ROBUST HUMAN-AI COLLABORATIVE BAYESIAN OPTIMIZATION UNDER SPATIALLY VARYING EXPERT RELIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Human knowledge can accelerate Bayesian optimization (BO), but such knowledge is typically qualitative, noisy, biased, and, more importantly, spatially varying in reliability. Although this spatial heterogeneity better reflects practical expertise, it is largely overlooked by existing human-augmented BO methods, which control human influence globally or according to a prespecified schedule. To address this limitation, we propose a novel trust-aware human–AI collaborative BO framework that adaptively modulates the contribution of human knowledge across the search space. Specifically, input-dependent trust is estimated from local human–objective ranking agreement using localized conformal calibration and then incorporated into the multi-fidelity kernel. Furthermore, we establish two theoretical guarantees: a calibration guarantee that controls the long-run frequency of assigning high trust to human guidance inconsistent with objective observations, and a sublinear cumulative regret bound for the resulting trust-aware BO procedure. Experiments with spatially varying emulated experts, as well as two real-world experiments, consistently demonstrate that the proposed method outperforms standard and human-augmented BO baselines under spatially varying expert reliability while remaining robust to misleading guidance.

1 INTRODUCTION

Bayesian optimization (BO) is an automated paradigm for optimizing expensive black-box functions and has been widely applied to scientific experimentation, including materials discovery (Kusne et al., 2020), chemical design (Griffiths & Hernández-Lobato, 2020), and experimental process optimization (Shields et al., 2021). However, standard BO only relies on objective observations, which implies that only limited observations are available during the early stage of the optimization. In many scientific fields, human experts possess domain knowledge that provides useful guidance for the optimization before sufficient observations have been collected. Incorporating such knowledge into BO offers huge potential to accelerate the optimization process (AV et al., 2022; Adachi et al., 2024; Xu et al., 2024).

Integrating human knowledge, however, remains challenging in various aspects. Expert knowledge is typically expressed through qualitative and indirect assessments rather than objective values (Wong & Lingras, 1994; Chu & Ghahramani, 2005), and the resulting feedback can be noisy and biased (Lun Chau et al., 2022). More importantly, its reliability can be input-dependent, meaning that it varies with the candidate configuration being assessed. For example, in materials formulation, an expert may provide accurate assessments within familiar composition regimes but unreliable guidance in regions that differ substantially from prior experience.

Existing human–AI collaborative BO methods incorporate expert information by either modifying the query-selection process or augmenting the objective surrogate (AV et al., 2022; Huang et al., 2023; Adachi et al., 2024; AV et al., 2024). Some methods further establish convergence, regret, or robustness guarantees for incorporating potentially inaccurate human knowledge (Hvarfner et al., 2021; Adachi et al., 2024; Xu et al., 2024). However, in the prior studies, human influence is generally regulated at a global or iteration level through fixed weights, predefined decay schedules, or model-level adjustments. These mechanisms fail to explicitly estimate whether human guidance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

agrees with objective observations at a particular input and therefore cannot directly adapt the level of trust to spatially varying human reliability of expert knowledge.

In this work, we regard human knowledge as an auxiliary low-fidelity information source. However, human feedback does not directly provide objective values; instead, it expresses the expert's preferences or relative assessments of queried configurations. Therefore, we infer a latent utility function from such feedback, which serves as a potentially biased and noisy proxy for the objective and induces an expert-preferred ordering over the queried configurations. Meanwhile, objective observations constitute the high-fidelity information source and induce a corresponding ordering based on observed outcomes. Consequently, the agreement between these two orderings provides empirical evidence of how reliably human knowledge reflects the optimization objective.

Based on this perspective, we propose a trust-aware human-AI collaborative BO framework to systematically address the spatially varying human reliability for the first time. We develop a trust-aware surrogate model that integrates the human-induced latent utility with objective observations through a human-dependent covariance component. We then quantify input-dependent human reliability from the local agreement between their induced orderings using localized conformal calibration. The trust weights gate the contribution of human information to the surrogate, allowing the surrogate to exploit reliable human knowledge locally while mitigating the influence of misleading feedback.

Our contributions. (1) We propose a first-of-its-kind trust-aware BO framework that incorporates human feedback as low-fidelity information and adaptively regulates its influence through input-dependent trust under spatially varying expert reliability. (2) We establish new theoretical results on a calibration guarantee that controls the long-run frequency of assigning high trust to human guidance inconsistent with objective observations, together with a sublinear cumulative regret bound. (3) Experiments on synthetic benchmarks and two scientific real-world datasets demonstrate consistent improved performance and robustness under the more realistic setting of spatially varying expert reliability.

2 Background and related work

2.1 Standard and multi-fidelity Bayesian optimization

Standard BO typically employs Gaussian process (GP) as the surrogate model of the objective function. Given the observations collected up to iteration t, the objective follows a Gaussian posterior $f(\boldsymbol{x} \mid \mathcal{D}_t) \sim \mathcal{N}(\mu_t(\boldsymbol{x}), \sigma_t^2(\boldsymbol{x}))$, and the predictive mean function $\mu_t(\boldsymbol{x})$ and standard deviation $\sigma_t(\boldsymbol{x})$ are given by (Shahriari et al., 2015; Seeger, 2004)

$$\mu_t(\boldsymbol{x}) = \mathbf{k}(\boldsymbol{x})^{\mathrm{T}} \left(\mathbf{K} + \epsilon^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}$$
$$\sigma_t^2(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}(\boldsymbol{x})^T \left(\mathbf{K} + \epsilon^2 \boldsymbol{I}\right)^{-1} \mathbf{k}(\boldsymbol{x})$$

where $\mathbf{k}(\boldsymbol{x}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), k(\boldsymbol{x}, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_{t-1})]^{\mathrm{T}}$, $\mathbf{K}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and ϵ^2 is the observation noise variance.

At each iteration, BO selects the next query point by maximizing an acquisition function, which uses the posterior mean and standard deviation to balance exploitation and exploration. A commonly used acquisition function is upper confidence bound (UCB) (Srinivas et al., 2009): $\alpha_{\mathrm{UCB}}(\boldsymbol{x}) = \mu_t(\boldsymbol{x}) + \beta_t^{1/2}\sigma_t(\boldsymbol{x})$, where β_t controls the exploration–exploitation trade-off.

Multi-fidelity Bayesian optimization (MFBO) extends standard BO by exploiting auxiliary information sources with different costs and accuracies (Kennedy & O'Hagan, 2000). A central challenge in MFBO is modeling the dependence between fidelity levels so that lower-fidelity information can improve prediction of the high-fidelity objective. Classical autoregressive models represent the high-fidelity function as a scaled version of a lower-fidelity function plus an independent discrepancy term (Kennedy & O'Hagan, 2000; Kandasamy et al., 2016), forming the basis of widely used Co-kriging models. Nonlinear autoregressive Gaussian processes further generalize this relationship by modeling the high-fidelity function as a nonlinear mapping of both the original input and the lower-fidelity output, $f_H(\boldsymbol{x}) = g(\boldsymbol{x}, f_L(\boldsymbol{x}))$, allowing more flexible cross-fidelity dependence (Perdikaris et al., 2017).

These approaches highlight the importance of modeling dependence between information sources when incorporating auxiliary information. Human knowledge, however, differs from conventional low-fidelity observations in that it is often qualitative and may not directly approximate the objective value. Moreover, its reliability can vary across the search space. Therefore, rather than assuming a globally valid cross-fidelity relationship, human-augmented BO requires a spatially adaptive mechanism that modulates the contribution of human information according to its local reliability.

2.2 Human knowledge-augmented Bayesian optimization

Human knowledge-augmented Bayesian optimization incorporates expert feedback as an auxiliary source of information to guide the optimization process. In addition to collecting objective observations, these methods solicit assessments from human experts of points or regions in the search space either before or during optimization process. The resulting feedback provides qualitative information about which solutions are likely to be promising and is used together with the objective observations to inform subsequent query selection. By exploiting such prior knowledge, human-augmented BO aims to identify high-performing regions more rapidly.

Human feedback forms and latent utility modeling. It is easier for humans to provide relative judgements than precise assessments of absolute magnitude (Kahneman & Tversky, 2013). Accordingly, preference-based feedback has been widely used to incorporate human knowledge into BO (González et al., 2017; Mikkola et al., 2020). In the pairwise setting, the expert is presented with two alternatives $(\boldsymbol{x}^{(a)}, \boldsymbol{x}^{(b)})$ and indicates which one is preferred.

Preference learning models these feedback through a latent utility function $\pi : \mathcal{X} \to \mathbb{R}$, where larger utility values correspond to stronger human preferences. For instance, an observed preference relation $\boldsymbol{x}^{(a)} \succ \boldsymbol{x}^{(b)}$ indicates that $\pi(\boldsymbol{x}^{(a)})$ is likely to exceed $\pi(\boldsymbol{x}^{(b)})$. By combining a probabilistic surrogate with an appropriate preference likelihood, preference learning infers a posterior distribution over π, thereby representing both the preference structure implied by qualitative feedback and its associated uncertainty (Chu & Ghahramani, 2005).

Incorporating human knowledge into BO. Existing methods incorporate human knowledge at either the acquisition or surrogate-model level. Acquisition-based approaches guide query selection by reweighting the acquisition function using user-specified beliefs (Hvarfner et al., 2021) or learned preference models (Adachi et al., 2024). Surrogate-based approaches use expert feedback to constrain GP model selection (AV et al., 2022), jointly learn the objective and human utility through shared representations (Huang et al., 2023), or augment the objective surrogate with latent representations of expert-assessed auxiliary properties (AV et al., 2024). Several methods further establish convergence or regret guarantees, including no-harm guarantees that ensure that incorporating inaccurate human knowledge does not asymptotically degrade optimization performance relative to standard BO (Hvarfner et al., 2021; AV et al., 2022; Adachi et al., 2024; Xu et al., 2024).

Trust in human knowledge. Existing methods adopt different mechanisms to limit the influence of potentially inaccurate human knowledge. In πBO, the user-specified belief decreases globally according to a predefined schedule (Hvarfner et al., 2021). CoExBO additionally accounts for uncertainty in the learned preference model, while retaining an iteration-dependent decay that causes the preference-augmented acquisition function to asymptotically recover standard UCB (Adachi et al., 2024). Other approaches control human influence through global model selection. For example, enhanced BO selects between an expert-augmented surrogate and an objective-only surrogate according to their predictive likelihoods (AV et al., 2024). More recently, Xu et al. (2024) introduced a global trust parameter for binary expert advice. These control mechanisms support robustness guarantees in several methods, including no-harm guarantees under which incorporating inaccurate human knowledge does not asymptotically worsen the convergence rate relative to standard BO.

However, existing trust mechanisms operate primarily at the level of the overall acquisition rule, surrogate model, or optimization iteration. While some of these mechanisms are data-dependent or input-dependent, they do not explicitly estimate how the reliability of human knowledge varies across the search space. In particular, input-dependent uncertainty in a learned preference model reflects confidence in the inferred human preference itself rather than the reliability of that preference with respect to the optimization objective. Consequently, existing methods lack an explicit mechanism for representing spatially varying human reliability, limiting their ability to adapt the influence of human knowledge according to where it is informative or misleading.

3 PROBLEM FORMULATION

We consider the black-box optimization problem with the collaboration of a human expert, where the goal is to maximize an unknown function f over a search space $\mathcal{X} \subset \mathbb{R}^d$:

$$\boldsymbol{x}^\star = \arg\max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \tag{1}$$

We assume that f is expensive to evaluate and can only observe a noisy estimate $y = f(\boldsymbol{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \epsilon^2)$ is zero-mean Gaussian noise. At iteration t, the high-fidelity objective observation is collected in $\mathcal{D}_t^{\mathrm{HF}} = \{(\boldsymbol{x}_i, y_i^{\mathrm{HF}})\}_{i=1}^t$.

Human knowledge as auxiliary information. Besides objective evaluations, we consider pairwise comparisons provided by an expert as auxiliary information. Such feedback typically exhibits region-dependent bias and therefore need not be uniformly reliable across the search space. The expert is presented with a pair of query points $(\boldsymbol{x}_i^{(a)}, \boldsymbol{x}_i^{(b)})$ in the i-th interaction, and indicate which one they believe is more promising with respect to the objective function. We denote the comparison label by $y_i^{\mathrm{P}} \in \{0, 1\}$ where $y_i^{\mathrm{P}} = 1$ if the expert prefers $\boldsymbol{x}_i^{(a)}$ and otherwise $y_i^{\mathrm{P}} = 0$. The pairwise feedback dataset collected by iteration t is denoted by $\mathcal{D}_t^{\mathrm{P}} = \{(\boldsymbol{x}_i^{(a)}, \boldsymbol{x}_i^{(b)}, y_i^{\mathrm{P}})\}_{i=1}^{n_t}$, where n_t is the number of comparisons collected up to iteration t.

Notation. The identity matrix is denoted by $\boldsymbol{I}$, and $\mathbf{1}$ denotes the indicator function. For a matrix $\boldsymbol{M}$, we use M_{ij} for the element in its i-th row and j-th column.

4 METHODOLOGY

In this section, we present our trust-aware human-AI collaborative Bayesian Optimization framework for integrating human knowledge with spatially varying reliability.

We regard human expert knowledge as an auxiliary low-fidelity information source that complements expensive objective observations. This requires addressing two essential challenges. First, human knowledge and objective observations have heterogeneous representations, and thus need to be incorporated into a unified probabilistic model. Second, the reliability of human knowledge typically varies across the design space, requiring a spatially adaptive mechanism that exploits informative knowledge while mitigating the impact of unreliable information.

To address these challenges, we develop a novel trust-aware modeling approach to establish the probabilistic relationship between human expert knowledge and objective observations, together with an input-dependent trust gating mechanism to characterize the spatially varying reliability of human expert. As illustrated in Figure 1, the ordering induced by the latent utility is compared with that supported by objective observations to construct a local trust function $\omega_t(\boldsymbol{x})$, which adaptively gates expert information and is subsequently used to select the next query point.

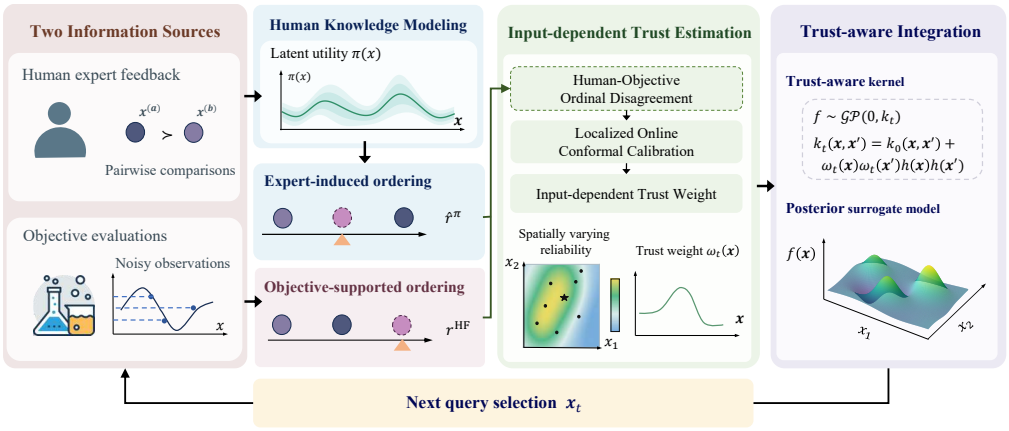


Figure 1: Overview of the proposed trust-aware human-AI collaborative BO framework

4.1 PROBABILISTIC MODELING OF HUMAN KNOWLEDGE AND OBJECTIVE OBSERVATIONS

Learning latent utility function from human feedback. An initial set of human feedback is collected before optimization and can be augmented as additional feedback becomes available. Let $\mathcal{D}_t^{\mathrm{P}}$ denote the human-feedback data available at iteration t. We introduce a latent utility function $\pi(\boldsymbol{x})$ to represent the expert's preference over the search space, where larger values indicate more preferred configurations. We use Gaussian process to model π and infer its posterior using a likelihood determined by the feedback form.

For pairwise-comparison feedback, various preference-learning models can be used to infer latent utilities from pairwise comparisons. In this work, we adopt the Bradley-Terry model (Bradley & Terry, 1952):

$$P(y_i^{\mathrm{P}} = 1 | \boldsymbol{x}_i^{(a)}, \boldsymbol{x}_i^{(b)}) = \sigma\left(\pi(\boldsymbol{x}_i^{(a)}) - \pi(\boldsymbol{x}_i^{(b)})\right) \tag{2}$$

where $\sigma(\cdot)$ is the logistic sigmoid function.

Approximate GP inference yields

$$\pi(\boldsymbol{x}) \mid \mathcal{D}_t^{\mathrm{P}} \approx \mathcal{N}\left(\mu_{\pi,t}(\boldsymbol{x}), \sigma_{\pi,t}^2(\boldsymbol{x})\right) \tag{3}$$

The posterior is updated whenever additional human feedback becomes available and remains unchanged at iterations where no new feedback is collected.

It is worth noting that, although we focus on pairwise comparisons, the proposed framework is not restricted to this feedback format. It is general enough to accommodate other forms of expert feedback by replacing the observation likelihood used to learn the latent utility model, while leaving the subsequent optimization procedure unchanged. As an example, we provide the formulation and experiments for binary-label feedback in Appendix B.5.

Trust-aware integration of human knowledge and objective observations. The inferred latent utility function $\pi(\boldsymbol{x})$ provides a low-fidelity representation of human knowledge. At iteration t, we transform the posterior mean of the latent utility into a bounded human-information feature $h_{\pi,t}(\boldsymbol{x}) = \sigma(\mu_{\pi,t}(\boldsymbol{x}))$, in which $\sigma(\cdot)$ is the logistic sigmoid function. The sigmoid transformation bounds the feature while preserving the ordering induced by the latent utility.

Becuase this utility is inferred from qualitative feedback and need not maintain a globally valid relationship with the objective, we therefore incorporate it through a new trust-gated covariance component.

$$k_t(\boldsymbol{x}_t, \boldsymbol{x}_t') = k_0(\boldsymbol{x}, \boldsymbol{x}') + \omega_t(\boldsymbol{x})\omega_t(\boldsymbol{x}')h_{\pi,t}(\boldsymbol{x})h_{\pi,t}(\boldsymbol{x}') \tag{4}$$

The first term, $k_0(\boldsymbol{x}, \boldsymbol{x}')$, is an objective-only kernel that models correlations supported solely by high-fidelity observations. The second term incorporates human information according to the local trust scores. As a result, human knowledge has a stronger effect in regions where it is assessed to be reliable, whereas its contribution is attenuated in regions of low trust. In particular, when the trust assigned to human knowledge approaches zero, the human-dependent component vanishes locally and the surrogate reduces toward the objective-only model. The construction of the input-dependent trust score $\omega_t(\boldsymbol{x})$ is presented in the following subsection.

4.2 INPUT-DEPENDENT TRUST ESTIMATION AND GATING

We estimate the local reliability of human knowledge through the ordinal agreement between the human-induced preference and the ordering supported by high-fidelity objective observations. To account for the uncertainty in such agreement, we employ conformal prediction (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023), which provides model-agnostic uncertainty calibration with provable coverage guarantees. We further adopt an online localized variant of conformal prediction, enabling the calibration to adapt sequentially to newly observed data and locally to different regions of the search space while providing long-run coverage control. The resulting uncertainty is then used to determine the local trust weight.

Human–objective ordinal disagreement. Using the posterior mean of the latent utility function, we define the expert-induced rank of a candidate $\boldsymbol{x}$ relative to previous observations as

$$\hat{r}_t^{\pi}(\boldsymbol{x}) = \sum_{j=1}^{t-1} \mathbf{1}\{\mu_{\pi,t}(\boldsymbol{x}) > \mu_{\pi,t}(\boldsymbol{x}_j)\} \tag{5}$$

Algorithm 1 Trust-aware human-AI collaborative BO

Input: Search space $\mathcal{X}$, initial high-fidelity dataset $\mathcal{D}_0^{\mathrm{HF}}$, human feedback dataset $\mathcal{D}^{\mathrm{LF}}$, miscoverage grid $\mathcal{A} = \{\alpha_1, \ldots, \alpha_K\}$, number of iterations T
Output: Best observed solution $\boldsymbol{x}^\star$
Learn the latent human utility function π from $\mathcal{D}^{\mathrm{LF}}$ and obtain its posterior mean $\mu_\pi(\boldsymbol{x})$
Initialize the localized online conformal calibrators c, ζ
for $t = 1, \ldots, T$ **do**
  Construct the conformal rank prediction sets $\Gamma_t^\alpha(\boldsymbol{x})$ for all α in $\mathcal{A}$
  Compute the input-dependent trust weight $\omega_t(\boldsymbol{x})$ by Eq. (10) and construct the trust-aware kernel k_t using $\omega_t(\boldsymbol{x})$
  Update the cross-fidelity GP posterior with kernel k_t
  Select the next query point $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \left[\mu_{t-1}(\boldsymbol{x}, \mu_{\pi,t}(\boldsymbol{x})) + \sqrt{\beta_t}\,\sigma_{t-1}(\boldsymbol{x}, \mu_{\pi,t}(\boldsymbol{x}))\right]$.
  Evaluate the high-fidelity objective and observe y_t^{HF}
  Update $\mathcal{D}_t^{\mathrm{HF}} = \mathcal{D}_{t-1}^{\mathrm{HF}} \cup \{(\boldsymbol{x}_t, y_t^{\mathrm{HF}})\}$
  Compute human-induced rankings $\hat{r}_t^\pi(\boldsymbol{x}_t)$ and objective-supported rankings r_i^{HF} on $\mathcal{D}_t^{\mathrm{HF}}$
  Update each localized online conformal calibrator $c_{t+1}^\alpha, \zeta_{t+1}^\alpha$ for all α in $\mathcal{A}$
end for
return $\boldsymbol{x}^\star \leftarrow \arg\max_{(\boldsymbol{x}_i, y_i^{\mathrm{HF}})\in\mathcal{D}_T^{\mathrm{HF}}} y_i^{\mathrm{HF}}$

For a candidate objective rank $r \in \{0, \ldots, t-1\}$, we define the bounded ordinal-disagreement score

$$s_t(\boldsymbol{x}, r) = \frac{|r - \hat{r}_t^\pi(\boldsymbol{x})|}{\max\{1, t-1\}} \in [0, 1]. \tag{6}$$

A smaller score indicates stronger agreement between the expert-induced ordering and the candidate objective ordering.

After evaluating x_t and observing y_t^{HF}, its realized objective-supported rank is

$$r_t^{\mathrm{HF}} = \sum_{j=1}^{t-1} \mathbf{1}\left\{y_t^{\mathrm{HF}} > y_j^{\mathrm{HF}}\right\}. \tag{7}$$

The disagreement score at round t is consequently $s_t(\boldsymbol{x}_t, r_t^{\mathrm{HF}})$. This score is used once to update the online calibration procedure and is not recomputed retrospectively at later iterations.

Localized online conformal trust certification. For any target miscoverage level $\alpha \in (0, 1)$, localized online conformal calibration maintains an input-dependent threshold $q_t^\alpha(\cdot) = c_t^\alpha + \zeta_t^\alpha(\cdot)$.

At iteration t, before observing the high-fidelity outcome, the threshold function maps the input $\boldsymbol{x}$ to define the prediction set

$$\Gamma_t^\alpha(\boldsymbol{x} \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}}) = \{r \in \{0, \ldots, t-1\} : s_t(x, r) \le q_t^\alpha(x)\}. \tag{8}$$

Following Zecchin & Simeone (2024), the calibration parameters are updated as

$$\begin{cases} c_{t+1}^\alpha & = c_t^\alpha + \eta_t \left(\mathbf{1}\{r_t^{\mathrm{HF}} \notin \Gamma_t^\alpha(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}})\} - \alpha\right) \\ \zeta_{t+1}^\alpha(\cdot) & = (1 - \lambda\eta_t)\zeta_t^\alpha(\cdot) + \eta_t \left(\mathbf{1}\{r_t^{\mathrm{HF}} \notin \Gamma_t^\alpha(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}})\} - \alpha\right) \cdot k_\zeta(\boldsymbol{x}_t, \cdot) \end{cases} \tag{9}$$

The resulting grid-valued trust certificate is

$$\alpha^\star(\boldsymbol{x}) = \inf_\alpha \{\alpha : \Gamma_t^\beta(\boldsymbol{x} \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}}) = \{\hat{r}^\pi(\boldsymbol{x})\}, \forall \beta > \alpha\} \tag{10}$$

$$\omega_t(\boldsymbol{x}) = 1 - \alpha^\star(x) \tag{11}$$

In practice, we use a fixed grid $\mathcal{A} = \{\alpha_1, \alpha_2, \cdots, \alpha_K\} \subset (0, 1)$ and maintain K localized online calibration processes in parallel.

The complete optimization procedure is summarized in Algorithm 1, which integrates the multi-fidelity surrogate with the proposed input-dependent trust estimation and gating mechanism.

4.3 THEORETICAL ANALYSIS

We theoretically establish new results on a calibration guarantee for the proposed trust gating mechanism and a cumulative regret bound for the resulting BO algorithm.

Assumption 1 (Bounded input domain)**.** $\forall \boldsymbol{x} \in \mathcal{X}$, $\|\boldsymbol{x}\| \leq D < \infty$

Assumption 2 (Stationary and bounded localization kernel)**.** *The localization kernel function is stationary, i.e.,* $k(\boldsymbol{x}, \boldsymbol{x}') = \tilde{k}(\|\boldsymbol{x} - \boldsymbol{x}'\|)$*, where* $\tilde{k}$ *is non-negative,* ρ*-Lipschitz for* $\rho > 0$*, bounded by* $k(\boldsymbol{x}, \boldsymbol{x}') \leq \kappa < \infty$ *and* $\lim_{z \to \infty} \tilde{k}(z) = 0$.

Theorem 4.1. *For any* $\tau \in (0, 1)$*, any regularization parameter* $\lambda > 0$ *and learning rate* $\eta_t = \eta_1 t^{-1/2} < 1/\lambda$*,*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{\omega_t(\boldsymbol{x}_t) \geq \tau, r_t^{\mathrm{HF}} \neq \hat{r}_t^{\pi}\} \leq 1 - \tau + \frac{C}{\sqrt{T}}$$

with $C = \dfrac{1}{\eta_1} + \dfrac{2\kappa + 4\sqrt{\rho \kappa D}}{\eta_1 \lambda} + 2(2\kappa + 1)$.

See Appendix A.1 for the proof. Theorem 4.1 gives a calibration guarantee for the proposed input-dependent trust score. In particular, it controls the long-run frequency of high-confidence errors, namely, cases in which the expert is assigned trust above a prescribed threshold while its induced ranking disagrees with the high-fidelity observation. As T increases, this frequency is bounded by $(1 - \tau)$ up to a vanishing finite-sample term of order $\mathcal{O}(T^{-1/2})$. Thus, the trust score admits an explicit reliability interpretation rather than serving merely as a heuristic weighting factor: a higher trust threshold corresponds to a tighter control on the frequency of objective-inconsistent expert assessments receiving high trust. Importantly, the result does not require the expert to have a uniform global reliability level, and therefore directly accommodates spatially varying expert reliability. This provides theoretical foundation for using the calibrated trust score to regulate the local influence of human knowledge in the surrogate model.

Theorem 4.2. *Suppose that the objective function* f *belongs to the RKHS* $\mathcal{H}_{k_0}$ *associated with the objective-only kernel* k_0*, with* $\|f\|_{\mathcal{H}_{k_0}} \leq B_f$*. Let Algorithm 1 employ the time-varying trust-aware kernel defined in Eq. (4).*

Then, with probability at least $1 - \delta$*, the cumulative regret after* T *iterations satisfies*

$$R_T = \mathcal{O}\left(\sqrt{T \gamma_T \log(T/\delta)}\right) + \mathcal{O}\left(\sqrt{T \log\left(1 + (T/\sigma_\epsilon^2)\right) \log(T/\delta)}\right).$$

where γ_T *denotes the maximum information gain associated with the objective-only kernel* k_0.

See Appendix A.2 for the proof of Theorem 4.2. Theorem 4.2 establishes that the proposed algorithm retains the no-regret property of GP-UCB despite incorporating potentially biased human knowledge. In particular, the average regret incurred at each iteration converges to zero, ensuring that the proposed algorithm eventually concentrates its evaluations around the global optimum rather than being persistently misled by the human-information source.

Remark. Standard GP-UCB with the objective-only kernel k_0 admits a cumulative regret bound of order $\mathcal{O}(\sqrt{T \gamma_T \log T})$. By comparison, the proposed algorithm incurs only an additional $\mathcal{O}(\sqrt{T})$ term. This additional term does not change the leading-order asymptotic regret rate. Therefore, even when the human feedback is misleading, the proposed algorithm retains the same asymptotic regret order as standard GP-UCB, establishing a rate-level no-harm guarantee.

Trust mechanism analysis. Beyond the above guarantees, we further illustrate how the proposed trust-aware kernel exploits informative human knowledge. Let $\phi_t(\boldsymbol{x}) = \omega_t(\boldsymbol{x}) h_{\pi,t}(\boldsymbol{x})$, so that the human-dependent kernel component can be written as $\phi_t(\boldsymbol{x})\phi_t(\boldsymbol{x}')$. This term introduces a human-induced feature direction whose local strength is controlled by the trust score $\omega_t(\boldsymbol{x})$, while its effective orientation is determined by the observed high-fidelity data rather than being imposed a priori. In particular, the learned contribution of this feature is governed by its alignment with the objective observations: informative human knowledge is reinforced when it is consistent with the observed objective trend, whereas inconsistent knowledge can be attenuated or assigned an opposite contribution. Meanwhile, as $\omega_t(\boldsymbol{x}) \to 0$, the human-induced component vanishes locally and the surrogate falls back to the objective-only kernel k_0. Therefore, the proposed kernel enables locally adaptive

exploitation of human knowledge while allowing high-fidelity observations to determine whether such knowledge should be positively utilized, suppressed, or reversed.

5 EXPERIMENTS

We evaluate the proposed trust-aware human-AI collaborative Bayesian optimization framework through both synthetic and real-world experiments.

Experimental Setup. In each experiment, the latent utility model is learned from the expert-feedback dataset before the optimization. We compare the proposed method with standard GP-UCB, πBO (Hvarfner et al., 2021) and CoExBO (Adachi et al., 2024). GP-UCB serves as the feedback-free baseline, while BOAP and CoExBO represent existing human-augmented BO methods that incorporate pairwise expert preferences.

Optimization performance is evaluated using simple regret $r_T = f(\boldsymbol{x}^\star) - \max_{1\le t\le T} f(\boldsymbol{x}_t)$ and cumulative regret $R_T = \sum_{t=1}^{T} (f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t))$. Each experiment is repeated over ten independent runs. Within each run, all methods are initialized with the same three high-fidelity observations, and the results are reported as the mean $\pm$ standard error.

5.1 SYNTHETIC EXPERIMENTS

We consider four commonly used multimodal benchmark functions: Ackley, Hölder Table, Styblinski–Tang and Levy (details provided in Appendix B.1).

Expert emulation. For each test function f, we construct a simulated human latent utility function $f_{\text{human}}(\boldsymbol{x}) = a(\boldsymbol{x})\bar{f}(\boldsymbol{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, 0.01^2)$, , where $\bar{f}(\boldsymbol{x})$ denotes the normalized objective value and $a(\boldsymbol{x})$ is an input-dependent multiplicative bias coefficient. The function f_{human} represents the underlying utility used to emulate the human expert.

To emulate spatially varying expert reliability, $a(\boldsymbol{x})$ and $\delta(\boldsymbol{x})$ are constructed such that the ordering induced by f_{human} is informative in some regions of the search space but unreliable or misleading in others. This setting directly reflects the spatially heterogeneous reliability considered in this work.

Based on the same simulated expert utility f_{human}, we generate pairwise-comparison feedback. Given a query pair $(\boldsymbol{x}, \boldsymbol{x}')$, the expert preference is generated according to the relative human utility: $P(\boldsymbol{x} \succ \boldsymbol{x}' \mid \boldsymbol{x}, \boldsymbol{x}') = \sigma\left(f_{\text{human}}(\boldsymbol{x}) - f_{\text{human}}(\boldsymbol{x}')\right)$. The simulated expert utility is used only to generate human feedback and is not observed by the optimization algorithm.

We evaluate all methods under spatially varying expert reliability on Ackley, Hölder Table, Styblinski–Tang, and Levy to assess whether the proposed trust mechanism remains effective across different objective landscapes. Results for the globally informative and globally misleading experts on these additional benchmarks are provided in Appendix B.3.

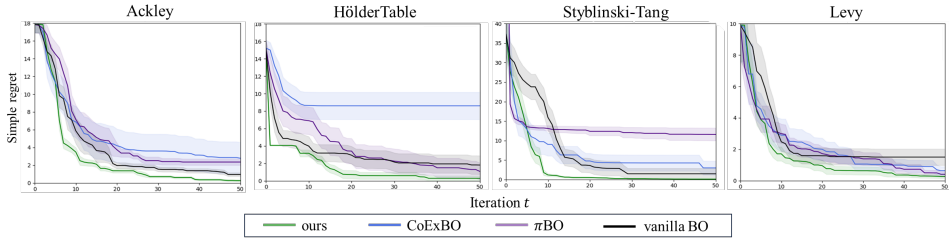


Figure 2: Optimization performance across test functions under spatially varying emulated expert.

Figure 2 further compares the methods on four test functions under spatially-varying expert reliability. Our method achieves lower final regret than all baselines on all three functions. In terms of sample efficiency, our method requires only 16, 8, and 15 evaluations to reduce the initial simple regret by 90% on Hölder Table, Styblinski–Tang, and Levy, respectively. In comparison, competing methods require substantially more evaluations or fail to reach this target within the evaluation budget. For example, on Hölder Table, πBO and UCB require 41 and 49 evaluations, respectively,

while CoExBO fails to reach the target. These results demonstrate that the proposed trust mechanism improves sample efficiency across different objective landscapes by exploiting locally informative human guidance while suppressing unreliable feedback.

5.2 Real-world experiments

Tasks. We evaluate the proposed method on two scientific optimization benchmarks.

Li-based battery (Zhu et al., 2024): This task aims to maximize the ionic conductivity of a $LiPF_6$ electrolyte with an EC-PC-DMC solvent system by optimizing its composition fractions.

P3HT-CNT composite (Bash et al., 2021): This task aims to maximize the electrical conductivity of a P3HT–CNT composite by optimizing its composition parameters.

Expert feedback generation. We use an AI agent to simulate expert feedback for both tasks. The agent is provided with the task description, the definitions and feasible ranges of the composition variables, and relevant domain knowledge, without access to the ground-truth objective values. For each task, we sample 100 pairs of candidate compositions and ask the agent to identify which candidate is more likely to exhibit higher conductivity. The resulting pairwise comparisons are used as auxiliary expert feedback during optimization.

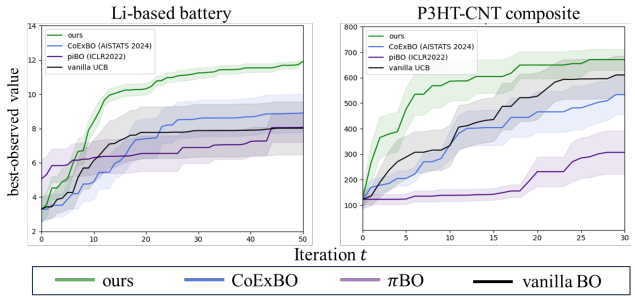


Figure 3: Optimization performance on the Li-based battery electrolyte and P3HT–CNT composite tasks.

Figure 3 presents the results on the Li-based battery and P3HT–CNT composite tasks. Our method consistently outperforms all baselines, achieving faster improvement and higher final best-observed values throughout the optimization process.

On the Li-based battery task, our method reaches a final best-observed conductivity of 11.8 mS/cm, compared with 8.9 mS/cm for CoExBO, 8.0 mS/cm for πBO, and 8.0 mS/cm for vanilla UCB. It also exceeds a value of 10 mS/cm within 15 iterations, while the baselines remain substantially lower.

On the P3HT–CNT composite task, our method achieves a final conductivity of 688.7 S/cm, outperforming CoExBO (533.5 S/cm), vanilla UCB (605 S/cm), and πBO (306.8 S/cm). These results demonstrate that the proposed trust-aware framework effectively leverages expert knowledge to accelerate optimization in practical scientific design problems.

6 Conclusion

We presented a trust-aware human–AI collaborative Bayesian optimization framework that treated human knowledge as an auxiliary low-fidelity information source, providing a unified perspective for incorporating different feedback modalities. The proposed method estimated input-dependent trust from local human–objective ranking agreement and used trust weighting to exploit reliable guidance while curbing locally misleading feedback. Theoretical analysis established a rate-level no-harm guarantee and characterized the benefit of increasing trust when human knowledge was aligned with objective observations. Experiments on synthetic and scientific optimization tasks demonstrated the superiority and robustness of the framework.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets used to emulate expert feedback in the real-world experiments, propose or refine research hypotheses, formulate mathematical claims, assist in the writing of proofs and assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, or interpret results. Support for qualitative and thematic data analysis is not applicable to this work.

Additionally, we used generative AI tools to suggest experimental parameter settings and assist with code editing. We reviewed all AI-assisted work. In particular, the generated synthetic expert-feedback data were checked for consistency with the intended experimental design, suggested experimental settings were independently assessed before use, and AI-assisted code modifications were reviewed and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings, benchmark definitions, expert-simulation configurations, implementation details, and additional results in the appendix, together with the sources of the real-world datasets. Source code will be made available upon acceptance of the paper.

REFERENCES

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*, pp. 505–513. PMLR, April 2024.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Arun Kumar AV, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Human-AI collaborative Bayesian optimisation. *Advances in neural information processing systems*, 35:16233–16245, 2022.

Arun Kumar AV, Alistair Shilton, Sunil Gupta, Santu Rana, Stewart Greenhill, and Svetha Venkatesh. Enhanced bayesian optimization via preferential modeling of abstract properties. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 234–250. Springer, 2024.

Daniil Bash, Yongqiang Cai, Vijila Chellappan, Swee Liang Wong, Xu Yang, Pawan Kumar, Jin Da Tan, Anas Abutaha, Jayce JW Cheng, Yee-Fun Lim, et al. Multi-fidelity high-throughput optimization of electrical conductivity in p3ht-cnt composites. *Advanced Functional Materials*, 31 (36):2102606, 2021.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pp. 137–144, 2005.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, pp. 1282–1291. PMLR, 2017.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science*, 11(2):577–586, 2020.

Daolang Huang, Louis Filstroff, Petrus Mikkola, Runkai Zheng, Milica Todorovic, and Samuel Kaski. Augmenting bayesian optimization with preference-based expert feedback. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. October 2021.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. *Advances in neural information processing systems*, 29, 2016.

Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

A Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, et al. On-the-fly closed-loop materials discovery via bayesian active learning. *Nature communications*, 11(1):5966, 2020.

Siu Lun Chau, Javier Gonzalez, and Dino Sejdinovic. Learning inconsistent preferences with gaussian processes. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2266–2281. PMLR, 28–30 Mar 2022.

Petrus Mikkola, Milica Todorović, Jari Järvi, Patrick Rinke, and Samuel Kaski. Projective preferential bayesian optimization. In *International Conference on Machine Learning*, pp. 6884–6892. PMLR, 2020.

Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings. Mathematical, physical, and engineering sciences*, 473(2198):20160751, 2017.

Matthias Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2015.

Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

S. K. Michael Wong and Pawan Lingras. Representation of qualitative user preference by quantitative belief functions. *IEEE Transactions on Knowledge and Data Engineering*, 6(1):72–78, 1994.

Wenjie Xu, Masaki Adachi, Colin N Jones, and Michael A Osborne. Principled bayesian optimization in collaboration with human experts. *Advances in Neural Information Processing Systems*, 37:104091–104137, 2024.

Matteo Zecchin and Osvaldo Simeone. Localized adaptive risk control. *Advances in Neural Information Processing Systems*, 37:8165–8192, 2024.

Shang Zhu, Bharath Ramsundar, Emil Annevelink, Hongyi Lin, Adarsh Dave, Pin-Wen Guan, Kevin Gering, and Venkatasubramanian Viswanathan. Differentiable modeling and optimization of non-aqueous li-based battery electrolyte solutions using geometric deep learning. *Nature Communications*, 15(1):8649, 2024.

A PROOF

A.1 PROOF OF THEOREM 4.1

Proof. For a fixed miscoverage level $\alpha \in (0,1)$, define

$$L_t^\alpha = \mathbf{1}\{r_t^{\mathrm{HF}} \notin \Gamma_t^\alpha(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}})\}$$

Under Assumptions 1-2 and $\eta_t < 1/\lambda$, the localized component is uniformly bounded as

$$\|\zeta_t^\alpha\|_\infty \leq \frac{\kappa}{\lambda}$$

Moreover, the stability of localized adaptive calibration gives

$$Q_{\min} \leq q_t^\alpha(\boldsymbol{x}) \leq Q_{\max}, \quad \forall t, \boldsymbol{x} \in \mathcal{X}$$

where $Q_{\max} = 1 + \dfrac{2\sqrt{\rho\kappa D}}{\lambda} + \eta_1(2\kappa+1)$ and $Q_{\min} = -\dfrac{2\sqrt{\rho\kappa D}}{\lambda} - \eta_1(2\kappa+1)$.

Since $c_t^\alpha = q_t^\alpha(\boldsymbol{x}) - \zeta_t^\alpha(\boldsymbol{x})$, we obtain

$$c_- \leq c_t^\alpha \leq c_+$$

where

$$c_+ = Q_{\max} + \frac{\kappa}{\lambda}, \quad c_- = Q_{\min} - \frac{\kappa}{\lambda}$$

Hence,

$$c_+ - c_- = 1 + \frac{4\sqrt{\rho\kappa D}}{\lambda} + 2\eta_1(2\kappa+1) + \frac{2\kappa}{\lambda}$$

From the update of c_t^α,

$$L_t^\alpha - \alpha = \frac{c_{t+1}^\alpha - c_t^\alpha}{\eta_t}$$

Since η_t is non-increasing,

$$\begin{aligned}
\sum_{t=1}^T (L_t^\alpha - \alpha) =& \sum_{t=1}^T \frac{c_{t+1}^\alpha - c_t^\alpha}{\eta_t} \\
=& \frac{c_{T+1}^\alpha}{\eta_T} - \frac{c_1^\alpha}{\eta_1} + \sum_{t=2}^T \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) c_t^\alpha \\
\leq& \frac{c_+}{\eta_T} + c_- \sum_{t=2}^T \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) \\
=& \frac{c_+ - c_-}{\eta_T} + \frac{c_-}{\eta_1} \\
\leq& \frac{c_+ - c_-}{\eta_T}
\end{aligned}$$

Since $\eta_T = \eta_1 T^{-1/2}$,

$$\frac{1}{T}\sum_{t=1}^T L_t^\alpha \leq \alpha + \frac{C}{\sqrt{T}}$$

where

$$C = \frac{1}{\eta_1} + \frac{4\sqrt{\rho\kappa D}}{\eta_1\lambda} + 2(2\kappa+1) + \frac{2\kappa}{\eta_1\lambda}$$

We now relate this calibration guarantee to the trust score. Recall that

$$\alpha_t^\star = \inf\left\{\alpha \in (0,1) : \Gamma_t^\alpha(\boldsymbol{x}_t) = \{\hat{r}_t^\pi(\boldsymbol{x}_t)\}\right\}.$$

By the definition of the trust score,

$$\omega_t(\boldsymbol{x}_t) = 1 - \alpha_t^\star.$$

Hence, for any $\tau \in (0,1)$,

$$\omega_t(\boldsymbol{x}_t) \geq \tau \quad \Longrightarrow \quad \alpha_t^\star \leq 1-\tau.$$

Since the conformal prediction sets are nested with respect to the miscoverage level α, i.e.,

$$\alpha_1 \leq \alpha_2 \quad \Longrightarrow \quad \Gamma_t^{\alpha_2}(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}}) \subseteq \Gamma_t^{\alpha_1}(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}})$$

and $\hat{r}_t^\pi(\boldsymbol{x}_t)$ always belongs to $\Gamma_t^\alpha(\boldsymbol{x}_t)$ because $s_t(\boldsymbol{x}_t, \hat{r}_t^\pi(\boldsymbol{x}_t)) = 0$, it follows that

$$\omega_t(\boldsymbol{x}_t) \geq \tau \quad \Longrightarrow \quad \Gamma_t^{1-\tau}(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}}) = \{\hat{r}_t^\pi(\boldsymbol{x}_t)\}.$$

Therefore, whenever the expert prediction is incorrect under the high-trust event,

$$\left\{\omega_t(\boldsymbol{x}_t) \geq \tau,\ r_t^{\mathrm{HF}} \neq \hat{r}_t^\pi(\boldsymbol{x}_t)\right\} \subseteq \left\{r_t^{\mathrm{HF}} \notin \Gamma_t^{1-\tau}(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}})\right\}.$$

Consequently,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{\omega_t(\boldsymbol{x}_t) \geq \tau,\ r_t^{\mathrm{HF}} \neq \hat{r}_t^\pi(\boldsymbol{x}_t)\right\} \leq & \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{r_t^{\mathrm{HF}} \notin \Gamma_t^{1-\tau}(\boldsymbol{x}_t \mid \mathcal{D}_{t-1}^{\mathrm{HF}}, \mathcal{D}_{t-1}^{\mathrm{P}})\right\} \\
= & \frac{1}{T}\sum_{t=1}^{T} L_t^{1-\tau} \\
\leq & 1-\tau+\frac{C}{\sqrt{T}}
\end{aligned}$$

□

A.2 Proof of theorem 4.2

Lemma A.1. *Let $k_t(\boldsymbol{x},\boldsymbol{x}') = k_0(\boldsymbol{x},\boldsymbol{x}') + \phi_t(\boldsymbol{x})\phi_t(\boldsymbol{x}'), \qquad |\phi_t(\boldsymbol{x})| \leq H_\pi$. For any sequence of queried points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, the information gain associated with the trust-aware covariance satisfies*

$$\Gamma_T \leq \gamma_T + \frac{1}{2}\log\left(1+\frac{TH_\pi^2}{\sigma_\epsilon^2}\right)$$

where γ_T denotes the maximum information gain associated with the objective-only kernel k_0.

Proof. Let $\phi_T = [\phi_1(\boldsymbol{x}_1), \ldots, \phi_T(\boldsymbol{x}_T)]^\top$ The Gram matrix of the trust-aware covariance is

$$\boldsymbol{K}_T = \boldsymbol{K}_{0,T} + \phi_T\phi_T^\top$$

Hence, by the matrix determinant lemma,

$$\begin{aligned}
2\Gamma_T &= \log\det\left(\boldsymbol{I} + \sigma_\epsilon^{-2}\boldsymbol{K}_T\right) \\
&= \log\det\left(\boldsymbol{I} + \sigma_\epsilon^{-2}\boldsymbol{K}_{0,T}\right) \\
&\quad + \log\left[1 + \sigma_\epsilon^{-2}{}_T^\top \left(\boldsymbol{I} + \sigma_\epsilon^{-2}\boldsymbol{K}_{0,T}\right)_T^{-1}\right]
\end{aligned}$$

Since $\left(\boldsymbol{I} + \sigma_\epsilon^{-2}\boldsymbol{K}_{0,T}\right)^{-1} \preceq \boldsymbol{I}$, and $|\phi_t(\boldsymbol{x}_t)| \leq H_\pi$,

$$\|\phi_T\|_2^2 \leq TH_\pi^2.$$

Therefore,

$$\Gamma_T \leq \gamma_T + \frac{1}{2}\log\left(1+\frac{TH_\pi^2}{\sigma_\epsilon^2}\right).$$

□

Proof of Theorem 4.2. On the standard GP confidence event, with probability at least $1-\delta$,

$$|f(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})| \leq \sqrt{\beta_t}\sigma_{t-1}(\boldsymbol{x})$$

simultaneously for all t and $\boldsymbol{x}$.

Since $\boldsymbol{x}_t$ maximizes the UCB acquisition function,

$$r_t = f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t) \le 2\sqrt{\beta_t}\sigma_{t-1}(\boldsymbol{x}_t).$$

Summing over t and applying Cauchy–Schwarz gives

$$R_T \le 2\sqrt{T\beta_T}\left(\sum_{t=1}^{T}\sigma_{t-1}^2(\boldsymbol{x}_t)\right)^{1/2}.$$

The standard GP information-gain argument yields

$$\sum_{t=1}^{T}\sigma_{t-1}^2(\boldsymbol{x}_t) = \mathcal{O}(\Gamma_T).$$

Applying Lemma A.1,

$$R_T = \mathcal{O}\left(\sqrt{T\beta_T\left[\gamma_T + \log\left(1 + \frac{TH_\pi^2}{\sigma_\epsilon^2}\right)\right]}\right).$$

Finally, using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and $\beta_T = \mathcal{O}(\log(T/\delta))$ gives

$$R_T = \mathcal{O}\left(\sqrt{T\gamma_T\log(T/\delta)}\right) + \mathcal{O}\left(\sqrt{T\log\left(1 + \frac{TH_\pi^2}{\sigma_\epsilon^2}\right)\log(T/\delta)}\right),$$

which proves the result. □

A.3 Trust mechanism analysis

We provide a more detailed interpretation of the proposed trust-aware kernel. At iteration t, define the trust-modulated human feature as

$$\phi_t(\boldsymbol{x}) = \omega_t(\boldsymbol{x})h_{\pi,t}(\boldsymbol{x}).$$

The kernel in Eq.(4) can then be written as

$$k_g^t(\boldsymbol{x}, \boldsymbol{x}') = k_0(\boldsymbol{x}, \boldsymbol{x}') + \phi_t(\boldsymbol{x})\phi_t(\boldsymbol{x}').$$

For a fixed iteration t, this kernel admits the equivalent latent representation

$$f(\boldsymbol{x}) = g_0(\boldsymbol{x}) + a_t\phi_t(\boldsymbol{x}),$$

where $g_0 \sim \mathcal{GP}(0, k_0)$, $\quad a_t \sim \mathcal{N}(0,1)$, and g_0 and a_t are independent.

Let $\phi_{\mathbf{t}} = [\phi_{\mathbf{t}}(\mathbf{x_1}), \ldots, \phi_{\mathbf{t}}(\mathbf{x_t})]^\top$, and $\boldsymbol{A}_t = \boldsymbol{K}_0(\boldsymbol{X}_t, \boldsymbol{X}_t) + \epsilon^2\boldsymbol{I}$.

Conditional on a_t, the observations satisfy

$$\boldsymbol{y}_t^{\mathrm{HF}} \mid a_t \sim \mathcal{N}\left(a_t\phi_t, \boldsymbol{A}_t\right).$$

Combining this likelihood with the prior $a_t \sim \mathcal{N}(0,1)$ gives

$$a_t \mid \mathcal{D}_t^{\mathrm{HF}} \sim \mathcal{N}\left(\bar{a}_t, \sigma_{a,t}^2\right),$$

where

$$\bar{a}_t = \frac{\phi_t^\top \boldsymbol{A}_t^{-1}\boldsymbol{y}_t^{\mathrm{HF}}}{1 + \phi_t^\top \boldsymbol{A}_t^{-1}\phi_t}$$

and

$$\sigma_{a,t}^2 = \frac{1}{1 + \phi_t^\top \boldsymbol{A}_t^{-1}\phi_t}.$$

Since the denominator of $\bar{a}_t$ is strictly positive, its sign is completely determined by

$$S_t = \phi_t^\top \boldsymbol{A}_t^{-1}\boldsymbol{y}_t^{\mathrm{HF}}.$$

Therefore,

$$S_t > 0 \quad \Longrightarrow \quad \bar{a}_t > 0,$$

meaning that the observed objective values support a positive contribution along the human-induced feature direction. Conversely,

$$S_t < 0 \quad \Longrightarrow \quad \bar{a}_t < 0,$$

so that the objective observations indicate an inverse relationship with the human-induced feature. When S_t is close to zero, the posterior coefficient is also close to zero, and the human-induced direction has little contribution to the posterior prediction.

More explicitly, let

$$\boldsymbol{k}_{0,t}(\boldsymbol{x}) = k_0(\boldsymbol{x}, \boldsymbol{X}_t).$$

The posterior mean can be written as

$$\mu_t(\boldsymbol{x}) = \boldsymbol{k}_{0,t}(\boldsymbol{x})\boldsymbol{A}_t^{-1}\boldsymbol{y}_t^{\mathrm{HF}} + \bar{a}_t \left[\phi_t(\boldsymbol{x}) - \boldsymbol{k}_{0,t}(\boldsymbol{x})\boldsymbol{A}_t^{-1}\phi_t\right].$$

Hence, the first term after the prior mean corresponds to the prediction supported by the objective-only kernel k_0, while the final term represents the additional contribution induced by human knowledge. Its orientation is determined by the data-adaptive coefficient $\bar{a}_t$, rather than being fixed by the human feedback a priori.

The trust score controls the local magnitude of this human-induced feature through

$$\phi_t(\boldsymbol{x}) = \omega_t(\boldsymbol{x})h_{\pi,t}(\boldsymbol{x}).$$

When $\omega_t(\boldsymbol{x})$ is large, the human feature is retained and can exert a stronger influence on the surrogate if it is supported by the high-fidelity observations. When $\omega_t(\boldsymbol{x})$ approaches zero, the direct human-dependent component of the kernel at $\boldsymbol{x}$ vanishes,

$$\phi_t(\boldsymbol{x})\phi_t(\boldsymbol{x}') \to 0,$$

and prediction at that location increasingly relies on the objective-only covariance structure k_0.

This analysis shows that the proposed kernel does not directly enforce the ordering suggested by human feedback. Instead, it introduces a locally trust-modulated feature direction and allows the high-fidelity observations to determine whether this direction should be positively utilized, negatively corrected, or effectively ignored. This provides a data-adaptive mechanism for exploiting informative human knowledge while limiting the influence of locally unreliable feedback.

B Experimental details

B.1 Test functions

Since the benchmark functions are conventionally minimized, we negate their values to formulate them as maximization problems.

Ackley The two-dimensional Ackley function is defined as

$$f(\boldsymbol{x}) = -20\exp\left(-0.2\sqrt{\frac{x_1^2 + x_2^2}{2}}\right) - \exp\left(\frac{\cos(2\pi x_1) + \cos(2\pi x_2)}{2}\right) + 20 + \mathrm{e},$$

where $\boldsymbol{x} \in [-32.768, 32.768]^2$. The global optimum is $\boldsymbol{x}^\star = (0, 0), \quad f(\boldsymbol{x}^\star) = 0.$

Hölder Table The two-dimensional Hölder-table function is defined as

$$f(\boldsymbol{x}) = -\left|\sin(x_1)\cos(x_2)\exp\left(\left|1 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi}\right|\right)\right|,$$

where $\boldsymbol{x} \in [-10, 10]^2$. It has four global minima at $\boldsymbol{x}^\star = (\pm 8.05502, \pm 9.66459)$, with $f(\boldsymbol{x}^\star) = -19.20850$.

Levy The three-dimensional Levy function is defined as

$$f(\boldsymbol{x}) = \sin^2(\pi w_1) + \sum_{i=1}^{2}(w_i - 1)^2 \left[1 + 10\sin^2(\pi w_i + 1)\right] + (w_3 - 1)^2 \left[1 + \sin^2(2\pi w_3)\right],$$

where

$$w_i = 1 + \frac{x_i - 1}{4}, \qquad \boldsymbol{x} \in [-10, 10]^3.$$

The global optimum is $\boldsymbol{x}^\star = (1, 1, 1)$, with $f(\boldsymbol{x}^\star) = 0$.

Styblinski-Tang The two-dimensional Styblinski–Tang function is defined as

$$f(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{2}\left(x_i^4 - 16x_i^2 + 5x_i\right)$$

where $\boldsymbol{x} \in [-5, 5]^2$. The global optimum is $\boldsymbol{x}^\star \approx (-2.903534, -2.903534)$, with $f(\boldsymbol{x}^\star) \approx -78.33233$.

B.2 Spatially varying expert reliability in synthetic experiments

To emulate spatially varying expert reliability, we introduce an input-dependent coefficient $a(\boldsymbol{x})$ in the simulated human utility. Positive values correspond to locally informative objective signals, negative values induce locally misleading signals, and small positive values represent weakly informative regions. The benchmark-specific constructions emulate different forms of partial expertise, such as knowledge of a particular input direction or familiarity with a promising region of the objective landscape, rather than assuming uniformly accurate knowledge over the entire search space.

For Hölder Table, Levy, and Styblinski-Tang, we use a smooth radial gate. Given a center $\boldsymbol{c}$, let $r(\boldsymbol{x}; \boldsymbol{c}) = \|\boldsymbol{x} - \boldsymbol{c}\|_2$ and define

$$g(\boldsymbol{x}; \boldsymbol{c}) = \begin{cases} 1, & r(\boldsymbol{x}; \boldsymbol{c}) \le r_{\text{in}}, \\ 1 - 3s^2 + 2s^3, & r_{\text{in}} < r(\boldsymbol{x}; \boldsymbol{c}) < r_{\text{out}}, \\ 0, & r(\boldsymbol{x}; \boldsymbol{c}) \ge r_{\text{out}}, \end{cases}$$

where $s = \frac{r(\boldsymbol{x};\boldsymbol{c}) - r_{\text{in}}}{r_{\text{out}} - r_{\text{in}}}$. The reliability coefficient is

$$a(\boldsymbol{x}) = a_{\text{weak}} + (a_{\text{good}} - a_{\text{weak}})g(\boldsymbol{x}; \boldsymbol{c}_{\text{good}}) + (a_{\text{bad}} - a_{\text{weak}})g(\boldsymbol{x}; \boldsymbol{c}_{\text{bad}})$$

with $a_{\text{weak}} = 0.08$, $a_{\text{good}} = 0.35$, $a_{\text{bad}} = -0.10$. The benchmark-specific parameters in the original input space are summarized in Table 1.

Table 1: Parameters of the spatially varying expert reliability regions.

Benchmark	$\boldsymbol{c}_{\text{good}}$	$\boldsymbol{c}_{\text{bad}}$	r_{in}	r_{out}
Hölder Table	(6.12, 7.34)	(0, 0)	2.12	3.80
Levy	(3.8, −1.8, 3.8)	(−4.0, 4.0, −4.0)	3.40	6.00
Styblinski–Tang	(−2.0, −1.8)	(1.8, 2.0)	1.00	1.80

For Ackley, reliability varies smoothly along the first input dimension:

$$a(\boldsymbol{x}) = 0.10 + 0.90\sigma\left(\frac{x_1 + 12}{2}\right)\sigma\left(\frac{10 - x_1}{2}\right).$$

These constructions represent different forms of partial human expertise. For Ackley, the dependence of $a(\boldsymbol{x})$ only on x_1 emulates an expert with knowledge primarily along a particular input direction. For Hölder Table, the informative region overlaps a globally optimal basin, representing familiarity with a promising local region rather than uniformly accurate knowledge. For Levy, the informative region covers the neighborhood of the global optimum without being centered at the optimizer, reflecting coarse knowledge of a promising region. For Styblinski-Tang, the informative region is placed near but not at the global optimum, while a distant region is misleading, modeling an expert whose knowledge is locally useful but unreliable elsewhere.

B.3 ADDITIONAL EXPERIMENTS

To complement the main experiments on spatially varying expert reliability, we additionally evaluate all methods under globally informative and adversarial experts on the four test functions.

Figure 4 summarizes both settings, with the first row corresponding to globally informative experts and the second row to adversarial experts. These experiments examine whether the proposed framework can exploit consistently useful human guidance while remaining robust to systematically misleading feedback.

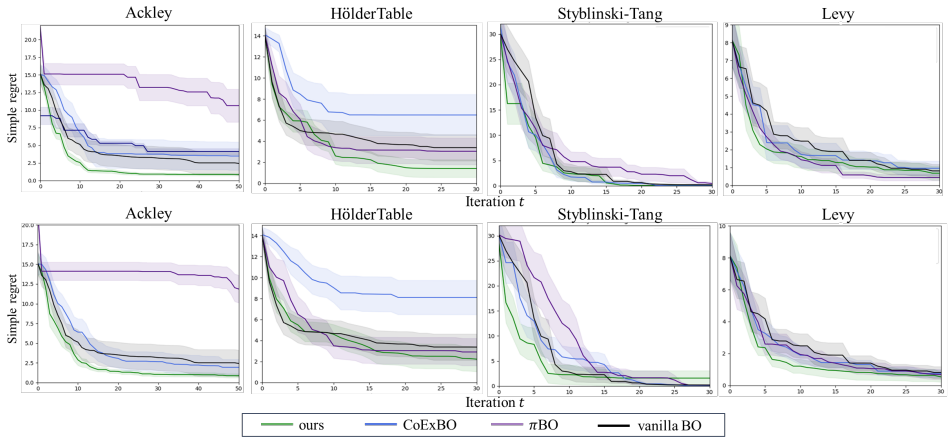


Figure 4: Optimization performance under globally informative and misleading experts on the test functions. The first row shows results with globally informative experts and the second row corresponds to adversarial experts.

As shown in Figure 4, our method achieves consistently competitive convergence across the four benchmarks under globally informative experts. Under adversarial experts, its performance remains close to vanilla UCB, with only limited degradation despite systematically misleading feedback. These results suggest that the trust mechanism can exploit useful guidance when it is reliable while limiting its adverse influence when it conflicts with objective observations.

B.4 REAL-WORLD TASKS

Li-based battery The Li-based battery dataset contains 104 measurements of ionic conductivity for $LiPF_6$ electrolytes with an EC–PC–DMC solvent system at 293.15 K (Zhu et al., 2024). Each electrolyte is represented by three inputs: the $LiPF_6$ mole fraction x_1, the EC fraction among all solvents x_2, and the PC fraction among the remaining PC–DMC mixture x_3. Their data-supported ranges are $x_1 \in [0.002, 0.1453]$ and $x_2, x_3 \in [0, 1]$. We construct a continuous benchmark from the 104 discrete observations by fitting a GP regression model.

P3HT–CNT composite The P3HT–CNT dataset records the electrical conductivity of composite films containing P3HT and four types of carbon nanotubes, denoted by D1, D2, D6, and D8 (Bash et al., 2021). After aggregating repeated measurements at identical compositions, the dataset contains 191 distinct composition–conductivity observations. The five input variables are the fractions of P3HT, D1, D2, D6, and D8, with respective ranges $[0.15, 1]$, $[0, 0.60]$, $[0, 0.70]$, $[0, 0.85]$, and $[0, 0.75]$. We fit a RBF kernel regression model to these observations and use its fixed posterior mean as the continuous benchmark objective.

B.5 EXPERIMENTS WITH BINARY LABEL FEEDBACK

We compare our method with standard GP-UCB and COBOL (Xu et al., 2024). All other settings, including the initial observations and the number of independent runs, remain unchanged. For each

real-world task, the AI agent generates 100 binary labels indicating whether individual candidate compositions are promising.

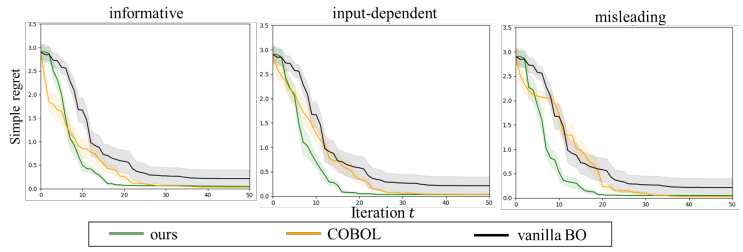


Figure 5: Performance on the Ackley function under informative, input-dependent and misleading emulated experts with binary label feedback.

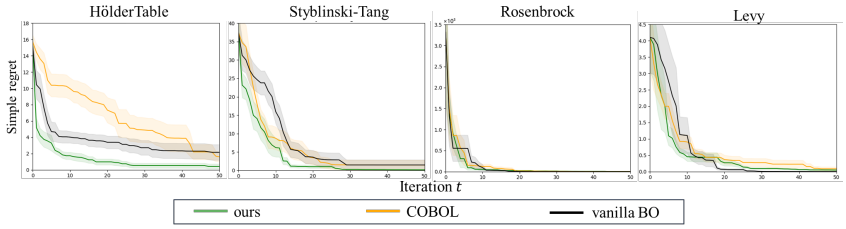


Figure 6: Performance on test functions under emulated input-dependent expert with binary label feedback.

Synthetic results. Figure 5 reports performance on the Ackley function under informative, input-dependent, and misleading experts. Our method remains effective across the three feedback settings, demonstrating that the trust mechanism can exploit informative binary labels while limiting the influence of unreliable ones. Figure 6 further shows that this behavior generalizes to the Hölder Table and Styblinski–Tang functions under input-dependent feedback.

Real-world results. Figure 7 reports the results under binary expert feedback on the Li-based battery and P3HT–CNT composite tasks. On the Li-based battery task, our method improves rapidly during the early stage and reaches a best-observed conductivity of approximately 9 mS/cm within about 15 iterations, substantially earlier than COBOL. It ultimately achieves a value of around 11 mS/cm, while vanilla BO plateaus near 8 mS/cm. On the P3HT–CNT composite task, our method similarly exhibits faster early improvement, reaching values above 500 S/cm within 15 iterations and converging to around 610 S/cm. Although vanilla BO approaches a comparable final value, it requires more evaluations to reach the same performance region, while COBOL remains lower at the end of the optimization budget. These results demonstrate that the proposed trust-aware framework can effectively exploit alternative feedback formats beyond pairwise comparisons while maintaining its sample-efficiency advantage in real-world optimization tasks.

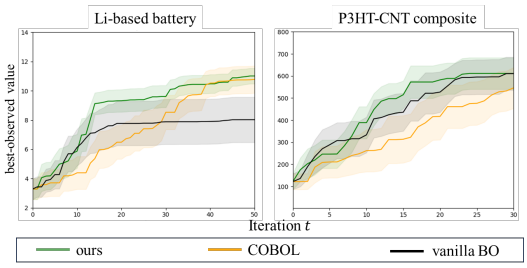


Figure 7: Performance on real-world scientific tasks with binary label feedback.