Language-Based Bayesian Optimization Research Assistant (BORA)*

Abdoulatif Cissé[1,2], **Xenophon Evangelopoulos**[1,2], **Vladimir V. Gusev**[3] and **Andrew I. Cooper**[1,2]
[1]Department of Chemistry, University of Liverpool, England, UK
[2]Leverhulme Research Centre for Functional Materials Design, University of Liverpool, England, UK
[3]Department of Computer Science, University of Liverpool, England, UK
{abdoulatif.cisse, evangx, vladimir.gusev, aicooper}@liverpool.ac.uk

Abstract

Many important scientific problems involve multivariate optimization coupled with slow and laborious experimental measurements. These high-dimensional searches can be defined by complex, non-convex optimization landscapes that resemble needle-in-a-haystack surfaces, leading to entrapment in local minima. Contextualizing optimizers with human domain knowledge is a powerful approach to guide searches to localized fruitful regions. However, this approach is susceptible to human confirmation bias. It is also challenging for domain experts to keep track of the rapidly expanding scientific literature. Here, we propose the use of Large Language Models (LLMs) for contextualizing Bayesian optimization (BO) via a hybrid optimization framework that intelligently and economically blends stochastic inference with domain knowledge-based insights from the LLM, which is used to suggest new, better-performing areas of the search space for exploration. Our method fosters user engagement by offering real-time commentary on the optimization progress, explaining the reasoning behind the search strategies. We validate the effectiveness of our approach on synthetic benchmarks with up to 15 variables and demonstrate the ability of LLMs to reason in four real-world experimental tasks where context-aware suggestions boost optimization performance substantially.

1 Introduction

Exploring large experimental design spaces requires intelligent navigation strategies because of the costly and time-consuming function evaluations involved. Bayesian optimization has been established as an optimal experimental design methodology across disciplines spanning chemistry [Häse *et al.*, 2021], solar energy production [Andrés-Thió *et al.*, 2024] and agronomy [Zhang *et al.*, 2024]. BO can be used to efficiently navigate combinatorially large landscapes, and to identify promising solutions in an active-learning setting. Typically, BO uses a probabilistic surrogate to approximate an expensive or unknown objective function f while iteratively searching for a maximizer,

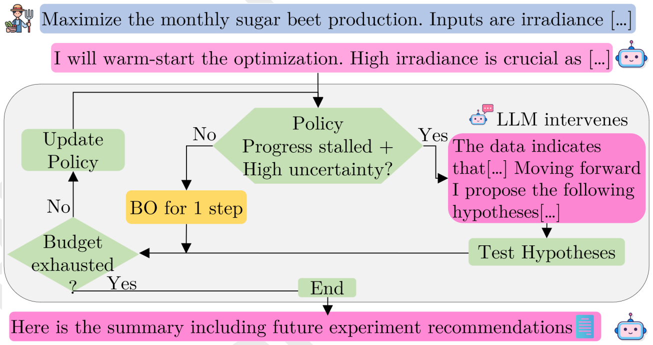

Maximize the monthly sugar beet production. Inputs are irradiance [...]
I will warm-start the optimization. High irradiance is crucial as [...]
Update Policy
No
Policy Progress stalled + High uncertainty?
Yes
LLM intervenes
The data indicates that[...] Moving forward I propose the following hypotheses[...]
BO for 1 step
No
Budget exhausted ?
Yes
Test Hypotheses
End
Here is the summary including future experiment recommendations

Figure 1: The BORA framework. Icons from [Flaticon, 2025].

$$x^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}} f(x), \qquad (1)$$

with $f : \mathcal{X} \to \mathbb{R}$ defined on the search domain $\mathcal{X} \subseteq \mathbb{R}^d$.

This surrogate – often a Gaussian Process (GP) [Garnett, 2023] – subsequently undergoes Bayesian updates as new data about the design space are acquired, according to a predefined acquisition policy or function $\alpha(\cdot)$, allowing for the refinement of predictions of the objective $f(x)$. In turn, this acquisition function suggests the next set of parameters for experiments with the highest expected utility, balancing exploration in new regions and exploitation in promising ones.

Despite its successful application to a plethora of scientific tasks, BO is frequently characterized by slow initial optimization phases due to random or Latin Hypercube [Poloczek *et al.*, 2016] selection of initial samples. This can slow the search for combinatorially large spaces substantially. This highlights a fundamental challenge for standard BO; that is, the lack of inherent domain knowledge and contextual understanding of the problem at hand. Recently, BO variants have been proposed that are capable of injecting domain-specific knowledge into the search, either through structural problem characteristics [Xie *et al.*, 2023] or by using human expert knowledge [Cissé *et al.*, 2024]. The latter approach, sometimes known as 'Human-in-the-Loop (HIL)', has drawn considerable recent attention, and it aims to infuse domain knowledge and human reasoning into BO workflows [Adachi *et al.*, 2024; Huang *et al.*, 2022]. By leveraging expert in-

*Supplementary Material (SM) for this work can be found in the extended version of the paper at https://arxiv.org/abs/2501.16224.

sights in the form of hypotheses [Cissé *et al.*, 2024], preferences [Kristiadi *et al.*, 2024] or specified priors over possible optima [Hvarfner *et al.*, 2022; Li *et al.*, 2020], it is possible to enrich the optimization process and to direct searches to fruitful regions much faster. Thus, HIL methods have shown increased effectiveness and efficiency compared with data-only approaches. In particular, hypothesis-based methods have shown gains in both performance and cost. Nonetheless, these HIL approaches can be human-capital resource-intensive because they require regular human interventions. Moreover, it is easy, even for domain experts, to lose track of the state of the art in fast-moving research areas and to ignore certain promising regions of the search space [Ou *et al.*, 2022].

To address these challenges, we propose the use of Large Language Models (LLMs) [Guo *et al.*, 2024] as a facilitating framework in black-box optimization to enrich searches with domain knowledge. Specifically, we have coupled an LLM with standard BO in a hybrid optimization framework that automatically monitors and regulates the amount of domain knowledge needed when the search becomes 'trapped' in local minima. The algorithm capitalizes on the LLM's inherent in-context learning (ICL) capacity and reasoning mechanism to suggest, in the form of hypotheses, promising areas in the search space from which to sample. LLMs have been employed recently to address limitations in core BO methodologies, as well as HIL variants [Liu *et al.*, 2024; Yin *et al.*, 2024]. LLMs have the capacity to encode vast amounts of domain-specific and general knowledge and have demonstrated the ability to reason about relatively complex tasks through in-context learning [Kroeger *et al.*, 2024; Xie *et al.*, 2022] as well as in multidisciplinary domains such as chemistry [Ramos *et al.*, 2024]. However, due to their numerically agnostic design, LLMs lag behind traditional BO methods in systematically balancing exploration versus exploitation, and have proved unreliable in many practical scenarios [Huang *et al.*, 2024]. Recent attempts have been made to integrate LLMs with BO frameworks [Mahammadli, 2024; Yin *et al.*, 2024] but thus far, these have been limited to small problem sizes, such as hyper-parameter optimization, or situations where the optimal solution is proximal to a special value [Huang *et al.*, 2024]. Also, LLM/BO hybrids could be prohibitively costly for more complex queries, particularly if the LLM is deployed for every iteration in the optimization.

Here, we propose a language-based Bayesian Optimization Research Assistant, BORA, that enriches BO with domain knowledge and contextual understanding across a range of scientific tasks. We frame BORA as a hybrid framework that augments surrogate-based optimizers with uncertainty estimates by localizing areas of interest in the search space, guided by a knowledge-enriched LLM (Figure 1). A heuristic policy regulates the LLM involvement in the optimization process, adaptively balancing rigorous stochastic inference with LLM-generated insights within a feasible budget of LLM computation and API usage limits. During the intervention stage, the LLM uses its domain knowledge and reasoning capabilities to comment on the optimization progress thus far, highlighting patterns observed and forming hypotheses that may yield more rewarding solutions. It then tests these hypotheses by proposing new samples that maximize the target objective. BORA is also designed to provide an effective user-optimizer interaction through its dynamic commentary on the optimization process. This promotes deeper insights from the user and, in the future, the option to intervene; for example, by either reinforcing or overriding certain insights from the LLM. To our knowledge, this is the first time that a rigorous, dynamic synergy of black-box BO with LLMs has been proposed in this context. We evaluated BORA on various synthetic functions and a pétanque gaming model, as well as real scientific tasks in chemical materials design, solar energy production, and crop production. BORA demonstrated significant improvements in search exploration, convergence speed, and optimization awareness. Compared to earlier techniques, our method shows significant efficiency gains and generalization beyond hyper-parameter optimization, emphasizing its potential for tackling real-world tasks.

The remainder of this paper is organized as follows. Section 2 presents recent works about domain knowledge and LLM integration in BO and Section 3 details our proposed methodology. Section 4 analyzes and compares the performance of our algorithm against state-of-the-art methods across diverse datasets, with Section 5 concluding our work and discussing future directions.

2 Related Works

To cope with non-convex optimization landscapes in science tasks, intelligent approaches have been proposed that focus on promising regions through adaptive exploration-exploitation strategies [Shoyeb Raihan *et al.*, 2024], or 'smooth out' the optimization landscape by enriching it with domain knowledge [Ramachandran *et al.*, 2020]. Notable examples include local BO methods that restrict the search space, such as ZoMBI [Siemenn *et al.*, 2023], which aims to improve efficiency by focusing on local regions assumed to contain the optimum. Similarly, TuRBO [Eriksson *et al.*, 2019] uses multiple independent GP surrogate models within identified trust regions and a multi-armed bandit strategy [Vermorel and Mohri, 2005] to decide which local optimizations to continue. These approaches are well-suited to handling high-dimensional problems, but their potential is perhaps more limited in small budgets and highly multimodal spaces due to a lack of built-in domain knowledge.

Incorporating domain knowledge into BO can improve both its efficiency and its performance [Adachi *et al.*, 2024; Häse *et al.*, 2021]. DKIBO [Xie *et al.*, 2023] enhances BO's acquisition function with structural knowledge from an additional deterministic surrogate model to enrich the GP's approximation power. Others, such as ColaBO [Hvarfner *et al.*, 2024] and HypBO [Cissé *et al.*, 2024], allow users to inject their beliefs at the start to guide the optimization process. However, those methods keep the users' beliefs static and cannot refine them as the optimization progresses, even if they are wrong. Meanwhile, other HIL methods rely on frequent user inputs [Savage and del Rio Chanona, 2023] and for robotic experiments [Burger *et al.*, 2020], for example, that run 24/7 in a closed-loop way, waiting for this human user input might become the rate-limiting step.

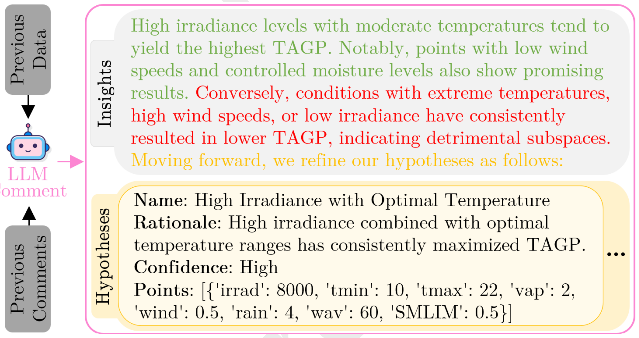
Recently, some studies have explored LLMs as standalone replacements for traditional optimizers due to their exceptional ability to solve complex problems in various domains [M. Bran *et al.*, 2024; Nguyen and Grover, 2024]. Methods like LLAMBO [Liu *et al.*, 2024] and OPRO [Yang *et al.*, 2024] use the generative and ICL capabilities of LLMs to propose solutions to optimization problems directly. LLAMBO mimics BO's structure and replaces its key components with LLMs. In OPRO, the LLM is iteratively prompted with the gathered optimization data as input and tasked to generate new solutions as output that are then evaluated. These methods are innovative but have focused so far on low-dimensional hyperparameter tuning and are not yet obviously suitable as a general framework for optimization tasks. Querying LLMs at all iterations also incurs a larger computational and financial footprint than traditional BO algorithms, particularly if reasoning models are used. Standalone LLM optimizers also lack the mathematical guarantees offered by traditional optimizers such as BO.

In response to the limitations of using LLMs as standalone optimizers, hybrid approaches such as BoChemian [Ranković and Schwaller, 2023] have emerged that combine the strengths of LLMs to featurize traditional optimization methods. SLLMBO [Mahammadli, 2024] integrates the strengths of LLMs in warm-starting optimization, and it loops between LLM-based parameter exploitation and Tree-structured Parzen Estimator (TPE)'s exploration capabilities to achieve a balanced exploration-exploitation trade-off. This reduces API costs and mitigates premature early stoppings for more effective parameter searches. However, SLLMBO, like LLaMEA-HPO [van Stein *et al.*, 2024], is limited to hyperparameter tuning. Moreover, its LLM exploration / TPE exploitation cycle lacks dynamic adjustment because it is an alternating process fixed at a 50:50 balance. Another limitation is the risk of premature optimization termination in complex search spaces due to a strict early stopping mechanism.

Our approach, BORA, shares similarities with the above studies by incorporating domain knowledge and adapting search mechanisms. However, BORA is distinguished by leveraging LLMs when they are most required, for online hypothesis generation *and* for real-time commentary on optimization progress. Unlike static methods such as HypBO, which assume fixed human-injected soft constraints, our method refines the optimization trajectory based on the contextual insights given by the LLM. Moreover, BORA extends beyond previous hybrid approaches such as SLLMBO by introducing adaptive heuristics that intelligently modulate LLM involvement with BO to maximize optimization performance.

3 Methodology

The BORA optimization framework is illustrated in Figure 1. It is an automated hybrid BO-LLM synergy operating under a common GP whose parameters are updated as new points are sampled, either from BO or the LLM. A user-provided experiment description is used to initially contextualize the LLM which then warm-starts the optimization with proposed samples through its ICL capabilities. The optimization progresses by alternating BO and LLM runs that are accordingly triggered by performance plateaus. Our proposed framework employs an adaptive heuristic policy to (a) assess the need to invoke the LLM, (b) determine the type of LLM intervention needed, and (c) update the frequency of those interventions as the optimization progresses. We use Reinforcement Learning (RL) terminology in the manuscript to describe our approach, but we use hand-crafted policy update rules because learning generalized rules in the traditional sense [Liu *et al.*, 2022; Volpp *et al.*, 2020] would be impractical in Bayesian scientific optimization settings [Lee *et al.*, 2020], which is the focus of this work. In its interventions, the LLM provides user interpretability via real-time commentary of the optimization progress and generates hypotheses to maximize the objective.



High irradiance levels with moderate temperatures tend to yield the highest TAGP. Notably, points with low wind speeds and controlled moisture levels also show promising results. Conversely, conditions with extreme temperatures, high wind speeds, or low irradiance have consistently resulted in lower TAGP, indicating detrimental subspaces. Moving forward, we refine our hypotheses as follows:

Name: High Irradiance with Optimal Temperature
Rationale: High irradiance combined with optimal temperature ranges has consistently maximized TAGP.
Confidence: High
Points: [{'irrad': 8000, 'tmin': 10, 'tmax': 22, 'vap': 2, 'wind': 0.5, 'rain': 4, 'wav': 60, 'SMLIM': 0.5}]

Figure 2: The LLM agent commenting and refining its hypotheses on the Sugar Beet Production experiment (complete comment in the SM). This experiment is detailed in Section 4.1.

3.1 LLM Comments and Hypotheses

The LLM is prompt-engineered to return structured JSON responses that we call *Comments* (for formatting details we refer the reader to the SM). The Comment object, illustrated in Figure 2, contains *insights* into the optimization progress and potential reasons for stagnation, as well as a list of *hypotheses* to remedy that stagnation. Each hypothesis includes a meaningful name, a rationale, a confidence level, and the corresponding input point to test it. Unlike in HypBO [Cissé *et al.*, 2024] where hypotheses are defined as rather static regions of interest, BORA dynamically builds hypotheses during the optimization process typically in the form of single search points through the LLM's ICL model. As demonstrated in LLAMBO [Liu *et al.*, 2024], LLMs tasked with regression inherently perform an implicit ICL modeling of the objective function, estimating the conditional distribution $p(y|x;\mathcal{D})$, where y is the target value at x. BORA extends this modeling by integrating all previously gathered data $\mathcal{D}$ and all the LLM's comments $\mathcal{C}$, enhancing the LLM surrogate to model $p(y|x;\mathcal{D};\mathcal{C})$. From this augmented model, the LLM proposes hypotheses, exploring regions likely to improve on the current best observation $y_{\max}$ and derived from the conditional probability $x^{\text{LLM}} \sim p(x|y > y_{\max};\mathcal{D};\mathcal{C})$.

3.2 LLM Initialization

User-Provided Experiment Card

To inform the LLM initially, the user prepares a comprehensive problem description following a standardized template that we refer to as the *Experiment Card*. This card includes any details or context about the black-box function f to be optimized, descriptions of its input variables, and the target variable to be maximized, along with any constraints that must be satisfied within the search space. From the experiment card, the LLM is prompted to generate n_{init} initial hypotheses for maximizing the target. This translates into n_{init} initial points that are evaluated to form the initial dataset $\mathcal{D}_0 = \big\{\big(x_i, y_i = f(x_i)\big)\big\}_{i=1}^{n_{\text{init}}}$.

3.3 Actions

BORA leverages an adaptive heuristic policy detailed in Section 3.4 to choose one action from a set of three possible actions defined in the following paragraphs. The chosen action suggests at least one next point for evaluation, which is evaluated and added to the dataset. While the Vanilla BO action a_1 appends one sample (x, y) to $\mathcal{D}_{t-1}$ at each step or iteration t, the LLM actions a_2 and a_3 add n_{LLM} and $n_{\text{LBO}} \geq 1$ samples, respectively. Hereon, we distinguish between t, the step number, and i, the sample index at step t, and denote with $\mathcal{S}_t = \{x_t^{(i)}\}_{i=1}^{k}$ the set of k points suggested by an action a at step t.

a_1 Continue with Vanilla BO

The acquisition function is maximized to get the next promising point $\mathcal{S}_t = \{x_t^{(1)}\}$, which is then evaluated and added to the dataset to form $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(x, f(x)) \mid x \in \mathcal{S}_t\}$.

a_2 LLM Comments and Suggests n_{LLM} Points

A prompt containing the gathered data up to $\mathcal{D}_{t-1}$ and any previous comments $\mathcal{C}$ is given to the LLM. The LLM is then tasked to comment on the optimization progress in light of the new data and to update any previous hypotheses. The returned Comment contains the next points $\mathcal{S}_t = \{x_t^{(i)}\}_{i=1}^{n_{\text{LLM}}} \sim p(x|y > y_{\max}; \mathcal{D}_{t-1}; \mathcal{C})$ chosen by the LLM, which are then evaluated and added to the dataset to form $\mathcal{D}_t$.

a_3 LLM Comments and Selects n_{LBO} BO Points

a_3 is a non-myopic, ICL step that focuses the LLM's attention on a high-quality set of n_{BO} candidate points $\{x_{\text{BO}}^{(j)}\}_{j=1}^{n_{\text{BO}}}$, which are generated by maximizing the acquisition function. A prompt containing $\mathcal{D}_{t-1}$, $\mathcal{C}$, and $\{x_{\text{BO}}^{(j)}\}_{j=1}^{n_{\text{BO}}}$ are given to the LLM, which is then tasked to comment on the optimization progress in light of the new data and constrained to select the n_{LBO} most promising BO candidates that best align with its hypotheses for maximizing the target objective. Setting $n_{\text{BO}} = 5$ and $n_{\text{LBO}} = 2$ empirically showed to offer enough diversity of hypothesized optima locations and ensure competitive performance overall. The returned Comment holds the n_{LBO} selected points $\mathcal{S}_t = \big\{x_t^{(i)}\big\}_{i=1}^{n_{\text{LBO}}} \sim p(x \in \{x_{\text{BO}}^{(j)}\}_{j=1}^{n_{\text{BO}}}|y > y_{\max}; \mathcal{D}_{t-1}; \mathcal{C})$ that are then evaluated and added to the dataset.

3.4 Adaptive Heuristic Policy

Action Selection

BORA's policy helps it to make informed choices about engaging the LLM without relying on data-hungry RL algorithms, thus maintaining BORA's practicality and effectiveness in real-life scenarios. The optimization starts with the Vanilla BO action a_1, and the subsequent action selection depends on (a) the average uncertainty $\sigma^{\text{GP}}_{\text{mean}}$ of the common GP over the search space $\mathcal{X}$ to determine the necessity and type of LLM intervention, and (b) the BO performance plateau detection, as well as the performance success (or trust in) of the previous LLM interventions. When the GP's uncertainty is high and above a pre-defined threshold ($\sigma^{\text{GP}}_{t,\text{mean}} > \sigma_{t,\text{upper}}$), BO needs significant guidance from the LLM, triggering a complete 'take-over' by the LLM in the search, suggesting new points informed by its own internal reasoning mechanism. As the GP's uncertainty decreases ($\sigma_{t,\text{lower}} \leq \sigma^{\text{GP}}_{t,\text{mean}} \leq \sigma_{t,\text{upper}}$), the LLM becomes less involved by relying only on BO suggested points, but still using its ICL capacity based on both $\mathcal{D}$ and $\mathcal{C}$ to select the most promising ones. When the GP's uncertainty is low enough ($\sigma^{\text{GP}}_{t,\text{mean}} < \sigma_{t,\text{lower}}$), BO has a better approximation of objective function's landscape and no longer needs guidance from the LLM. The rationale behind remark (b) is that the LLM should gain more trust as the LLM suggestions exhibit better performance, which in turn triggers the plateau duration to be re-defined as shorter, allowing the LLM to intervene more frequently. Conversely, if the LLM's so-far observed performance declines, its trust in itself diminishes and, consequently, its interventions are reduced, which results in longer plateau duration adjustments before invoking its assistance. In short, the action selection at each step t follows the policy π described below, where the GP parameters are updated after every action accordingly:

- If $\sigma^{\text{GP}}_{t,\text{mean}} < \sigma_{t,\text{lower}}$ or 'no plateau' → **action** a_1,
- Else if $\sigma^{\text{GP}}_{t,\text{mean}} > \sigma_{t,\text{upper}}$ → **action** a_2,
- Else → **action** a_3.

Selection Mechanism

Uncertainties The above action selection is realized by calculating and updating in every step the uncertainties from a set of fixed q monitoring points $x_{\text{mon}}^{(i)}$ that are randomly sampled before the optimization starts. Specifically,

$$\sigma^{\text{GP}}_{t,\text{mean}} = \frac{1}{q}\sum_{i=1}^{q} \sigma_t\big(x_{\text{mon}}^{(i)}\big), \tag{2}$$

$$\sigma^{\text{GP}}_{t,\max} = \max\big(\sigma^{\text{GP}}_{t-1,\max}, \max_{1 \leq i \leq q} \sigma_t(x_{\text{mon}}^{(i)})\big), \tag{3}$$

$$\sigma_{t,\text{upper}} = 0.5 \times \sigma^{\text{GP}}_{t,\max} \text{ and } \sigma_{t,\text{lower}} = 0.3 \times \sigma^{\text{GP}}_{t,\max}, \tag{4}$$

where $\sigma_t(\cdot)$ represents the uncertainty of the GP at a given point in iteration t. Here, the 50% and 30% fractions serve as empirically tuned bounds that consistently balance BO exploitation with LLM exploration across diverse tasks.

Plateau Detection Another important part of the action selection mechanism in the proposed framework is the detection of performance plateauing in BO. A performance plateau is detected at step t when

$$y_j^{\max} < y_{j-1}^{\max} \times \big(1 + \text{sign}(y_{j-1}^{\max}) \times \gamma\big), \text{for all } j \in [t-m+1, t], \tag{5}$$

where $y_t^{\max} = \max\big(\{y|(x, y) \in \mathcal{D}_t\}\big)$. That is, if for the past m consecutive BO steps, there is not enough performance improvement (w.r.t a set percentage γ), then the LLM involvement is triggered. The plateau duration m is initialized at $m_{\text{init}} = \lceil 2\sqrt{d} \rceil$, set to vary between $m_{\text{min}} = 0$ and $m_{\text{max}} = 3m_{\text{init}}$, and is automatically adjusted at every LLM intervention step l (here l counts the number of times actions a_2 or a_3 are invoked). The adjustment depends on the current 'trust' $T_l \in [0, 1]$ BORA has on the LLM, which in turn relies on the LLM performance observed so far. Specifically

$$m_{\text{adjustment}} = \lfloor (T_l - T_{l-1}) \times \Delta_{\max} \rfloor \,, \tag{6}$$

$$m \leftarrow \text{clip}\left(m - m_{\text{adjustment}},\, m_{\text{min}},\, m_{\text{max}}\right), \tag{7}$$

where $\Delta_{\max}$ is the maximum allowed adjustment per step, here set to 15, and clip(x, a, b) is a function that restricts x to be within the bounds $[a, b]$.

Trust Mechanism As noted above, the plateau adjustment relies on an adaptive trust mechanism that regulates the trust in the LLM as defined by a 'trust score' calculated on previous performances. That is, at each step t where the LLM suggests (or selects) $\mathcal{S}_t = \{x_t^{(i)}\}_{i=1}^{k}$, the trust score is updated based on the following reward function

$$r_l = \max\bigl(\{f(x) | x \in \mathcal{S}_t\}\bigr) - y_{t-1}^{\max} \tag{8}$$

First, an intervention score, ranging in $[0, 1]$, reflects the utility of those LLM suggestions in finding a new optimum with respect to the reward function in Eq. (8) as

$$\text{score}_{\text{interv}}^{(l)} = \begin{cases} 1, & \text{if } r_l > 0, \\ \frac{1}{1+e^{-\frac{r_l}{|y_{t-1}^{\max}|+\epsilon}}}, & \text{if } r_l \leq 0, \end{cases} \tag{9}$$

where $\epsilon = 10^{-6}$ is a small constant to handle cases where $y_{t-1}^{\max} = 0$. By normalizing r_l with $|y_{t-1}^{\max}|$, the trust score becomes more sensitive to relative changes rather than absolute changes. This is particularly useful in domains where the magnitude of y varies widely, making it robust across scales. Then, this intervention score is added to $\mathcal{H} \leftarrow \mathcal{H} \cup \{\text{score}_{\text{interv}}^{(l)}\}$, keeping track of the previous intervention scores. Note that to reflect an initially optimistic view of the LLM, $\mathcal{H}$ is initialized as $\{0.9\}$, i.e., $\text{score}_{\text{interv}}^{(0)} = 0.9$. Finally, an average rolling trust score T_l is subsequently calculated as the average of the intervention scores in $\mathcal{H}$ over a sliding window W of the three most recent intervention scores as

$$T_l = \frac{1}{W} \sum_{i=|\mathcal{H}|-W}^{l} \text{score}_{\text{interv}}^{(i)} \text{ where } W = \min(|\mathcal{H}|, 3). \tag{10}$$

The complete BORA framework is described in Algorithm 1. Details about the LLM prompt engineering, reflection strategies, and fallback mechanisms can be found in the SM.

4 Experiments

We validated BORA's performance against current state-of-the-art methods on both synthetic functions and various real-world tasks, with dimensionality ranging from 4 to 15 independent variables. Section 4.1 outlines the experimental setup while Section 4.2 highlights the results. Details on the benchmarks, the method implementations, and the reproducibility details can be found in the SM. The source code is available at https://github.com/Ablatif6c/bora-the-explorer.

Algorithm 1 BORA

Input: Experiment card, Number of initial samples n_{init}, Maximum number of samples $i_{\max}$ **Output**: $y_{\max}$, LLM comments $\mathcal{C}$ and final report

```
LLM generates initial samples $\mathcal{D}_0 \leftarrow \{(x_i, f(x_i))\}_{i=1}^{n_{\text{init}}}$;
Initialize the GP surrogate model with $\mathcal{D}_0$;
Initialize policy parameters $\sigma_{0,\text{mean}}^{\text{GP}}$, $\sigma_{0,\max}^{\text{GP}}$, $\sigma_{0,\text{upper}}$, $\sigma_{0,\text{lower}}$, $m$, $\mathcal{H} = \{0.9\}$, $\gamma = 0.05$, $n_{\text{BO}} = 5$, $n_{\text{LBO}} = 2$;
Initialize sample index $i = n_{\text{init}}$, step $t = 1$, $\mathcal{C} = \{\}$;
while $i < n_{\text{init}} + i_{\max}$ do
    if $\sigma_{t,\text{mean}}^{\text{GP}} < \sigma_{t,\text{lower}}$ or 'no plateau' then
        $a = a_1$, $\mathcal{S}_t = \{x_t^{(1)}\}$;
    else if $\sigma_{t,\text{mean}}^{\text{GP}} > \sigma_{t,\text{upper}}$ then
        $a = a_2$, $\mathcal{S}_t = \{x_t^{(k)}\}_{k=1}^{\min(n_{\text{LLM}}, i_{\max}+n_{\text{init}}-i)}$;
    else
        $a = a_3$, $\mathcal{S}_t = \{x_t^{(k)}\}_{k=1}^{\min(n_{\text{LBO}}, i_{\max}+n_{\text{init}}-i)}$;
    end if
    Update dataset $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(x, f(x)) | x \in \mathcal{S}_t\}$;
    Update $y_{\max}$ as the maximum $y$ value in $\mathcal{D}_t$;
    Update GP, policy parameters and trust mechanism;
    $t \leftarrow t + 1$ and $i \leftarrow i + k$;
end while
LLM generates a final report.
```

4.1 Experimental Setup

Synthetic Function Benchmarks

- **Branin (2D)**: A function with a global maximum occurring in three distinct locations as shown in Figure 3. The input space bounds are $x_0 \in [-5, 10]$ and $x_1 \in [0, 15]$.
- **Levy (10D)**: A function with a highly rugged landscape. All inputs are bounded in $[-10, 10]$ with the maximum at $[1, \ldots, 1]$.
- **Ackley (15D)**: A challenging high dimensional function with several local maxima. Input bounds are $[-30, 20]$ with the maximum at $[0, \ldots, 0]$.

Note that the names of these functions in the experiment card were anonymized to 'mathematical function' to prevent the LLM from recognizing them by name.

Real-World Application Benchmarks

- **Solar Energy Production (4D)**: Maximizing the daily energy output of a solar panel by optimizing panel tilt, azimuth, and system parameters [Anderson *et al.*, 2023].
- **Pétanque Game (7D)**: A ball is thrown to hit a target. The goal is to maximize the score, which is inversely proportional to the target distance miss, by tuning the throw position, angles, velocity, and spins.
- **Sugar Beet Production (8D)**: Maximizing the monthly sugar beet Total Above Ground Production (TAGP) in a greenhouse by tuning the irradiance, and other weather and soil conditions [de Wit and contributors, 2024].
- **Hydrogen Production (10D)** Maximizing the hydrogen evolution rate (HER) for a multi-component catalyst mixture by tuning discrete chemical inputs under the constraint that the total volume of the chemicals must

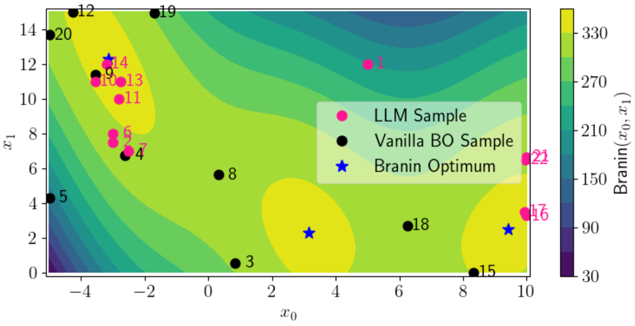


Figure 3: Visualization of BORA maximizing Branin 2D (which contains three global maxima) under a budget of 22 optimization steps (numbered points). Black samples were suggested by the BO action a_1, while pink ones came from the LLM actions a_2 and a_3.

not exceed 5 mL. Note that due to the discrete and constrained nature of the problem, we adapted all compared methods accordingly to account for this, by employing the bespoke implementation of [Burger *et al.*, 2020]. Dataset acquired from [Cissé *et al.*, 2024].

BORA

We implemented BORA using OpenAI's most cost-effective model at the time, GPT-4o-mini [OpenAI, 2025], which was not fine-tuned in our effort to make BORA more accessible to users with limited resources. For the BO action implementation, the GP uses a Matérn kernel, and the acquisition function is EI. We set $q = 5{,}000$ for $\sigma^{\text{GP}}_{t,\text{mean}}$.

Baselines

- **Random Search**: Unbiased exploration baseline.
- **BayesOpt** [Nogueira, 2014]: Example of vanilla BO.
- **TuRBO** [Eriksson *et al.*, 2019] with a single trust region.
- **ColaBO** [Hvarfner *et al.*, 2024] that uses a single static expert given-prior over the optimum to guide the optimization process.
- **HypBO** [Cissé *et al.*, 2024] that uses multiple static expert-given-promising regions to guide the optimization.
- **LAEA** [Hao *et al.*, 2024], a hybrid LLM-Evolutionary Algorithm method.

For ColaBO and HypBO, to avoid the impracticality of relying on humans to provide inputs for multiple trials across all experiments, we used the LLM GPT-4o-mini to generate the 'human' inputs. Likewise, we used the same task description prompts as used for BORA, to ensure consistency. For HypBO on the Hydrogen Production experiment, we employed the most realistic hypothesis used in [Cissé *et al.*, 2024], namely 'What They Knew', which encapsulates any human knowledge available prior to the execution of the experiment.

Experimental Protocol

The optimization performance was measured using the maximum objective value found so far, the cumulative regret, and statistical tests to measure significance. The maximum number of samples was set to 105 to account for realistic budgets with expensive functions. Average results of 10 repeated trials with distinct random seeds are reported. All methods were initialized with $n_{\text{init}} = 5$ initial samples apart from LAEA, for which we used 15 initial samples to keep the same number of evaluations to population size ratio as in [Hao *et al.*, 2024].

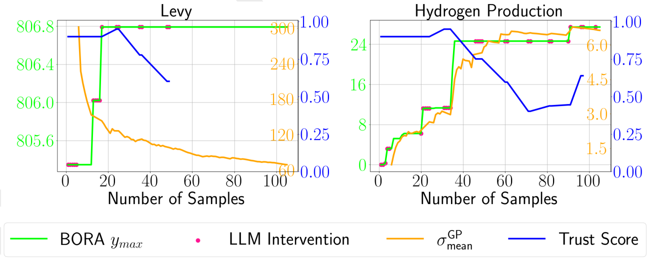


Figure 4: LLM intervention monitoring during a BORA run on 10D Levy (left) and Hydrogen Production (right). Mean uncertainty and Trust scores are also overlayed to highlight their interrelationships.

4.2 Results

Synthetic Functions

Figure 3 illustrates the exploration strategy of BORA on the Branin (2D) function. The LLM interventions helped to uncover two out of the three possible locations of the global maximum of Branin. This is further illustrated in Figure 5, which shows that BORA outperforms the baseline comparisons for the higher-dimension functions Levy (10D) and Ackley (15D). A key advantage of BORA is its LLM-informed initial sampling. For mathematical functions, BORA systematically suggests initializing at critical points such as the edges, central points, or other remarkable points like $[0, \ldots, 0]$. This strategy is particularly well-suited for the Levy function, whose search space bounds are symmetric, and its optimum is at $[1, \ldots, 1]$, almost always converging in its initialization stage. However, that strategy is less beneficial on the Ackley function because its bounds are asymmetric. Despite that, the LLM's ability to reflect and learn from the previous samples appears to help mitigate any unfavorable initializations. While HypBO demonstrates a similar benefit through its initial sampling in hypothesized regions, its performance is comparatively weaker because it relies on random initial sampling within these regions, resulting in less effective exploration of the search space. For ColaBO, which only works with a single input prior, the prior tended to be around one of the edges, which overall limits its convergence speed. Additionally, the left panel of Figure 4 shows how BORA's iterative hypothesis generation, informed by previous data, helps mitigate stagnant optimization, and discards the LLM when it is no longer needed. Notably, a sharp drop in the GP uncertainty is evident when vanilla BO is used due to is proved exploration-exploitation guarantees, as opposed to the less rigorous LLM where the uncertainty is bound to its inherent sampling strategy. Nevertheless, the dynamic synergistic effect of BO coupled with updated LLM hypotheses allows for faster convergence overall, in comparison to other baselines as illustrated in the last two bar plots of Figure 6.

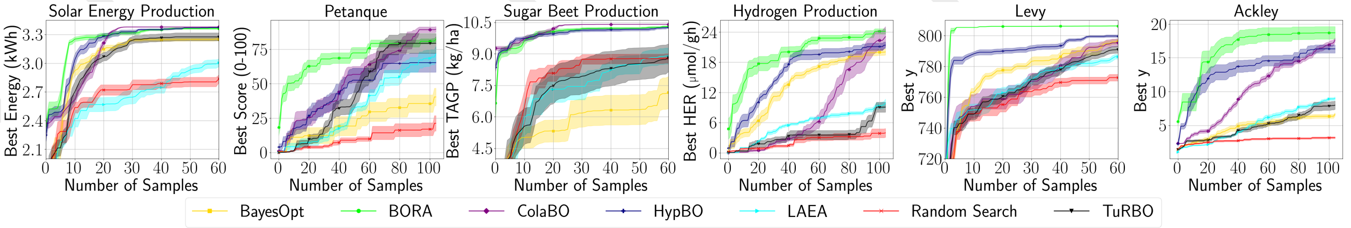


Figure 5: BORA vs Baselines on six experiments. Solid lines show average values while shaded areas indicate ±0.25 standard error. For visual clarity, some plots are zoomed in to show results up to 60 iterations, as the trends mostly stabilize beyond this point. Full results in SM.

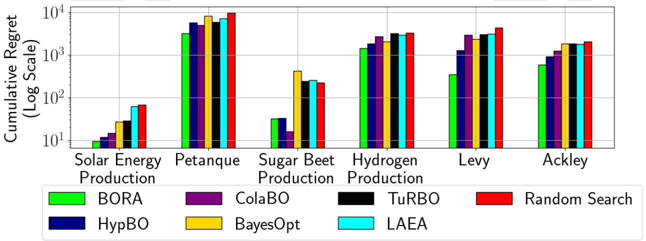


Figure 6: Comparison of BORA vs. Baselines' cumulative regrets in the six experiments.

Real-World Applications

BORA also exhibits superior performance across diverse real-world optimization problems, following the trends observed in the synthetic benchmarks. As shown in Figure 5, while BORA's initial sampling is on par with other input-based baselines for the Solar Energy and Sugar Beet Production experiments, its overall performance on all experiments surpasses the baselines significantly. This is particularly evident in the 7D Pétanque experiment, where BORA's diverse initial hypotheses based on trajectory dynamics led to a remarkable gain in score of 35 in the early stages compared to the baselines. This knowledge and context-based input bridges the knowledge gap typically encountered in early-stage optimization, providing BORA with a critical advantage, as shown in Figure 6. A similar effect is also observed in the Hydrogen Production experiment as illustrated in the right panel of Figure 4. In addition, in the later stage of the optimization the LLM further pushes the optimization to uncover new optima after the progress had stalled, thus gaining more trust. The increasing trend in the GP uncertainty here is a side-effect of the continual interventions of the LLM, which translates to a rather explorative and less exploitative strategy based on its inherent domain reasoning around cumulatively accrued scientific data. This goes beyond some of the near-instant convergence noted in the optimization of the synthetic functions because they are proximal to a special value. Figure 2 illustrates this by showing how the LLM reflects on the progress and generates hypotheses on the Sugar Beet Production experiment. While other baselines, particularly knowledge-based methods such as HypBO and local BO approaches such as TuRBO, demonstrate improved performance as the optimization progresses and more data is gathered, they often struggle to match BORA's sustained performance as the 105-sample budget mark is approached. In the Hydrogen Production experiment, this adaptive strategy ultimately achieved a 47% reduction in cumulative regret compared to ColaBO, demonstrating BORA's faster convergence and robustness in navigating complex, high-dimensional search spaces. To assess the significance of the performance difference w.r.t mean cumulative regret between BORA and its best two competitors, we performed a sign test which revealed that BORA performs consistently better than HypBO with a p-value of 0.02 at a 95% confidence level with a Bonferroni correction [Bonferroni, 1936], but not against ColaBO with a p-value of 0.20, yet still outscoring it in 5 out of 6 tasks. The superior performance of the hybrid approach in BORA was further validated by ablation studies that used only the LLM (action a_2) for optimizing Hydrogen Production (10D) and the Ackley function (15D) (see SM). While performing quite well in the initial stages for these two problems, the use of the LLM alone was ultimately less effective than the dynamic hybrid BO/LLM approach in BORA. We emphasize that these results do not mean that LLMs are 'smarter' than domain experts. Rather, they highlight BORA's ability to update and refine its hypotheses based on new data, which is not possible in the HypBO implementation [Cissé *et al.*, 2024], while also fostering user engagement by generating real-time optimization progress commentary and a final summary report (see SM). One potential limitation of BORA, however, is the stochastic nature of the LLM reasoning, which can diverge considerably even with identical prompts.

5 Conclusions

This work introduces BORA, the first optimization framework to integrate BO with LLMs in a cost-effective dynamic way for scientific applications. BORA leverages the reasoning capabilities of LLMs to inject domain knowledge into the optimization process, warm-starting the optimization and enabling hypothesis-driven exploration and adaptive strategies to navigate complex, non-convex search spaces. It addresses key limitations in traditional BO methods, including slow initialization, local minimum entrapment, and the lack of contextual understanding. Notably, BORA outperformed BO with the addition of static expert-knowledge-derived hypotheses in a challenging 10D chemistry experiment, Hydrogen Production, highlighting its potential as a collaborative AI tool to support and enhance expert decision making. Future directions will include refining BORA's meta-learning strategies using multi-agent LLMs and exploring its effectiveness in multi-objective, multi-fidelity optimization scenarios.

Acknowledgments

The authors acknowledge financial support from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design. The authors also acknowledge the AI for Chemistry: AIchemy hub for funding (EPSRC grant EP/Y028775/1 and EP/Y028759/1). This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 856405). AIC thanks the Royal Society for a Research Professorship (RSRP\S2\232003).

References

[Adachi *et al.*, 2024] Masaki Adachi, Brady Planden, David Howey, Michael A. Osborne, Sebastian Orbell, Natalia Ares, Krikamol Muandet, and Siu Lun Chau. Looping in the human: Collaborative and explainable Bayesian optimization. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 505–513. PMLR, 02–04 May 2024.

[Anderson *et al.*, 2023] Kevin S Anderson, Clifford W Hansen, William F Holmgren, Adam R Jensen, Mark A Mikofski, and Anton Driesse. pvlib python: 2023 project update. *Journal of Open Source Software*, 8(92):5994, 2023.

[Andrés-Thió *et al.*, 2024] Nicolau Andrés-Thió, Charles Audet, Miguel Diago, Aimen E Gheribi, Sébastien Le Digabel, Xavier Lebeuf, Mathieu Lemyre Garneau, and Christophe Tribes. solar: A solar thermal power plant simulator for blackbox optimization benchmarking. *arXiv:2406.00140*, 2024.

[Bonferroni, 1936] Carlo E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936.

[Burger *et al.*, 2020] Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiao yan Wang, Xiaobo Li, Ben M. Alston, Buyin Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583:237–241, 2020.

[Cissé *et al.*, 2024] Abdoulatif Cissé, Xenophon Evangelopoulos, Sam Carruthers, Vladimir V. Gusev, and Andrew I. Cooper. HypBO: Accelerating black-box scientific experiments using experts' hypotheses. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 3881–3889, 2024.

[de Wit and contributors, 2024] A.J.W. de Wit and contributors. Pcse: Python crop simulation environment, 2024. Accessed: 2024-12-27.

[Eriksson *et al.*, 2019] David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Flaticon, 2025] Contributors Flaticon. Icons by flaticon. Accessed on January 4, 2025., 2025. Available at: https://www.flaticon.com/.

[Garnett, 2023] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

[Guo *et al.*, 2024] Pei-Fu Guo, Ying-Hsuan Chen, Yun-Da Tsai, and Shou-De Lin. Towards optimizing with large language models. In *Fourth Workshop on Knowledge-infused Learning*, 2024.

[Hao *et al.*, 2024] Hao Hao, Xiaoqun Zhang, and Aimin Zhou. Large language models as surrogate models in evolutionary algorithms: A preliminary study. *arXiv preprint arXiv:2406.10675*, 2024.

[Huang *et al.*, 2022] Daolang Huang, Louis Filstroff, Petrus Mikkola, Runkai Zheng, and Samuel Kaski. Bayesian optimization augmented with actively elicited expert knowledge. *arXiv:2208.08742*, 2022.

[Huang *et al.*, 2024] Beichen Huang, Xingyu Wu, Yu Zhou, Jibin Wu, Liang Feng, Ran Cheng, and Kay Chen Tan. Exploring the true potential: Evaluating the black-box optimization capability of large language models. *arXiv:2404.06290*, 2024.

[Hvarfner *et al.*, 2022] Carl Hvarfner, Danny Stoll, Artur Souza, Luigi Nardi, Marius Lindauer, and Frank Hutter. πBO: Augmenting acquisition functions with user beliefs for Bayesian optimization. In *10th International Conference on Learning Representations, ICLR'22*, pages 1–30, April 2022.

[Hvarfner *et al.*, 2024] Carl Hvarfner, Frank Hutter, and Luigi Nardi. A general framework for user-guided Bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[Häse *et al.*, 2021] Florian Häse, Matteo Aldeghi, Riley J. Hickman, Loïc M. Roch, and Alán Aspuru-Guzik. Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge. *Applied Physics Reviews*, 8(3):031406, 07 2021.

[Kristiadi *et al.*, 2024] Agustinus Kristiadi, Felix Strieth-Kalthoff, Sriram Ganapathi Subramanian, Vincent Fortuin, Pascal Poupart, and Geoff Pleiss. How useful is intermittent, asynchronous expert feedback for Bayesian optimization? In *Sixth Symposium on Advances in Approximate Bayesian Inference - Non Archival Track*, 2024.

[Kroeger *et al.*, 2024] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. In-context explainers: Harnessing LLMs for explaining black box models. *arXiv:2310.05797*, 2024.

[Lee *et al.*, 2020] Eric Hans Lee, Valerio Perrone, Cédric Archambeau, and Matthias Seeger. Cost-aware Bayesian optimization. In *ICML 2020 Workshop on AutoML*, 2020.

[Li *et al.*, 2020] Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Antonio Robles-Kelly, and Svetha Venkatesh.

Incorporating expert prior knowledge into experimental design via posterior sampling. arxiv:2002.11256, 2020.

[Liu *et al.*, 2022] Zijing Liu, Xiyao Qu, Xuejun Liu, and Hongqiang Lyu. Robust Bayesian optimization with reinforcement learned acquisition functions. *arXiv:2210.00476*, 2022.

[Liu *et al.*, 2024] Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance Bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[M. Bran *et al.*, 2024] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

[Mahammadli, 2024] Kanan Mahammadli. Sequential large language model-based hyper-parameter optimization. *arXiv:2410.20302*, 2024.

[Nguyen and Grover, 2024] Tung Nguyen and Aditya Grover. Lico: Large language models for in-context molecular optimization. *arXiv:2406.18851*, 2024.

[Nogueira, 2014] Fernando Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python. https://github.com/bayesian-optimization/BayesianOptimization, 2014.

[OpenAI, 2025] OpenAI. Gpt-4o-mini model card. https://platform.openai.com/docs/models/gpt-4o-mini, 2025.

[Ou *et al.*, 2022] Changkun Ou, Daniel Buschek, Sven Mayer, and Andreas Butz. The human in the infinite loop: A case study on revealing and explaining human-AI interaction loop failures. In *Mensch und Computer 2022*, MuC '22, page 158–168. ACM, September 2022.

[Poloczek *et al.*, 2016] Matthias Poloczek, Jialei Wang, and Peter I Frazier. Warm starting Bayesian optimization. In *2016 Winter simulation conference (WSC)*, pages 770–781. IEEE, 2016.

[Ramachandran *et al.*, 2020] Anil Ramachandran, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Incorporating expert prior in Bayesian optimisation via space warping. *Knowledge-Based Systems*, 195:105663, 2020.

[Ramos *et al.*, 2024] Mayk Caldas Ramos, Christopher Collison, and Andrew D White. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 2024.

[Ranković and Schwaller, 2023] Bojana Ranković and Philippe Schwaller. Bochemian: Large language model embeddings for Bayesian optimization of chemical reactions. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023.

[Savage and del Rio Chanona, 2023] Tom Savage and Ehecatl Antonio del Rio Chanona. Expert-guided Bayesian optimisation for human-in-the-loop experimental design of known systems. *arXiv:2312.02852*, 2023.

[Shoyeb Raihan *et al.*, 2024] Ahmed Shoyeb Raihan, Hamed Khosravi, Srinjoy Das, and Imtiaz Ahmed. Accelerating material discovery with a threshold-driven hybrid acquisition policy-based Bayesian optimization. *Manufacturing Letters*, 41:1300–1311, 2024. 52nd SME North American Manufacturing Research Conference (NAMRC 52).

[Siemenn *et al.*, 2023] Alexander E. Siemenn, Zekun Ren, Qianxiao Li, and Tonio Buonassisi. Fast Bayesian optimization of needle-in-a-haystack problems using zooming memory-based initialization (zombi). *npj Computational Materials*, 9(1), May 2023.

[van Stein *et al.*, 2024] Niki van Stein, Diederick Vermetten, and Thomas Bäck. In-the-loop hyper-parameter optimization for LLM-based automated design of heuristics. *arXiv:410.16309*, 2024.

[Vermorel and Mohri, 2005] Joannès Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of the 16th European Conference on Machine Learning*, ECML'05, page 437–448, Berlin, Heidelberg, 2005. Springer-Verlag.

[Volpp *et al.*, 2020] Michael Volpp, Lukas P. Fröhlich, Kirsten Fischer, Andreas Doerr, Stefan Falkner, Frank Hutter, and Christian Daniel. Meta-learning acquisition functions for transfer learning in Bayesian optimization. In *International Conference on Learning Representations*, 2020.

[Xie *et al.*, 2022] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

[Xie *et al.*, 2023] Zikai Xie, Xenophon Evangelopoulos, Joseph C. R. Thacker, and Andrew I. Cooper. Domain knowledge injection in Bayesian search for new materials. In *26th European Conference on Artificial Intelligence*. IOS Press, September 2023.

[Yang *et al.*, 2024] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024.

[Yin *et al.*, 2024] Yuxuan Yin, Yu Wang, Boxun Xu, and Peng Li. Ado-llm: Analog design Bayesian optimization with in-context learning of large language models. *arXiv:2406.18770*, 2024.

[Zhang *et al.*, 2024] Jun Zhang, Jinpeng Cheng, Cuiping Liu, Qiang Wu, Shuping Xiong, Hao Yang, Shenglong Chang, Yuanyuan Fu, Mohan Yang, Shiyu Zhang, et al. Enhanced crop leaf area index estimation via random forest regression: Bayesian optimization and feature selection approach. *Remote Sensing*, 16(21):3917, 2024.