COM-BOM: Bayesian Exemplar Search for Efficiently Exploring the Accuracy-Calibration Pareto Frontier

Gaoxiang Luo
University of Minnesota
luo00042@umn.edu

Aryan Deshwal
University of Minnesota
adeshwal@umn.edu

Abstract

Selecting an optimal set of exemplars is critical for good performance of in-context learning. However, prior exemplar search methods narrowly optimize for predictive accuracy, critically neglecting model calibration—a key determinant of trustworthiness and safe deployment. In this paper, we formulate exemplar selection as a multi-objective optimization problem, explicitly targeting both the maximization of predictive accuracy and the minimization of expected calibration error. We solve this problem with a sample-efficient Combinatorial Bayesian Optimization algorithm (COM-BOM) to find the Pareto front that optimally trades off the two objectives of accuracy and calibration. We evaluate COM-BOM on multiple tasks from unsaturated MMLU-Pro benchmark and find that COM-BOM beats or matches the baselines at jointly optimizing the two objectives, while requiring a minimal number of LLM API calls.

1 Introduction

In-context learning (ICL) has emerged as a powerful paradigm, enabling large language models (LLMs) to solve new tasks by conditioning on a prompt containing an instruction and a set of exemplars. While a large body of work focuses on instruction optimization, recent findings (Wan et al., 2024a; Ajith et al., 2024) reveal that exemplar selection contributes substantially more to ICL's performance than instructions alone. Despite this critical importance, principled approaches to exemplar selection remain surprisingly under-explored, particularly concerning the two objectives of predictive accuracy and model reliability.

The paper addresses this gap by re-evaluating the core goals of exemplar selection in ICL. Prior methods for exemplar selection predominantly pursue a single objective: maximize predictive accuracy. While accuracy is undoubtedly important, many high-stakes real-world applications (such as finance, healthcare, and legal settings) require not only high accuracy but also *well-calibrated* confidence. Poorly calibrated models may exhibit overconfidence in erroneous predictions or under confidence in correct ones, diminishing their practical value. Indeed, current exemplar selection strategies, by largely ignoring calibration, risk scenarios where gains in accuracy are achieved at the expense of degradation in calibration (Zhang et al., 2024b).

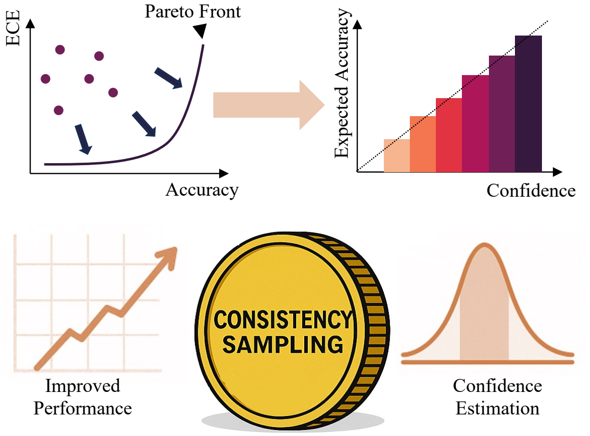

Figure 1: Optimizing for accuracy and calibration error leads to better reliability (top). Two sides of the same coin for self-consistency sampling (bottom).

Consequently, in this work, we propose that exemplar selection for LLMs should be reframed as a multi-objective optimization problem where the goal is to jointly optimize for both accuracy and calibration by identifying a set of exemplars that optimally trade off these two conflicting objectives. However, our multi-objective optimization formulation of exemplar selection in ICL presents several technical challenges: 1) combinatorial nature of the search space in selecting an optimal subset of exemplars, 2) the need to identify a Pareto frontier of solutions that represent varying trade-offs between accuracy and calibration, and 3) the expensive and noisy nature of evaluating these metrics, which often involve multiple LLM API calls.

In order to address these challenges, we propose

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20339–20352
November 4-9, 2025 ©2025 Association for Computational Linguistics

a principled Bayesian optimization algorithm titled: Combinatorial Bayesian Optimization of Multiple ICL Metrics (COM-BOM). The key idea is to leverage probabilistic surrogate models defined over combinatorial inputs with multi-objective acquisition functions to intelligently guide the search towards the pareto-optimal frontier of accuracy and calibration in a sample-efficient manner i.e. minimizing the number of calls to the expensive LLM API calls. The key contributions of this paper are:

1. We introduce a new formulation of exemplar selection in ICL as multi-objective optimization problem, explicitly targeting two important metrics of accuracy and calibration.
2. We propose COM-BOM, a combinatorial multi-objective Bayesian optimization algorithm, specifically designed to tackle the key technical challenges of multiple objectives, combinatorial search spaces, and expensive and noisy evaluations specific to this problem.
3. We empirically validate our approach, showcasing its ability to identify exemplar sets that achieve better Pareto frontier of the two objectives compared to existing methods, while requiring fewer LLM evaluations.

2 Background and Related Work

Exemplar Selection. The simplest form of methods for exemplar selection in ICL use existing lexical or embedding-based similarity metrics to retrieve examples most similar to the test input from a pool of candidates (Wang et al., 2024a). Another line of work trains bespoke retrievers specifically to pick in-context examples (Rubin et al., 2022; Cheng et al., 2023). One drawback of these methods is the amount of redundancy in their selected exemplars. Diversity based approaches aim to overcome this drawback by encouraging diversity in the selected exemplars (Ye et al., 2023). However, all of these approaches focus on predictive accuracy as the primary performance metric. Recent works demonstrate that exemplar selection contributes more to ICL's accuracy compared to instruction optimization (Wan et al., 2024a; Ajith et al., 2024). Recently, Wan et al. (2025) proposed BRIDGE algorithm for exemplar selection in the many-shot ICL setting which alternates between generating and optimizing a set of exemplars. The optimization step in BRIDGE leverages Bayesian optimization to solve a bi-objective (accuracy and sparsity) problem by converting it to a single objective via random scalarization. Our approach differs from BRIDGE in that we focus on optimizing calibration and accuracy as the two objectives and directly reason about the Pareto frontier using a hypervolume based multiobjective acquisition function.

Multi-Objective Optimization. Multi-objective optimization (MOO) (Emmerich and Deutz, 2018; Belakaria et al., 2019; Daulton et al., 2020) addresses problems where several, often conflicting, objective functions must be optimized simultaneously. Unlike single-objective problems that typically seek a single optimal solution, MOO aims to identify a set of solutions representing the best possible trade-offs, known as the *Pareto optimal set.* A solution is considered Pareto optimal if no objective function can be improved without degrading at least one other objective. The values of this set in the objective space forms the *Pareto front*, which characterizes the optimal attainable trade-off between the competing objectives. The core challenge in MOO lies in efficiently exploring the input space to discover or approximate this Pareto front.

3 Methodology

3.1 Problem Definition: Multiobjective Formulation for Exemplar Selection

In this work, we study the problem of exemplar selection in ICL where our objective is to identify an optimal subset of exemplars from a given pool $\mathcal{E} = \{e_1, e_2 ... e_m\}$ that maximizes the LLM's performance on a given task. We formalize this as a combinatorial search problem where the search space is defined over binary indicator vectors $\mathbf{z} \in \{0, 1\}^m$. Each element $\mathbf{z}_i$ indicates whether exemplar e_i is included in the prompt or not.

While most existing work on exemplar selection focuses solely on optimizing *predictive accuracy* as performance metric, many real-world applications such as clinical question answering (Agrawal et al., 2022) or legal decision making (Lai et al., 2024), often demand not just correctness but also reliable model confidence. The reliability of model confidence is quantified by expected calibration error (ECE) which measures how much its confidence, i.e. the predicted probability of correctness, diverges from its accuracy, i.e. the empirical probability of correctness. In this paper, we posit that exemplar selection should be formalized as a multi-objective optimization problem that seeks to jointly optimizes for high predictive accuracy and low calibration error. Accordingly, we define LLM performance using two (often conflicting) objectives:

▶ **Objective 1: Maximizing Predictive Accuracy** ($f_{\textbf{acc}}(\mathbf{z})$) :The ability of the LLM, conditioned on the exemplar set $\mathbf{z}$, to generate correct outputs.

▶ **Objective 2: Minimizing Expected Calibration Error (**$f_{\textbf{ECE}}(\mathbf{z})$**):** A metric to estimate LLM's miscalibration based on its confidence (derived from its output distribution conditioned on $\mathbf{z}$) and its observed accuracy.

Goal: Our overall problem setup in this paper is succinctly formalized as the following multi-objective optimization problem:

$$\max_{\mathbf{z}\in\{0,1\}^m} \quad (f_{\text{acc}}(\mathbf{z}), -f_{\text{ECE}}(\mathbf{z})) \tag{1}$$

i.e., we seek to find a set of pareto-optimal exemplar sets $\mathbf{z}^*$ that offer the best possible accuracy-calibration trade-off. Please note f_{ECE} is negated to maintain a consistent maximization formulation.

3.1.1 Evaluation of Accuracy and Calibration Error Objectives

To evaluate the effectiveness of a chosen exemplar set $\mathbf{z}$ according to our multi-objective formulation (1), we need robust methods for estimating predictive accuracy ($f_{\text{acc}}(\mathbf{z})$) and ECE ($f_{\text{ECE}}(\mathbf{z})$). Our evaluation procedure is designed for black-box LLM access and leverages the key principle of *semantic consistency* (Farquhar et al., 2024; Zhong et al., 2023) from multiple generated outputs. Interestingly, most popular techniques for estimating predictive accuracy, such as self-consistency decoding methods (Wang et al., 2023), already rely on generating multiple output samples from the LLM. We describe below that ECE can be cheaply computed from these same set of samples already being generated from the LLM, an opportunity overlooked by existing exemplar selection methods. The evaluation process for a given exemplar set $\mathbf{z}$ assumes a small validation set $\mathcal{D}_{\text{val}}$ and involves the following steps, logically grouped into *per-instance estimation* and *aggregate metric calculation*:

Per-Instance Confidence, and Accuracy Estimation For each input query x (with ground truth y) from $\mathcal{D}_{\text{val}}$, and the chosen exemplar set $\mathbf{z}$:

(a) *Diverse Output Generation:* We construct a `Prompt`$[x, \mathbf{z}]$ and prompt the LLM to generate M diverse output sequences (potential answers) $S_{x,\mathbf{z}} = \{s_1, s_2, ..., s_M\}$.

(b) *Semantic Clustering and Answer Determination:* The generated samples $S_{x,\mathbf{z}}$ are clustered based on semantic equivalence (Farquhar et al., 2024). Sequences s_i and s_j are grouped into the same semantic cluster c if they convey the same core meaning (e.g., via bidirectional entailment), despite lexical variations. This results in a set of semantic clusters $\mathcal{C}_{x,\mathbf{z}} = \{c_1, c_2, ..., c_L\}$ (where $L \leq M$). The predicted answer, $\text{ans}(x, \mathbf{z})$, is taken from the largest (most frequent) semantic cluster, $c^* = \arg\max_{c_l\in\mathcal{C}_{x,\mathbf{z}}} |c_l|$.

(c) *Confidence Estimation:* The confidence in the prediction $\text{ans}(x, \mathbf{z})$ is defined as the probability of its corresponding semantic cluster c^*: $\text{conf}(x, \mathbf{z}) = \max_{c_i\in\mathcal{C}_{x,\mathbf{z}}} \frac{|c_i|}{M}$. This score reflects the LLM's internal consistency in generating the chosen semantic output.

(d) *Per-Instance Accuracy:* The accuracy for this single input x given $\mathbf{z}$, denoted $\text{acc}(x, \mathbf{z})$, is 1 if $\text{ans}(x, \mathbf{z})$ matches the ground truth label y, and 0 otherwise.

Aggregate Metric Calculation over the Validation Set After processing all instances in $\mathcal{D}_{\text{val}}$:

(a) *Overall Predictive Accuracy* ($f_{\text{acc}}(\mathbf{z})$): The overall predictive accuracy for the exemplar set $\mathbf{z}$ is the average of the per-instance accuracies:

$$f_{\text{acc}}(\mathbf{z}) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x,y)\in\mathcal{D}_{\text{val}}} \text{acc}(x, \mathbf{z}) \tag{2}$$

(b) *Expected Calibration Error* ($f_{\text{ECE}}(\mathbf{z})$): ECE (Naeini et al., 2015) quantifies the mismatch between the model's confidence and its empirical accuracy. It is computed as follows:

(i) Collect all $(\text{conf}(x, \mathbf{z}), \text{acc}(x, \mathbf{z}))$ pairs for $(x, y) \in \mathcal{D}_{\text{val}}$.

(ii) Divide the confidence interval $[0, 1]$ into K equally spaced bins.

(iii) For each bin k ($1 \leq k \leq K$):

- Let $\mathcal{B}_k$ be the set of input-output pairs (x, y) whose $\text{conf}(x, \mathbf{z})$ falls into bin k.
- Calculate the average confidence in bin k: $\text{conf}_{\mathcal{B}_k} = \frac{1}{|\mathcal{B}_k|} \sum_{(x,y)\in\mathcal{B}_k} \text{conf}(x, \mathbf{z})$.

- Calculate the average accuracy in bin k: $\text{acc}_{\mathcal{B}_k} = \frac{1}{|\mathcal{B}_k|}\sum_{(x,y)\in\mathcal{B}_k}\text{acc}(x,\mathbf{z})$.

(iv) The ECE for exemplar set $\mathbf{z}$ is the weighted average of the absolute differences between average accuracy and average confidence across all bins:

$$f_{\text{ECE}}(\mathbf{z}) = \sum_{k=1}^{K}\frac{|\mathcal{B}_k|}{|\mathcal{D}_{\text{val}}|}\cdot|\text{acc}_{\mathcal{B}_k} - \text{conf}_{\mathcal{B}_k}| \quad (3)$$

A lower $f_{\text{ECE}}(\mathbf{z})$ indicates better calibration, meaning the LLM's confidence $\text{conf}(x,\mathbf{z})$ more accurately reflects its likelihood of being correct. This evaluation methodology works with only black-box access to API LLMs.

3.1.2 Key Optimization Challenges

There are two technical challenges that arise because of our evaluation procedure. *First,* each objective evaluation of the objective function is black-box and expensive (both monetary costs of LLM API calls and latency costs of multiple samples) which necessitates *sample-efficient* methods that can find high-quality exemplar sets with a minimal number of objective function evaluations. *Second,* each objective evaluation is black-box noisy estimate of the true ground truth. Therefore, we want to develop black-box optimization algorithms that are robust to noisy observations. In the subsequent sections, we detail our proposed Combinatorial Bayesian Optimization of Multiple ICL Metrics COM-BOM algorithm to address these challenges.

3.2 Bayesian Optimization Solution: COM-BOM

Bayesian optimization (BO) (Garnett, 2023) is an effective and principled framework for black-box optimization, particularly when objective function evaluations are expensive, as is the case in our problem setting where each evaluation requires multiple LLM calls for a given input $\mathbf{z}$ (exemplar set). The key idea in BO is to construct a probabilistic surrogate model and using it to guide sequential function evaluations. Our approach COM-BOM instantiates the Bayesian optimization framework with its three core components, each tailored to address the unique challenges of our problem setup.

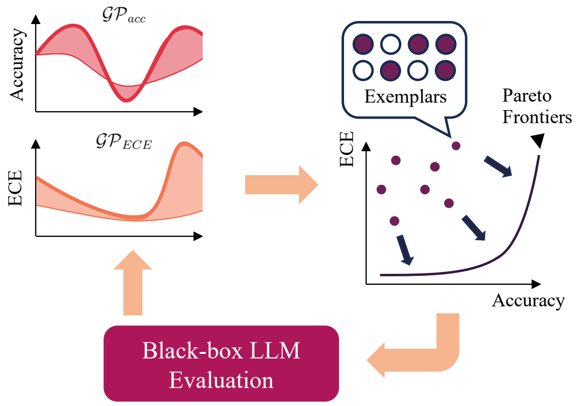

Figure 2: The BO loop with a single-task GP for each objective and multi-objective acquisition function.

Algorithm 1 Pseudocode for COM-BOM

```
Require: Exemplar pool $\mathcal{E}$, validation set $D_{val}$, evaluation budget $t_{\max}$
Randomly sample $t_0$ initial $\mathbf{z}$'s and evaluate the two objective functions to initialize $D_{t_0}, \mathcal{P}_{t_0}$;
    $D_{t_0} = \{(\mathbf{z}_i, o_i)\}_{i=1}^{t_0}, \quad o_i \in \{f_{\text{acc}}, f_{\text{ECE}}\}$,
    $\mathcal{P}_{t_0}$: Pareto Frontier$(D_{t_0})$
for $n = t_0$ to $t_{\max}$ do
    Fit Gaussian Process surrogate models $\tilde{f}_{\text{acc}}$, $\tilde{f}_{\text{ECE}}$ on $D_n$    ▷ Section 3.2.1
    Optimize the multiobjective acquisition function $\alpha(\mathbf{z} \mid D_n, \mathcal{P}_n)$ to get next point to evaluate $\mathbf{z}_{n+1}$    ▷ Section 3.2.2, 3.2.3
    Evaluate objectives on $\mathbf{z}_{n+1}$ by generating samples from LLM to get $o_{n+1} = (f_{\text{acc}}(\mathbf{z}), f_{\text{ECE}}(\mathbf{z}))$    ▷ Section 3.1.1
    $D_{n+1} = D_n \cup \{(\mathbf{z}_{n+1}, (o_{n+1}))\}$
    $\mathcal{P}_{n+1} \leftarrow$ Pareto Frontier$(\mathcal{D}_{n+1})$
end for
Ensure: Approximate Pareto frontier $\mathcal{P}_{t_{\max}}$ and corresponding Pareto set of the two objectives: accuracy and expected calibration error.
```

First, two Gaussian process surrogate models are constructed, which represents our belief about how different exemplar set choices ($\mathbf{z}$) influence both predictive accuracy f_{acc} and expected calibration error f_{ECE}. *Second*, a multi-objective acquisition function is employed which quantifies the utility of evaluating the objective at different exemplar sets by balancing exploitation of regions where the surrogate predicts high function values against exploration of regions with high predictive uncertainty. *Third*, an acquisition function optimization procedure is used to efficiently identify the most promising exemplar set to evaluate next, thereby guiding the iterative search process towards optimal inputs with minimum objective function calls. We describe the details of each component in the subsequent sections. Our approach is implemented in the BoTorch library (Balandat et al., 2020).

3.2.1 Gaussian Process Surrogate Models with Exponentiated Hamming Kernel

Gaussian Processes (GPs) (Williams and Rasmussen, 1995) are an effective class of probabilistic model for surrogate modeling in Bayesian optimization due to their ability to provide principled uncertainty quantification. GP models are mainly characterized by the choice of the kernel or covariance function ($k_H(\mathbf{z}, \mathbf{z}')$) and a mean function ($\mu(\mathbf{z})$; parametrized as a learnable constant). We use an independent Gaussian-process model, one for each objective, with an exponentiated hamming distance kernel (Wan et al., 2021) which is suitable for combinatorial space of exemplar selection, i.e.,

$$\tilde{f}_{\text{acc}}, \tilde{f}_{\text{ECE}} \sim \mathcal{GP}\big(\mu(\mathbf{z}),\, k_H(\mathbf{z}, \mathbf{z}')\big)$$

with kernel/covariance function

$$k_H(\mathbf{z}, \mathbf{z}') = \exp\big(-\, d_{\text{H}}(\mathbf{z}, \mathbf{z}')\big)\,, \tag{4}$$

$$d_{\text{H}}(\mathbf{z}, \mathbf{z}') = \sum_{j=1}^{m} \frac{1}{\ell_j} \cdot \mathbf{1}[z_j \neq z_j']. \tag{5}$$

where ℓ_j is a separate lengthscale parameter for each input dimension, typically referred as automatic relevance determination. These dimension-specific lengthscales allow the GP to learn the relative importance of different input dimensions. Given observations:

$$D_n = \big\{(\mathbf{z}_i, o_i)\big\}_{i=1}^{n}, \quad o_i \in \{f_{\text{acc}},\, f_{\text{ECE}}\},$$

define the kernel matrix and cross-covariances

$$[K_H]_{ij} = k_H(\mathbf{z}_i, \mathbf{z}_j), \quad \mathbf{k}_H(\mathbf{z}) = \big[k_H(\mathbf{z}, \mathbf{z}_i)\big]_{i=1}^{n}.$$

Then for each objective $\tilde{f} \in \{\tilde{f}_{\text{acc}}, \tilde{f}_{\text{ECE}}\}$, the posterior predictive at a new point $\mathbf{z}$ is computed in closed form, providing both a mean prediction and uncertainty estimate:

$$p\big(\tilde{f}(\mathbf{z}) \mid \mathbf{z}, D_n\big) = \mathcal{N}\big(\mu_n(\mathbf{z}),\, s_n^2(\mathbf{z})\big), \text{with}$$

$$\mu_n(\mathbf{z}) = \mu(\mathbf{z}) + \mathbf{k}_H(\mathbf{z})^\top \big[K_H + \sigma_n^2 I\big]^{-1} \big(\mathbf{o} - \mu(\mathbf{Z})\big),$$

$$s_n^2(\mathbf{z}) = k_H(\mathbf{z}, \mathbf{z}) - \mathbf{k}_H(\mathbf{z})^\top \big[K_H + \sigma_n^2 I\big]^{-1} \mathbf{k}_H(\mathbf{z}),$$

where σ_n^2 is the observation noise variance, $\mathbf{o} = [o_i]_{i=1}^n$, and $\mu(\mathbf{Z}) = [\mu(\mathbf{z}_i)]_{i=1}^n$.

Notation Remark: As mentioned in problem definition 3.1, our goal is to *minimize* ECE and *maximize* accuracy. In practice, we negate the ECE observations to transform the problem into a consistent maximization formulation for both objectives. Thus, our approach's implementation operates on $\{f_{\text{acc}},\, -f_{\text{ECE}}\}$. However, in our discussion, we refer to ECE in its original, unnegated form (f_{ECE}) where lower values are better.

3.2.2 Hypervolume based Multi-objective Acquisition Function

The acquisition function quantifies the utility of evaluating a candidate exemplar set $\mathbf{z}$ based on the current probabilistic models of the objectives ($p\big(\tilde{f}(\mathbf{z}) \mid \mathbf{z}, D_n\big)$). In multi-objective setting, an effective acquisition function must jointly reason about the two objectives and guide the search towards improving the incumbent Pareto frontier $\mathcal{P}_n$.

One such effective choice for a multiobjective acquisition function is the Expected Improvement in the Hypervolume Indicator (EHVI) (Daulton et al., 2020). The hypervolume itself is the Lebesgue measure of the objective space region dominated by a given Pareto frontier and bounded by a reference point (Emmerich and Deutz, 2018). Intuitively, for a set of non-dominated solutions (the Pareto frontier), the hypervolume quantifies the coverage of the objective space that these solutions collectively achieve i.e. gives us a global view of the solution space that is typically missed by the naive baseline of scalarizing the two objectives. EHVI, therefore, measures the expected increase in this dominated hypervolume if a new candidate $\mathbf{z}$ were to be evaluated and added to the set of known solutions (see Figure 3 for illustration). Maximizing EHVI aims to expand the coverage and quality of the approximated Pareto front.

Let $\mathcal{P}_n = \{\mathbf{o}^i\}_{i=1}^{|\mathcal{P}_n|}$ be the current Pareto frontier and $HV(\cdot)$ its hypervolume relative to a fixed reference point. Then the Expected Hypervolume Improvement is defined as:

$$\alpha(\mathbf{z} \mid D_n, \mathcal{P}_n) = \mathbb{E}_{\mathbf{F}(\mathbf{z})|D_n}\Big[HV\big(\mathcal{P}_n \cup \{\mathbf{F}(\mathbf{z})\}\big) - HV(\mathcal{P}_n)\Big] \tag{6}$$

$$\mathbf{F}(\mathbf{z}) \;=\; \begin{pmatrix} \tilde{f}_{\text{acc}}(\mathbf{z}) \\ \tilde{f}_{\text{ECE}}(\mathbf{z}) \end{pmatrix} \sim \mathcal{N}\Big(\boldsymbol{\mu}_n(\mathbf{z}),\, \boldsymbol{\Sigma}_n(\mathbf{z})\Big)$$

$$\boldsymbol{\mu}_n(\mathbf{z}) = \begin{pmatrix} \mu_n^{\text{acc}}(\mathbf{z}) \\ \mu_n^{\text{ECE}}(\mathbf{z}) \end{pmatrix},$$

$$\boldsymbol{\Sigma}_n(\mathbf{z}) = \text{diag}\big(s_{n,\text{acc}}^2(\mathbf{z}),\, s_{n,\text{ECE}}^2(\mathbf{z})\big)$$

where each mean and variance is given by the GP posterior. In this work, we utilize the related Noisy Expected Hypervolume Improvement (NEHVI) acquisition function, as introduced by (Daulton et al.,

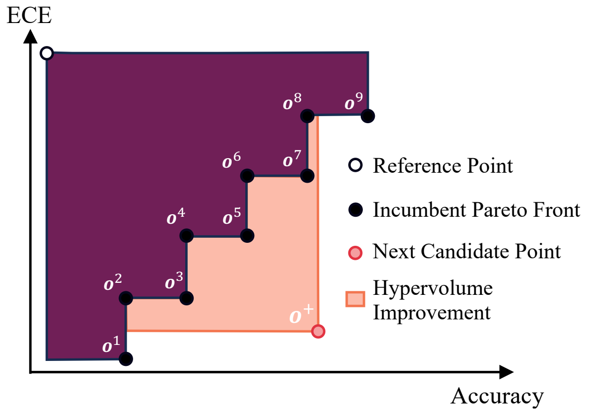

Figure 3: Illustration of Hypervolume improvement acquisition function for a candidate point (Section 3.2.2).

2021), which is suitable for settings with noisy observations. NEHVI extends EHVI by marginalizing out the uncertainty over the true Pareto frontier given the history of noisy evaluations.

3.2.3 Acquisition Function Optimization with Greedy Hill-Climbing

Following prior work in Bayesian optimization, we pick the next candidate exemplar set z_{t+1} by optimizing the acquisition function over the combinatorial space $\{0, 1\}^m$ using a trust-region based (Eriksson et al., 2019; Wan et al., 2021) heuristic local search strategy, which we found to work well in practice. The key idea is to optimize the acquisition function using a local search based greedy hill-climbing procedure within a trust region defined as a Hamming ball centered around a promising candidate z_{center}. In our work, the center point z_{center} is selected by identifying the point in the non-dominated set that contributes most to the hypervolume of the current approximate Pareto front (Daulton et al., 2022; Oh et al., 2019). Multiple random restarts are used to improve the performance of the greedy hill-climbing optimizer.

4 Experiments

Models and datasets. We evaluate `COM-BOM` using Qwen models (`Qwen3-8B`) (Yang et al., 2025) and LLaMA models (`LlaMA-3.3-70B`) (Grattafiori et al., 2024) on an extensive collection of tasks from MMLU-Pro (Wang et al., 2024b), a challenging benchmark where LLMs have not yet achieved performance saturation (as of May 2025). In addition, MMLU-Pro covers a diverse set of tasks beyond math and coding problems where data contamination is a known issue (Balunović et al., 2025). This makes it a suitable benchmark for evaluating methods that leverage exemplars to improve few-shot learning of LLMs on novel tasks, without requiring fine-tuning. For each task, we construct an exemplar pool by randomly selecting 32 samples, with the remaining samples divided equally between validation and test, the latter of which is held-out and unavailable to LLMs at search time. We refer readers to App. A for implementation details, including prompt template and sampling parameters.

Experimental setup and baselines. We benchmark `COM-BOM` against multiple optimization-based techniques, including random search (RS), genetic algorithm (GA), simulated annealing (SA), and hill climbing (HC) with scalarization, because they represent standard approaches, frequently employed in practice, for combinatorial black-box optimization problems (Deshwal et al., 2021; Dreczkowski et al., 2023). Following Wan et al. (2024b), we also evaluate against optimization-free retrieval baselines: Nearest and Diversity. The Nearest retrieves the k exemplars with highest text-embedding cosine similarity to the input query, while Diversity selects k input-output pairs closest to centroids identified through k-means clustering in the embedding space (Zhang et al., 2023). For the text-embedding model, we used `stella_en_400M_v5` (Zhang et al., 2024a) as it achieved the highest overall score among lightweight models (<1B parameters) on MTEB benchmark (Enevoldsen et al., 2025), as of May 2025. Additionally, we also include the two simplest baselines: using no exemplars and using all the exemplars.

Results and discussion. Nearly all tasks show improvement from ICL with exemplars (Tab. 1) as `Qwen3` is pretrained on 36T tokens spanning diverse domains including (non-)STEM fields. These supervised exemplars effectively focus the model's pre-training distribution toward domain-specific parametric knowledge (Lu et al., 2025). However, simply increasing number of exemplars does not guarantee monotonically increasing performance (Agarwal et al., 2024), as individual exemplars contribute differently to task outcomes, emphasizing the importance of strategic exemplar selection.

Our results reveal that online retrieval-based approaches underperform compared to `COM-BOM` which is an offline (which only leverage a fixed validation set) search method (Tab. 1 and Fig. 4). This finding persists across various values of k in the retrieval system (App. B). While offline search has initial computational costs, these can be amortized during deployment, yielding greater

Domain	No Exemplars		All Exemplars		Nearest$_{k=10}$		Diversity$_{k=10}$		Ours	
	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
biology	77.19	19.10	79.24	18.40	76.60	20.29	76.90	24.32	79.24	18.37
business	40.74	32.97	41.27	27.25	39.68	30.72	38.89	33.00	41.01	21.54
chemistry	34.79	37.45	39.34	29.50	37.34	31.79	36.61	33.26	37.34	28.38
computer science	48.94	35.84	50.53	32.42	48.94	37.16	51.06	35.05	55.85	32.08
economics	63.95	30.24	62.96	30.89	61.48	30.98	62.47	33.20	63.70	30.74
engineering	44.44	30.67	46.15	21.13	47.01	21.00	42.95	23.38	47.01	19.32
health	60.20	33.91	60.20	32.01	56.12	32.55	60.71	31.93	62.24	28.49
history	52.30	34.20	53.45	36.48	49.43	35.10	50.00	34.51	52.30	29.80
law	29.40	39.76	30.15	45.61	31.84	45.29	30.71	48.24	31.08	37.58
math	37.18	36.68	37.33	30.98	35.81	33.23	36.42	31.83	39.00	27.19
philosophy	45.49	43.53	46.78	36.62	45.92	41.23	44.64	43.20	46.78	33.84
physics	41.23	36.59	40.60	28.77	41.71	27.46	40.13	33.15	42.97	25.78
psychology	65.45	33.09	69.37	27.56	68.32	28.75	67.28	31.78	69.89	25.83
Average	43.48	31.73	44.55	28.06	43.62	29.11	43.32	30.87	45.31	25.24

Table 1: The test accuracy and ECE of optimization-free approaches on MMLU-Pro dataset with `Qwen3-8B`. While sometimes simply using all exemplars or setting up a retrieval system for online search obtains a better accuracy, our calibration-aware search outperforms all baselines with respect to ECE, establishing confidence in LLM predictions. The results for `LLaMA-3.3-70B` can be found in App. B.

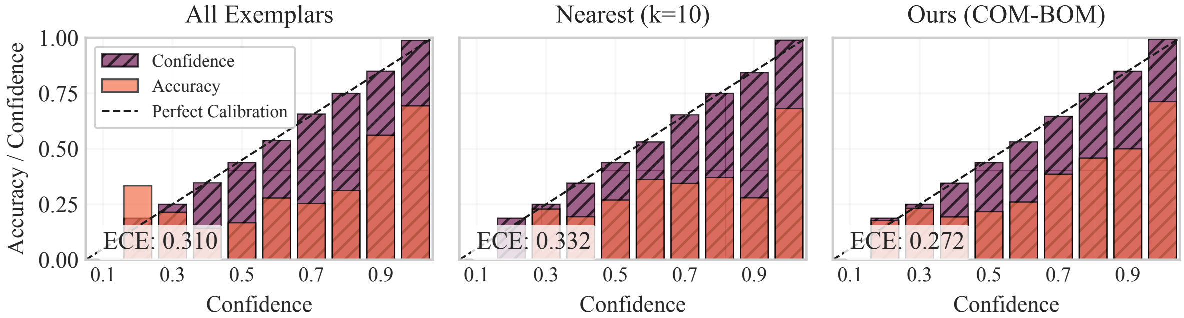

Figure 4: The reliability diagram of test accuracy and ECE for `Math` task from MMLU-Pro. It highlights the necessity of optimizing for calibration error that is paramount for deploying trustworthy LLM systems, in order to minimize over-confident wrong predictions and under-confident right predictions. Compared to online retrieval systems for exemplar search, `COM-BOM` is more cost-effective at inference time due to its offline search.

Domain	Genetic Algorithm		Simulated Annealing		Random Search		Hill Climbing		Ours	
	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
biology	76.61	20.37	77.49	22.90	78.65	20.32	77.19	20.00	79.24	18.37
business	37.83	23.89	38.10	23.13	37.57	33.44	41.53	28.26	41.01	21.54
chemistry	34.06	29.37	36.07	30.18	34.24	30.41	35.70	33.04	37.34	28.38
computer science	48.94	38.71	51.06	34.71	48.94	36.44	51.60	35.62	55.85	32.08
economics	62.22	31.29	62.72	31.24	62.22	31.75	62.72	30.35	63.70	30.74
engineering	45.51	21.42	44.23	22.57	43.80	20.22	48.50	24.72	47.01	19.32
health	61.22	32.79	58.67	32.78	60.20	34.84	59.95	29.11	62.24	28.49
history	52.87	23.72	47.70	30.73	51.15	38.74	51.15	27.49	52.30	29.80
law	31.09	38.68	31.46	42.58	32.58	44.95	31.84	39.10	31.08	37.58
math	36.87	33.92	37.02	28.48	37.32	31.51	37.48	30.23	39.00	27.19
philosophy	45.06	39.28	44.63	41.41	44.21	40.01	43.34	39.12	46.78	33.84
physics	40.28	31.32	39.34	29.33	40.13	33.20	40.76	29.19	42.97	25.78
psychology	67.28	28.02	67.80	29.33	68.06	31.51	68.84	24.94	69.89	25.83
Average	43.36	27.98	43.26	28.02	43.41	29.92	44.24	27.73	45.31	25.24

Table 2: The test accuracy and ECE of optimization-based approaches on MMLU-Pro dataset with `Qwen3-8B`. The results for `LLaMA-3.3-70B` can be found in Appendix B.

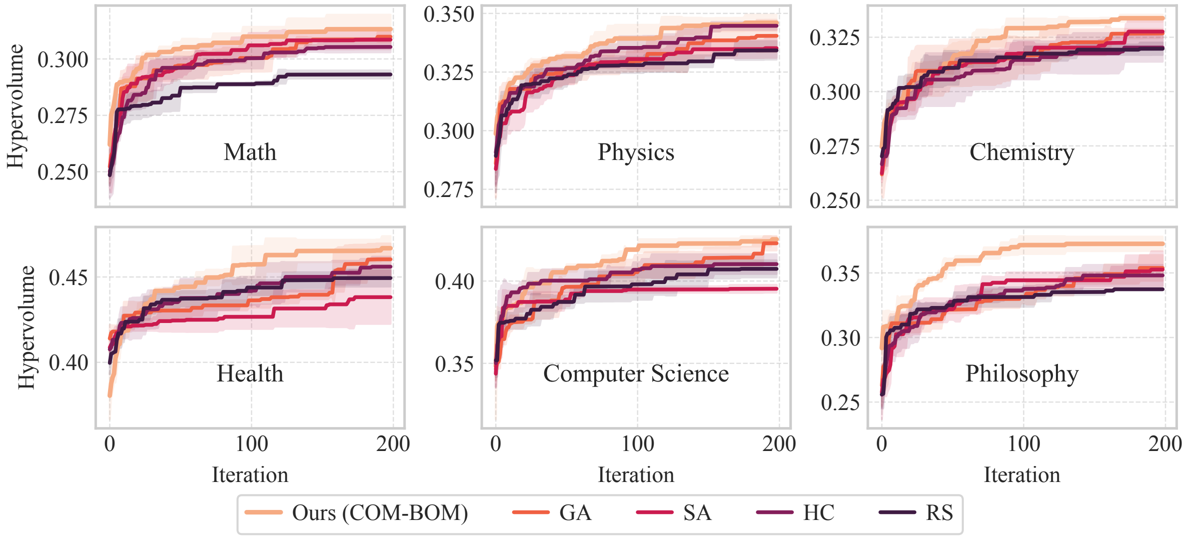

Figure 5: Evolution of best observed hypervolume on the validation data across `STEM`, `Medical` and `Humanity` tasks for optimization approaches. The hypervolume is measured against the reference point (accuracy=0%, ECE=100%). The evolution plots the average of three runs. Please see App. B for results on rest of the tasks.

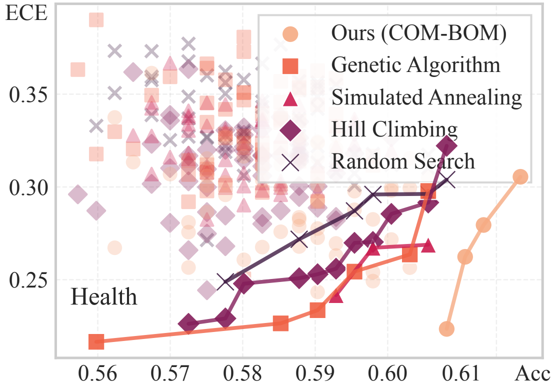

Figure 6: The observed validation accuracy and ECE with Pareto frontiers on `Health` task for optimization-based approaches with fixed evaluation iterations. `COM-BOM` identifies better Pareto-optimal solutions.

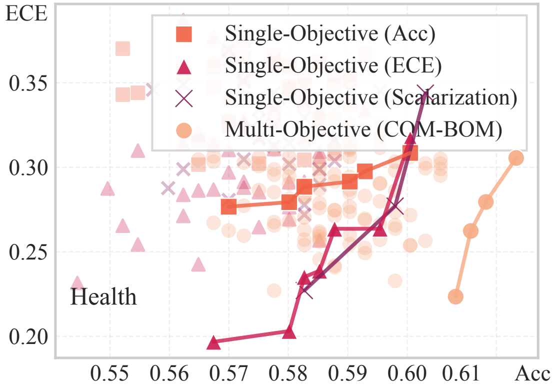

Figure 7: The observed validation accuracy and ECE with Pareto frontiers on `Health` task for single versus multi-objective formulation. Interestingly, optimizing for calibration error yields better Pareto frontiers than optimizing for accuracy alone. `COM-BOM` leverages both objectives more effectively than scalarization.

cost-effectiveness at inference time. Additionally, we observe consistently lower performance with diversity-based exemplar selection, likely due to the inherent difficulty of clustering in the high-dimensional space.

Our results also demonstrate that optimization approaches consistently outperform optimization-free baselines (Tab. 2), with `COM-BOM` exhibiting superior sample efficiency in solving the multi-objective optimization problem (Fig. 5). This efficiency translates to an improved accuracy-calibration trade-off while requiring fewer LLM API calls. Notably, `COM-BOM` offers a distinct advantage over other optimization baselines, that rely on scalarization, as it directly reasons about the Pareto front in terms of its hypervolume. Consequently, `COM-BOM` shows faster convergence toward Pareto-optimal solutions within a small number of evaluation iterations (Fig. 6).

Necessity of MOO formulation. To our knowledge, we are the first to propose calibration-aware exemplar selection as a multi-objective optimization formulation. To demonstrate the importance of such formulation, we also compare with a single-objective combinatorial BO baseline. Our results demonstrate that optimizing for accuracy alone often compromises calibration, resulting in predictors with sub-optimal reliability (Fig. 7). Conversely optimizing exclusively for ECE compromises accuracy, yielding less effective predictors. While scalarization provides some benefit, `COM-BOM` finds much better Pareto frontiers of the two objectives.

Ablative study of BO components. Our ablation analysis (Fig. 8) demonstrates that incorporating both noisy observations (Daulton et al., 2021) and trust region (Eriksson et al., 2019) yields the most favorable trade-offs, achieving lower ECE at higher accuracy. Removing either or both components results in degraded performance highlighting the complementary benefits of local search and handling noisy observations.

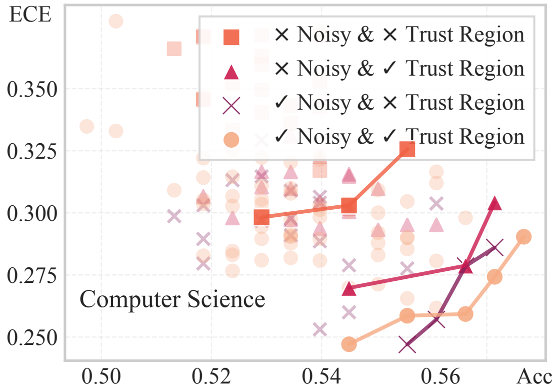

Figure 8: The observed validation accuracy and ECE with Pareto frontiers on `Computer Science` task if accounting for noisy observation and trust region.

5 Conclusion

We introduce a new formulation of exemplar selection as a multi-objective optimization problem where the key idea is to ensure reliability by jointly optimizing for calibration and accuracy metrics rather than optimizing the latter alone. Additionally, we propose a sample-efficient combinatorial Bayesian optimization algorithm `COM-BOM` for this black-box optimization problem that involves multiple objectives and is expensive to evaluate. Extensive experiments show that `COM-BOM` achieves better Pareto front with fewer LLM evaluations compared to existing methods.

Future work. There are multiple avenues for future work in this problem space. In our Bayesian optimization approach, we used a Gaussian process model with Exponentiated Hamming Kernel as the surrogate model, a simple yet effective choice for our search space. For even larger search spaces, this kernel can be replaced with more sophisticated ones that operate on learned embeddings for exemplars, allowing the model to leverage semantic similarity and scale more effectively. Another fertile ground for exploration is to consider the additional input space of the order in which exemplars are included in the prompt. Recent work in Bayesian optimization over permutation spaces (Deshwal et al., 2022) can be an effective direction in order to tackle such an input space of exemplar orderings.

6 Limitations

Our experiments are based on an exemplar pool of 32 candidates. We consider this pool size to be practical for novel tasks where only a limited number of high-quality, human-annotated demonstrations can be sourced. Furthermore, using a large number of exemplars during inference can introduce undesirable latency and serving costs for diminishing gains in performance. However, Bayesian Optimization is known to face challenges with scalability as the dimensionality of the search space (in our case, number of candidate exemplars) increases. Due to computational resource constraints, our evaluations were limited to dense LLMs with fewer than <100B parameters. We specifically choose `Qwen3-8B` and `LLaMA-3.3-70B` since they are pretrained on the largest corpora among models of the same size to elicit in-context learning behaviors. Additionally, our evaluation was restricted to English-language multiple-choice QA tasks from the MMLU-Pro dataset. Expanding this methodology to open-ended reasoning tasks (e.g., code generation) would be a valuable extension, though it would require more nuanced metrics for quantifying accuracy (e.g., pass@k) and confidence (e.g., semantic uncertainty). To ensure consistent output formatting, especially with models of this scale, we reserved one exemplar exclusively for demonstrating the desired format. We empirically observe that the output format is almost always enforced for multiple-choice QA tasks employed in our paper.

7 Ethics Statement

We are using open-source models (`Qwen3-8B` and `LLaMA-3.3-70B`) and data (MMLU-Pro). This work essentially modifies the input to an LLM by selecting better exemplars to include in the few-shot prompt, and thus does not introduce additional risk to LLMs themselves. Instead, this work implicitly finds exemplars that cause an LLM to be faithful to its own confidence in its prediction, making it more reliable at deployment. However, if an adversary collects a pool of popular malicious prompts such as jailbreaking exemplars, with a well-defined performance metric such as success rate, our method can potentially be used to find the Pareto-optimal sets of malicious exemplars that maximize the success rate and minimize the calibration errors (i.e., confidently break the system) with few black-box LLM evaluations only.

Acknowledgment

This work is supported by NSF grant IIS-2313174 and a Data Science Graduate Assistantship with funding made available by the MnDRIVE initiative through the University of Minnesota Data Science Initiative. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the results reported within this paper.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, and 1 others. 2024. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anirudh Ajith, Chris Pan, Mengzhou Xia, Ameet Deshpande, and Karthik Narasimhan. 2024. InstructEval: Systematic evaluation of instruction selection methods. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4336–4350, Mexico City, Mexico. Association for Computational Linguistics.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. 2020. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions.

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. 2019. Max-value entropy search for multi-objective bayesian optimization. *Advances in neural information processing systems*, 32.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337, Singapore. Association for Computational Linguistics.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2020. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. 2021. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, 34:2187–2200.

Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. 2022. Multi-objective bayesian optimization over high-dimensional search spaces. In *Uncertainty in Artificial Intelligence*, pages 507–517. PMLR.

Aryan Deshwal, Syrine Belakaria, and Janardhan Rao Doppa. 2021. Mercer features for efficient combinatorial bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7210–7218.

Aryan Deshwal, Syrine Belakaria, Janardhan Rao Doppa, and Dae Hyun Kim. 2022. Bayesian optimization over permutation spaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6515–6523.

Kamil Dreczkowski, Antoine Grosnit, and Haitham Bou Ammar. 2023. Framework and benchmarks for combinatorial and mixed-variable bayesian optimization. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Michael TM Emmerich and André H Deutz. 2018. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural computing*, 17:585–609.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. 2019. Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, 32.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Roman Garnett. 2023. *Bayesian optimization*. Cambridge University Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2024. Large language models in law: A survey. *AI Open*.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. 2025. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Changyong Oh, Jakub Tomczak, Efstratios Gavves, and Max Welling. 2019. Combinatorial bayesian optimization using the graph cartesian product. *Advances in Neural Information Processing Systems*, 32.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Xingchen Wan, Vu Nguyen, Huong Ha, Binxin Ru, Cong Lu, and Michael A Osborne. 2021. Think global and act local: Bayesian optimisation over high-dimensional categorical and mixed search spaces. *International Conference on Machine Learning (ICML) 38*.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Sercan Arik. 2024a. Teach better or show smarter? on instructions and exemplars in automatic prompt optimization. *Advances in Neural Information Processing Systems*, 37:58174–58244.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Sercan O Arik. 2024b. Teach better or show smarter? on instructions and exemplars in automatic prompt optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xingchen Wan, Han Zhou, Ruoxi Sun, and Sercan O Arik. 2025. From few to many: Self-improving many-shot reasoners through iterative optimization and generation. In *The Thirteenth International Conference on Learning Representations*.

Liang Wang, Nan Yang, and Furu Wei. 2024a. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Christopher Williams and Carl Rasmussen. 1995. Gaussian processes for regression. *Advances in neural information processing systems*, 8.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024a. Jasper and stella: distillation of SOTA embedding models. *arXiv*.

Hanlin Zhang, YiFan Zhang, Yaodong Yu, Dhruv Madeka, Dean Foster, Eric Xing, Himabindu Lakkaraju, and Sham Kakade. 2024b. A study on the calibration of in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6118–6136, Mexico City, Mexico. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Ruiqi Zhong, Charlie Victor Snell, Dan Klein, and Jason Eisner. 2023. Non-programmers can label programs indirectly via active examples: A case study with text-to-SQL. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

A Implementation Details

MMLU-Pro Prompt Template

The following are multiple choice questions (with examples) about {{ DOMAIN }}. When you provide the answer to the last question, please use the option letter without any modification, and provide the answer directly, with no formatting, no bolding, and no markup. For example, (A). The final answer must only be the letter corresponding to the correct answer.

{{ EXEMPLARS }}

Question:
{{ QUESTION }}
Options:
A. {{ OPTION_A }}
B. {{ OPTION_B }}
C. {{ OPTION_C }}
D. {{ OPTION_D }}
E. {{ OPTION_E }}
F. {{ OPTION_F }}
G. {{ OPTION_G }}
H. {{ OPTION_H }}
I. {{ OPTION_I }}
J. {{ OPTION_J }}
The answer is:

Hyper-parameters	
# Exemplars	32
LLM 1	Qwen3-8B
LLM 2	LLaMA-3.3-70B-Instruct
Temperature	0.7
# Samples	16
Top P	0.8
Top K	20
Eval Budget	200 Iterations
# BO Initial Points	20
Trust Region	Local
Acquisition Function	qNEHVI
Seeds	0,1,2

Table 3: Implementation Details of MMLU-Pro Experiments.

The MMLU-Pro dataset (Wang et al., 2024b) is licensed under the MIT License, Qwen-8B model (Yang et al., 2025) is publicly accessible under Apache 2.0 and LLaMA-3.3-70B model (Grattafiori et al., 2024) is licensed under their own community license agreement. The models and dataset used in this paper are for research purposes only. The dataset does not contain personal identifiable information or offensive content. The dataset consists of complex questions in English in various disciplines, including the domains described in Tab. 6, where each question has at most 10 options. The size of the validation set has to be reasonably large to obtain a distinguishable accuracy and ECE over search iterations. Specifically, for validation we use 174 samples in History, 188 samples in Computer Science, 233 samples in Philosophy, 342 samples in Biology, 378 samples in Business, 382 samples in Psychology, 392 samples in Health, 485 samples in Economics, 468 samples in Engineering, 534 samples in Law, 549 samples in Chemistry, 633 samples in Physics and 659 samples in Math. In addition, we have the same number of non-overlapping test samples for each task. In addition to validation and test data, we have 32 samples for the exemplar pool and 1 sample set aside for output formatting.

All Qwen3-8B experiments were performed on 2x NVIDIA A100-SXM4-40GB because an 8B model in full precision roughly occupies 56GB GPU memory. The total number of runtime is approximately 126 GPU hours for all Qwen3-8B experiments. All optimization-based methods are repeated 3 times with 3 different random seeds on the search, and non-dominated Pareto frontier points are used to evaluate on the test data respectively. All optimization-free baselines are repeated 3 times before recording the mean on the test data.

B More Experimental Results

To demonstrate the generality of COM-BOM, we conducted additional experiments on LLaMA-3.3-70B, a representative larger LLM. Similar to our experiments with Qwen3-8B, using all exemplars does not necessarily lead to optimal performance, indicating that exemplar selection is critical. Among all the optimization-based and optimization-free methods, COM-BOM consistently finds an optimal exemplar set that maximizes accuracy while minimizing calibration error. We find online retrieval-based methods can sometimes be competitive in accuracy but consistently compromise calibration error. In contrast, through its multi-objective formulation, COM-BOM aims to identify exemplars with optimal tradeoff between accuracy and calibration error.

	No Exemplars		All Exemplars		Nearest$_{k=10}$		Diversity$_{k=10}$		Ours	
Domain	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
biology	81.29	14.94	78.65	16.48	79.23	16.94	78.36	19.05	79.23	13.37
business	38.62	29.99	40.21	31.87	40.74	32.35	42.86	33.00	47.09	29.87
chemistry	35.52	38.53	42.26	32.95	41.35	35.72	42.62	37.35	47.18	30.74
computer science	47.87	19.14	55.32	30.68	56.91	32.18	52.66	29.98	57.45	26.73
economics	67.16	23.16	67.16	24.53	64.44	23.90	64.69	25.34	67.40	18.52
engineering	37.82	29.88	35.68	28.06	39.10	29.90	35.89	30.69	40.81	27.18
health	71.17	24.04	69.64	24.67	68.62	25.12	70.66	26.16	71.17	22.56
history	70.69	20.75	67.82	24.39	70.11	22.08	68.96	24.25	71.26	22.95
law	50.56	41.46	51.31	40.02	53.74	38.88	51.68	40.70	52.24	36.24
math	27.16	39.59	35.36	36.53	31.87	40.68	31.41	39.58	37.18	34.37
philosophy	62.23	27.79	57.51	30.58	62.66	29.82	60.01	32.38	64.81	25.92
physics	36.49	39.55	45.50	36.12	43.75	35.59	41.70	40.06	47.86	32.47
psychology	73.56	22.70	78.01	20.17	78.27	20.48	76.96	20.83	77.25	18.09
Average	46.10	28.71	48.66	27.95	48.64	28.69	47.91	29.90	51.12	25.30

Table 4: The test accuracy and ECE of optimization-free approaches on MMLU-Pro dataset with `LLaMA3-3.3-70B`.

	Genetic Algorithm		Simulated Annealing		Random Search		Hill Climbing		Ours	
Domain	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
biology	81.28	15.64	79.53	17.56	80.11	19.20	80.11	15.54	79.23	13.37
business	43.12	32.52	44.70	32.23	42.33	32.42	43.12	31.69	47.09	29.87
chemistry	44.08	32.54	44.08	34.32	45.53	31.99	45.17	30.97	47.18	30.74
computer science	55.85	29.53	56.91	30.60	52.66	31.26	54.79	28.01	57.45	26.73
economics	67.65	21.60	66.66	21.23	66.91	25.58	65.67	24.20	67.40	18.52
engineering	39.31	29.00	36.75	32.12	36.11	33.93	39.95	29.74	40.81	27.18
health	69.38	22.39	68.36	19.92	69.64	26.13	68.88	23.23	71.17	22.56
history	69.54	28.08	68.39	25.84	68.96	22.70	70.11	26.22	71.26	22.95
law	51.87	39.50	51.68	37.88	51.68	41.65	52.05	38.80	52.24	36.24
math	33.54	37.80	34.90	39.85	35.66	38.08	34.46	40.04	37.18	34.37
philosophy	59.23	26.01	59.66	26.55	58.80	30.35	61.80	26.54	64.81	25.92
physics	45.18	34.61	45.33	37.23	42.33	39.35	45.33	34.07	47.86	32.47
psychology	76.43	18.90	75.13	24.30	77.74	22.01	75.39	19.68	77.25	18.09
Average	49.34	27.36	49.07	28.40	48.92	29.47	49.40	27.55	51.12	25.30

Table 5: The test accuracy and ECE of optimization-based approaches on MMLU-Pro dataset with `LLaMA-3.3-70B`.

	Nearest$_{k=5}$		Nearest$_{k=10}$		Nearest$_{k=20}$		Ours	
Domain	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
biology	77.19	20.47	76.60	20.29	78.95	19.11	79.24	18.37
business	42.06	29.53	39.68	30.72	39.95	30.15	41.01	21.54
chemistry	36.79	33.28	37.34	31.79	36.43	31.61	37.34	28.38
computer science	51.06	36.21	48.94	37.16	51.60	34.62	55.85	32.08
economics	63.70	31.91	61.48	30.98	62.47	29.10	63.70	30.74
engineering	48.08	20.97	47.01	21.00	46.79	19.71	47.01	19.32
health	59.18	30.21	56.12	32.55	59.69	30.20	62.24	28.49
history	52.87	34.43	49.43	35.10	50.57	34.46	52.30	29.80
law	32.96	42.96	31.84	45.29	29.78	45.65	31.08	37.58
math	36.27	34.48	35.81	33.23	37.48	30.99	39.00	27.19
philosophy	44.64	42.73	45.92	41.23	43.35	42.61	46.78	33.84
physics	39.81	30.83	41.71	27.46	39.49	27.39	42.97	25.78
psychology	67.80	27.23	68.32	28.75	69.63	27.44	69.89	25.83
Average	44.24	29.29	43.62	29.11	43.84	28.22	45.31	25.24

Table 6: The test accuracy and ECE of online retrieval-based approach on MMLU-Pro dataset with `Qwen3-8B`. With $k \in \{5, 10, 20\}$, `COM-BOM` demonstrates superior performance (especially ECE) while being cost-effective at inference time due to its offline search before deployment with only a fixed validation set.

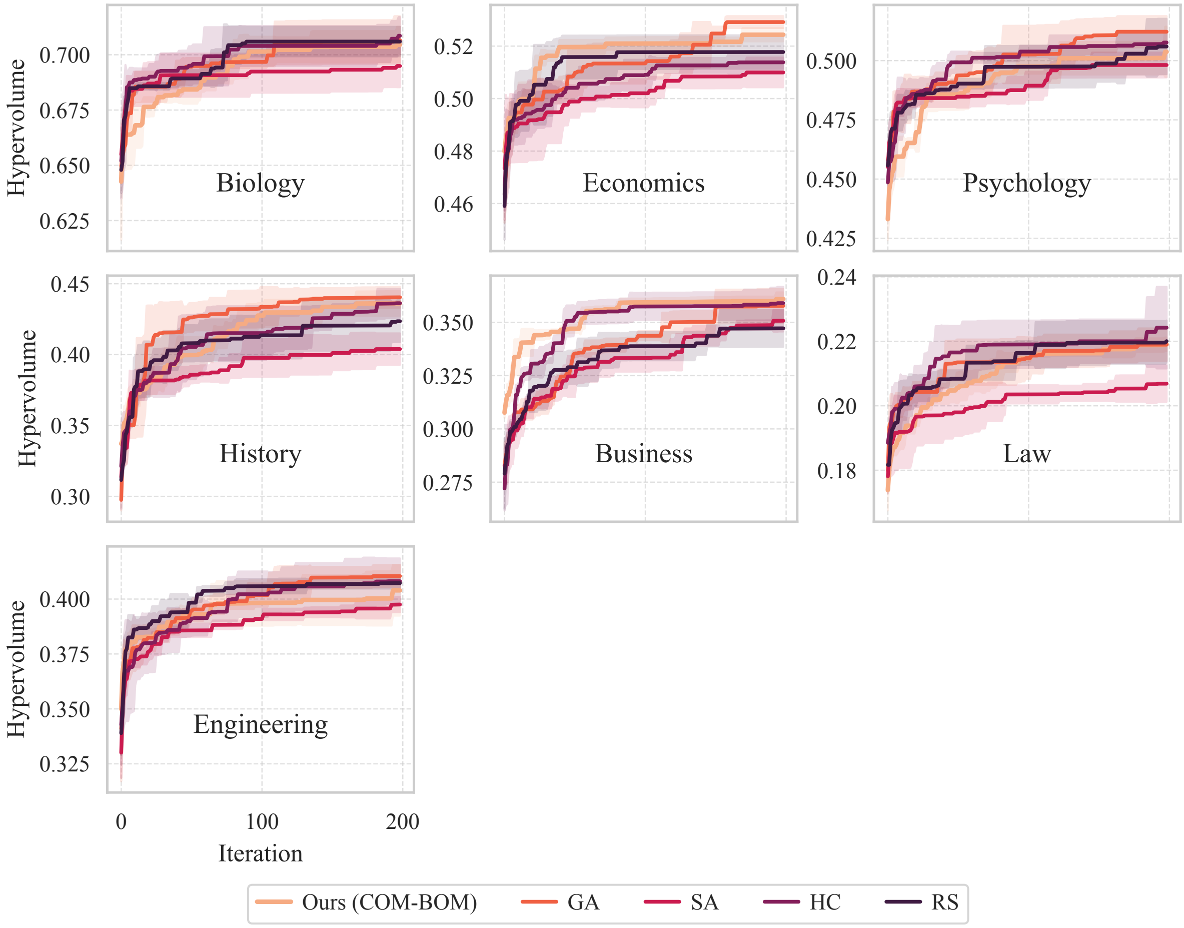

Figure 9: Evolution of best observed hypervolume on the validation data across the remaining MMLU-Pro tasks for optimization approaches. The hypervolume is measured against the reference point (accuracy=0%, ECE=100%).