces. Crucially, both N and L are larger for ranking queries than for exemplar selection queries. Consequently, both bounds on the stopping time are lower for ranking queries, theoretically confirming that their higher information content necessitates fewer user interactions.

B. Empirical Sample Complexity Reduction

We empirically validate Algorithms 2 and 3 on word and image classification tasks with existing crowdsourced datasets. To facilitate further exploration and validation by the research community, we provide the code for replicating our experiments[1] [37].

[1]https://github.com/BelenMU/HiTL-SentimentClassify/

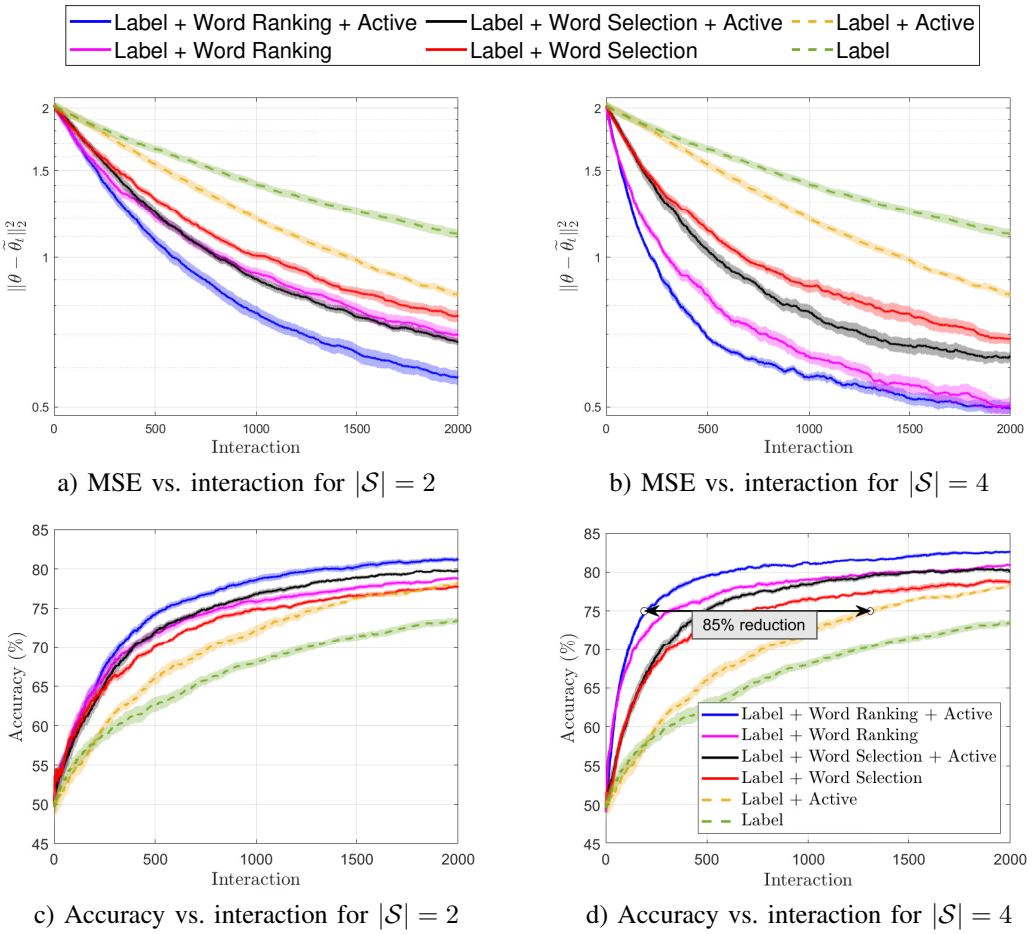
1) Word Sentiment Classification: We first focus on the binary word sentiment classification task [38]. While humans can intuitively label words according to their connotation as positive (e.g., healthy) or negative (e.g., scary) [39], justifying the categorization in machine-interpretable terms proves challenging for most. We test the performance of our algorithms in learning this implicit knowledge from humans. We use the list of most frequent words in the decade of the 2000s [16]. For every word w, we simulate the implicit human score by sampling from $\mathcal{N}(\mu_w, \sigma_w^2)$, where μ_w and σ_w^2 are the mean and variance of the valence score as given by the dataset [16]. This simulation approximates a realistic distribution of human sentiment scores for each word and captures inter-subject variability in valence assessments.

We use existing 300-dimensional monolingual word embeddings [13] to which we prepend a 1; this way, the parameter $\boldsymbol{\theta}$ accounts for both the direction and the offset of the hyperplane characterizing the classifier. We define the ground-truth classifier $\boldsymbol{\theta} \in \mathbb{R}^{301}$ as the hyperplane that minimizes the labeling error over all words in the dataset. Figures 3a and 3b show how the distance from the estimator to the ground truth decreases as more queries are collected. All the Bayesian strategies lead to an MSE reduction, but the convergence speed markedly differs. The richer the query, the faster the decrease of the error: q_{rank} outperforms q_{high} and q_{low}, which in turn outperform traditional labeling queries. Additionally, actively choosing the items in the queries boosts performance. In fact, word selection between two actively chosen items surpasses ranking two randomly selected words. Algorithms 2 and 3 thus facilitate a faster reduction in MSE, effectively reducing sample complexity.

The lower bound in Theorem III.5 suggests that larger word sets should accelerate learning; we observe this behavior empirically when comparting Figures 3a and 3b. Figure 4 confirms this effect for word selection queries: after 1000 iterations, the MSE is about 20% lower when the annotator chooses from 10 options instead of 2.

Beyond MSE, we also assess the word classification accuracy. Given a word w with embedding $\mathbf{x}_w$, we report the prediction $\text{sign}(\widetilde{\boldsymbol{\theta}}^T \mathbf{x}_w)$ is accurate when it matches its label $\text{sign}(\mu_w)$. To measure the predictor $\widetilde{\boldsymbol{\theta}}$ accuracy, we consider the words with $|P[Y_w = 1] - 0.5| = |\int_0^\infty f_{\mathcal{N}}(\mu_w, \sigma_w^2) - 0.5| \leq 0.1$, where $f_{\mathcal{N}}$ represents the probability density function of a normal distribution. This range is selected to avoid words that have a completely neutral score, like "branch" or "mouth." We focus on words with stronger sentiment scores, for which the prediction accuracy of our algorithm can be most meaningfully assessed.

Figures 3c and 3d show how classification accuracy evolves with the number of interactions. As a baseline, using only randomly chosen labeling queries leads to a slow increase in accuracy, requiring more than 2000 interactions to reach 75% accuracy. When labels are collected using active learning, the same accuracy is achieved after about 1300 labeling queries. Allowing the annotator to also select the most positive or most negative word among four candidates yields further gains: with randomly chosen word sets, 75% accuracy is reached in roughly



a) MSE vs. interaction for $|\mathcal{S}| = 2$

b) MSE vs. interaction for $|\mathcal{S}| = 4$

c) Accuracy vs. interaction for $|\mathcal{S}| = 2$

d) Accuracy vs. interaction for $|\mathcal{S}| = 4$

Fig. 3: Performance of the human in the loop learning algorithms with human data on the word sentiment analysis task. All configurations are run with 10 different random initializations. The lines represent the mean of those experiments, while the shaded areas represent the standard error. Adding word selection or ranking to the queries together with actively selecting the word set reduces the number of iterations needed to achieve a good performance.

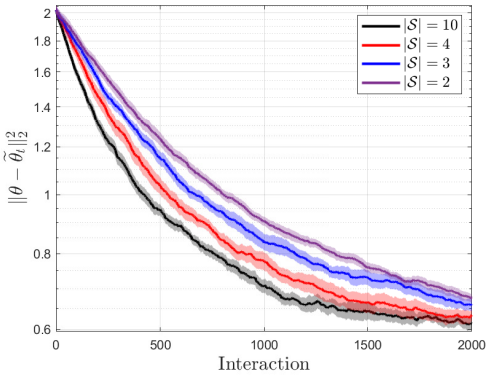


Fig. 4: Performance of Algorithm 2 with human data on word sentiment classification. The larger the word set, the faster the decrease of MSE, as suggested by the lower bound in Theorem III.5.

700 interactions, dropping to about 500 interactions when the word sets are chosen actively. Ranking queries q_{rank} offer the most substantial acceleration, requiring just 315 rankings of four random words or 196 rankings of actively chosen words to match the 75% accuracy. This represents a notable reduction of approximately 85% in the number of interactions needed compared to active labeling queries. These results demonstrate that our method not only improves estimator alignment metrics but also enhances the performance of the downstream classification task. Our results confirm the efficiency and practicality of Algorithms 2 and 3 for valence classification.

2) Image Aesthetic Classification: We further evaluate our approach on the task of binary image aesthetic classification. While humans naturally perceive the beauty of an image, automatically quantifying this aesthetic quality remains a challenging computational task [40]. To test the efficiency of our algorithms in learning this subjective quality, we utilize the Aesthetic Visual Analysis (AVA) dataset [18]. We focus on a subset of over 20,000 landscape images, which we embed into a 768 dimensional space with CLIP [14] and prepend a one. Each image is associated with a distribution of scores between 1 and 10 collected from an online photography community. We define the ground-truth binary labels by using the dataset median mean score as a threshold $\tau = 5.58$, creating two balanced classes. To simulate human

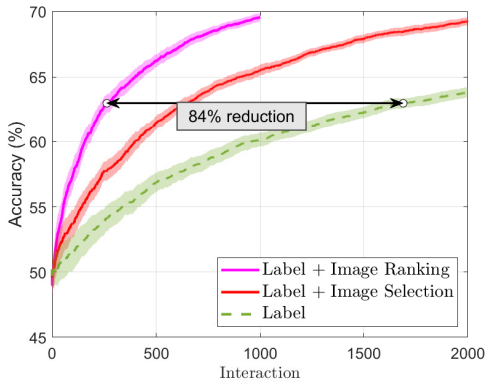


Fig. 5: Performance of the human in the loop learning algorithms with human data on the image aesthetic classification task across 10 initializations. There are $|\mathcal{S}| = 4$ candidate images for ranking and selection questions. The accuracy increases faster when asking richer queries.

evaluations, we leverage the crowdsourced score distributions: for every queried image, we sample a score from its empirical distribution and deterministically map these scores to query responses. Namely, we compare the sampled scores against the threshold τ to simulate human labeling; we select the image with the maximum or minimum sampled score to answer q_{high} and q_{low}, respectively; and we order the images according to their sampled scores to simulate a human ranking. Figure 5 shows how the classification accuracy of the learned classifier evolves as more feedback is gathered. Consistent with our previous experiments, richer queries require substantially fewer interactions to achieve a given accuracy. In particular, q_{rank} yields the fastest accuracy gains, followed by the selection queries q_{high} and q_{low}, while traditional labeling leads to the slowest improvement. In fact, solely labeling does not reach a 65% accuracy within 2000 interactions. In contrast, augmenting labels with image selection or ranking reduces the number of required interactions to approximately 400 and 900, respectively. Similarly, achieving a 63% accuracy with q_{rank} requires 84% less interactions than traditional labeling.

C. Empirical Time Savings in Cost Aware Query Selection

1) Human Cost Model: To accurately maximize the information rate defined in Equation (6), we require a model of the human cost. Since the data collection costs are typically driven by collection times [41], [42], we define the cost as the expected time in seconds required for an annotator to answer a query. This response time depends on both the question type q (ranking vs. selection) and the set size $|\mathcal{S}|$.

We estimated this cost model in the context of the word sentiment classification task using crowdsourced experiments. We recruited participants through Prolific[2] and recorded their response times for different combinations of question type and set size. Because we expect word selection questions q_{low} and q_{high} to incur similar times, we restricted data collection

[2]The study was categorized as minimal risk research qualified for exemption status by the Institutional Review Board (IRB).

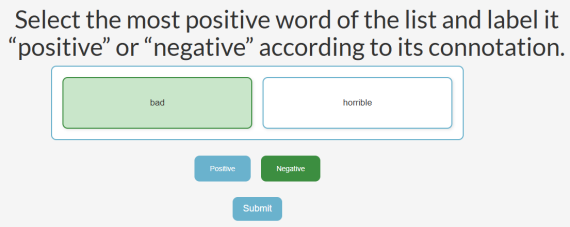


Fig. 6: User interface for selection query. User must select the most positive word in the list, in this case "bad" and label the word according to its connotation, in this case "negative".

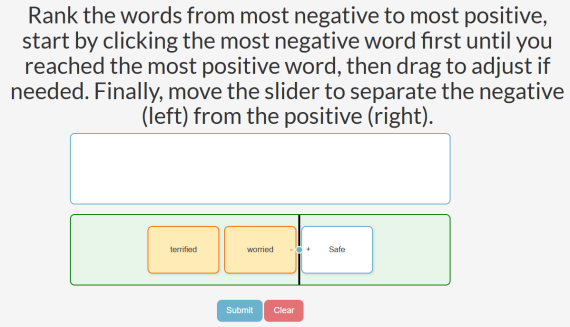


Fig. 7: User interface for ranking query. The user must rank the words from the most negative word in the list "terrified" to the most positive "safe", and then use the vertical bar to separate words with positive and negative connotations.

for selection queries to q_{high}. Each participant answered 20 q_{rank} queries and 20 q_{high} queries (in addition to 5 gold-standard questions per question type, used as attention checks and excluded from the analysis). We followed A/B testing guidelines and randomly assigned annotators to start with either the word selection or the word ranking queries. After excluding the participants who failed the attention checks, data from 83 participants remained for selection queries and 63 for ranking queries. Note that there is substantial overlap, most participants contributed to both question types. The graphic user interfaces used to record response times are shown in Figures 6 and 7. Additional details of the study may be found in Section C of the supplemental material.

To select an appropriate parametric form for response time, we compared candidate models using a Vuong closeness test, a likelihood-ratio based test for comparing non-nested models. See Appendix C-B for further details on the candidate models and statistical results. Under our experimental conditions, a linear model $\widehat{t} = \beta_0 + \beta_1|\mathcal{S}|$ describes the time response to selection queries significantly better than a logarithmic model $\widehat{t} = \beta_0 + \beta_1 \log|\mathcal{S}|$ (p-value $= 10^{-5}$). Similarly, the response times collected for ranking queries are significantly better explained by a linear model than by a purely quadratic model $\widehat{t} = \beta_0 + \beta_1|\mathcal{S}|^2$ (p-value $= 4 \times 10^{-5}$). Motivated by these results, we fitted linear models with least-squares regression on individual response times and modeled the human response times as

$$\widehat{t}_{\text{high}} = \widehat{t}_{\text{low}} = 4.01 + 0.63|\mathcal{S}| \text{ and } \widehat{t}_{\text{rank}} = -0.32 + 4.41|\mathcal{S}|.$$

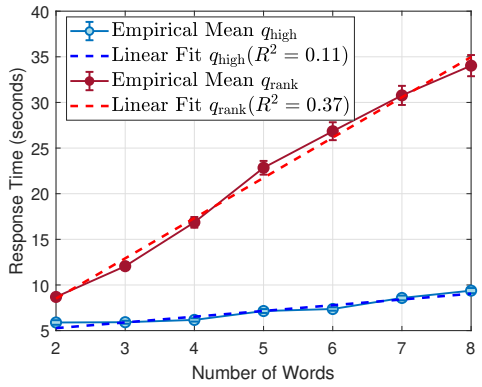


Fig. 8: Mean response time from crowdsourced experiments and linear model for different question types and word set sizes. The points show the empirical mean of the data, and the bars show the standard error among all the combined response times collected. We observe that the linear models closely track the empirical means across word set sizes.

Figure 8 shows the fitted predictions against the empirical mean response times for each set size. The linear models capture the overall increase in response time with set size. Although substantial trial-to-trial variability remains, this is typical in human response-time data. As expected, the slope for ranking queries is steeper than for word selection queries, indicating that adding items to a set incurs a substantially higher time penalty when participants must produce a full ranking and labeling rather than identify and label a single most positive word from the set.

2) Information Gain Ratios: Maximizing the information rate in Equation (6) also requires estimates of the expected information gained for each combination of question type and set size. Directly recomputing these quantities at every iteration would be computationally expensive. Instead, we exploit an empirical regularity in the information gain ratios across queries.

As illustrated in Figure 9, the ratios of information gain between different queries remain approximately constant across interactions. In particular, more complex queries consistently yield higher information, and in all our experiments, we observe that these relative advantages are stable not only across iterations but also across different initializations of the algorithm. This behavior allows us to estimate the ratios of expected information gain for each question type and set size relative to labeling from a small number of computations. Concretely, we compute the expected information gain as in Subsection II-C2 for a few initial conditions and iterations. Then, we use the resulting average information gain ratio relative to the labeling query as a proxy for the proportional information gain in Equation (6). The values of these proxies strongly depend on the annotator noise σ, as well as on the slope of the linear relationship between item distance to the classifier and score, i.e., a in Equation (2). This dependence is shown in Figure 10. When a/σ is large, it is easier to distinguish the score ordering of items that are close in the embedding space, thereby increasing the benefit of richer queries.

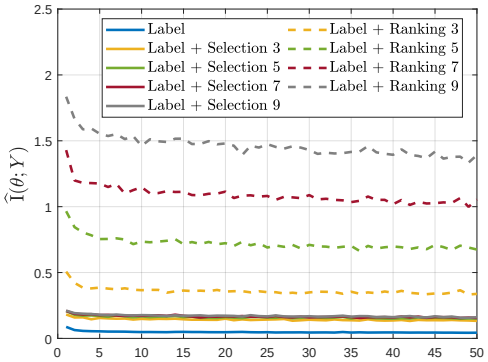


Fig. 9: Estimated information gain for several query types. We empirically observe that the ratio of information gain between query types stays approximately constant with interactions.

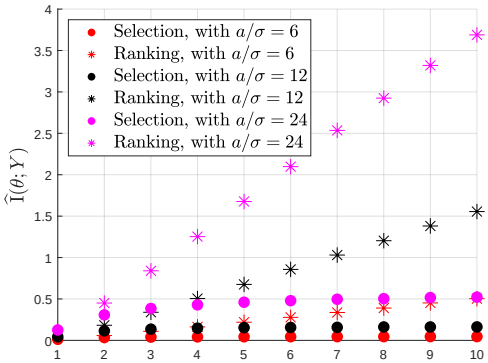
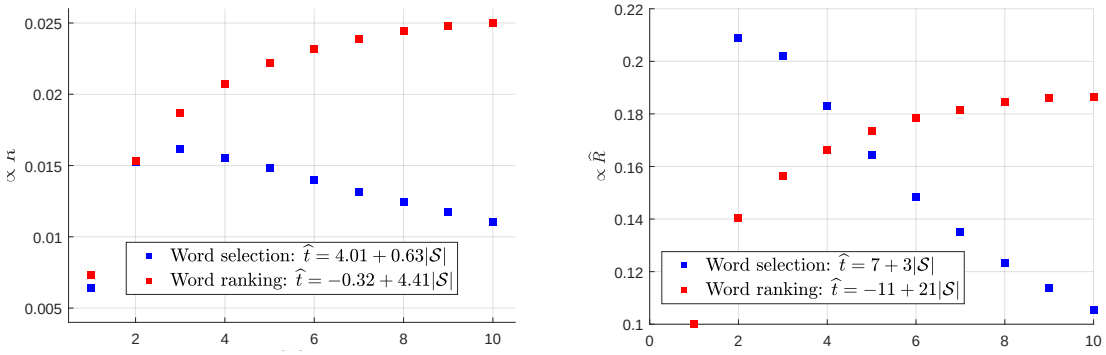


Fig. 10: Estimated information gain for different slope and noise factor ratios a/σ.

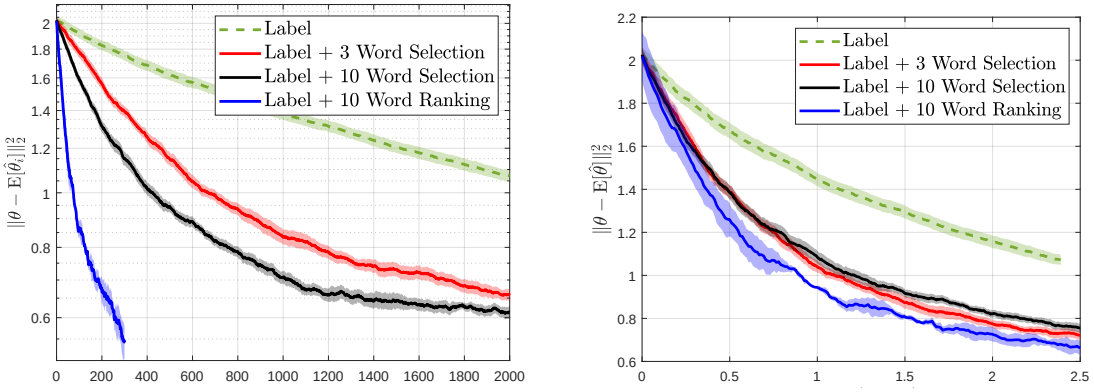
For a fixed set size $\mathcal{S}$, ranking queries consistently have a higher estimated information gain than selection queries. For a fixed question type, increasing $|\mathcal{S}|$ raises the expected information gain. However, these improvements exhibit diminishing returns, especially when a/σ is small.

3) Question and Set Size Selection: Our goal is to select both the question type and the set size $|\mathcal{S}|$ that maximize information rate, by balancing information gain and human effort. Once the response time and the relative information gains of different query configurations have been estimated, we select the combination according to Equation (6). Figure 11 summarizes this trade-off for the word sentiment classification task. Figure 11a shows the predicted information rate for each question and set size when using the response time models fitted from human experiments. For any fixed $|\mathcal{S}|$, ranking queries achieve a higher information rate than selection queries. The two question types exhibit different behavior as $|\mathcal{S}|$ increases. For ranking, the information rate grows monotonically with $|\mathcal{S}|$ over the feasible range ($|\mathcal{S}| \leq 10$). For selection questions, however, the information rate peaks at $|\mathcal{S}| = 3$ and decreases for both smaller and larger set sizes. This optimal query depends strongly on the underlying information gain estimation and cost models. For example, Figure 11b illustrates the information rates when the response time models are altered; in this case, the optimal query becomes a selection question with two words.



a) Response time modeled from experiments b) Synthetic response time model

Fig. 11: Predicted ratio of information gain over response time on the word sentiment classification task for different response time models. To maximize the rate with the interface we tested, we should query ranking questions with 10 words. However, if an interface resulted on response time model in (b) we should query selection questions with 2 words.



a) Sample complexity b) Time cost

Fig. 12: MSE evolution on the word sentiment classification task. The gains in sample complexity with complex queries is not so prominent in time cost. In fact, while the sample complexity decreases faster with word selections of size 10 vs. 3, the time cost decreases slower.

Different query selections translate into distinct learning behaviors. Figure 12a compares the reduction in MSE as a function of the number of interactions on the word sentiment classification task. Larger and richer queries exhibit a strong advantage in sample complexity. When measured in wall-clock time, however, the picture changes. As shown in Figure 12b, the additional time required for larger sets can offset their sample-complexity gains, so that selection from 10 words reduces error more slowly in time than selection from 3 words. Consistent with our analysis, ranking queries with 10 items offer clear benefits in both sample complexity and time, but the improvements in time are less pronounced than in interaction count. These results underscore the importance of optimizing for information rate rather than sample complexity alone when costs are query dependent.

IV. Conclusion and Future Work

We have presented a framework for incorporating nuanced expert feedback into interactive learning. By exploiting embedding geometries, we have designed human-in-the-loop algorithms that use exemplar selection and ranking queries, providing richer supervision than standard label queries. We have shown, both theoretically and empirically, that these queries reduce sample complexity and accelerate learning. We have also proposed a query-selection strategy that accounts for query-dependent costs and demonstrated time savings on a word sentiment classification task, *moving towards more cost-effective alignment between models and human expertise*.

Several directions remain open. The relationship between scores and embeddings suggests new query types, such as asking which item a user can label most confidently or is most uncertain about. As in much of the literature, we assume human responses are conditionally independent given latent parameters, which may fail due to context effects or fatigue. Future work includes learning user-specific behavior models or adapting query policies online to user state.

REFERENCES

[1] A. Helbling, C. J. Rozell, and M. O'shaughnessy, "PrefGen: Preference Guided Image Generation with Relative Attributes."

[2] X. Chen, Y. As, and A. Krause, "Learning Safety Constraints for Large Language Models," in *International Conference on Machine Learning*, 5 2025.

[3] J. Zhu and E. Hovy, "Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem," in *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague: Association for Computational Linguistics, 2007, pp. 783–790.

[4] S. Roy, T. Meena, and S.-J. Lim, "Demystifying Supervised Learning in Healthcare 4.0: A New Reality of Transforming Diagnostic Medicine," *Diagnostics*, vol. 12, no. 10, p. 2549, 10 2022.

[5] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell, "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," no. 3, 7 2023.

[6] G. H. Canal, A. K. Massimino, M. A. Davenport, and C. J. Rozell, "Active embedding search via noisy paired comparisons," *Proc. of International Conference on Machine Learning*, vol. 2019-June, pp. 1493–1512, 2019.

[7] S. Shekhar, M. Ghavamzadeh, and T. Javidi, "Active Learning for Classification With Abstention," *Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 705–719, 6 2021.

[8] B. Settles, "Active Learning Literature Survey," University of Wisconsin, Madison, Tech. Rep., 2009.

[9] Z. Ashktorab, M. Desmond, J. Andres, M. Muller, N. N. Joshi, M. Brachman, A. Sharma, K. Brimijoin, Q. Pan, C. T. Wolf, E. Duesterwald, C. Dugan, W. Geyer, and D. Reimer, "AI-Assisted Human Labeling: Batching for Efficiency without Overreliance," *Proc. of Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–27, 2021.

[10] G. H. Canal, M. Bloch, and C. J. Rozell, "Feedback Coding for Active Learning," in *AISTATS Artificial Intelligence and Statistics 2021*, 2021.

[11] S. Shekhar, M. Ghavamzadeh, and T. Javidi, "Active Learning for Binary Classification with Abstention," in *IEEE International Symposium on Information Theory*. IEEE, 2020, pp. 2801–2806.

[12] B. Martin-Urcelay, C. J. Rozell, and M. R. Bloch, "Enhancing Human-in-the-Loop Learning for Binary Sentiment Word Classification," in *Conference on Decision and Control (CDC)*. IEEE, 12 2024, pp. 2293–2298.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. of International Conference on Machine Learning*, 2 2021.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. of Conference on Automatic Face & Gesture Recognition*, 5 2018.

[16] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 595–605.

[17] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, I. Gurevych and Y. Miyao, Eds., vol. 1. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 174–184.

[18] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, 2012.

[19] Z. Zhang, Y. Song, and H. Qi, "Age Progression/Regression by Conditional Adversarial Autoencoder," *Computer Vision and Pattern Recognition (CVPR)*, 3 2017.

[20] Q. Liu, H. Huang, Y. Gao, X. Wei, Y. Tian, and L. Liu, "Task-oriented word embedding for text classification," in *Proc. of International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Santa Fe, 2018, pp. 2023–2032.

[21] H.-O. Georgii, *Gibbs Measures and Phase Transitions*. DE GRUYTER, 5 2011.

[22] K. E. Train, *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press, 2003, vol. 9780521816.

[23] R. D. Luce *et al.*, *Individual choice behavior*. Wiley New York, 1959, vol. 4.

[24] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 24, no. 2, pp. 193–202, 1975.

[25] C. Tiferet-Dweck, A. Keegan, and K. Unger, "Constraints on multi-item working memory access: performance costs and retrieval dynamics," *Frontiers in Psychology*, vol. 16, p. 1558689, 2025.

[26] B. Settles, M. Craven, and L. Friedland, "Active Learning with Real Annotation Costs," in *NIPS Workshop on Cost-Sensitive Learning*, vol. 227, 1 2008.

[27] A. Kapoor, E. Horvitz, and S. Basu, "Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning," in *Proc. of international joint conference on Artifical intelligence*, 1 2007.

[28] G. Canal, S. Fenu, and C. Rozell, "Active ordinal querying for tuplewise similarity learning," *Proc. of Conference on Artificial Intelligence*, pp. 3332–3340, 2020.

[29] R. Ranganath, S. Gerrish, and D. M. Blei, "Black box variational inference," in *Journal of Machine Learning Research*, vol. 33, 2014, pp. 814–822.

[30] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, pp. 25–37, 2000.

[31] Kevin P. Murphy, *Probabilistic Machine Learning Advanced Topics*. MIT Press, 2022.

[32] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 324–335, 12 2007.

[33] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 6 2012.

[34] A. Kachites McCallum and K. Nigam, "Employing EM and Pool-Based Active Learning for Text Classification," in *Proc. of International Conference on Machine Learning*, 1998, pp. 350–358.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 9 2005.

[36] D. Williams, *Probability with martingales*. Cambridge university press, 1991.

[37] B. Martin-Urcelay, "Beyond labels: Information-efficient human-in-the-loop learning via ranking and selection queries," 2024, gitHub repository. [Online]. Available: https://github.com/BelenMU/HiTL-SentimentClassify

[38] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," *Proc. of International Conference on Knowledge Capture*, pp. 70–77, 2003.

[39] M. R. R. Rana, A. Nawaz, and J. Iqbal, "A survey on sentiment classification algorithms, challenges and applications," *Acta Universitatis Sapientiae, Informatica*, vol. 10, no. 1, pp. 58–72, 8 2018.

[40] A. Anwar, S. Kanwal, M. Tahir, M. Saqib, M. Uzair, M. K. I. Rahmani, and H. Ullah, "A survey on image aesthetic assessment," *arXiv preprint arXiv:2103.11616*, 2021.

[41] S. Clancy, S. Bayer, and R. Kozierok, "Active learning with a human in the loop," The MITRE Corporation, Technical Report MTR 12-0603, 11 2012.

[42] G. Sigurdsson, O. Russakovsky, A. Farhadi, I. Laptev, and A. Gupta, "Much ado about time: Exhaustive annotation of temporal data," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 4, 2016, pp. 219–228.

[43] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 9 1977.

[44] L. Lovász and S. Vempala, "The geometry of logconcave functions and sampling algorithms," *Random Structures and Algorithms*, vol. 30, no. 3, pp. 307–358, 5 2007.

[45] R. Durrett, *Probability: Theory and Examples*, 5th ed. Cambridge University Press, 2019.

[46] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. of International Conference on Machine Learning*. New York: ACM Press, 2009.

APPENDIX A
LINEAR RELATIONSHIP BETWEEN EMBEDDINGS AND CLASSIFIERS

Assumption II.1 is motivated by the observed relationship between implicit human scores and the geometry of existing embedding spaces. As Figure 1 shows, there exists a linear relationship between human scores and the distance from items to the MMSE classifier in several standard embedding spaces.

A. Empirical Evidence Across Domains

Previous influential work [43] identifies three fundamental dimensions of meaning: *valence*, representing the spectrum between positivity and negativity; *arousal*, representing the contrast between active and passive emotions; and *dominance*, capturing the power dynamics from submissive to dominant. Figures 1a, 13a and 13b demonstrate that the scores across all three dimensions, as provided by the National Research Council Canada (NRC) dataset [17], vary linearly with the distance to the MMSE classifier. This linear behavior is consistent across distinct and independently gathered datasets [16] of valence scores: Figure 1b shows the learned human scores of the top-5000 most frequent non-stop words in the decade of the 2000s, the same dataset used for the word sentiment classification tasks in Sections III-B and III-C. Figure 13c examines the adjectives appearing more than 100 times in the data from the 2000s, we observe a similar linear behavior in their mean score in these distinct datasets.

We construct our word embedding representations through three steps. First, each word is mapped to a 300-dimensional vector using the standard *word2vec* mapping[3] [13]. Second, the embedded words are normalized to unit norm. Finally, following standard practice for linear classification, we prepend a constant value of 1 to each normalized vector, yielding 301-dimensional feature vectors. This augmentation allows the classifier $\boldsymbol{\theta} \in \mathbb{R}^{301}$ to capture both the decision boundary direction and offset.

Next, we show that the linear relationship holds in the visual domain. Figure 1c shows average aesthetic scores from human raters for 21,979 landscape images in the AVA dataset [18] versus their distance to the MMSE classifier in the 769-dimensional (we prepend a 1) ViT-L/14 embedding space pretrained on CLIP [14]. The classifier separates high and low aesthetic images using the median score (5.48) as threshold. We observe a strong correlation between score and distance.

Figure 1d compares the ages of 23,625 aligned and cropped face images from the UTKFace dataset [19] against their distance to a classifier separating faces under 30 years from older faces. Face images are embedded using InceptionResnetV1 pretrained on VGGFace2 [15], yielding 512-dimensional vectors to which we prepend a constant value of 1. We observe a linear score-distance relationship again in this image embedding space.

B. Gumbel Noise Model Validation

We model the noise in Equation (2) as extreme-value distributed because it yields the Boltzmann choice model, a canonical framework in behavioral economics and discrete choice theory [21]. To validate this modeling assumption empirically, we examine the residuals $\delta_i = \text{score}(\mathbf{x}_i) - (a\mathbf{x}_i^T\boldsymbol{\theta} + b)$ against both $\mathrm{Gumbel\text{-}Max}$ and $\mathrm{Gumbel\text{-}Min}$ distributions.

Figure 14 shows diagnostic plots for the word valence (a) and image aesthetic (b) datasets. The $\mathrm{Gumbel\text{-}Max}$ distribution shows a superior fit, with Kolmogorov-Smirnov statistics of 0.095 for the word valence dataset and 0.030 for the image aesthetic dataset (indicating maximum CDF deviations of 3% and 9.5%, respectively). Visually, the residual histograms align well with the fitted PDFs, the empirical and theoretical CDFs nearly overlap, and the Q-Q plot points closely follow the theoretical line across the central range of the distribution. While some misspecification appears in the extreme tails, the $\mathrm{Gumbel\text{-}Max}$ distribution provides both tractable human decision-making models and a practical approximation for 60-80% of the data distribution in both datasets.

The $\mathrm{Gumbel\text{-}Min}$ distribution shows acceptable but slightly weaker fit, with Kolmogorov-Smirnov statistics of 0.136 for word valence and 0.123 for image aesthetics. While these values indicate larger maximum deviations compared to $\mathrm{Gumbel\text{-}Max}$, they remain within conventional bounds for acceptable model fit in behavioral data analysis.

APPENDIX B
AUXILIARY LEMMAS AND PROOFS

Lemma B.1. *The posterior distribution of the classifier given the history $p(\boldsymbol{\theta}|\mathcal{F}_t)$ is log-concave (LCC).*

Proof:

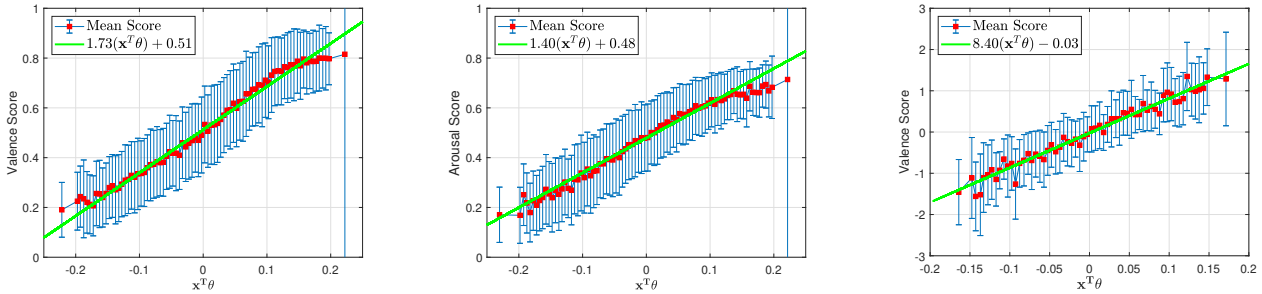
$$
\begin{aligned}
p_t(\boldsymbol{\theta}) :&= p(\boldsymbol{\theta}|\mathcal{F}_t) = p(\boldsymbol{\theta}|o_t, q_t, \mathcal{S}_t, \mathcal{F}_{t-1}) \\
&\overset{(1)}{=} \frac{p(o_t|q_t, \mathcal{S}_t, \boldsymbol{\theta}, \mathcal{F}_{t-1}) p(\boldsymbol{\theta}|q_t, \mathcal{S}_t, \mathcal{F}_{t-1})}{p(o_t|q_t, \mathcal{S}_t, \mathcal{F}_{t-1})} \\
&\overset{(2)}{=} \frac{p(o_t|q_t, \mathcal{S}_t, \boldsymbol{\theta})}{p(o_t|\mathcal{F}_{t-1})} p(\boldsymbol{\theta}|\mathcal{F}_{t-1}) \\
&\overset{(3)}{=} \prod_{j=1}^{t} \frac{p(o_j|q_j, \mathcal{S}_j, \boldsymbol{\theta})}{p(o_j|\mathcal{F}_{j-1})} p_0(\boldsymbol{\theta}) \\
&\overset{(4)}{=} \frac{\prod_{j=1}^{t} p(o_j|q_j, \mathcal{S}_j, \boldsymbol{\theta})}{p(\mathcal{F}_t)} p_0(\boldsymbol{\theta})
\end{aligned}
$$

where

(1) follows from Bayes theorem,
(2) follows from Assumption III.1, and because given the past history the query and word set are selected deterministically by the algorithm,
(3) follows by induction,
(4) follows from the law of total probability.

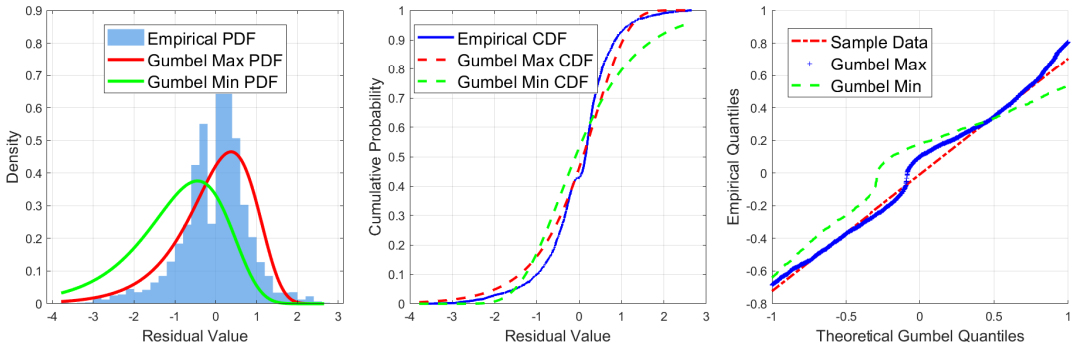
Combining Equations (1)-(5) and Assumption III.2, we deduce that the likelihood of the human response, $p(o_j|q_j, \mathcal{S}_j, \boldsymbol{\theta})$, is given by a product of softmax and logistic functions, both of which are LCC. From Assumption III.4, the prior $p_0(\boldsymbol{\theta})$ is sampled from a uniform distribution, which is also LCC. The product of LCC functions is also LCC, thus the posterior $p_t(\boldsymbol{\theta})$ is LCC. ■

[3] http://ixa2.si.ehu.es/martetxe/vecmap/en.emb.txt.gz

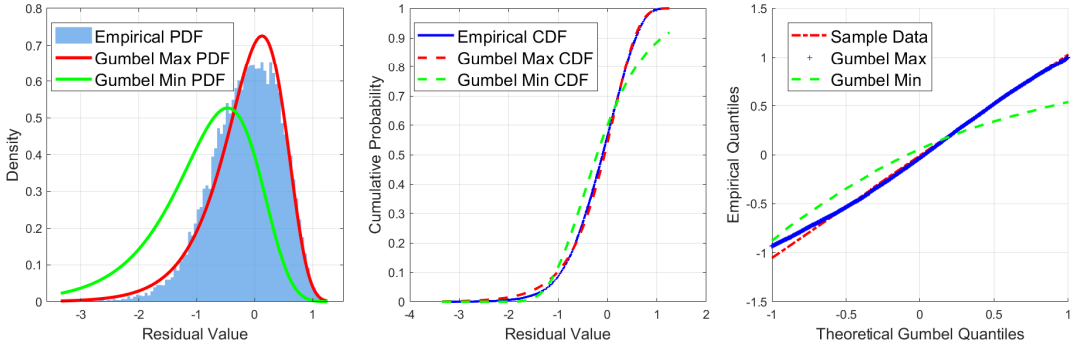


a) Valence in NRC b) Arousal in NRC Valence of adjectives in SocialSent

Fig. 13: Relationship between the empirical mean score of words, given by the NRC [17] or SocialSent [16] lexicons, and the distance of their word2vec embeddings to the ground truth classifier. We observe there is an approximately linear relationship.



a) Word Valence



b) Image Aesthetic

Fig. 14: Goodness-of-fit diagnostics for the Gumbel-Max and Gumbel-Min noise assumption in the linear score model. Each row shows, from left to right, the empirical residual density vs. fitted Gumbel PDFs, the empirical CDF vs. fitted Gumbel CDFs, and a Q–Q plot comparing empirical and theoretical Gumbel quantiles.

Lemma B.2. *The likelihood of any word and label pair* $\mathbf{x} \in \mathcal{X}$ *and* $y \in \{-1, 1\}$ *is lower bounded as*

$$\min_{\mathbf{x},y,\mathcal{S},\boldsymbol{\theta},q\in\{q_{high},q_{low}\}} \mathbb{P}\left[\mathbf{x}, y|\mathcal{S}, \boldsymbol{\theta}, q\right] \geq \gamma_1\gamma_2 > 0,$$

with $\gamma_1 := \frac{\exp\left(-\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right)}{\exp\left(-\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right)+(|\mathcal{S}|-1)\exp\left(\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right)}$ *and* $\gamma_2 = \frac{1}{1+\exp\left(wM\sqrt{P_{\mathbf{x}}d}\right)}$.

Proof: By construction, all word embeddings are bounded, i.e., $\|\mathbf{x}\|_2^2 \leq P_{\mathbf{x}}\}$. From Assumption III.4, we know the norm of the classifier is bounded by $\|\boldsymbol{\theta}\|_2^2 \leq dM^2$. We leverage these constraints to bound the likelihood.

Note that

$$\exp\left(-\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right) \leq \exp\left(k\mathbf{x}^T\boldsymbol{\theta}\right) \leq \exp\left(\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right),$$

where $k = \frac{a}{\sigma}$ when $q = q_{\text{high}}$ and $k = -\frac{a}{\sigma}$ when $q = q_{\text{low}}$. Therefore,

$$\min_{\mathbf{x},\mathcal{S},\boldsymbol{\theta},q} \mathbb{P}\left[\mathbf{x}|\mathcal{S}, \boldsymbol{\theta}, q\right]$$
$$\geq \frac{\exp\left(-\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right)}{\exp\left(-\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right) + (|\mathcal{S}|-1)\exp\left(\left|\frac{a}{\sigma}\right| M\sqrt{P_{\mathbf{x}}d}\right)} =: \gamma_1.$$

In a similar fashion, we use Assumption III.2 to bound

$$\min_{y,\mathbf{x},\mathcal{S},\boldsymbol{\theta},q} \mathbb{P}\left[y|\mathbf{x},\boldsymbol{\theta},q,\mathcal{S}\right]$$
$$= \min_{\mathbf{x},\boldsymbol{\theta}} \min_{y\in\{-1,1\}} \left(\frac{1}{1+\exp\left(yw(\boldsymbol{\theta}^T\mathbf{x})\right)}\right)$$
$$\geq \frac{1}{1+\exp\left(wM\sqrt{P_{\mathbf{x}}d}\right)} =: \gamma_2.$$

We conclude the proof by combining the bounds

$$\min_{\mathbf{x},\mathcal{S},y,\boldsymbol{\theta},q} \mathbb{P}\left[\mathbf{x},y|\mathcal{S},\boldsymbol{\theta},q\right]$$
$$= \min_{\mathbf{x},\mathcal{S},y,\boldsymbol{\theta},q} \mathbb{P}\left[\mathbf{x}|\mathcal{S},\boldsymbol{\theta},q\right]\mathbb{P}\left[y|\mathbf{x},\mathcal{S},\boldsymbol{\theta},q\right] \geq \gamma_1\gamma_2.$$

■

Corollary B.3. *The likelihood of any answer to a query in $q \in \mathcal{Q} = \{q_{high}, q_{low}, q_{rank}\}$ is lower bounded as*

$$\min_{o,\mathcal{S},\boldsymbol{\theta},q} \mathbb{P}\left[o|\mathcal{S},\boldsymbol{\theta},q\right] \geq 2^{\gamma_L} > 0,$$

with $\gamma_L = \log_2\left(\gamma_1^{|\mathcal{S}|-1}\gamma_2^{|\mathcal{S}|}\right)$.

Proof: The likelihood of any answer to a ranking query, i.e., ranking order and threshold pair $(\mathbf{r}, l)$, is lower bounded as

$$\min_{\mathbf{r},l,\mathcal{S},\boldsymbol{\theta},q} \mathbb{P}\left[\mathbf{r},l|\mathcal{S},\boldsymbol{\theta},q_{\text{rank}}\right]$$
$$= \min_{\mathbf{r},l,\mathcal{S},\boldsymbol{\theta},q} \mathbb{P}\left[\mathbf{r}|\mathcal{S},\boldsymbol{\theta},q_{\text{rank}}\right]\mathbb{P}\left[l|\mathbf{r},\mathcal{S},\boldsymbol{\theta},q_{\text{rank}}\right]$$
$$= \min_{\mathbf{r},l,\mathcal{S},\boldsymbol{\theta},q} \prod_{j=1}^{|\mathcal{S}|-1} \mathbb{P}\left[r_j|\mathcal{R}_j,\boldsymbol{\theta},q_{\text{high}}\right] \times$$
$$\mathbb{P}\left[y(\mathbf{x}_{r_1},....r_{l-1}) = 1 \cap y(\mathbf{x}_{r_l},...,r_{|\mathcal{S}|}) = -1|\mathbf{r},\mathcal{S},\boldsymbol{\theta},q_{\text{rank}}\right]$$
$$\geq \gamma_1^{|\mathcal{S}|-1}\gamma_2^{|\mathcal{S}|} > 0,$$

which is strictly lower than the lower bound from Lemma B.2, because $\gamma_1\gamma_2 < 1$. ■

Lemma B.4. *The expected difference in entropy from one iteration to the next is bounded as*

$$\mathbb{E}_{o_i}\left[|h(\boldsymbol{\theta};\mathcal{F}_i) - h(\boldsymbol{\theta};\mathcal{F}_{i-1})|\right] \leq \gamma < \infty,$$

with $\gamma = 16d + d\log_2 2\pi ed - 2\gamma_L$.

Proof: We extend [6, Lemma A.2.] to bound the expected posterior entropy difference for non-equiprobable and non-binary query schemes. We rewrite

$$|h(\boldsymbol{\theta};\mathcal{F}_i) - h(\boldsymbol{\theta};\mathcal{F}_{i-1})| = (h(\boldsymbol{\theta};\mathcal{F}_{i-1}) - h(\boldsymbol{\theta};\mathcal{F}_i))^+ + (h(\boldsymbol{\theta};\mathcal{F}_i) - h(\boldsymbol{\theta};\mathcal{F}_{i-1}))^+, \quad (18)$$

where $x^+ := \max(x, 0)$ denotes the positive part.

Note that

$$\begin{aligned}
-h(\boldsymbol{\theta};\mathcal{F}_i) &\overset{(1)}{\leq} \log_2 \mathbb{E}_{\boldsymbol{\theta}|\mathcal{F}_i}[p(\boldsymbol{\theta}|\mathcal{F}_i)] \\
&\overset{(2)}{=} \log_2 \mathbb{E}_{\boldsymbol{\theta}|\mathcal{F}_i}\left[\frac{p(o_i|\boldsymbol{\theta},q_i,\mathcal{S}_i,\mathcal{F}_{i-1})}{p(o_i|q_i,\mathcal{S}_i,\mathcal{F}_{i-1})}p(\boldsymbol{\theta}|\mathcal{F}_{i-1})\right] \\
&\overset{(3)}{\leq} \log_2 \mathbb{E}_{\boldsymbol{\theta}|\mathcal{F}_i}\left[\frac{p(\boldsymbol{\theta}|\mathcal{F}_{i-1})}{p(o_i|q_i,\mathcal{S}_i,\mathcal{F}_{i-1})}\right] \\
&\overset{(4)}{\leq} -\frac{1}{2}\log_2\left|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_i}\right| + 8d + \frac{d}{2}\log_2 d \\
&\quad - \log_2 p(o_i|q_i,\mathcal{S}_i,\mathcal{F}_{i-1}) \\
&\overset{(5)}{=} -\frac{1}{2}\log_2\left((2\pi e)^d\left|\Sigma_{\boldsymbol{\theta}|\mathcal{F}_i}\right|\right) + \frac{1}{2}\log_2(2\pi e)^d \\
&\quad + 8d + \frac{d}{2}\log_2 d - \log_2 p(o_i|q_i,\mathcal{S}_i,\mathcal{F}_{i-1}) \\
&\overset{(6)}{\leq} -h(\boldsymbol{\theta};\mathcal{F}_{i-1}) + 8d + \frac{d}{2}\log_2 2\pi ed - \gamma_L \\
&:= -h(\boldsymbol{\theta};\mathcal{F}_{i-1}) + \frac{1}{2}\gamma.
\end{aligned}$$

Thus, we bound the first summand in (18) as $(h(\boldsymbol{\theta};\mathcal{F}_{i-1}) - h(\boldsymbol{\theta};\mathcal{F}_i))^+ \leq \frac{1}{2}\gamma$. The inequalities follow

(1) from Jensen's inequality,
(2) from Bayes Theorem and because the query is determined by the history,
(3) because o_i is discrete and its likelihood is a valid probability distribution, thus the likelihood is at most 1, and the logarithm is monotonically increasing,
(4) applying the bound in [44, Theorem 5.14] to the LCC isotropic $V = \Sigma_{\boldsymbol{\theta}|\mathcal{F}_i}^{-1/2}W$, where $W \sim p(\boldsymbol{\theta}|\mathcal{F}_i)$, together with the density of a linear transformation of a random variable.
(5) adding and subtracting $\frac{1}{2}\log_2(2\pi e)^d$,
(6) from the maximum entropy distribution [35, Theorem 8.6.5.] and Corollary B.3.

To bound the second summand in (18), we recall the non-negativity of mutual information

$$0 \leq \mathrm{I}\left(\boldsymbol{\theta}; o_i, q_i, \mathcal{S}_i|\mathcal{F}_{i-1}\right) = \mathbb{E}_{o_i}\left[h(\boldsymbol{\theta};\mathcal{F}_{i-1}) - h(\boldsymbol{\theta};\mathcal{F}_i)\right]$$
$$= \mathbb{E}_{o_i}\left[(h(\boldsymbol{\theta};\mathcal{F}_{i-1}) - h(\boldsymbol{\theta};\mathcal{F}_i))^+ - (h(\boldsymbol{\theta};\mathcal{F}_i) - h(\boldsymbol{\theta};\mathcal{F}_{i-1}))^+\right]$$
$$\leq \frac{1}{2}\gamma - \mathbb{E}_{o_i}\left[(h(\boldsymbol{\theta};\mathcal{F}_i) - h(\boldsymbol{\theta};\mathcal{F}_{i-1}))^+\right].$$

Therefore, $\mathbb{E}_{o_i}\left[(h(\boldsymbol{\theta};\mathcal{F}_i) - h(\boldsymbol{\theta};\mathcal{F}_{i-1}))^+\right] \leq \frac{1}{2}\gamma$. Combining the bounds we conclude

$$\mathbb{E}_{o_i}\left[|h(\boldsymbol{\theta};\mathcal{F}_{i-1}) - h(\boldsymbol{\theta};\mathcal{F}_i)|\right] \leq \frac{1}{2}\gamma + \frac{1}{2}\gamma = \gamma.$$

■

Lemma B.5. *The random variable $U_i := \frac{-h(\boldsymbol{\theta};\mathcal{F}_i)}{L} - i$ is a submartingale that fullfils the conditions of the optional stopping theorem.*

Proof: The expectation of U_i given the previous values of the sequence is

$$\mathbb{E}[U_i|U^{i-1}] = \frac{\mathbb{E}[Z_i|Z^{i-1}]}{L} - i \geq \frac{\mathbb{E}[Z_{i-1}|Z^{i-1}] + L}{L} - i$$
$$= \frac{Z_{i-1}}{L} - (i-1) = U_{i-1}, \quad (19)$$

where $Z_i = -h(\boldsymbol{\theta}; \mathcal{F}_i)$. The first inequality follows from Assumption III.3.

We bound the expected increment per step with Lemma B.4,

$$\mathbb{E}\left[|U_{i+1} - U_i|\right] = \mathbb{E}\left[\left|\frac{Z_{i+1}}{L} - i - 1 - \frac{Z_i}{L} + i\right|\right] = \frac{\mathbb{E}\left[|Z_{i+1} - Z_i|\right]}{L} + 1 \leq \frac{\gamma}{L} + 1. \quad (20)$$

Lastly, we want to show that $\mathbb{E}[T] < \infty$. Note that a bound on the determinant of the posterior covariance implies a bound on the posterior entropy, so we introduce a threshold $0 < \tau < \infty$ dependent on ϵ such that the stopping time becomes $T := \min\{i : -h(\boldsymbol{\theta}; \mathcal{F}_i) > \tau\} = \min\{i : U_i > \frac{\tau}{L} - i\}$.

From [45, Theorem 4.2.9.] we know that $-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})$ is also submartingale, where $i \wedge T := \min\{i, T\}$. Additionally, it is bounded by

$$-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T}) \leq \tau + \frac{\gamma}{L} + 1,$$

by definition of T and (20). Therefore, by the Martingale convergence theorem [45, Theorem 4.2.11.], as $i \to \infty$, $-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})$ converges a.s. to a limit H with $\mathbb{E}[|H|] < \infty$. Analogously, $U_{i \wedge T}$ also converges a.s. to a limit U with $\mathbb{E}[|U|] < \infty$ as $i \to \infty$. Putting this together

$$\begin{aligned} i \wedge T &= \left|(i \wedge T) - \frac{-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})}{L} + \frac{-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})}{L}\right| \\ &\leq \left|(i \wedge T) - \frac{-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})}{L}\right| + \left|\frac{-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})}{L}\right| \\ &= |U_{i \wedge T}| + \frac{|-h(\boldsymbol{\theta}; \mathcal{F}_{i \wedge T})|}{L} \xrightarrow{\text{a.s.}} |U| + \frac{|H|}{L} < \infty. \end{aligned}$$

For large enough i, $i \wedge T = T$, which implies $T < \infty$ a.s. and therefore $\mathbb{E}[|T|] < \infty$. Combining this fact with (20), we conclude that the conditions for the optional stopping theorem are fulfilled. ■

APPENDIX C
DETAILS ON HUMAN RESPONSE MODELING

To model annotator response times for word selection and word ranking queries, we conducted a crowdsourced study on Prolific. The study was categorized as minimal risk research qualified for exemption status under 45 CFR 46 104d.2 by the Institutional Review Board (IRB).

A. Participant Selection and Demographics

To promote data quality, we restricted participation to Prolific users with an approval rate of at least 95% and at least 1,000 prior submissions. To ensure strong English proficiency, we limited eligibility to participants located in the United Kingdom or United States who reported completing an undergraduate degree and listed English as their primary language. All eligible participants received detailed instructions on the task and the annotation interface. Within these instructions, we included multiple-choice attention check questions to verify comprehension. Twenty-two individuals did not pass the preliminary attention checks and were excluded from the study.

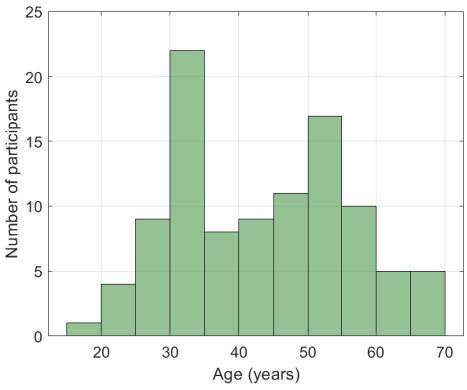


Fig. 15: Age distribution of participants. The mean age is 43.25 years and median 44 years.

After applying eligibility filters and attention checks, **101 annotators** participated in our study. The demographic breakdown was as follows: 42 female and 59 male participants; 41 residing in the UK and 60 in the US. The majority of participants were students (69 yes, 17 no, 15 no response). The age distribution is shown in Figure 15, with a median age of 44 years. The breakdown by ethnicity and country of birth is provided in Tables II and III, respectively. Most participants self-identified as White and were born in either the UK or USA.

TABLE II: Ethnicity distribution

Ethnicity	Count
White	68
Asian	15
Black	11
Mixed	4
Other	1
Not available	2
Total	**101**

TABLE III: Country of birth distribution

Country	Count
United States	55
United Kingdom	36
Nigeria	2
Bulgaria	1
China	1
Hungary	1
Indonesia	1
Japan	1
Korea	1
Malta	1
Philippines	1

To mitigate potential confounds caused by interface familiarity and learning effects, we followed standard A/B testing practice and counterbalanced the query order. Half of the participants completed the word-ranking queries first and then the word-selection queries, while the remaining participants completed the two query types in the reverse order.

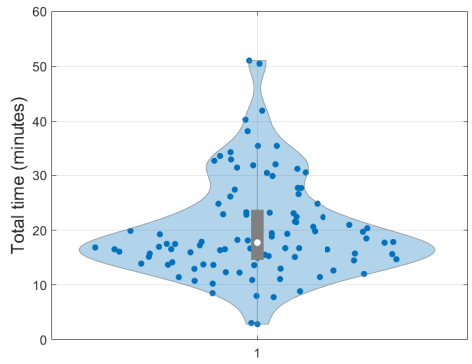


Fig. 16: Distribution of total study completion time per participant (including instructions and all word-selection and word-ranking queries).

Figure 16 reports the total time taken by participants to go through the instructions and answer all 50 queries (25 q_{rank} queries and 25 q_{high} queries). On average, participants required approximately 20 minutes. Based on the observed completion times and the study compensation, the hourly wage was on average 12.23 USD, with a median of 10.60 and a standard deviation of 9.53.

Beyond the eligibility criteria and pre-task attention checks described above, we applied additional post hoc quality control to identify inattentive annotators before modeling response times. In particular, we embedded the following five gold-standard sentiment queries with clear and unambiguous sentiment polarity:

- *amazing, rejection*
- *bad, horrible*
- *terrified, worried*
- *inspiring, boring*
- *happy, sad*

These gold-standard queries were intentionally distributed across the task at fixed positions (queries 1, 2, 11, 12, and 21), ensuring that attention was assessed at the beginning, middle, and end of the experiment rather than only at a single point. We excluded any participant who failed to exactly match the expected response on at least one gold-standard query, since these items were constructed to have an unambiguous correct answer for annotators who understood the task and responded attentively.

After applying all quality-control filters, including gold-standard checks, the final dataset contained 1648 observations for the word selection task and 1256 observations for the ranking task. The breakdown by query type is shown in Table IV. A total of 83 participants remained for the word selection task and 63 participants remained for the ranking task. These cleaned datasets were used for all subsequent analyses.

B. Data Analysis

Before collecting the data, we hypothesized the following models could accurately describe the response time:

TABLE IV: Counts by word set size after preprocessing (gold-standard queries removed).

N	Ranking	Selection
2	142	147
3	184	272
4	175	269
5	204	234
6	176	276
7	191	225
8	184	225

- **Hypothesis 1 Linear model for selection queries**: Drawing from the literature on serial scanning, we hypothesize that the response time increases linearly with the number of options: $\widehat{t}_{\text{selection}} = \beta_0 + \beta_1 |\mathcal{S}|$.
- **Hypothesis 2 Logarithmic model for selection queries**: Based on Hick's law, we hypothesize that the response time increases logarithmically with the number of options: $\widehat{t}_{\text{selection}} = \beta_0 + \beta_1 \log |\mathcal{S}|$.
- **Hypothesis 3 Linear model for ranking queries**: Assuming reading or viewing the options dominates the burden, we hypothesize that the response time increases linearly with the number of options: $\widehat{t}_{\text{rank}} = \beta_0 + \beta_1 |\mathcal{S}|$.
- **Hypothesis 4 Quadratic model for ranking queries**: Inspired by the complexity of simple sorting algorithms like Bubble sort, we hypothesis the response time increases quadratically with the number of options: $\widehat{t}_{\text{rank}} = \beta_0 + \beta_1 |\mathcal{S}|^2$.

We evaluated these candidate parametric forms on the cleaned dataset. Table V presents the complete regression results for all candidate models. For both question types, the linear models achieved higher R^2 values and lower MSE, confirming superior fit as concluded by the Vuong tests. Notably, the slope for ranking queries (4.41) is substantially steeper than for selection queries (0.63), indicating that each additional word imposes a much greater time burden when participants must produce a complete ranking rather than identify a single word.

C. Item difficulty stratification and effect on response time

Curriculum learning [46] suggests that optimal learning strategies often progress from simple examples to more complex ones. We hypothesized that our information-gain-based active learning strategy might exhibit a curriculum-like progression: selecting relatively easy queries at first, and transitioning to more challenging boundary cases as the uncertainty about the classifier decreases. If true, learning stage could confound the response time model beyond the effect of set size $|\mathcal{S}|$.

To test this, we analyzed queries from three stages of the learning process: early (first 5 queries), mid-stage (around iteration 200), and late-stage (around iteration 1000). Using selection stage as a proxy for difficulty, we examined whether queries selected at different points in training exhibited different response time patterns after controlling for set size.

Figure 17 presents the response time as a function of word set size $|\mathcal{S}|$ for these three difficulty tiers. To formally test whether difficulty explains additional variance beyond

TABLE V: Comparison of candidate response time models fitted using least-squares regression.

Query Type	**Model**	β_0 **(SE)**	β_1 **(SE)**	R^2	**Adjusted** R^2	**MSE**	N
Selection	Linear	4.01 (0.24)	0.63 (0.04)	0.112	0.111	11.22	1648
Selection	Logarithmic	3.10 (0.32)	1.84 (0.14)	0.097	0.097	11.36	1648
Ranking	Linear	−0.32 (0.88)	4.41 (0.16)	0.373	0.372	125.44	1256
Ranking	Quadratic	9.69 (0.57)	0.42 (0.02)	0.356	0.356	127.69	1256

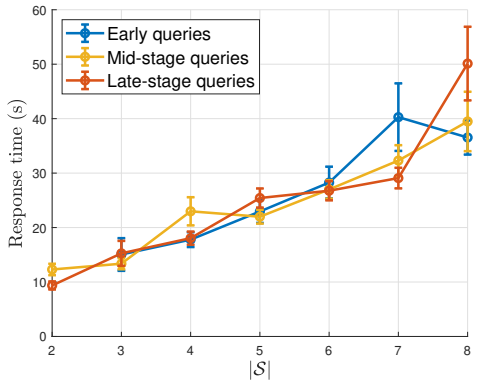


Fig. 17: Query Response time versus query set size stratified by stage of the learning process. Queries were sampled from early (first 5 queries), mid (iteration 200), and late stages (iteration 1000). The shaded area represents the standard error. The three categories show highly overlapping trends, indicating that learning stage contributes minimal variance beyond set size.

set size, we conducted an analysis of covariance (ANCOVA) comparing a baseline model containing only $|\mathcal{S}|$ to one that also included difficulty as a categorical predictor. The extended model did not improve fit ($p = 0.90$), indicating that once query length is accounted for, difficulty contributes no measurable additional variance to response time. Consequently, we model response time solely as a function of $|\mathcal{S}|$ in the main analysis.