WHICH VARIABLES MATTER? LLM-INFORMED COMPETITIVE RELEVANCE LEARNING FOR HIGH-DIMENSIONAL BAYESIAN OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional design problems are common in scientific discovery and formulation optimization, where effective designs must be identified under tight experimental budgets. However, existing high-dimensional BO methods often struggle to uncover useful structural biases from sparse observations alone. To address this challenge, we exploit the pronounced hierarchy of variable relevance and propose `CoRe-BO`, a framework for LLM-informed competitive relevance learning. Rather than asking the LLM for precise relevance scores or direct kernel modifications, `CoRe-BO` uses coarse judgments about variable importance to initialize a task-aware relevance prior, with variables competing for a shared sensitivity budget. As experiments accumulate, this structure is continually adapted through bounded LLM reflection and posterior inference, progressively concentrating modeling and search on influential variables. A regret analysis characterizes how the quality of LLM relevance guidance affects search complexity and the cost of imperfect priors. Experiments on six scientific optimization tasks with 29–150 variables show consistent gains in both convergence and final optimization performance over high-dimensional and LLM-assisted BO baselines. Our results suggest that `CoRe-BO` provides a practical and theoretically grounded approach to integrating LLMs into high-dimensional scientific optimization.

1 INTRODUCTION

Many scientific optimization problems involve dozens or even hundreds of design variables with markedly different relevance to the target property. In multicomponent systems, a relatively small subset of variables often exerts greater influence on the objective, while many others have secondary or marginal effects (Box & Meyer, 1986; Clackson & Wells, 1995; Peng et al., 2024). For example, superconducting materials have complex multielement compositions, yet the temperature at which superconductivity emerges may be governed largely by a few dopants that alter carrier density and lattice strain (Smith et al., 1991; Attfield et al., 1998). Bayesian optimization (BO) provides a principled way to navigate such scientific design spaces. However, its effectiveness often deteriorates in high dimensions, where many variables must be tuned jointly under a limited budget of costly experiments (Wang et al., 2016; Shields et al., 2021).

Existing high-dimensional BO methods address this challenge by introducing inductive biases that reduce effective search complexity. Some assume limited interactions across variable groups and decompose the objective function into tractable components (Kandasamy et al., 2015). Others exploit an intrinsic low-dimensional effective structure in the search space (Wang et al., 2016; Eriksson & Jankowiak, 2021) or adaptively restrict the search to promising regions (Eriksson et al., 2019). These strategies are effective in many settings, but the structures they impose are largely task-agnostic, either specified through generic assumptions or inferred from limited observations. Since scientific objectives are governed by diverse physicochemical mechanisms, such generic structures may not guide the search efficiently across tasks. Moreover, identifying useful structure from data alone in high dimensions can consume a substantial fraction of the experimental budget.

High-dimensional scientific design provides a natural source of prior structure. Variables often carry meaningful physicochemical context, and their mechanistic roles can reveal the differences in rele-

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.

vance, indicating how strongly each variable influences the objective. Such knowledge offers early clues about which dimensions deserve greater attention during optimization (Grinberg et al., 2005). Large language models (LLMs), pretrained on broad scientific corpora, provide a convenient interface for eliciting such knowledge by judging which variables are likely to matter more (Taylor et al., 2022; Jablonka et al., 2024). However, directly translating raw LLM outputs into a coherent relevance structure for BO remains challenging, as their judgments are often coarse and inconsistent across variables. Moreover, LLMs may have incomplete knowledge of specialized scientific systems, leading to imperfect guidance. These challenges raise a central question: *How can coarse and imperfect LLM knowledge about variable importance be translated into an informative yet revisable structural bias for high-dimensional BO?*

To this end, we propose `CoRe-BO` (Competitive Relevance BO), which introduces an LLM-informed prior over variable relevance for high-dimensional BO. At its foundation lies the `CoRe` Prior, a hierarchical Dirichlet model that makes variables compete for a shared budget of surrogate sensitivity, concentrating modeling capacity on the more influential dimensions. By exploiting the geometry of Dirichlet distribution, we establish an interpretable and calibrated mapping from coarse LLM judgments to the competitive relevance allocation, providing an informative starting point for relevance learning. As experimental observations accumulate, the LLM revises the relevance structure through KL-constrained updates, while its influence relative to data-driven inference is calibrated with predictive evidence. Together, these mechanisms enable `CoRe-BO` to identify important variables early and progressively focus the search on them. Our regret analysis establishes when LLM-guided relevance improves optimization efficiency and how its benefits degrade under misspecification. Extensive evaluations across high-dimensional optimization tasks in physics, chemistry, and materials science show that `CoRe-BO` finds high-performing solutions efficiently under fixed experimental budgets. Our contributions are threefold:

1. We propose `CoRe-BO`, which translates coarse LLM-derived variable importance into a competitive relevance prior. This prior makes variables compete for a shared surrogate sensitivity budget and is jointly refined through experimental evidence and LLM-guided revision.
2. We theoretically characterize how LLM-informed relevance shapes optimization complexity. Our analysis shows that well-aligned guidance reduces effective search complexity and tightens regret bounds, while quantifying the sensitivity to imperfect guidance.
3. We demonstrate the practical utility and generality of `CoRe-BO`. Across six scientific optimization tasks spanning 29–150 variables, it consistently converges faster and achieves stronger final performance than representative high-dimensional and LLM-assisted BO methods.

2 Related Work

Bayesian Optimization for Scientific Discovery. Bayesian optimization (BO) provides a sample-efficient framework for costly experimental design by fitting a probabilistic surrogate to observed data and using an acquisition function to balance exploitation of promising regions with informative exploration. It has been successfully applied to scientific problems such as reaction optimization, materials discovery, and self-driving laboratories (Shields et al., 2021; MacLeod et al., 2020; Burger et al., 2020). However, standard BO is most effective in low-dimensional settings, often with fewer than 20 variables (Eriksson & Jankowiak, 2021). In high-dimensional multi-component systems, the limited experimental budget must support learning an increasing number of variable interactions and hyperparameters. This data scarcity often leads to unreliable surrogate modeling, with weakened kernel correlations, overfitting, and difficulty in uncovering meaningful structural patterns (Eriksson & Jankowiak, 2021; Hvarfner et al., 2024; Jansen et al., 2025).

High-Dimensional Bayesian Optimization. High-dimensional BO improves scalability by introducing inductive biases that reduce the effective search or modeling complexity. REMBO searches within low-dimensional random embeddings (Wang et al., 2016), while additive methods decompose the objective into low-dimensional components (Kandasamy et al., 2015). TuRBO exploits locality via adaptive trust regions and local Gaussian process models (Eriksson et al., 2019), whereas SAASBO uses a shrinkage prior to favor axis-aligned subspaces (Eriksson & Jankowiak, 2021). More recently, Hvarfner et al. (2024) show that dimension-scaled lengthscale priors can make vanilla

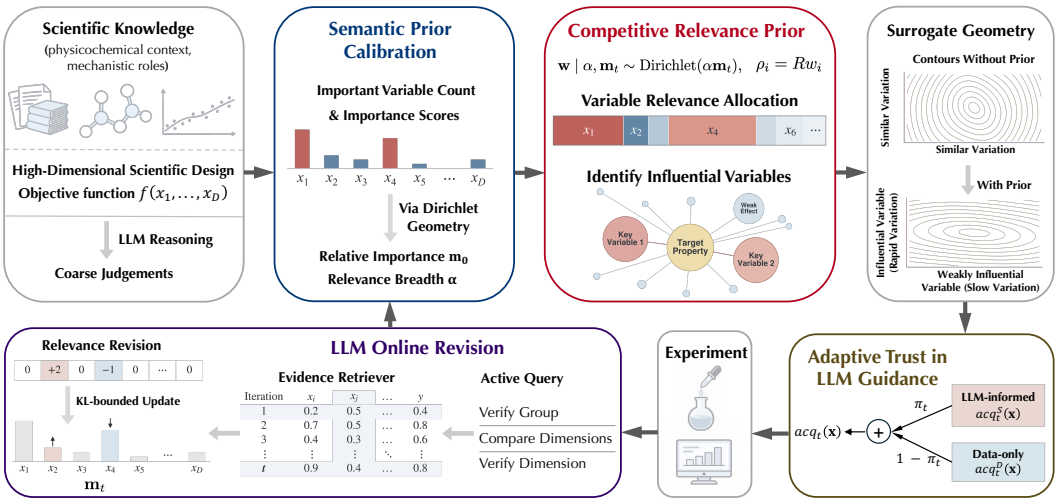


Figure 1: Overview of CoRe−BO. At initialization, the LLM judges variable importance and estimates the number of influential variables. Semantic prior calibration uses the outputs to inform a competitive relevance prior, where variables compete for a finite sensitivity budget. The prior shapes the surrogate geometry to allow faster variation along influential dimensions. At each round, LLM-informed and data-only acquisition functions are adaptively weighted to select the next experiment. Periodically, LLM issues queries to reassess uncertain judgments and refines the relevance structure.

BO competitive. Despite their effectiveness, these inductive biases generally do not leverage task-specific scientific semantics, leading to potential model misspecification and misleading exploration.

LLMs for Bayesian Optimization. Large language models (LLMs) provide a convenient interface for incorporating task descriptions and pretrained knowledge into BO. LLAMBO supports warm-starting, surrogate modeling, and candidate sampling with LLMs (Liu et al., 2024), whereas BOPRO uses latent-space BO to guide LLM candidate generation through retrieved in-context examples (Agarwal et al., 2025). Other methods integrate language-derived information into surrogate models. LGBO translates regional preferences into shifts of the surrogate mean (Yuan et al., 2026), LABO combines LLM predictions with experimental measurements through a dual-fidelity model (Chen et al., 2026), and CAKE evolves GP kernels through LLM-guided search (Suwandi et al., 2025). However, directly asking LLMs to provide numerical predictions or modify kernel structures becomes increasingly difficult and unreliable as dimensionality grows, leaving how to harness their prior knowledge for effective structural priors in high-dimensional BO underexplored.

3 Method

We present CoRe−BO, which turns scientific knowledge into a revisable prior over variable relevance. Section 3.1 reviews BO and ARD-based relevance modeling. Section 3.2 presents the CoRe prior and how to calibrate the prior using LLM judgments. Section 3.3 presents online LLM reflection to accelerate the identification of an effective relevance structure. Section 3.4 introduces adaptive trust calibration to balance LLM guidance with data-driven inference. The overall optimization workflow of CoRe−BO is illustrated in Figure 1.

3.1 Preliminary

Standard Bayesian optimization. We consider optimizing an expensive black-box objective $f : \mathcal{X} \to \mathbb{R}$ over a potentially high-dimensional domain $\mathcal{X} \subseteq [0,1]^D$. BO places a Gaussian process (GP) prior $f \sim \mathrm{GP}(\mu, k)$ over the objective, where μ denotes the mean and the kernel k encodes similarity between inputs. Conditioning this prior on observations $\mathcal{D}_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^t$ yields a posterior that provides updated predictions and uncertainty. An acquisition function $a_t(\mathbf{x})$, like EI or UCB, then uses the posterior to select the next evaluation by balancing exploration and exploitation.

ARD kernels and variable relevance. To model unequal relevance across variables, a common choice is an RBF kernel with automatic relevance determination (ARD), which assigns each dimension a separate lengthscale (Rasmussen & Williams, 2006):

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{D} \rho_i (x_i - x_i')^2\right), \qquad \rho_i = \ell_i^{-2}. \tag{1}$$

Here σ_f^2 is the signal variance and ℓ_i is the lengthscale of variable i. Its inverse square ρ_i controls how strongly the GP responds to changes along that dimension, with larger values indicating greater relevance. The ARD parameters $\{\rho_i\}_{i=1}^D$ thus provide a natural representation of variable relevance (Hvarfner et al., 2024; Eriksson & Jankowiak, 2021). A prior over these parameters expresses initial beliefs about variable importance, which posterior inference refines using observations.

3.2 Core Tool: Competitive Relevance Prior

High-dimensional scientific design often exhibits a pronounced hierarchy of variable relevance, where a small number of variables dominate the target response and warrant greater modeling capacity under limited experimental evaluations. Meanwhile, relevance is relative rather than absolute, and greater emphasis on one variable implies less emphasis on others. This motivates a competitive view of high-dimensional structure, in which variables compete for a shared relevance budget.

Competitive relevance allocation. To instantiate this principle, we introduce the *Competitive Relevance (`CoRe`) prior*. Instead of assigning variable relevance independently, `CoRe` lets all variables share a finite relevance budget and learns how this budget should be distributed:

$$\begin{aligned}
\text{[relevance budget]} \quad & \log R \sim \mathcal{N}(\mu_R, \sigma_R^2), \\
\text{[allocation concentration]} \quad & \log \alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \\
\text{[relevance allocation]} \quad & \mathbf{w} \mid \alpha, \mathbf{m}_t \sim \mathrm{Dirichlet}(\alpha \mathbf{m}_t), \\
\text{[ARD sensitivities]} \quad & \rho_i = R w_i, \qquad \ell_i = (R w_i)^{-1/2}, \quad i = 1, \ldots, D.
\end{aligned} \tag{2}$$

Here, $R > 0$ sets the total relevance budget, while the simplex-valued vector $\mathbf{w} \in \Delta^{D-1}$ competitively allocates this budget across variables. The Dirichlet mean $\mathbf{m}_t$ determines the expected relevance allocation, whereas α controls its concentration, with smaller α favoring mass concentrated on fewer coordinates. Together, R, $\mathbf{m}_t$, and α govern the overall sensitivity scale, its distribution across variables, and the concentration of that distribution, respectively. Since R captures the global sensitivity scale rather than relative variable relevance, we place a weakly informative prior on it and infer it from observations instead of requiring the LLM to calibrate absolute kernel lengthscales.

The shared budget design preserves meaningful similarity estimates as dimensionality grows. For two normalized inputs, $\sum_i \rho_i = R$ implies a bounded ARD distance, $\sum_{i=1}^D \rho_i (x_i - x_i')^2 \leq R$. Thus, increasing the ambient dimension redistributes a fixed relevance budget across coordinates rather than mechanically enlarging the overall ARD distance.

Crucially, `CoRe` remains structurally flexible. It neither hard-codes an active set nor assumes predefined structures such as additivity. Instead, the relevance allocation $\mathbf{w}$ and distributional parameters R and α remain latent and are refined through posterior inference. This allows relevance to concentrate on a small subset of variables while preserving the flexibility to assign substantial relevance to any variable when supported by data.

From semantic judgments to the `CoRe` prior. Before observing any experimental data, we use scientific knowledge encoded in the LLM to initialize an effective relevance model. Rather than asking the LLM to directly specify abstract distributional hyperparameters, which is difficult to calibrate reliably in high dimensions, we elicit two coarse but semantically natural judgments: *which variables are likely to matter more* and *how many variables are likely to be important*. Given the task description, optimization objective, and variable semantics, the LLM returns ordinal importance scores $\mathbf{s} = (s_1, \ldots, s_D)$, with $s_i \in \{1, \ldots, 10\}$, together with an estimated range $[d_L, d_U]$ for the number of important variables. We then translate these judgments into the parameters of the `CoRe` prior by exploiting the geometry of the Dirichlet distribution.

- Encoding relative importance. The ordinal scores provide reliable information about the *relative* importance of variables. We use them to shape the mean direction of the relevance allocation:

$$\mathbf{m}_0 = \frac{1-\lambda}{D}\mathbf{1} + \lambda \,\mathrm{softmax}\left(\frac{\mathbf{s}}{\tau_s}\right), \qquad 0 \le \lambda < 1, \tag{3}$$

where $\tau_s > 0$ controls the contrast between semantic scores and λ controls the strength of semantic guidance. Through modulating the Dirichlet mean $\mathbf{m}_0$, LLM outputs shape the expected relevance allocation while leaving each variable's relevance free to adapt to observations. The uniform component preserves non-negligible prior mass for variables that the LLM may underestimate.

- Encoding relevance breadth. Compared with $\mathbf{m}_0$, whose entries directly represent the relative relevance of individual variables, the concentration parameter α has a less intuitive semantic interpretation and is therefore harder to elicit directly from an LLM. The Dirichlet distribution, however, provides a natural connection between α and a more interpretable quantity, the effective number of variables receiving substantial relevance mass, defined as

$$d_{\mathrm{eff}}(\alpha, \mathbf{m}_0) := \frac{1}{\mathbb{E}[\|\mathbf{w}\|_2^2]} = \frac{1+\alpha}{1+\alpha\|\mathbf{m}_0\|_2^2}. \tag{4}$$

The monotonic relationship above maps the LLM-estimated range $[d_L, d_U]$ of influential-variable counts to a corresponding interval $[\alpha_L, \alpha_U]$. Rather than fixing α to a single semantic estimate, we use this interval to define a log-normal hyperprior through quantile matching, with α_L and α_U assigned to its lower and upper quantiles, respectively. This provides an interpretable prior over relevance breadth while allowing experimental evidence to refine it.

Thus, by exploiting the analytical properties of the Dirichlet distribution, we establish a principled bridge between coarse LLM judgments and a prior over the surrogate's lengthscales. The resulting probabilistic model induces a relevance hierarchy across all variables, providing an informative warm start for early search while remaining open to further revision throughout optimization.

3.3 Online Relevance Revision

The LLM-informed CoRe prior provides guidance for early search, but its relevance hierarchy may be imperfect. In high-dimensional BO, correcting such errors through posterior inference from data alone can be slow, as sparse observations often support multiple plausible relevance structures. We therefore complement posterior inference with online LLM reflection, refining semantic judgments through targeted evidence queries and controlled belief revisions.

Active evidence querying. We let the LLM reassess its relevance belief using evidence accumulated during optimization. However, directly exposing it to raw high-dimensional trajectories, as in LLAMBO (Liu et al., 2024) and LGBO (Yuan et al., 2026), can bury evidence about variable relevance in a large numerical context. Fixed statistical summaries may instead be unreliable with sparse observations and duplicate relevance estimation outside the surrogate. Since different relevance judgments require different evidence, we tailor the LLM's context to its current uncertainties. At each reflection, the LLM draws on its current belief and accumulated experimental findings to identify variables requiring further examination and issue targeted queries. We provide a coarse-to-fine query interface: `VERIFY_GROUP`$(\{x_1, x_2, \ldots\})$ to assess group relevance, `COMPARE_DIM`(x_1, x_2) to compare relative importance, and `VERIFY_DIM`(x_1) to examine an individual variable. A deterministic retriever then returns a small set of query-aligned observation pairs from previously evaluated designs, keeping the context compact and the reflection grounded in relevant experimental evidence.

Stable belief revision. To prevent noisy LLM feedback from causing large, inconsistent drift in the relevance hierarchy, we revise rather than regenerate the current semantic belief. The LLM outputs sparse ordinal corrections $z_{t,i} \in \{-2, -1, 0, 1, 2\}$, assigning zero to variables that are unexamined or lack sufficient evidence for revision. We incorporate these corrections through a deterministic KL-regularized update:

$$\mathbf{m}_t = \arg\max_{\mathbf{m} \in \Delta^{D-1}} \left\{ \langle \mathbf{m}, \mathbf{z}_t \rangle - \frac{1}{\eta_t} D_{\mathrm{KL}}(\mathbf{m} \| \mathbf{m}_{t-1}) \right\}. \tag{5}$$

Here, η_t controls the revision magnitude and is reduced when necessary to satisfy a fixed KL-divergence cap. The previous belief remains the anchor while new evidence induces only a bounded correction. This consistent state-transition rule prevents each reflection from independently redefining the relevance structure and limits abrupt changes in semantic guidance. The revised $\mathbf{m}_t$ sets the Dirichlet mean of the adaptive `CoRe` prior for subsequent semantic surrogate fits. At each BO iteration, we perform joint posterior inference over $\mathbf{w}$, α, R, and the remaining GP hyperparameters using the full observation history $\mathcal{D}_t$ and the current relevance prior. Adaptive evidence querying and bounded belief revision harness the LLM's scientific knowledge to reassess variable importance in light of observed outcomes, helping BO identify influential variables with fewer evaluations.

3.4 Adaptive Trust in LLM Guidance

CoRe-BO combines LLM knowledge with data-driven inference to learn variable relevance. A natural question is how to balance their influence, benefiting from useful semantic guidance while limiting the impact of incorrect judgments. We therefore calibrate LLM trust by testing whether semantic guidance improves predictions of new experimental outcomes. Specifically, we compare the semantic GP, fitted under the LLM-modulated `CoRe` prior, with a matched data-only GP using a neutral relevance prior. Their probabilistic predictions of experimental outcomes at candidate designs, $p_t^S(y \mid \mathbf{x})$ and $p_t^D(y \mid \mathbf{x})$, provide a basis for calibrating the trust weight π_t. Following sequential Bayesian model comparison, we update this weight after observing y_{t+1}:

$$\operatorname{logit} \pi_{t+1} = \operatorname{logit} \pi_t + \log \frac{p_t^S(y_{t+1} \mid \mathbf{x}_{t+1})}{p_t^D(y_{t+1} \mid \mathbf{x}_{t+1})}. \tag{6}$$

Both predictions are made before observing y_{t+1}. Trust increases when the semantic GP better predicts the new outcome and decreases when the data-only GP does. At round t, the current trust weight controls the next evaluation through a weighted acquisition:

$$\mathbf{x}_{t+1} \in \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \pi_t a_t^S(\mathbf{x}) + (1 - \pi_t) a_t^D(\mathbf{x}) \right\}, \tag{7}$$

where a_t^S and a_t^D are the acquisition functions of the semantic and data-only GPs, respectively. We initialize a large π_0 value to favor the LLM's pretrained scientific knowledge when data are scarce. Together, belief revision and trust calibration allow `CoRe-BO` to adapt both the relevance belief and its influence on search, retaining useful knowledge while reducing reliance on misleading guidance.

4 Theoretical Analysis

We provide a theoretical characterization of how an LLM-derived relevance prior influences high-dimensional Bayesian optimization by deriving a cumulative regret bound. Since each LLM intervention modifies the surrogate through the same relevance-guidance mechanism, analyzing the full adaptive process would conflate the effect of a single intervention with that of repeated online revisions. Therefore, in our theoretical analysis, we consider only a fixed-prior specialization, where the initial `CoRe` prior defines an ARD kernel k_S, referred to as the semantic kernel, which remains fixed throughout optimization. This formulation isolates the effect of the semantically guided relevance prior, where allocating sensitivity across variables reshapes the surrogate geometry and thereby changes the effective complexity of the optimization problem.

Semantic alignment and search complexity. Under this specialization, the LLM affects optimization through the relevance allocation it induces, which determines the ARD sensitivities ρ_i and hence the semantic kernel k_S. We characterize the quality of this guidance through two complementary quantities:

$$\epsilon_{\text{sem}} = \inf_{\substack{g \in \mathcal{H}_{k_S} \\ \|g\|_{\mathcal{H}_{k_S}} \le B}} \|f - g\|_\infty, \qquad \gamma_T^S = \max_{\substack{A \subset \mathcal{X} \\ |A| = T}} I(\mathbf{y}_A; f_A), \tag{8}$$

where $\mathcal{H}_{k_S}$ denotes the RKHS induced by k_S. The approximation error ϵ_{sem} measures how well the semantic hypothesis class represents the objective, while the maximum information gain γ_T^S measures the complexity of exploring that class. An informative semantic prior should therefore be both well aligned (small ϵ_{sem}) and usefully restrictive (small γ_T^S). Formal definitions and assumptions are given in Appendix A.1.

Theorem 1 (Regret with LLM-guided relevance)**.** *Suppose k_S approximates the objective up to uniform error ϵ_{sem} with an RKHS function of norm at most B. Under the standard GP-UCB assumptions of a compact domain, a bounded continuous kernel, and sub-Gaussian observation noise, consider fixed-prior* `CoRe-UCB` *whose confidence width uses a valid upper bound on ϵ_{sem}. When the supplied bound equals ϵ_{sem}, with probability at least $1-\delta$, the cumulative regret R_T satisfies*

$$R_T \le 2\beta_T\sqrt{C_\nu T\gamma_T^S} + \epsilon_{\text{sem}} T\left(2+\sqrt{\frac{2C_\nu\gamma_T^S}{\nu}}\right), \qquad C_\nu = \frac{2}{\log(1+\nu^{-1})}, \tag{9}$$

where β_T is the standard GP-UCB confidence parameter and ν is the regularization parameter.

Remark. The bound exposes an explicit accuracy–complexity trade-off. The first term captures the search complexity induced by the semantic kernel, so a well-aligned relevance structure with smaller γ_T^S yields a tighter regret bound. Any loss of alignment with the objective is captured separately by ϵ_{sem}, explicitly quantifying its effect on the guarantee. Thus, LLM guidance can be beneficial by simplifying the search while preserving sufficient fidelity to the objective.

5 Experiments

We evaluate `CoRe-BO` on six high-dimensional optimization tasks across physics, chemistry, and materials science, comparing its efficiency and stability with representative high-dimensional BO and LLM-assisted methods under limited evaluation budgets. Extensive analyses and ablations further demonstrate the effectiveness and robustness of our design.

5.1 Experimental Setup

Tasks. We conduct our experiments on six continuous black-box scientific optimization tasks ranging from 29 to 150 dimensions: **1) Perovskite band gap (29D)** (Pilania et al., 2016) adjusts the fractions of 29 candidate elements in a double perovskite to maximize its electronic band gap; **2) Superhard shear modulus (62D)** (Mansouri Tehrani et al., 2018) searches over elemental fractions to find materials with greater resistance to shear deformation; **3) Formation stability (79D)** (Kim et al., 2017) optimizes elemental compositions for thermochemical stability, measured by negative formation enthalpy; **4) Superconductor (86D)** (Hamidieh, 2018) seeks compositions with higher superconducting critical temperatures by tuning 86 scaled elemental coefficients; **5) Ferroelectric (118D)** (Zhang & Chen, 2025) varies elemental fractions to raise the symmetry-breaking transition temperature, below which the crystal leaves its high-symmetry cubic phase; and **6) Nutrition (150D)** (U.S. Department of Agriculture, Agricultural Research Service, 2018) maximizes a dietary health score over 150 ingredient quantities under a calorie budget.

Baselines. We compare against the following methods: **1) Vanilla BO** (Hvarfner et al., 2024) uses dimension-scaled length-scale priors to support high-dimensional search without explicit structural assumptions; **2) SAASBO** (Eriksson & Jankowiak, 2021) uses hierarchical shrinkage priors to favor sparse axis-aligned structure; **3) TuRBO** (Eriksson et al., 2019) fits local GP models within adaptive trust regions; **4) BOPRO** (Agarwal et al., 2025) uses latent-space BO to guide LLM candidate generation through retrieved in-context examples; and **5) LGBO** (Yuan et al., 2026) incorporates LLM-derived regional preferences into the surrogate mean to guide search. Comparisons with additional baselines, including Add-GP and LLAMBO, are provided in Appendix E.

Experimental details. All methods employ the GP surrogate and Log-qEI acquisition function. `CoRe-BO` and the LLM-assisted baselines are initialized with 20 LLM-proposed points, whereas the others use 20 Sobol points. All LLM-based methods use Intern-S1, a large language model pretrained on scientific knowledge (Bai et al., 2025). `CoRe-BO` invokes LLM reflection every 10 iterations, allowing sufficient evidence to accumulate between reflections. We run 80 BO iterations after initialization and report the mean best-so-far performance and standard error over five seeds.

5.2 Main Results

Across all tasks, `CoRe-BO` converges faster and achieves stronger final performance than the baselines (Figure 2). On Formation Stability and Superconductor in particular, it continues to improve

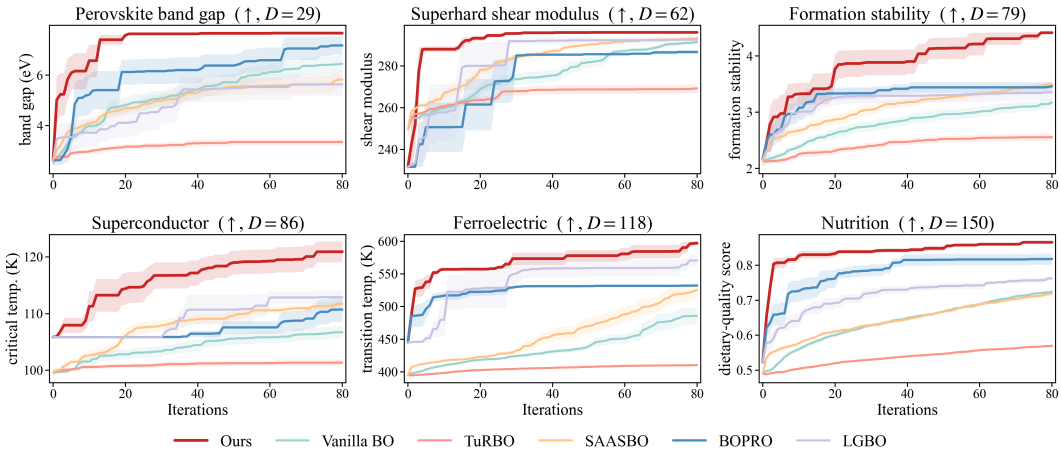


Figure 2: Optimization trajectories across six scientific tasks. The x-axis reports the iteration number, and the y-axis reports the best objective value found so far. Lines show the mean, and shaded regions indicate one standard error (s.e.) over five random seeds.

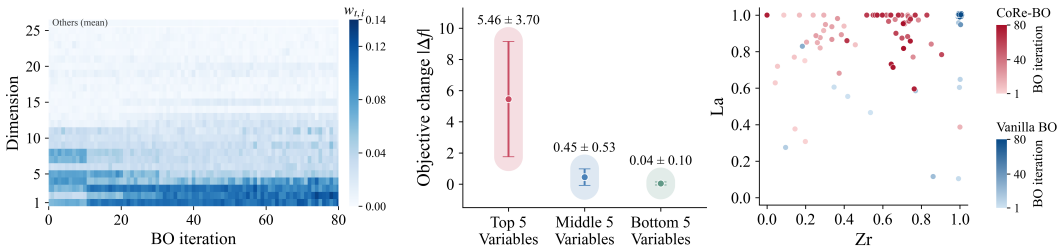
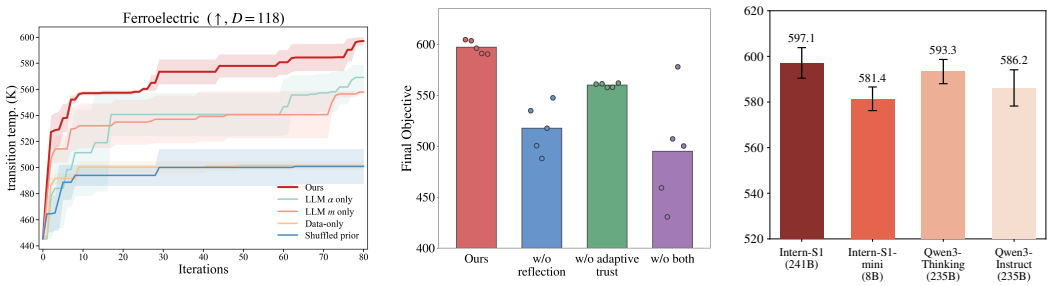


Figure 3: Relevance learning and sampling patterns on Ferroelectric. (a) Posterior relevance of the top 25 variables ranked by final posterior relevance; "Others" denotes the mean over the remaining variables. (b) Objective changes under equal-sized, single-variable perturbations around the best solution found, grouped by final relevance rank: top five, middle five (ranks 16–20), and bottom five. Results in (a) and (b) are averaged over five independent seeds. (c) CoRe-BO and Vanilla BO sampling locations in the pre-normalization search coordinates of Zr and La, the two variables with the highest posterior relevance. Darker colors indicate later iterations.

late in optimization, whereas most competing methods largely plateau. Vanilla BO generally makes slower progress, highlighting the cold-start bottleneck in high dimensions. SAASBO and TuRBO impose task-agnostic structural biases that may not align with the underlying mechanism and require extra evaluations to adapt. LLM-assisted methods such as BOPRO and LGBO may fail to sustain performance gains from LLM guidance throughout optimization. By learning a competitive relevance allocation informed by LLM knowledge, `CoRe-BO` more effectively identifies promising search directions and improves sample efficiency, a key requirement in costly scientific optimization.

Learned Relevance and Search Behavior. We analyze how CoRe-BO translates learned variable relevance into targeted exploration on Ferroelectric. As optimization proceeds, the posterior relevance progressively concentrates on a small subset of variables (Figure 3(a)). This hierarchy is consistent with empirical objective sensitivity. As shown in Figure 3(b), applying equal-sized perturbations to individual variables around the best solution reveals substantially larger objective variation for the top-ranked variables than for those in the middle and bottom relevance groups. The learned structure also changes where CoRe-BO searches. When the sampling locations are projected onto Zr and La, the two variables with the highest posterior relevance, CoRe-BO first explores broadly across the influential subspace and then concentrates evaluations in a promising local region (Figure 3(c)). In contrast, Vanilla BO repeatedly samples near a boundary location and leaves



(a) (b) (c)

Figure 4: Ablation results on the Ferroelectric task over 5 seeds. (a) Optimization trajectories under different semantic prior injection strategies. (b) Final performance of CoRe−BO with online reflection, adaptive trust, or both removed. (c) Final performance of CoRe−BO with different LLMs.

much of this subspace unexplored. Together, these results show that CoRe-BO identifies influential dimensions and uses them to guide diverse exploration followed by focused search.

5.3 Ablation Study

Semantic Prior Calibration. To assess the value of LLM-informed prior calibration, we compare our full design with variants that use LLM judgement only for the concentration α or allocation mean m, remove it from the prior entirely, or shuffle the entries of m across variables. As shown in Figure 4(a), calibrating either component improves over the data-only prior, while calibrating both yields further gains. The data-only variant shares the same LLM-proposed initial candidates, confirming that the improvement is not solely due to initialization. Shuffling the allocation slows early optimization, but performance eventually recovers to the data-only level, suggesting that adaptive trust can mitigate misaligned guidance. These results show that jointly calibrating the allocation mean and concentration translates scientific knowledge into an effective relevance hierarchy.

Online Reflection and Adaptive Trust. Online reflection revises LLM judgments about variable importance, while adaptive trust controls the influence of LLM-informed relevance structure on search. To confirm their contributions, we remove each component separately and then both. Ablating either component reduces final performance, with a larger drop when reflection is disabled. When both are removed, performance degrades further and variability across seeds increases substantially (Figure 4(b)). These results highlight that continually refining LLM guidance and regulating its influence allows CoRe−BO to exploit LLM capacity judiciously as observations accumulate.

Different LLMs. We further study the effect of the LLM backbone by comparing Intern-S1-241B and Intern-S1-mini (Bai et al., 2025) with the Thinking and Instruct variants of Qwen3-235B (Yang et al., 2025). As shown in Figure 4(c), CoRe−BO achieves competitive final performance with all models, showing its robustness to the choice of LLM. Intern-S1 outperforms its smaller variant, while Qwen3-Thinking performs better than Qwen3-Instruct, suggesting that larger model scale and stronger reasoning capabilities provide more precise relevance estimates for optimization. Intern-S1-241B attains the best performance, indicating that pretraining on large-scale scientific corpora provides more informative priors by better capturing task-relevant properties and mechanisms.

6 Conclusion

We introduce CoRe−BO, a framework for high-dimensional scientific design that learns and exploits a variable-importance hierarchy through competitive relevance learning. CoRe−BO turns coarse LLM judgments into a revisable relevance prior that evolves with experimental evidence, facilitating the identification of influential dimensions and directing the search toward them. Our theoretical and empirical results support the effectiveness and robustness of this design. Overall, CoRe−BO provides an interpretable and stable way to harness LLM capabilities throughout high-dimensional optimization, enabling sample-efficient experimentation and accelerating scientific discovery.

AI USE STATEMENT

In this work, we used generative AI tools to interpret results. We have not used generative AI tools to generate synthetic datasets; help develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in the writing of proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; assist with translation; clean or reformat datasets; or support qualitative or thematic data analysis. Additionally, we used generative AI tools to edit the research paper to improve readability. We have reviewed all AI-assisted work: the authors independently checked AI-assisted interpretations against the original experimental results and verified all AI-assisted edits for accuracy and consistency with the intended technical meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 5 specifies the baselines, initialization, evaluation budgets, LLM backbone, reflection frequency, and reporting protocol across independent runs. Appendix B describes the datasets, input representations, and optimization objectives. Appendix C provides implementation details, including prior configuration and concentration calibration, bounded relevance updates, posterior sampling, acquisition optimization, random seed handling, computational settings, and deterministic evidence retrieval. Appendix D provides the LLM prompt templates. Appendices E–I report additional baseline comparisons, component and initialization ablations, parameter and LLM sensitivity analyses, runtime measurements, and AutoML experiments. Section 4 states the theoretical setting and main result for the fixed-prior specialization, with detailed assumptions and complete proofs in Appendix A. The source code will be made publicly available in the final version.

REFERENCES

Dhruv Agarwal, Manoj Ghuhan Arivazhagan, Rajarshi Das, Sandesh Swamy, Sopan Khosla, and Rashmi Gangadharaiah. Searching for optimal solutions with LLMs via Bayesian optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aVfDrl7xDV.

JP Attfield, AL Kharlanov, and JA McAllister. Cation effects in doped la2cuo4 superconductors. *Nature*, 394(6689):157–159, 1998.

Lei Bai et al. Intern-S1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025. doi: 10.48550/arXiv.2508.15763.

Ilija Bogunovic and Andreas Krause. Misspecified gaussian process bandit optimization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy S. Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3004–3015. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/177db6acfe388526a4c7bff88e1feb15-Abstract.html.

George E. P. Box and R. D. Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18, 1986. doi: 10.1080/00401706.1986.10488093. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1986.10488093.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583(7815): 237–241, 2020. doi: 10.1038/s41586-020-2442-2.

Zhuo Chen, Xinzhe Yuan, Jianshu Zhang, Jinzong Dong, Ruichen Zhou, Yingchun Niu, Tianhang Zhou, Yu Yang Fredrik Liu, Yuqiang Li, Nanyang Ye, and Qinying Gu. LABO: LLM-accelerated Bayesian optimization through broad exploration and selective experimentation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.22054.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 2017. URL https://proceedings.mlr.press/v70/chowdhury17a.html.

Tim Clackson and James A Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

David Eriksson and Martin Jankowiak. High-dimensional Bayesian optimization with sparse axis-aligned subspaces. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 493–503. PMLR, 2021. URL https://proceedings.mlr.press/v161/eriksson21a.html.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D. Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5496–5507, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/6c990b7aca7bc7058f5e98ea909e924b-Abstract.html.

Ilya Grinberg, Matthew R. Suchomel, Peter K. Davies, and Andrew M. Rappe. Predicting morphotropic phase boundary locations and transition temperatures in Pb- and Bi-based perovskite solid solutions from crystal chemical data and first-principles calculations. *Journal of Applied Physics*, 98(9):094111, 2005. doi: 10.1063/1.2128049. URL https://doi.org/10.1063/1.2128049.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018. doi: 10.1016/j.commatsci.2018.07.052.

Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla Bayesian optimization performs great in high dimensions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20793–20817. PMLR, 2024. URL https://proceedings.mlr.press/v235/hvarfner24a.html.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6:161–169, 2024. doi: 10.1038/s42256-023-00788-1. URL https://www.nature.com/articles/s42256-023-00788-1.

Stef A. H. Jansen, Albert J. Markvoort, Freek V. de Graaf, Martin G. T. A. Rutten, Patricia Y. W. Dankers, Ghislaine Vantomme, Tom F. A. de Greef, and E. W. Meijer. Bayesian optimization for multicomponent supramolecular systems. *Journal of the American Chemical Society*, 147(37):33607–33614, 2025. doi: 10.1021/jacs.5c08539.

Kirthevasan Kandasamy, Jeff Schneider, and Barnabas Poczos. High dimensional Bayesian optimisation and bandits via additive models. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 295–304. PMLR, 2015. URL https://proceedings.mlr.press/v37/kandasamy15.html.

George Kim, S. V. Meschel, Philip Nash, and Wei Chen. Experimental formation enthalpies for intermetallic phases and other inorganic compounds. *Scientific Data*, 4:170162, 2017. doi: 10.1038/sdata.2017.162.

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. URL https://openreview.net/forum?id=OOxotBmGol.

Benjamin P. MacLeod, Fraser G. L. Parlane, Thomas D. Morrissey, Florian Häse, Loïc M. Roch, Kevan E. Dettelbach, Raphaell Moreira, Lars P. E. Yunker, Michael B. Rooney, Joseph R. Deeth, Veronica Lai, Gordon J. Ng, Henry Situ, Ray H. Zhang, Michael S. Elliott, Ted H. Haley, David J. Dvorak, Alán Aspuru-Guzik, Jason E. Hein, and Curtis P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6:eaaz8867, 2020. doi: 10.1126/sciadv.aaz8867. URL https://arxiv.org/abs/1906.05398.

Aria Mansouri Tehrani, Anton O. Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D. Sparks, and Jakoah Brgoch. Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society*, 140(31): 9844–9853, 2018. doi: 10.1021/jacs.8b02717.

Leonard Papenmeier, Luigi Nardi, and Matthias Poloczek. Increasing the scope as you learn: Adaptive Bayesian optimization in nested subspaces. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4b7439a4ab0b8e4bcb4e2412c6a10a58-Abstract-Conference.html.

Huadong Peng, Alexander PS Darlington, Eric J South, Hao-Hong Chen, Wei Jiang, and Rodrigo Ledesma-Amaro. A molecular toolkit of cross-feeding strains for engineering synthetic yeast communities. *Nature microbiology*, 9(3):848–863, 2024.

Florian Pfisterer, Lennart Schneider, Julia Moosbauer, Martin Binder, and Bernd Bischl. YAHPO Gym: An efficient multi-objective multi-fidelity benchmark for hyperparameter optimization. In *Proceedings of the First International Conference on Automated Machine Learning*, volume 188, pp. 3/1–39. PMLR, 2022. URL https://proceedings.mlr.press/v188/pfisterer22a.html.

G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman. Machine learning bandgaps of double perovskites. *Scientific Reports*, 6:19375, 2016. doi: 10.1038/srep19375.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 9780262182539. URL https://gaussianprocess.org/gpml/.

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*, 590:89–96, 2021. doi: 10.1038/s41586-021-03213-y. URL https://www.nature.com/articles/s41586-021-03213-y.

MG Smith, A Manthiram, J Zhou, JB Goodenough, and JT Markert. Electron-doped superconductivity at 40 k in the infinite-layer compound sr1–yndycu02. *nature*, 351(6327):549–551, 1991.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In Johannes Fürnkranz and Thorsten Joachims (eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1015–1022, Haifa, Israel, June 2010. Omnipress. URL https://icml.cc/Conferences/2010/papers/422.pdf.

Richard Suwandi, Feng Yin, Juntao Wang, Renjie Li, Tsung-Hui Chang, and Sergios Theodoridis. Adaptive kernel design for Bayesian optimization is a piece of CAKE with LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4421. URL https://papers.nips.cc/paper_files/paper/2025/hash/c03a2610bca2712b984b331fd4f7bb6f-Abstract-Conference.html.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022. URL https://arxiv.org/abs/2211.09085.

U.S. Department of Agriculture, Agricultural Research Service. USDA national nutrient database for standard reference, legacy release (SR legacy). FoodData Central, 2018. URL https://fdc.nal.usda.gov/download-datasets/.

Amanda Wang, Ryan Kingsbury, Matthew McDermott, Matthew Horton, Anubhav Jain, Shyue Ping Ong, Shyam Dwaraknath, and Kristin A. Persson. A framework for quantifying uncertainty in DFT energy corrections. *Scientific Reports*, 11:15496, 2021. doi: 10.1038/s41598-021-94550-5. URL https://www.nature.com/articles/s41598-021-94550-5.

Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando de Freitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016. doi: 10.1613/jair.4806. URL https://jair.org/index.php/jair/article/view/10983.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Xinzhe Yuan, Zhuo Chen, Jianshu Zhang, Huan Xiong, Nanyang Ye, Yuqiang Li, and Qinying Gu. Unleashing LLMs in Bayesian optimization: Preference-guided framework for scientific discovery. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2605.17976.

Chenbo Zhang and Xian Chen. FerroAI: a deep learning model for predicting phase diagrams of ferroelectric materials. *npj Computational Materials*, 11:282, 2025. doi: 10.1038/s41524-025-01778-0.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Method **3**
3.1 Preliminary . . . 3
3.2 Core Tool: Competitive Relevance Prior . . . 4
3.3 Online Relevance Revision . . . 5
3.4 Adaptive Trust in LLM Guidance . . . 6

4 Theoretical Analysis **6**

5 Experiments **7**
5.1 Experimental Setup . . . 7
5.2 Main Results . . . 7
5.3 Ablation Study . . . 9

6 Conclusion **9**

A Theoretical Details and Proofs **15**
A.1 Fixed-Prior Setting . . . 15
A.2 Confidence under Semantic Misspecification . . . 16
A.3 Information-Gain Control . . . 17
A.4 Proof of Theorem 1 . . . 17
A.5 Dirichlet Calibration Identity . . . 18

B Dataset Details **20**

C Implementation Details **23**
C.1 Model Configuration and Optimization . . . 23
C.2 Evidence Retriever Implementation Details . . . 23

D Prompt Templates **25**

E Additional Baseline Comparisons **28**
E.1 Additional Baselines . . . 28
E.2 Results with Additional Baselines . . . 28

F Core Design Ablations **29**
F.1 Semantic Prior Injection . . . 29
F.2 Competition under a Shared Relevance Budget . . . 29
F.3 Effect of LLM-Proposed Initialization . . . 30
F.4 Parameter Sensitivity . . . 30
F.5 Effect of LLM Reflection Frequency . . . 31
F.6 Evidence Retrieval for LLM Reflection . . . 32
F.7 Adaptive Trust in Semantic Guidance . . . 32
F.8 Online Reflection and Adaptive Trust on Superhard . . . 34

G LLM Robustness Analysis **34**
G.1 Different LLMs . . . 34
G.2 Robustness to Prompt . . . 34
G.3 Sensitivity to LLM Decoding Stochasticity . . . 35

H Runtime Analysis **37**

I Generalization to AutoML Tasks **38**

A Theoretical Details and Proofs

We provide the complete proof of Theorem 1. The analysis focuses on the fixed-prior specialization introduced in Section 4: a single instantiation of the initial `CoRe` prior induces a semantic kernel, which remains fixed throughout GP-UCB optimization. This isolates the effect of LLM-guided relevance from the online reflection and trust-calibration mechanisms of the full algorithm.

A.1 Fixed-Prior Setting

Let $\mathcal{X} \subseteq [0,1]^D$ be compact and let $f : \mathcal{X} \to \mathbb{R}$ be a continuous objective. A fixed relevance allocation $\rho_i = Rw_i$ induces the semantic ARD kernel

$$k_S(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\sum_{i=1}^{D} \rho_i (x_i - x_i')^2\right). \tag{10}$$

We treat k_S as fixed before the first analyzed BO round. Assume that it is continuous and positive semidefinite, and normalize it such that $k_S(\mathbf{x}, \mathbf{x}) \leq 1$. Its RKHS is denoted by $\mathcal{H}_{k_S}$.

For a radius $B > 0$, the semantic approximation error is

$$\epsilon_{\text{sem}} := \inf_{\substack{g \in \mathcal{H}_{k_S} \\ \|g\|_{\mathcal{H}_{k_S}} \leq B}} \|f - g\|_\infty. \tag{11}$$

Let $\widetilde{f} \in \mathcal{H}_{k_S}$ be a comparator attaining this error, so that

$$\|\widetilde{f}\|_{\mathcal{H}_{k_S}} \leq B, \qquad \|f - \widetilde{f}\|_\infty = \epsilon_{\text{sem}}. \tag{12}$$

Such a comparator exists for a continuous kernel on a compact domain; equivalently, Equation 12 can be taken directly as the approximation assumption.

At round t, the algorithm selects an $\mathcal{F}_{t-1}$-measurable point $\mathbf{x}_t$ and observes

$$y_t = f(\mathbf{x}_t) + \xi_t, \tag{13}$$

where the noise is conditionally σ_ξ-sub-Gaussian:

$$\mathbb{E}[\exp(u\xi_t) \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\sigma_\xi^2 u^2}{2}\right) \qquad \text{for all } u \in \mathbb{R}. \tag{14}$$

For queried points $\mathbf{x}_{1:n}$, define

$$\begin{aligned} K_n &:= [k_S(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n, \\ \mathbf{k}_n(\mathbf{x}) &:= [k_S(\mathbf{x}_1, \mathbf{x}), \ldots, k_S(\mathbf{x}_n, \mathbf{x})]^\top. \end{aligned} \tag{15}$$

With regularization $\nu > 0$, the kernel posterior quantities are

$$\begin{aligned} \mu_n(\mathbf{x}) &= \mathbf{k}_n(\mathbf{x})^\top (K_n + \nu I)^{-1} \mathbf{y}_{1:n}, \\ s_n^2(\mathbf{x}) &= k_S(\mathbf{x}, \mathbf{x}) - \mathbf{k}_n(\mathbf{x})^\top (K_n + \nu I)^{-1} \mathbf{k}_n(\mathbf{x}). \end{aligned} \tag{16}$$

The maximum information gain of k_S is

$$\gamma_n^S := \max_{\mathbf{z}_{1:n} \in \mathcal{X}^n} \frac{1}{2} \log\det\left(I + \nu^{-1} K_n(\mathbf{z}_{1:n})\right), \qquad \gamma_0^S := 0, \tag{17}$$

where $K_n(\mathbf{z}_{1:n})$ is the Gram matrix evaluated at $\mathbf{z}_{1:n}$. Under the corresponding Gaussian observation model, this log-determinant is equivalent to the mutual-information definition in Equation 8.

For $\delta \in (0,1)$, define the standard GP-UCB confidence parameter

$$\beta_t := B + \frac{\sigma_\xi}{\sqrt{\nu}} \sqrt{2\left(\gamma_{t-1}^S + \log\frac{1}{\delta}\right)}, \qquad C_\nu := \frac{2}{\log(1 + \nu^{-1})}. \tag{18}$$

To account for semantic misspecification, the analyzed UCB rule uses a valid upper bound $\bar{\epsilon} \geq \epsilon_{\text{sem}}$ and selects

$$\mathbf{x}_t \in \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{\mu_{t-1}(\mathbf{x}) + c_t s_{t-1}(\mathbf{x})\right\}, \qquad c_t := \beta_t + \frac{\bar{\epsilon}\sqrt{t-1}}{\sqrt{\nu}}. \tag{19}$$

This enlarged width follows standard misspecified GP-bandit analysis (Bogunovic & Krause, 2021). Theorem 1 considers the tight choice $\bar{\epsilon} = \epsilon_{\text{sem}}$.

Let $\mathbf{x}^\star \in \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. The cumulative regret is

$$R_T := \sum_{t=1}^{T} \left[f(\mathbf{x}^\star) - f(\mathbf{x}_t)\right]. \tag{20}$$

A.2 Confidence under Semantic Misspecification

Introduce clean pseudo-observations

$$\widetilde{y}_i := \widetilde{f}(\mathbf{x}_i) + \xi_i \tag{21}$$

and let $\widetilde{\mu}_n$ denote the posterior mean obtained by replacing $\mathbf{y}_{1:n}$ with $\widetilde{\mathbf{y}}_{1:n}$ in Equation 16. The posterior variance remains s_n^2 because it depends only on the query locations.

The standard kernelized confidence result (Chowdhury & Gopalan, 2017) gives, with probability at least $1-\delta$, simultaneously for all $t \geq 1$ and $\mathbf{x} \in \mathcal{X}$,

$$\left|\widetilde{f}(\mathbf{x}) - \widetilde{\mu}_{t-1}(\mathbf{x})\right| \leq \beta_t s_{t-1}(\mathbf{x}). \tag{22}$$

The following lemma quantifies how the semantic approximation error perturbs the posterior mean.

Lemma 2 (Posterior perturbation)**.** *For every $n \geq 0$ and $\mathbf{x} \in \mathcal{X}$,*

$$|\mu_n(\mathbf{x}) - \widetilde{\mu}_n(\mathbf{x})| \leq \frac{\epsilon_{\text{sem}}\sqrt{n}}{\sqrt{\nu}} s_n(\mathbf{x}). \tag{23}$$

Proof. Let $\mathbf{b}_n = [f(\mathbf{x}_i) - \widetilde{f}(\mathbf{x}_i)]_{i=1}^n$. Equation 12 implies $\|\mathbf{b}_n\|_2 \leq \epsilon_{\text{sem}}\sqrt{n}$, and

$$\mu_n(\mathbf{x}) - \widetilde{\mu}_n(\mathbf{x}) = \mathbf{k}_n(\mathbf{x})^\top (K_n + \nu I)^{-1}\mathbf{b}_n. \tag{24}$$

It remains to control the coefficient multiplying $\mathbf{b}_n$. Write $K_n = U\,\mathrm{diag}(\kappa_i)U^\top$ and $\mathbf{q} = U^\top \mathbf{k}_n(\mathbf{x})$. Positive semidefiniteness of the augmented Gram matrix implies $q_i = 0$ whenever $\kappa_i = 0$ and

$$k_S(\mathbf{x},\mathbf{x}) \geq \sum_{\kappa_i>0} \frac{q_i^2}{\kappa_i}. \tag{25}$$

Consequently,

$$\begin{aligned} s_n^2(\mathbf{x}) &\geq \sum_{\kappa_i>0} \frac{\nu q_i^2}{\kappa_i(\kappa_i+\nu)} \\ &\geq \nu \sum_{\kappa_i>0} \frac{q_i^2}{(\kappa_i+\nu)^2} \\ &= \nu \left\|(K_n+\nu I)^{-1}\mathbf{k}_n(\mathbf{x})\right\|_2^2. \end{aligned} \tag{26}$$

Applying Cauchy–Schwarz to Equation 24 and using Equation 26 proves the result. □

Combining Eqs. (22) and (23), together with $\bar{\epsilon} \geq \epsilon_{\text{sem}}$, gives

$$\left|\widetilde{f}(\mathbf{x}) - \mu_{t-1}(\mathbf{x})\right| \leq c_t s_{t-1}(\mathbf{x}) \tag{27}$$

simultaneously for all $t \geq 1$ and $\mathbf{x} \in \mathcal{X}$ on the same high-probability event.

A.3 Information-Gain Control

The posterior variances satisfy

$$\frac{1}{2}\sum_{t=1}^{T}\log\bigl(1+\nu^{-1}s_{t-1}^2(\mathbf{x}_t)\bigr)\le\gamma_T^S \tag{28}$$

(Srinivas et al., 2010). Since $0\le s_{t-1}^2(\mathbf{x}_t)\le 1$, concavity of $u\mapsto\log(1+u/\nu)$ implies

$$\begin{aligned}\sum_{t=1}^{T}s_{t-1}^2(\mathbf{x}_t)&\le C_\nu\gamma_T^S,\\ \sum_{t=1}^{T}s_{t-1}(\mathbf{x}_t)&\le\sqrt{C_\nu T\gamma_T^S},\\ \sum_{t=1}^{T}\sqrt{t-1}\,s_{t-1}(\mathbf{x}_t)&\le\sqrt{\frac{C_\nu T(T-1)\gamma_T^S}{2}}.\end{aligned} \tag{29}$$

The last two inequalities follow directly from Cauchy–Schwarz.

A.4 Proof of Theorem 1

Proof. We work throughout on the event in Equation 27, which holds with probability at least $1-\delta$ and is uniform over all rounds and query points. The proof has three steps. We first bound regret with respect to the best function described by the semantic kernel, then transfer this bound to the true objective, and finally control the accumulated posterior uncertainty by information gain.

First, let $\widetilde{\mathbf{x}}^\star\in\arg\max_{\mathbf{x}\in\mathcal{X}}\widetilde{f}(\mathbf{x})$ denote a maximizer of the RKHS comparator. The upper side of the confidence interval makes the UCB optimistic at this point. Moreover, $\mathbf{x}_t$ maximizes the same UCB. Combining these two facts with the lower side of the confidence interval at $\mathbf{x}_t$ gives

$$\begin{aligned}\widetilde{f}(\widetilde{\mathbf{x}}^\star)&\le\mu_{t-1}(\widetilde{\mathbf{x}}^\star)+c_t s_{t-1}(\widetilde{\mathbf{x}}^\star)\\ &\le\mu_{t-1}(\mathbf{x}_t)+c_t s_{t-1}(\mathbf{x}_t)\\ &\le\widetilde{f}(\mathbf{x}_t)+2c_t s_{t-1}(\mathbf{x}_t).\end{aligned} \tag{30}$$

The first and third lines use Equation 27; the middle inequality follows from the acquisition rule in Equation 19. Rearranging therefore yields the one-step comparator regret

$$\widetilde{f}(\widetilde{\mathbf{x}}^\star)-\widetilde{f}(\mathbf{x}_t)\le 2c_t s_{t-1}(\mathbf{x}_t). \tag{31}$$

This is the usual optimism argument, but it is deliberately applied to $\widetilde{f}$ rather than directly to f: the semantic kernel is guaranteed to describe the comparator, whereas the true objective may lie outside its RKHS ball.

We next transfer Equation 31 to the true objective. Adding and subtracting the comparator at the true maximizer and at the selected point gives

$$\begin{aligned}f(\mathbf{x}^\star)-f(\mathbf{x}_t)&=\bigl[f(\mathbf{x}^\star)-\widetilde{f}(\mathbf{x}^\star)\bigr]+\bigl[\widetilde{f}(\mathbf{x}^\star)-\widetilde{f}(\mathbf{x}_t)\bigr]\\ &\quad+\bigl[\widetilde{f}(\mathbf{x}_t)-f(\mathbf{x}_t)\bigr]\\ &\le\widetilde{f}(\mathbf{x}^\star)-\widetilde{f}(\mathbf{x}_t)+2\epsilon_{\mathrm{sem}}\\ &\le\widetilde{f}(\widetilde{\mathbf{x}}^\star)-\widetilde{f}(\mathbf{x}_t)+2\epsilon_{\mathrm{sem}}\\ &\le 2c_t s_{t-1}(\mathbf{x}_t)+2\epsilon_{\mathrm{sem}}.\end{aligned} \tag{32}$$

Here, the first inequality follows from the uniform approximation guarantee in Equation 12, and the second uses the optimality of $\widetilde{\mathbf{x}}^\star$ for $\widetilde{f}$. The additive $2\epsilon_{\mathrm{sem}}$ has a transparent origin: approximation error may be incurred once at the unknown optimum and once at the point selected by the algorithm.

It remains to sum the one-step bounds. Substituting the definition of c_t into Equation 32 gives

$$R_T\le 2\sum_{t=1}^{T}\beta_t s_{t-1}(\mathbf{x}_t)+\frac{2\bar{\epsilon}}{\sqrt{\nu}}\sum_{t=1}^{T}\sqrt{t-1}\,s_{t-1}(\mathbf{x}_t)+2\epsilon_{\mathrm{sem}}T. \tag{33}$$

The three terms have distinct roles. The first is the standard cost of exploration under k_S; the second is the additional confidence width needed because observations come from f rather than $\widetilde{f}$; and the third transfers performance from the comparator back to the true objective.

Because γ_t^S is nondecreasing, so is β_t. Hence, using Cauchy–Schwarz and the first inequality in Equation 29,

$$\begin{aligned}\sum_{t=1}^{T} \beta_t s_{t-1}(\mathbf{x}_t) &\le \beta_T \sum_{t=1}^{T} s_{t-1}(\mathbf{x}_t) \\ &\le \beta_T \sqrt{T \sum_{t=1}^{T} s_{t-1}^2(\mathbf{x}_t)} \\ &\le \beta_T \sqrt{C_\nu T \gamma_T^S}.\end{aligned} \tag{34}$$

The misspecification-weighted sum is controlled in the same way:

$$\begin{aligned}\sum_{t=1}^{T} \sqrt{t-1}\, s_{t-1}(\mathbf{x}_t) &\le \sqrt{\sum_{t=1}^{T}(t-1)} \sqrt{\sum_{t=1}^{T} s_{t-1}^2(\mathbf{x}_t)} \\ &\le \sqrt{\frac{C_\nu T(T-1)\gamma_T^S}{2}},\end{aligned} \tag{35}$$

where we used $\sum_{t=1}^{T}(t-1) = T(T-1)/2$. Substituting Eqs. (34) and (35) into Equation 33 yields the more general bound

$$R_T \le 2\beta_T \sqrt{C_\nu T \gamma_T^S} + \frac{2\bar{\epsilon}}{\sqrt{\nu}} \sqrt{\frac{C_\nu T(T-1)\gamma_T^S}{2}} + 2\epsilon_{\text{sem}} T. \tag{36}$$

Finally, Theorem 1 uses the tight confidence-width choice $\bar{\epsilon} = \epsilon_{\text{sem}}$. Since $\sqrt{T(T-1)} \le T$, Equation 36 simplifies to

$$R_T \le 2\beta_T \sqrt{C_\nu T \gamma_T^S} + \epsilon_{\text{sem}} T \left(2 + \sqrt{\frac{2C_\nu \gamma_T^S}{\nu}}\right). \tag{37}$$

The first term is the usual information-gain-controlled GP-UCB regret under the semantic kernel; every additional term is proportional to the semantic approximation error. Equation 37 therefore implies the order stated in Equation 9. Since the confidence event is simultaneous over all rounds, the conclusion holds with probability at least $1 - \delta$. This completes the proof. □

Equation 37 separates the two effects of the semantic prior. When $\epsilon_{\text{sem}} = 0$, the misspecification term vanishes and the regret is governed by the information gain of the semantic kernel. When $\epsilon_{\text{sem}} > 0$, the additional cost grows linearly with the approximation error. For a fixed nonzero error, the result does not claim asymptotic no-regret.

A.5 Dirichlet Calibration Identity

For completeness, we derive the identity used to translate the LLM-estimated number of important variables into a prior over the Dirichlet concentration. This calibration result is independent of the regret proof above.

Proposition 3 (Effective breadth of the `CoRe` allocation)**.** *Let* $\mathbf{w} \sim \text{Dirichlet}(\alpha\mathbf{m})$*, where* $\alpha > 0$ *and* $\mathbf{m}$ *lies in the interior of the probability simplex. Then*

$$\mathbb{E}[w_i] = m_i, \qquad \mathbb{E}\big[\|\mathbf{w}\|_2^2\big] = \frac{1 + \alpha\|\mathbf{m}\|_2^2}{1+\alpha}, \tag{38}$$

and hence

$$d_{\text{eff}}(\alpha, \mathbf{m}) := \frac{1}{\mathbb{E}[\|\mathbf{w}\|_2^2]} = \frac{1+\alpha}{1+\alpha\|\mathbf{m}\|_2^2}. \tag{39}$$

For any feasible target $1 < d < \|\mathbf{m}\|_2^{-2}$*, the corresponding concentration is uniquely given by*

$$\alpha(d) = \frac{d-1}{1-d\|\mathbf{m}\|_2^2}. \tag{40}$$

Proof. The second moment of a Dirichlet coordinate is

$$\mathbb{E}[w_i^2] = \frac{\alpha m_i(\alpha m_i + 1)}{\alpha(\alpha+1)} = \frac{m_i + \alpha m_i^2}{1+\alpha}. \tag{41}$$

Summing over i and using $\sum_i m_i = 1$ gives

$$\mathbb{E}[\|\mathbf{w}\|_2^2] = \frac{1+\alpha\|\mathbf{m}\|_2^2}{1+\alpha}. \tag{42}$$

Taking the reciprocal proves Equation 39. Solving $d = (1+\alpha)/(1+\alpha\|\mathbf{m}\|_2^2)$ for α gives Equation 40. The stated range makes both its numerator and denominator positive, and strict monotonicity of d_{eff} in α ensures uniqueness. □

For a feasible LLM range $[d_L, d_U]$, define $\alpha_L = \alpha(d_L)$ and $\alpha_U = \alpha(d_U)$. Matching these endpoints to the 2.5% and 97.5% quantiles of $\log\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ gives

$$\mu_\alpha = \frac{\log\alpha_L + \log\alpha_U}{2}, \qquad \sigma_\alpha = \frac{\log\alpha_U - \log\alpha_L}{2\Phi^{-1}(0.975)}. \tag{43}$$

The quantity d_{eff} is used only to calibrate the Dirichlet prior; it is not identified with the information gain γ_T^S in the regret analysis.

B DATASET DETAILS

This section provides detailed descriptions of the six experiments used in this paper. The selected datasets cover materials science, solid state physics, thermochemistry, and nutrition. Each task focuses on optimizing a particular material or system by adjusting its composition or formulation. The tasks are presented in increasing order of dimensionality, with representative input and output variables summarized in the accompanying tables.

Perovskite Band Gap (Pilania et al., 2016). This task aims to maximize the electronic band gap of double perovskites by adjusting their elemental composition. The dataset contains 1,306 compositions with band gaps calculated using the GLLB-SC potential in GPAW. Each composition is described by the atomic fractions of 29 elements, which vary continuously in [0,1]. The output is the electronic band gap in eV.

Table 1: Input and Output variables for Perovskite Band Gap task

Input/Output	Variable	Description
Input	Ba	Barium atomic fraction
Input	Ti	Titanium atomic fraction
Input	La	Lanthanum atomic fraction
Input	O	Oxygen atomic fraction
Input	Na	Sodium atomic fraction
Input	K	Potassium atomic fraction
Input	Ca	Calcium atomic fraction
Input	Sr	Strontium atomic fraction
Input	Nb	Niobium atomic fraction
Input	Pb	Lead atomic fraction
⋮	⋮	⋮
Output	Band gap	Electronic band gap in eV

Superhard Shear Modulus (Mansouri Tehrani et al., 2018). This task aims to identify materials with greater resistance to shear deformation by maximizing their shear modulus. The dataset contains 2,574 materials and defines a search space of 62 composition variables. These variables specify the atomic proportions of the constituent elements and are bounded between 0 and 1. The output is shear modulus in GPa, with larger values indicating greater resistance to shear deformation.

Table 2: Input and Output variables for Superhard Shear Modulus task

Input/Output	Variable	Description
Input	B	Boron atomic fraction
Input	C	Carbon atomic fraction
Input	N	Nitrogen atomic fraction
Input	Si	Silicon atomic fraction
Input	Al	Aluminum atomic fraction
Input	Ti	Titanium atomic fraction
Input	Cr	Chromium atomic fraction
Input	Mo	Molybdenum atomic fraction
Input	W	Tungsten atomic fraction
Input	Fe	Iron atomic fraction
⋮	⋮	⋮
Output	Shear modulus	Resistance to shear deformation in GPa

Formation Stability (Kim et al., 2017). This task aims to improve thermochemical stability by optimizing elemental composition. We use the `expt_formation_enthalpy_kingsbury` dataset distributed through Matminer, which contains experimental standard formation enthalpies for 2,135 solid compounds. The data were compiled primarily from Kim et al., Kubaschewski, and the NIST-JANAF tables and deduplicated to retain one formation enthalpy per material; the compilation procedure is described by Wang et al. (2021). The inputs describe the relative atomic content of 79 elements on a continuous scale in [0,1]. The objective is negative formation enthalpy in eV per atom, so larger values favor more exothermic formation from the constituent elements. This quantity measures formation energetics relative to the elements rather than stability against all competing compounds.

Table 3: Input and Output variables for Formation Stability task

Input/Output	Variable	Description
Input	Al	Aluminum atomic fraction
Input	Fe	Iron atomic fraction
Input	Ni	Nickel atomic fraction
Input	Cu	Copper atomic fraction
Input	Mg	Magnesium atomic fraction
Input	Ca	Calcium atomic fraction
Input	Ti	Titanium atomic fraction
Input	O	Oxygen atomic fraction
Input	S	Sulfur atomic fraction
Input	Li	Lithium atomic fraction
⋮	⋮	⋮
Output	$-\Delta H_{\mathrm{f}}$	Negative formation enthalpy in eV per atom

Superconductor (Hamidieh, 2018). This task aims to maximize superconducting critical temperature by adjusting elemental composition. The UCI Superconductor dataset contains 21,263 records represented by 86 elemental coefficients. Column scaling maps the coefficients to the unit interval, yielding continuous inputs that need not sum to one. The output is superconducting critical temperature in kelvin.

Table 4: Input and Output variables for Superconductor task

Input/Output	Variable	Description
Input	Cu	Scaled copper coefficient
Input	Ba	Scaled barium coefficient
Input	Y	Scaled yttrium coefficient
Input	O	Scaled oxygen coefficient
Input	Fe	Scaled iron coefficient
Input	La	Scaled lanthanum coefficient
Input	Sr	Scaled strontium coefficient
Input	Ca	Scaled calcium coefficient
Input	Bi	Scaled bismuth coefficient
Input	Nb	Scaled niobium coefficient
⋮	⋮	⋮
Output	T_{c}	Superconducting critical temperature in K

Ferroelectric (Zhang & Chen, 2025). We optimize 118 elemental weights ($x \in [0,1]^{118}$) to maximize the predicted symmetry-breaking transition temperature. Before evaluation, the weights are normalized into atomic fractions.

Table 5: Input and Output variables for Ferroelectric task

Input/Output	Variable	Description
Input	Ba	Barium weight before normalization
Input	Ti	Titanium weight before normalization
Input	Pb	Lead weight before normalization
Input	O	Oxygen weight before normalization
Input	Sr	Strontium weight before normalization
Input	Ca	Calcium weight before normalization
Input	Zr	Zirconium weight before normalization
Input	Nb	Niobium weight before normalization
Input	K	Potassium weight before normalization
Input	Na	Sodium weight before normalization
⋮	⋮	⋮
Output	Transition temperature	Temperature marking departure from the cubic phase

Nutrition (U.S. Department of Agriculture, Agricultural Research Service, 2018). This task aims to maximize dietary quality by adjusting 150 ingredient quantities under a 700 kcal energy budget. The foods are selected from the USDA Standard Reference Legacy release and span 23 categories. Ingredient quantities are expressed as continuous variables on a normalized scale. Nutrient amounts are recorded per 100 grams, with missing entries filled with zero. The dietary score rewards adequacy targets of 30 g protein, 12 g fiber, 500 mg calcium, 7 mg iron, 1,400 mg potassium, 45 mg vitamin C, 300 μg vitamin A in retinol activity equivalents, and 140 mg magnesium. It also penalizes intake exceeding 900 mg sodium, 8 g saturated fat, and 30 g total sugars.

Table 6: Input and Output variables for Nutrition task

Input/Output	Variable	Description
Input	Beet greens	Amount of cooked beet greens
Input	Chili with beans	Amount of canned chili with beans
Input	Walnuts	Amount of English walnuts
Input	Sardines	Amount of sardines canned in tomato sauce
Input	Barley malt flour	Amount of barley malt flour
Input	Lemon juice	Amount of lemon juice from concentrate
Input	Curry powder	Amount of curry powder
Input	White chocolate	Amount of white chocolate
Input	Apple pie	Amount of commercially prepared Dutch apple pie
Input	Egg drop soup	Amount of egg drop soup
⋮	⋮	⋮
Output	Dietary score	Nutrient adequacy balanced against excess intake and energy

C Implementation Details

C.1 Model Configuration and Optimization

Prior configuration and calibration. Across all tasks, we use $\log R \sim \mathcal{N}(\log 20, 1)$ and infer R jointly with the remaining GP hyperparameters. Initial relevance scores are transformed using a softmax temperature $\tau_s = 1$ and mixed with the uniform allocation using $\lambda = 0.5$. For concentration calibration, we intersect the LLM-provided interval with $[1 + \epsilon, S_m^{-1} - \epsilon]$, where $S_m = \|\mathbf{m}_0\|_2^2$ and $\epsilon = 10^{-3}$. The mapped concentration endpoints are clipped to $[10^{-3}, 10^3]$ before log-normal quantile matching, with α_L and α_U matched to the 2.5% and 97.5% quantiles, respectively. Missing, invalid, or degenerate intervals, including those resulting from clipping, trigger the fallback prior $\log \alpha \sim \mathcal{N}(\log D, 1)$, while retaining $\mathbf{m}_0$. Fallback events and their reasons are logged and account for 0.1% of concentration-calibration attempts.

Semantic revision. We use the multiplicative sequential update $m_{t,i} \propto m_{t-1,i} \exp(\eta_t z_{t,i})$, normalized to sum to one, with a base step size $\eta = 0.5$ and ordinal revisions $z_{t,i} \in \{-2, -1, 0, 1, 2\}$. The step size η_t is reduced by bisection when necessary so that the resulting update satisfies the fixed KL-divergence cap $D_{\mathrm{KL}}(\mathbf{m}_t \| \mathbf{m}_{t-1}) \leq 0.5$.

Posterior inference. We refit the GPs at each BO iteration using Pyro NUTS through BoTorch's `fit_fully_bayesian_model_nuts`, with 256 warm-up steps and 128 post-warm-up samples. Thinning by a factor of 16 retains eight posterior hyperparameter samples. We use a full mass matrix, a maximum tree depth of six, and disable JIT compilation. All runs use double precision on CPU with one thread per run. The random seed is reset at each iteration to $100003s + n$, where s is the run seed and n is the current number of observations.

Acquisition optimization. We use `qLogExpectedImprovement` with $q = 1$ and optimize over $[0, 1]^D$ using `optimize_acqf`, with eight restarts and 512 raw initialization samples. The trust weights in Equation (7) are applied on the original qEI scale. We optimize the logarithm of this weighted sum, computed via log-sum-exp from the two Log-qEI outputs for numerical stability.

C.2 Evidence Retriever Implementation Details

Directly including the full optimization history in the prompt can bury query-relevant contrasts among many observations and coordinates, making it harder for the LLM to identify and compare evidence about the queried variables. We therefore use a deterministic retriever to construct a compact evidence view for each reflection. It selects observation pairs aligned with the LLM query, and retains the dimensions needed to interpret these contrasts.

From queries to observation pairs. A query specifies its type and target variables. For `VERIFY_DIM` and `VERIFY_GROUP`, let S denote the queried variable or variable group. Given normalized historical inputs $\mathbf{x}_a, \mathbf{x}_b \in [0, 1]^D$, we measure target and residual variation as

$$C_S(a, b) = \sqrt{\frac{1}{|S|} \sum_{i \in S} (x_{a,i} - x_{b,i})^2}, \qquad N_S(a, b) = \sqrt{\frac{1}{D - |S|} \sum_{j \notin S} (x_{a,j} - x_{b,j})^2}. \tag{44}$$

We partition candidate pairs into local, medium, and global variation bins according to the empirical distribution of C_S. From each nonempty bin $\mathcal{B}_s$, we retrieve

$$p_s^\star = \arg \min_{(a,b) \in \mathcal{B}_s} N_S(a, b), \qquad s \in \{L, M, G\}. \tag{45}$$

This exposes the queried variables at several variation scales while preferring smaller changes in the remaining coordinates within each scale. For `COMPARE_DIM`(i, j), we apply the same procedure separately to $S = \{i\}$ and $S = \{j\}$ and present their contrasts by scale. The pairs need not share an anchor observation. Pair selection depends only on input locations, with objective values attached afterwards for interpretation.

Combining evidence. Let E_t^{query} contain the observations belonging to the retrieved pairs. We supplement these with the current best observation E_t^{best} and the $K_R = 2$ most recent observations,

$$E_t = E_t^{\text{query}} \cup E_t^{\text{best}} \cup E_t^{\text{recent}}, \qquad E_t^{\text{recent}} = \{(\mathbf{x}_n, y_n)\}_{n=N_t-K_R+1}^{N_t}, \tag{46}$$

where N_t is the number of available observations at reflection t. The best observation provides a reference for the optimization objective, while the recent observations preserve context from the latest evaluations. Overlapping observations are deduplicated before constructing the evidence view.

Selecting displayed dimensions. Within a dimension budget $B_D = 20$, we first retain every query target, then add a small number of variables with the largest current relevance weights $m_{t-1,j}$. The remaining slots are filled with previously unselected dimensions having the largest spans across E_t,

$$v_j = \max_{\mathbf{x} \in E_t} x_j - \min_{\mathbf{x} \in E_t} x_j. \tag{47}$$

The resulting set J_t therefore includes the queried variables, dimensions emphasized by the current belief, and other dimensions that vary substantially in the selected observations.

Constructing the evidence view. The LLM receives the selected observations restricted to J_t, their objective values, and the corresponding pair comparisons. To expose variation hidden by dimension selection, each pair is accompanied by two summaries over the omitted set $\bar{J}_t = \{1, \ldots, D\} \setminus J_t$,

$$r_{\text{RMS}}(a, b) = \sqrt{\frac{1}{|\bar{J}_t|} \sum_{j \in \bar{J}_t} (x_{a,j} - x_{b,j})^2}, \qquad r_{\max}(a, b) = \max_{j \in \bar{J}_t} |x_{a,j} - x_{b,j}|. \tag{48}$$

By selecting both observations and displayed dimensions, the retriever limits the amount of historical data placed in the prompt and keeps the evidence relevant to the current query easy to compare. The omitted-variation summaries preserve context for interpreting these contrasts, helping the LLM assess relevance without treating observational differences as isolated variable effects. Retrieval and GP inference continue to use the full dimensional observations.

D PROMPT TEMPLATES

We now provide the prompt templates for semantic initialization and online reflection in CoRe-BO. A shared system prompt specifies the LLM's role and response requirements, while task-specific user prompts provide the scientific context, variable descriptions, and available evidence. The initialization prompt elicits relative relevance scores and an estimate of the number of influential variables. The reflection prompt revisits these judgments as observations accumulate. Angle brackets denote placeholders that are filled before each call. The JSON examples illustrate the response structure, with placeholders replaced by values of the indicated types.

Shared system prompt. The system prompt asks the LLM to assess how strongly each variable may affect the objective over its allowed range. It distinguishes variable relevance from the direction of a favorable change and requires structured responses with stable variable identifiers. The same instructions apply across all tasks.

```
You are a scientific assistant supporting Bayesian optimization.
Assess the relative relevance of the supplied variables to the
objective using their scientific meaning and the task context.
Relevance concerns how strongly a variable can affect the objective
across its allowed range. It does not indicate whether increasing
that variable improves the objective.

Use only the supplied variable identifiers. Treat observations as
read-only evidence and do not invent measurements. Distinguish
scientific hypotheses from conclusions supported by observations.

Return only a valid JSON object matching the requested structure.
Use double quotes, with no markdown, comments, or trailing commas.
Keep explanations brief and place them only in the requested fields.
Do not return objective predictions, experimental candidates, GP
lengthscales, or a numerical trust weight.
```

Semantic initialization. Before observing optimization outcomes, the LLM receives the task description and the names, meanings, units, ranges, and constraints of all D variables. It returns a relative score $s_i \in \{1, \dots, 10\}$ for each variable, an interval $[d_L, d_U]$ for the number of influential variables, and an initial evidence query q_1. These outputs parameterize the initial relevance prior through the numerical transformations defined in the method. The LLM is not asked to calibrate absolute kernel lengthscales.

```
Scientific task
<task description and relevant scientific background>

Optimization objective
<target property, units, and optimization direction>

Design variables
<ordered list of variable IDs, names, meanings, units, and ranges>

Feasibility constraints
<composition constraints and other restrictions, if any>

No optimization outcomes are available at this stage.

Assess the relative influence of every variable on the objective.
Assign an integer score from 1 to 10, where a larger score indicates
greater expected relevance. Allow ties and weak effects. A low score
does not declare a variable irrelevant.

Estimate a plausible interval [d_L, d_U] for the number of variables
with substantial influence. Require 1 <= d_L <= d_U <= D.
```

```
Identify one uncertainty in this assessment to examine once
observations become available. Select one query type.
VERIFY_DIM examines one variable.
COMPARE_DIM compares two variables.
VERIFY_GROUP examines a scientifically related group.
The query requests evidence from observations, not a new experiment.

Return this JSON structure. Include every variable exactly once in
scores, in the supplied order, and select one query type.
{
  "scores": [
    {"variable": "<variable ID>", "score": <integer from 1 to 10>}
  ],
  "influential_count": [<d_L>, <d_U>],
  "rationale": "<brief scientific basis and main uncertainty>",
  "next_query": {
    "type": "<VERIFY_DIM or COMPARE_DIM or VERIFY_GROUP>",
    "targets": ["<variable ID>"],
    "reason": "<relevance uncertainty to examine>"
  }
}
```

Online reflection. At each scheduled reflection, the user prompt combines the original scientific context $\mathcal{K}$, the current relevance belief $\mathbf{m}_{t-1}$, the historical evidence summary M_{t-1}, and the retrieved evidence view $\widetilde{E}_t$. The view displays a subset of dimensions J_t and includes the root mean square and maximum changes among omitted dimensions. These summaries help the LLM recognize when changes outside the displayed coordinates complicate interpretation. Dimension compression affects only the information shown to the LLM. Numerical inference continues to use the full observations.

Meanwhile, since each query provides only a local view of the optimization history, the LLM maintains a compact summary M_t of accumulated evidence and empirical lessons as a proxy for the global context. This summary and the relevance belief are updated jointly at each reflection, allowing new local evidence to be interpreted against the broader optimization history.

The response contains sparse ordinal revisions $z_{t,i} \in \{-2, -1, 0, 1, 2\}$, an updated summary M_t, and the next query q_{t+1}. Nonzero revisions are restricted to displayed dimensions. The default is to preserve the current belief when the evidence is insufficient. The summary records supported, downweighted, mixed, and underexplored relevance judgments within a fixed text budget.

```
Scientific context
<task description, objective, variable semantics, and constraints>

Current relevance belief
<variable IDs and their current relevance weights m_previous>

Historical evidence summary
<supported, downweighted, mixed, and underexplored judgments>

Current evidence query
<query type, target variable IDs, and scientific reason>

Visible dimensions
<set J and the meanings of its variables>

Critical evidence
<query-relevant observations, the two most recent observations,
and the current best observation,
showing objective values and a selected subset of input dimensions>
<query-aligned comparisons across available perturbation scales>
<root mean square and maximum changes among omitted dimensions per
    pair>
```

```
Summary budget
<maximum summary length>

Reconsider the queried variables in the following order.
1. Review their scientific roles and the current relevance belief.
2. Assess whether the evidence supports, challenges, or leaves
   those judgments unresolved.
3. Consider variation in other variables, including omitted ones,
   before attributing an objective difference to a queried variable.
4. Revise only when the evidence provides a credible reason to
   reconsider the scientific judgment. Otherwise preserve it.

Use the following ordinal revisions.
-2 strongly decreases relevance.
-1 moderately decreases relevance.
 0 preserves the current judgment.
 1 moderately increases relevance.
 2 strongly increases relevance.

Return only nonzero revisions, using displayed variable IDs only.
Unlisted variables receive revision 0. Return an empty revisions
array if no change is justified. Do not output a new weight vector.
A large objective difference alone does not justify a revision.
Conflicting or strongly confounded evidence generally calls for
preserving the belief and requesting further evidence.

Update the compact evidence summary within the supplied budget.
Retain useful earlier findings, revise outdated judgments, and
record unresolved questions. Do not reproduce full observations or
numerical relevance weights in the summary.

Request one next evidence query using VERIFY_DIM, COMPARE_DIM, or
VERIFY_GROUP. It may concern a variable not currently displayed.
Select target variable IDs, not observation IDs or new candidates.

Return this JSON structure. Repeat revision entries as needed.
{
  "revisions": [
    {
      "variable": "<displayed variable ID>",
      "change": <one of -2, -1, 1, 2>,
      "reason": "<brief scientific and evidential justification>"
    }
  ],
  "summary": {
    "supported": ["<brief relevance finding>"],
    "downweighted": ["<brief relevance finding>"],
    "mixed": ["<brief unresolved conflict>"],
    "underexplored": ["<brief evidence gap>"]
  },
  "next_query": {
    "type": "<VERIFY_DIM or COMPARE_DIM or VERIFY_GROUP>",
    "targets": ["<variable ID>"],
    "reason": "<relevance uncertainty to examine next>"
  }
}
```

E Additional Baseline Comparisons

We provide additional comparisons of CoRe-BO with representative high-dimensional BO and LLM-assisted BO methods. All methods use the same task oracles and 80-iteration budget as in the main experiments. We retain method-specific surrogate models and acquisition rules where these are part of the original design; otherwise, we use a GP surrogate and the Log-qEI acquisition function. LLM-assisted methods use 20 Intern-S1-proposed points, and Add-GP uses 20 Sobol points. BAxUS follows its native initialization, lifting a Latin-hypercube design from an initial low-dimensional target space through its sparse random embedding (Papenmeier et al., 2022).

E.1 Additional Baselines

LLM-assisted BO. LLAMBO (Liu et al., 2024) uses in-context learning for surrogate modeling, warm-start initialization, and candidate generation, allowing task semantics and observed evaluations to inform search. LABO (Chen et al., 2026) treats LLM predictions as an auxiliary low-fidelity information source and uses a gating rule to decide when real-fidelity evaluation is warranted. CAKE (Suwandi et al., 2025) uses LLM-driven kernel evolution to adapt the GP covariance structure to the observed optimization history. Together with BOPRO and LGBO above, these methods cover the principal ways recent work injects LLM knowledge into BO: initialization and candidate generation, implicit priors, regional preferences, auxiliary fidelity, and adaptive kernel design.

High-dimensional BO. Add-GP (Kandasamy et al., 2015) models the objective as a sum of lower-dimensional components, thereby reducing the dimension of each constituent optimization problem. BAxUS (Papenmeier et al., 2022) combines trust-region search with a sparse random embedding and expands the target dimension when progress stalls, reusing observations across nested subspaces. These baselines complement SAASBO, TuRBO, and dimension-scaled Vanilla BO by testing additive and embedding-based structural assumptions.

E.2 Results with Additional Baselines

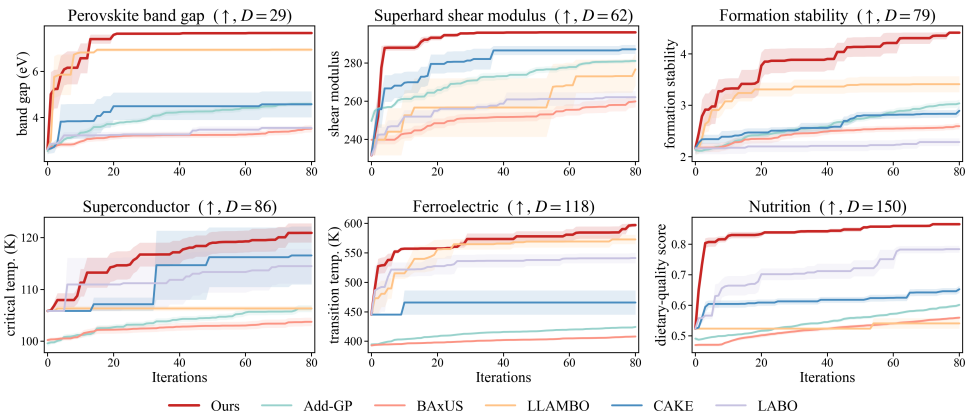


Figure 5: Comparison with additional LLM-assisted and high-dimensional BO baselines. Curves show the mean incumbent, and shaded regions indicate one standard error over 5 independent seeds.

Figure 5 shows that CoRe-BO converges faster and achieves the best final performance on all six tasks. Individual baselines can be competitive on selected problems: LLAMBO performs strongly on perovskite, formation stability, and ferroelectric optimization, CAKE on superhard materials and superconductors, and LABO on nutrition. However, their gains do not transfer consistently across tasks. Add-GP and BAxUS also make steady progress but remain limited by fixed additive or random-subspace assumptions. These results further show that LLM-guided competitive relevance learning provides a more reliable search bias across heterogeneous scientific objectives.

F CORE DESIGN ABLATIONS

This section studies semantic prior injection, competition under a shared relevance budget, LLM-proposed initialization, parameter sensitivity, the LLM reflection interval, evidence retrieval for LLM reflection, and adaptive trust in semantic guidance. These ablations assess the rationale for our default experimental settings and the robustness of `CoRe-BO` to alternative choices.

F.1 SEMANTIC PRIOR INJECTION

Following the comparison in Section 5.3, we evaluate the same five variants on the Superhard dataset (Figure 6). The variants are constructed to isolate the two semantic signals used to parameterize the CoRe prior. "LLM α only" retains the LLM-estimated relevance breadth for calibrating the concentration prior, but replaces the allocation mean with the uniform vector $\mathbf{m} = \frac{1}{D}\mathbf{1}$. "LLM m only" retains the LLM-derived allocation mean and uses a dimension-scaled concentration prior, $\log \alpha \sim \mathcal{N}(\log D, 1)$, shared with "Data-only". This prior maintains a consistent per-dimension concentration scale across tasks. "Data-only" removes both semantic signals, using a uniform allocation mean together with the same dimension-scaled prior on α.

For "Shuffled prior", we preserve the numerical relevance values produced by the LLM but randomly reassign them across variable identities. For example, a high relevance assigned to dimension i may instead be associated with dimension j. We apply the same permutation to the initial allocation $\mathbf{m}_0$ and the per-dimension relevance outputs of every subsequent online reflection. This preserves the magnitude and temporal evolution of LLM guidance while deliberately breaking its alignment with the scientific meaning of the variables.

The full CoRe prior reaches high objective values earlier and achieves the best final performance, while shuffling the prior delays early progress. This further supports combining semantic guidance for relevance allocation and concentration to construct an effective prior.

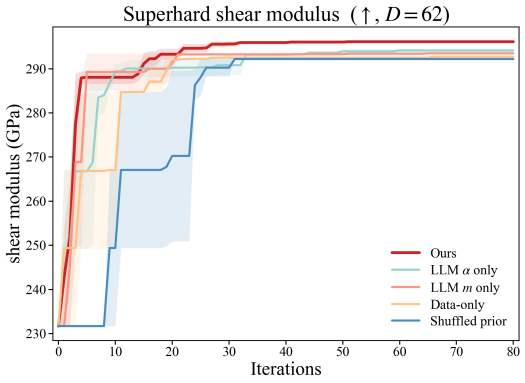


Figure 6: Ablation of semantic prior injection on the Superhard dataset.

F.2 COMPETITION UNDER A SHARED RELEVANCE BUDGET

To isolate the effect of relevance competition, we compare the full CoRe prior with a control that preserves each variable's marginal prior but removes the shared budget constraint. The full prior samples $\mathbf{w} \sim \mathrm{Dirichlet}(\alpha\mathbf{m})$ and sets $\rho_i = Rw_i$, enforcing $\sum_i \rho_i = R$. The matched no-budget variant instead samples

$$\tilde{w}_i \mid \alpha, \mathbf{m} \overset{\text{ind}}{\sim} \mathrm{Beta}(\alpha m_i, \alpha(1-m_i)), \qquad \rho_i = R\tilde{w}_i. \tag{49}$$

Conditional on α, $\mathbf{m}$, and R, both variants have identical marginal distributions for each ρ_i and the same expected total sensitivity R. The control removes the fixed-sum constraint and its induced negative dependence across dimensions, while retaining the marginal semantic information.

We choose Superhard ($D = 62$) and Ferroelectric ($D = 118$) as representative lower- and higher-dimensional tasks. We evaluate both variants using paired seeds and identical initial designs. The

initial semantic allocation, hyperpriors on α and R, LLM, reflection interval of 10 rounds, and trust mechanism are unchanged. Only the prior family differs.

The shared-budget prior yields faster convergence and substantially better final performance on both tasks (Figure 7). On Superhard, it rapidly reaches high objective values, whereas the no-budget control improves more slowly and plateaus at a markedly lower level. On Ferroelectric, the full prior continues to improve after the control has largely stalled, widening the performance gap over the optimization horizon. Because the two variants have matched marginal relevance priors, these results show that informative semantic guidance alone does not account for the gains. Coupling dimensions through a shared sensitivity budget provides an effective structure for allocating relevance and improving sample efficiency under limited observations.

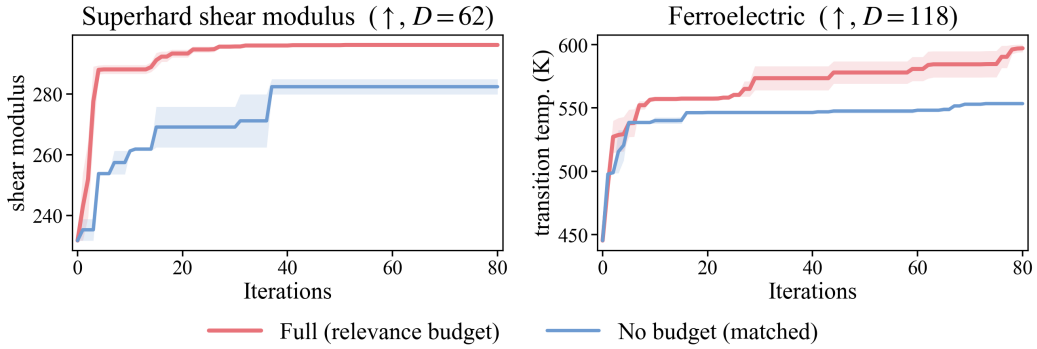


Figure 7: Effect of relevance budget competition on Superhard and Ferroelectric. The no-budget control preserves the marginal relevance priors while removing the shared budget constraint.

F.3 Effect of LLM-Proposed Initialization

To further assess whether the gains of `CoRe-BO` arise solely from LLM-proposed initial points, we initialize SAASBO and TuRBO with LLM-proposed points on Ferroelectric. Both methods improve over their standard initialization, but still fall short of `CoRe-BO` (Figure 8). This comparison shows that the benefit extends beyond initialization and supports using LLM outputs to construct a structural prior that captures the hierarchy of variable relevance.

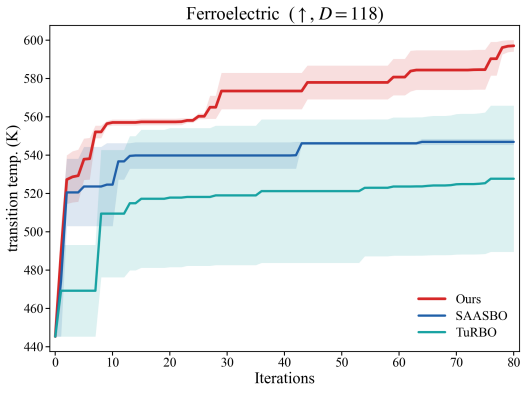


Figure 8: Comparison of CoRe-BO, SAASBO, and TuRBO with LLM-proposed initialization on Ferroelectric. Lines show the mean best objective value, and shaded regions indicate one standard error over five independent seeds.

F.4 Parameter Sensitivity

We evaluate the sensitivity of `CoRe-BO` to three parameters governing its initial prior and semantic guidance. The prior median of R sets the typical total surrogate sensitivity before observing data,

while R itself is learned through posterior inference. The coefficient λ controls the contribution of LLM-derived relevance scores to the initial allocation mean, with the remaining weight assigned uniformly across dimensions. The initial trust π_0 determines the relative weight of the semantic surrogate in acquisition decisions before trust is updated using predictive evidence. The main experiments use a prior median of 20 for R, $\lambda = 0.5$, and $\pi_0 = 0.75$.

To assess robustness to these choices, we vary one parameter at a time on the Superhard and Ferroelectric tasks, keeping the remaining settings unchanged. We consider prior medians of R in $\{5, 20, 80\}$, $\lambda \in \{0.25, 0.5, 0.9\}$, and $\pi_0 \in \{0.50, 0.75, 0.99\}$, with five seeds for every setting. Table 7 reports the area under the incumbent curve (AUC) and the final best objective value, capturing optimization efficiency and final performance, respectively.

Table 7: Parameter sensitivity on Superhard and Ferroelectric over five seeds. Higher AUC and final objective values are better. Default settings are marked with an asterisk.

Parameter	Value	Superhard		Ferroelectric	
		AUC	Final	AUC	Final
R prior median	5	294.19 ± 0.72	296.14 ± 0.30	559.57 ± 15.22	596.00 ± 5.82
	20^*	293.12 ± 1.24	296.12 ± 0.33	569.89 ± 12.45	597.12 ± 6.71
	80	292.29 ± 1.96	296.08 ± 0.38	576.88 ± 3.28	589.23 ± 4.15
λ	0.25	289.87 ± 4.38	295.75 ± 0.02	548.50 ± 8.27	573.88 ± 21.18
	0.5^*	293.12 ± 1.24	296.12 ± 0.33	569.89 ± 12.45	597.12 ± 6.71
	0.9	283.41 ± 4.80	296.48 ± 0.32	577.59 ± 9.14	595.26 ± 6.39
π_0	0.50	293.04 ± 2.21	295.94 ± 0.30	573.91 ± 16.85	595.62 ± 9.63
	0.75^*	293.12 ± 1.24	296.12 ± 0.33	569.89 ± 12.45	597.12 ± 6.71
	0.99	293.53 ± 2.48	295.87 ± 0.18	577.67 ± 16.43	597.42 ± 6.62

Performance remains broadly stable across the tested settings, particularly when varying the prior median of R or the initial trust π_0. Even a sixteenfold change in the prior median produces nearly identical final values on Superhard, while Ferroelectric retains similar overall performance. Varying π_0 also yields comparable AUC and final values on both tasks, suggesting limited dependence on the initial trust assignment.

The clearest degradation occurs at $\lambda = 0.25$ on Ferroelectric, where AUC decreases from 569.89 to 548.50 and the final value falls from 597.12 to 573.88. This suggests that overly diluting the initial semantic allocation can weaken useful guidance when observations are scarce. Giving greater weight to informative LLM judgments can therefore accelerate progress on this task. However, stronger guidance is not uniformly better. On Superhard, $\lambda = 0.9$ reduces AUC to 283.41, although the final value remains comparable to the default. Thus, the strength of initial guidance can affect convergence speed even when final performance is stable, and the default $\lambda = 0.5$ provides a balanced choice across the two tasks, reducing the effort and evaluation cost of task-specific tuning.

Overall, these results support the robustness of `CoRe-BO` to a broad range of parameter settings. The observed stability is consistent with its adaptive design, in which posterior inference learns the sensitivity scale and relevance allocation, LLM reflection revises the semantic belief, and predictive evidence calibrates trust as optimization proceeds.

F.5 Effect of LLM Reflection Frequency

The reflection interval K determines how many BO iterations occur between successive LLM updates to the relevance belief. Our default setup invokes reflection every $K = 10$ iterations over an 80-iteration optimization horizon. Below, we explain this choice and compare it with more frequent, less frequent, and no reflection.

Choice of reflection interval. We invoke LLM reflection every 10 BO iterations, allowing new evidence to accumulate while providing regular opportunities to revise semantic beliefs. Over the 80-iteration optimization horizon, this yields seven reflection steps. We use the same interval across all tasks and evaluate its effect on optimization performance below.

Frequency ablation. We evaluate $K \in \{1, 5, 10, 20, \infty\}$ on the Superhard shear modulus and Ferroelectric tasks, where $K = \infty$ denotes a static semantic belief. All variants use the same initial semantic prior and differ only in how often the LLM is allowed to revise its relevance belief. As shown in Figure 9, $K = 10$ consistently outperforms sparse ($K = 20$) or absent reflection, while remaining close to the more frequent $K = 5$ and $K = 1$ settings. This indicates diminishing returns from increasingly frequent LLM intervention. Thus, $K = 10$ provides a favorable operating point on the performance–cost trade-off, retaining most of the benefit of online revision while substantially reducing API usage and inference latency.

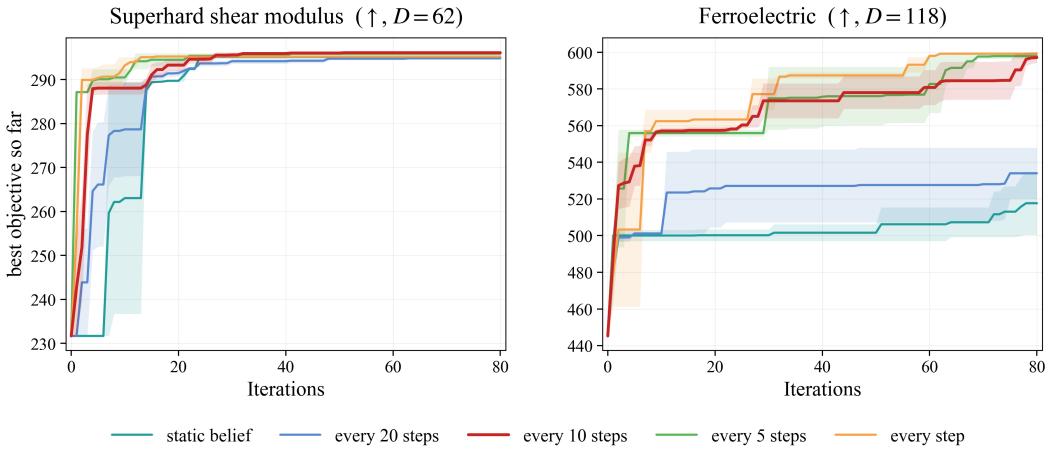


Figure 9: Effect of LLM reflection frequency on the superhard shear modulus and ferroelectric tasks. Lines show the mean best objective value found so far, and shaded regions indicate one standard error (s.e.).

F.6 Evidence Retrieval for LLM Reflection

To assess the effectiveness of active evidence querying, we compare our retriever with two alternatives. "Random Evidence" retains the LLM query but replaces the corresponding local, medium, and global comparison pairs with an equal number of random historical pairs, keeping the scale and target labels, displayed dimension budget, and prompt size unchanged. "Recent + Best" provides only the best observed point and the two most recent observations.

Our approach achieves the highest final objective values on both tasks (Figure 10). "Random Evidence" reduces performance from 296.1 to 285.3 on Superhard and from 597.1 to 559.0 on Ferroelectric, highlighting the importance of query alignment beyond context size alone. "Recent + Best" outperforms random retrieval but remains below our approach, suggesting that recency and objective quality alone do not provide sufficient evidence for effective belief revision. Together, these results support allowing the LLM to actively query evidence that helps assess and refine its relevance judgments.

F.7 Adaptive Trust in Semantic Guidance

The trust weight π_t controls the relative contribution of the semantic and data-only surrogates to acquisition decisions. In `CoRe-BO`, it is updated using their predictive likelihoods for each newly observed outcome, increasing when the semantic surrogate predicts better and decreasing otherwise.

To assess the value of adapting semantic guidance to predictive evidence, we compare our trust mechanism with two alternatives on Superhard and Ferroelectric. "Fixed trust" keeps $\pi_t = 0.75$ throughout optimization, while "Time-decay trust" linearly reduces it from 0.75 to 0.5. Both alternatives replace the adaptive trust update, with the remaining components unchanged.

Adaptive trust achieves the highest mean AUC on both tasks (Figure 11, top row). On Superhard, it reaches 293.1, compared with 288.0 for fixed trust and 290.8 for time decay. On Ferroelectric, the corresponding values are 569.9, 557.7, and 538.0. The relative ordering of the two alternatives also

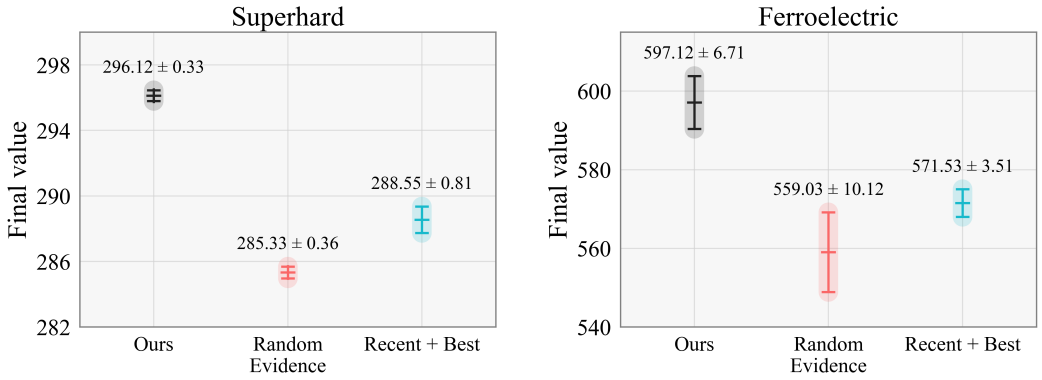


Figure 10: Ablation of evidence retrieval for LLM reflection on Superhard and Ferroelectric.

changes across tasks, suggesting that a predetermined trust schedule does not consistently capture when semantic guidance is useful. These results support calibrating its influence through predictive evidence to improve optimization efficiency.

The trajectories in the bottom row show distinct trust dynamics across tasks. On Superhard, trust drops sharply from 0.75 to around 0.4, then gradually recovers and fluctuates near 0.6, remaining below its initial level. On Ferroelectric, a brief initial dip is followed by an increase toward 0.9, with subsequent fluctuations around that level. These patterns illustrate how predictive evidence can lead to a sustained reduction or increase in semantic influence depending on the task. Neither a fixed weight nor a monotonic decay schedule captures both behaviors, supporting the use of evidence-based trust calibration.

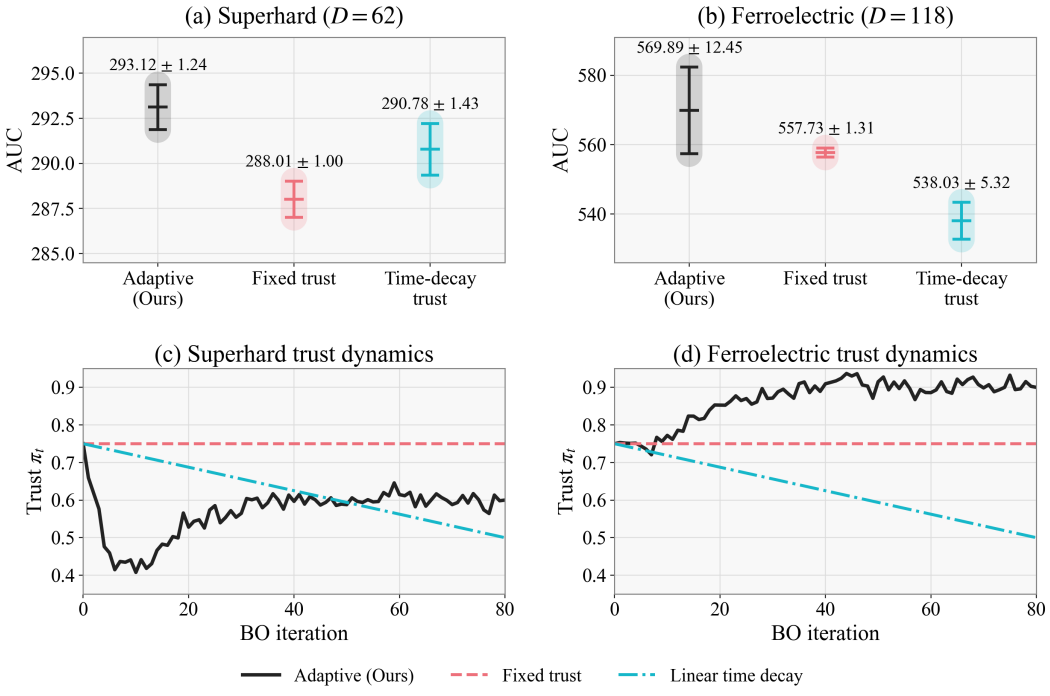


Figure 11: Trust mechanism ablation on Superhard and Ferroelectric.

F.8 Online Reflection and Adaptive Trust on Superhard

We further assess the contributions of online reflection and adaptive trust on the Superhard shear modulus dataset by removing either component or both, with five seeds per variant. Although the final objective values are similar, the AUC results reveal differences in how quickly the methods reach high-performing solutions (Figure 12). The full method achieves a mean AUC of 293.1. Removing either online reflection or adaptive trust lowers the AUC, while removing both yields the lowest mean AUC of 284.6. Thus, the benefits of these components extend beyond final performance to faster progress within the evaluation budget. Together, belief revision and trust calibration help turn semantic guidance into more sample-efficient search.

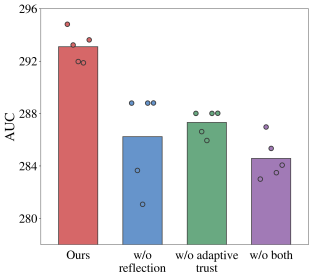


Figure 12: AUC comparison on Superhard with online reflection, adaptive trust, or both removed.

G LLM Robustness Analysis

G.1 Different LLMs

We repeat the LLM ablation on the superhard shear modulus task using the same four models as in Section 5.3. Figure 13 shows competitive final performance across all four models, further supporting the robustness of `CoRe-BO` to differences in LLM capacity and the specific structure guidance. Intern-S1 outperforms its smaller variant, while Qwen3-Thinking improves over Qwen3-Instruct. These results again suggest that greater model capacity and stronger reasoning capabilities can improve optimization performance.

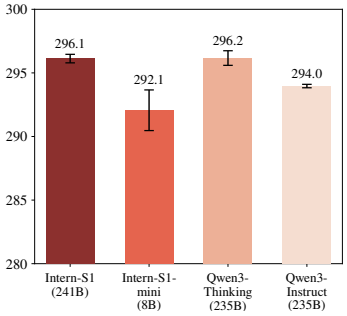


Figure 13: Effect of different LLM backbones on the superhard shear modulus task.

G.2 Robustness to Prompt

The prompt ablation tests how the instructions used to assess and revise variable relevance affect `CoRe-BO`. We independently simplify the prompts for semantic initialization and online revision, resulting in a 2×2 ablation. Across all variants, the task description, variable semantics, available experimental evidence, and output format are kept unchanged; we simplify the instructions for assessing variable relevance during initialization and revising it during optimization. The gray boxes compare selected instructions from the full prompts with the complete minimal instructions that replace them. Shared task information and output schemas are omitted from the comparison. The complete full prompts are provided in Appendix D.

Initialization: full prompt (selected excerpts)

```
Assign an integer score from 1 to 10, where a larger score indicates
greater expected relevance.  Allow ties and weak effects.  A low
score does not declare a variable irrelevant.

Identify one uncertainty in this assessment to examine once
observations become available.  Select one query type.

The query requests evidence from observations, not a new experiment.
```

Initialization: minimal prompt

```
Given the scientific task, optimization objective, and variable
descriptions, assign an importance score to each variable, estimate
a plausible interval for the number of influential variables, and
propose one relevance question for later empirical examination.
Return the required structured output.
```

Revision: full prompt (selected excerpts)

```
Reconsider the queried variables in the following order.
1.  Review their scientific roles and the current relevance belief.
2.  Assess whether the evidence supports, challenges, or leaves
those judgments unresolved.
3.  Consider variation in other variables, including omitted ones,
before attributing an objective difference to a queried variable.
4.  Revise only when the evidence provides a credible reason to
reconsider the scientific judgment.  Otherwise preserve it.

A large objective difference alone does not justify a revision.
Conflicting or strongly confounded evidence generally calls for
preserving the belief and requesting further evidence.
```

Revision: minimal prompt

```
Given the task semantics, current relevance belief, accumulated
summary, and retrieved observations, decide whether the displayed
variables should increase, decrease, or maintain their relevance.
Update the summary, propose the next evidence query, and return the
required structured output.
```

We evaluate the four prompt combinations on the Superhard shear modulus ($D = 62$) and Ferroelectric ($D = 118$) tasks. As shown in Figure 14, the full prompts achieve the best overall performance, indicating that explicit instructions for assessing and revising relevance can improve LLM guidance. Nevertheless, substantially simplified prompts preserve most of the optimization performance. On Superhard, the minimal variants retain 96.0–97.4% of the AUC and 98.5–98.9% of the final objective relative to the full configuration. On Ferroelectric, they retain 97.5–98.8% of the AUC and 94.4–96.8% of the final objective.

These results indicate that `CoRe-BO` does not rely critically on a particular detailed prompt design. More structured instructions provide a consistent benefit, while the method remains effective under much simpler prompts.

G.3 Sensitivity to LLM Decoding Stochasticity

Stochastic decoding can produce different LLM guidance even when the task and prompt remain unchanged. This experiment tests how that variation affects optimization behavior. We run `CoRe-BO` five times on the superhard shear modulus task, keeping the task, prompt, initialization procedure, BO random seeds, and surrogate and acquisition configurations fixed. Each run independently samples LLM outputs for the initial design, the initial relevance prior, and subsequent proposals during optimization. The realized initial points may therefore differ, but the source of variation is restricted to LLM generation.

Our main experiments use greedy decoding with `temperature=0` and `do_sample=False`. Repeated calls to Intern-S1 return identical outputs under this configuration in our implementation, so repeating it would yield identical trajectories. For this analysis, we instead use `do_sample=True`

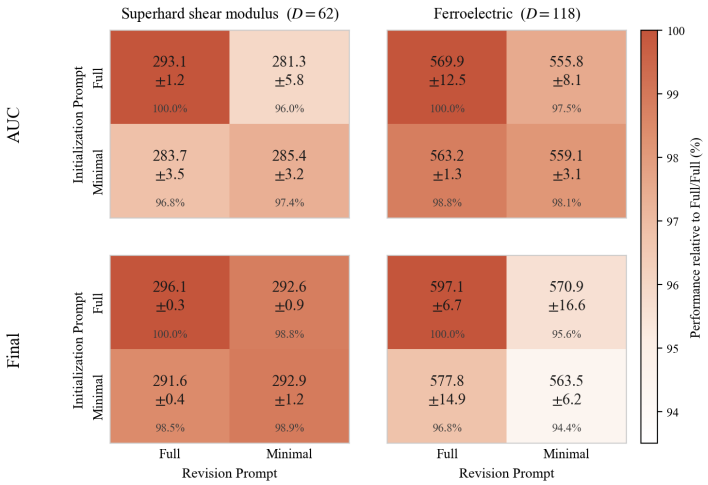


Figure 14: Prompt ablation on the Superhard shear modulus and Ferroelectric tasks. We independently replace the full initialization and revision prompts with minimal versions while preserving the same task information, experimental evidence, and output format. Cells report mean ± standard deviation; percentages and background intensity indicate performance relative to the full/full configuration for each task and metric.

and `temperature=0.7`. The experiment thus probes sensitivity to stochastic decoding and is not directly comparable with the main experiments.

Figure 15 shows that the five runs follow somewhat different early search trajectories but soon reach a similar high-performing regime and converge to a narrow range of final objective values. Their similar convergence patterns suggest that sample efficiency is largely preserved across different LLM generations. These results support the robustness of incorporating LLM knowledge through a structured, revisable relevance prior, which allows optimization to remain stable despite variation in the generated guidance.

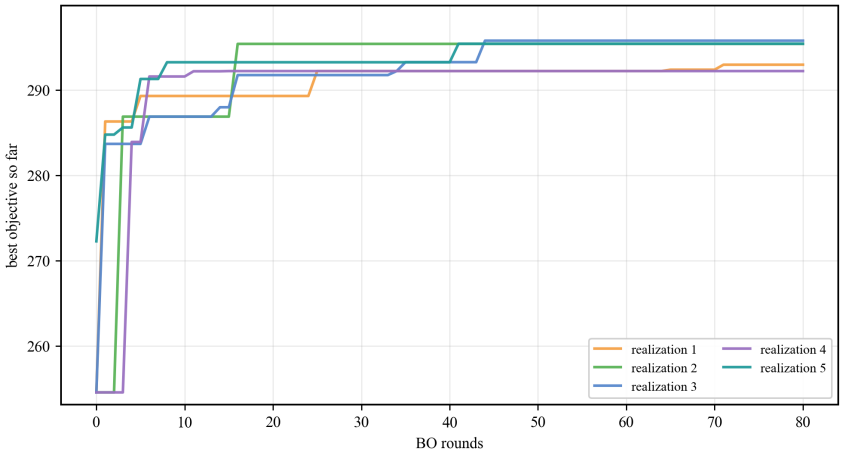


Figure 15: Sensitivity to LLM decoding stochasticity on the superhard shear modulus task. Each realization denotes one complete independent run with newly sampled LLM outputs, while all non-LLM configurations are held fixed. The runs show similar convergence behavior and final best-so-far performance.

H RUNTIME ANALYSIS

We measure the wall-clock runtime of different BO methods on the Superhard task with 62 input dimensions, using an Intel Xeon Gold 6530 CPU and 16 GB of RAM. Table 8 reports the average time per BO round, computed by dividing the total runtime of the 80 acquisition rounds by 80, excluding the initial design of 20 points. For methods that use LLMs, the reported runtime includes LLM calls during optimization.

Table 8: Average wall-clock time per BO round on Superhard, excluding initialization.

Method	Seconds per BO round
CoRe-BO	130.9
LGBO	199.9
LABO	240.2
LLAMBO	386.0
CAKE	289.1
SAASBO	81.9
Vanilla BO	3.62
TuRBO	0.57
Add-GP	26.3
BAxUS	5.68

CoRe-BO takes 130.9 seconds per round on average, with posterior inference, acquisition optimization, and LLM calls accounting for approximately 94%, 1%, and 5% of the runtime, respectively. It is faster than all other LLM-based BO methods evaluated here, which require 199.9–386.0 seconds per round. Its modest LLM overhead is consistent with the use of compact retrieved context and sparse relevance revisions, avoiding lengthy numerical outputs across all input dimensions. Most of the computational cost instead comes from posterior inference.

Although CoRe-BO is slower than the methods without LLMs, its average overhead of roughly two minutes per round can be small in scientific optimization loops dominated by lengthy physical experiments. When each wet-lab evaluation takes hours or longer, this computational cost becomes a minor part of the overall experimental cycle. In such settings, enhancing the sample efficiency and accelerating convergence can be more consequential than minimizing optimizer runtime alone.

I GENERALIZATION TO AUTOML TASKS

To assess whether CoRe-BO generalizes beyond scientific composition optimization, we evaluate it on three AutoML tasks from the YAHPO Gym rbv2_super benchmark (Pfisterer et al., 2022).

Tasks and setup. The tasks are Amazon Commerce Reviews (OpenML 1457), a 50-class authorship classification problem with 1,500 reviews; Breast-W (OpenML 15), a binary benign-versus-malignant classification problem with 699 samples; and CNAE-9 (OpenML 1468), a nine-class classification problem with 1,080 company activity descriptions. All three share a 38-dimensional mixed, hierarchical AutoML search space comprising learner selection, preprocessing, and conditional hyperparameters for six learner families: approximate nearest neighbors, elastic-net models, random forests, decision trees, SVMs, and XGBoost. This dimensionality refers to pipeline configuration, rather than the number of raw dataset features. The objective is to minimize log loss predicted by the YAHPO surrogate. All methods share 20 initial configurations, followed by 80 optimization steps.

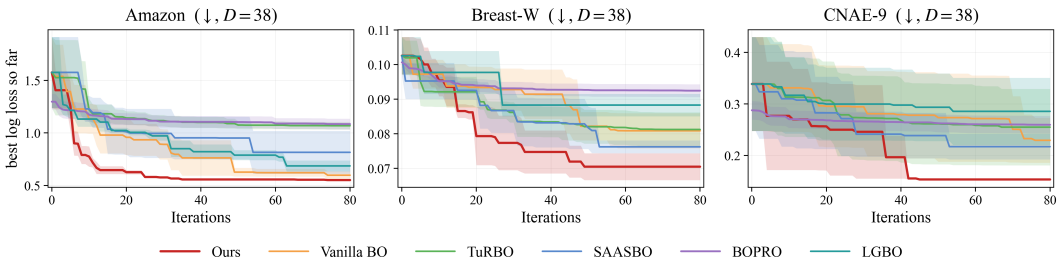


Figure 16: Optimization trajectories on three AutoML tasks with representative baselines. Lines show mean best-so-far log loss, and shaded regions indicate one standard error (s.e.). The x-axis counts the 80 optimization steps after initialization; lower is better.

Table 9: AutoML results for all methods (mean ± standard deviation). Final is the best log loss after 100 evaluations; AUC is the mean best-so-far log loss over evaluations 21–100. Lower is better for both metrics. Best means are bold.

	Amazon		Breast-W		CNAE-9	
Method	Final ↓	AUC ↓	Final ↓	AUC ↓	Final ↓	AUC ↓
CoRe-BO	$\mathbf{0.5512 \pm 0.0003}$	$\mathbf{0.6538 \pm 0.0268}$	$\mathbf{0.0705 \pm 0.0067}$	$\mathbf{0.0786 \pm 0.0059}$	$\mathbf{0.1532 \pm 0.0021}$	$\mathbf{0.2074 \pm 0.0770}$
CoRe-BO w/o LLM	0.6323 ± 0.0129	1.0092 ± 0.3378	0.0815 ± 0.0026	0.0887 ± 0.0054	0.1805 ± 0.0355	0.2525 ± 0.0773
LGBO	0.6861 ± 0.0894	0.9044 ± 0.0562	0.0883 ± 0.0056	0.0918 ± 0.0007	0.2855 ± 0.1129	0.2996 ± 0.1219
LABO	0.9288 ± 0.2333	1.0480 ± 0.1048	0.0925 ± 0.0043	0.0989 ± 0.0059	0.2586 ± 0.0751	0.2704 ± 0.0848
LLAMBO	1.0831 ± 0.5222	1.1446 ± 0.5217	0.0938 ± 0.0080	0.0990 ± 0.0063	0.2483 ± 0.0958	0.2708 ± 0.0924
CAKE	0.8526 ± 0.3490	0.9740 ± 0.2615	0.0833 ± 0.0030	0.0959 ± 0.0069	0.2138 ± 0.0672	0.2650 ± 0.1221
BOPRO	1.0840 ± 0.2287	1.1239 ± 0.2486	0.0925 ± 0.0083	0.0937 ± 0.0075	0.2594 ± 0.0764	0.2659 ± 0.0761
SAASBO	0.8152 ± 0.3540	1.0040 ± 0.3466	0.0762 ± 0.0032	0.0842 ± 0.0074	0.2170 ± 0.0428	0.2535 ± 0.0879
BAxUS	1.0263 ± 0.2034	1.1187 ± 0.1929	0.0892 ± 0.0107	0.0933 ± 0.0108	0.1766 ± 0.0438	0.2394 ± 0.0971
TuRBO	1.0683 ± 0.0607	1.1529 ± 0.1288	0.0812 ± 0.0086	0.0860 ± 0.0043	0.2549 ± 0.1285	0.2795 ± 0.1358
Add-GP	1.0037 ± 0.0770	1.0851 ± 0.0187	0.0874 ± 0.0047	0.0934 ± 0.0053	0.1916 ± 0.0255	0.2117 ± 0.0283
Vanilla BO	0.5976 ± 0.0676	0.8400 ± 0.1875	0.0809 ± 0.0088	0.0886 ± 0.0088	0.2293 ± 0.0754	0.2860 ± 0.1339
Random	1.0041 ± 0.1251	1.2055 ± 0.1951	0.0935 ± 0.0072	0.0978 ± 0.0077	0.1944 ± 0.0312	0.2416 ± 0.0597

Results. Figure 16 shows that CoRe-BO reaches low-loss configurations more quickly and maintains a strong advantage as optimization proceeds. Table 9 confirms the lowest mean Final and AUC on all three tasks among the 13 methods and variants. The full method also improves both metrics over its no-LLM variant on every task, supporting the value of semantic guidance in learning useful relevance priors for AutoML. Together, these results demonstrate the potential of competitive relevance learning beyond materials design, where the relative importance of pipeline choices and hyperparameters must also be inferred from limited evaluations.