EVIDENCE-GROUNDED BAYESIAN OPTIMIZATION WITH COLLECTIVE LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Expensive black-box optimization is challenging under tight evaluation budgets, particularly when reliable prior knowledge is limited. In such settings, context-informed warm starts can be ineffective or misleading. We propose a Bayesian optimization (BO) framework that grounds collective large language model (LLM) reasoning in experimental evidence. The framework uses design of experiments (DoE) to guide evaluations of the target system and derive statistical evidence that complements available prior information. Multiple LLM agents reason over this shared evidence, and their aggregated proposals guide search-region refinement and sequential BO. Our theoretical analysis establishes robustness properties of proposal aggregation, while experiments show that aggregated proposals achieve a lower mean normalized distance to reference optima than individual-agent proposals. On complex yet structured multimodal benchmarks with limited prior knowledge, our method reaches stringent regret targets with 8.8%–70% fewer evaluations than Gaussian process (GP)-based and LLM-assisted baselines. In FCNet hyperparameter-tuning ablations, evidence-based variants often reach top 0.5% configurations within 38 evaluations, earlier than non-evidence-based variants. These results support combining structured experimental evidence with collective LLM reasoning to improve the evaluation efficiency of black-box optimization.

1 INTRODUCTION

General Introduction. Bayesian Optimization (BO) is widely used for expensive black-box optimization when only a limited number of evaluations are available. Classical BO combines a probabilistic surrogate, often a Gaussian Process (GP), with an acquisition function that balances exploration and exploitation (Srinivas et al., 2010; Shahriari et al., 2016). Its performance, however, can depend strongly on initialization, prior information, and how the search space is represented and explored (Feurer et al., 2015; Hvarfner et al., 2025).

Many experimental optimization problems involve heterogeneous decision variables, including categorical, integer, and continuous factors, so the configuration space can grow rapidly. Design of Experiments (DoE) provides statistically efficient evaluations of such systems (Montgomery, 2017; Box & Wilson, 1951). Depending on factor types and inferential goals, the design may use full or fractional factorial runs for screening, central composite or other response-surface designs for curvature, or constrained D-optimal designs. The resulting evidence can characterize main effects, interactions, ordered trends, and local response surfaces, while subsequent optimization still determines which admissible regions should receive additional evaluations. However, identifying important factors is only part of the optimization problem: one must still decide which admissible subdomains should receive subsequent evaluations and how that budget should be allocated.

A related line of work combines classical DoE with Bayesian optimization. For example, Neumayr et al. (2026) use a fractional-factorial design to screen important variables and interactions before applying GP-based Bayesian optimization to a battery optimization problem. This demonstrates that statistically designed screening experiments can provide useful structure for subsequent BO. However, the transition from screening to optimization is still largely determined by the statistical procedure and predefined search space. Our interest is in whether the information obtained from DoE can instead be used to efficiently guide the subsequent optimization space.

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.

Large Language Models (LLMs) can complement Bayesian optimization by using semantic knowledge to interpret factor relationships and guide search. However, LLM guidance based primarily on prior knowledge may be unreliable when the target system is poorly represented in the model's pretrained knowledge, and direct LLM optimization may respond inconsistently to numerical feedback or concentrate prematurely on promising regions (Gupta et al., 2025).

We therefore use an evidence-grounded design with two initialization paths, as illustrated in Figure 1. When budget permits, Path A selects a DoE compatible with the factor types and a prespecified inferential goal, then applies design-matched statistical analysis. Multiple LLM agents interpret this evidence with limited factor descriptions, and their aggregated proposals refine the admissible search domain. When DoE is unaffordable, Path B instead uses collectively aggregated LLM proposals to initialize the search directly. Both paths subsequently enter the same sequential optimization loop, combining LLM-guided proposals, GP-based acquisition, and random exploration.

Contributions. Our contributions are the followings:

- We address unreliable LLM warm starts under scarce prior knowledge by grounding optimization decisions in statistical evidence, using only about 30% of the evaluation budget for initial target-system evaluations.
- We introduce collective LLM reasoning with aggregation mechanisms that reduce dependence on individual agents, supported by theoretical guarantees for medoid and projected-mean aggregation.
- We empirically characterize when statistical grounding is most beneficial. When prior information about the optimization objective and variable interactions is limited, our method generally overcomes its initial screening cost within 50–80% of the total evaluation budget, after which it achieves lower regret than LLM- and GP-based BO baselines and finishes with lower normalized regret on most benchmarks. Across our synthetic benchmarks, our method reaches the 1% normalized-regret target in 98/120 runs (81.7%), compared with 61–73/120 runs (50.8–60.8%) for the baselines.

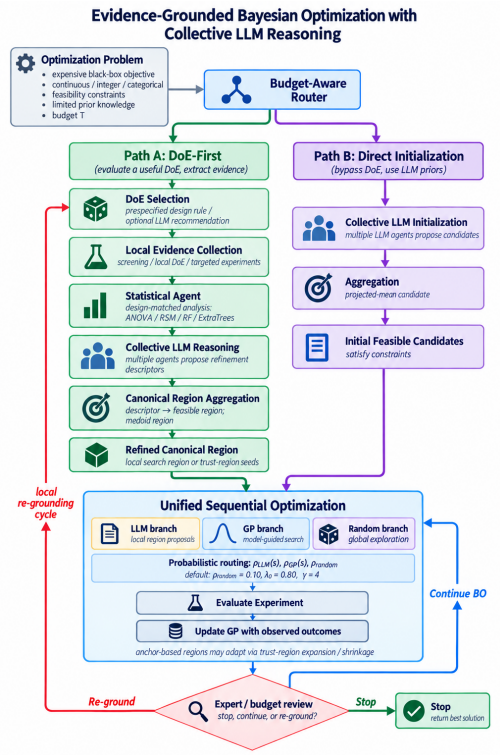


Figure 1: Evidence-grounded Bayesian optimization with collective LLM reasoning.

Related Work. Our work lies at the intersection of Bayesian optimization, Design of Experiments, LLM-assisted optimization, and sequential experimental design. Classical BO commonly uses GP surrogates and acquisition functions to optimize expensive black-box objectives (Srinivas et al., 2010; Shahriari et al., 2016). Prior work has separately studied improved initialization, surrogate modeling, and experimental design. For example, Hvarfner et al. (2025) study informed initialization for BO, while classical factorial, response-surface, and optimal designs efficiently estimate effects, interactions, trends, and curvature (Montgomery, 2017; Box & Wilson, 1951). Recent applications have also combined screening designs with subsequent GP-based optimization (Neumayr et al., 2026). Our work builds on this sequential DoE–optimization logic but introduces LLM reasoning as an intermediate mechanism for determining how the subsequent admissible domain should be refined.

A growing literature incorporates LLMs directly into optimization. OPRO (Yang et al., 2024) treats an LLM as a standalone optimizer by iteratively conditioning on previous solution–score pairs.

LLAMBO (Liu et al., 2024) uses LLMs for several components of BO, including informative initialization and candidate generation. LLINBO (Chang et al., 2025) introduces mechanisms for combining in-loop LLM proposals with statistical optimization. LGBO (Yuan et al., 2026) incorporates LLM-derived regional preferences into BO, while other approaches use language feedback or meta-level reasoning to modify optimization behavior (Austin et al., 2024; Zeng et al., 2025; van Stein & Bäck, 2025). These methods demonstrate that LLMs can provide useful prior structure, but their primary focus is not on grounding LLM reasoning in design-matched statistical evidence obtained from target-system experiments and using that evidence to adaptively refine the admissible optimization domain.

More recently related to our setting are methods that use LLMs to guide the structure or localization of the Bayesian optimization search. LGBO (Yuan et al., 2026), GOLLuM (Ranković et al., 2026), and ChemBOMAS (Han et al., 2025) integrate LLMs into Bayesian optimization through distinct mechanisms. LGBO uses LLM-derived regional preferences to guide search, GOLLuM adapts language-model representations using experimental outcomes within GP optimization, and ChemBOMAS combines multi-agent reasoning, chemistry-specific knowledge, external information, and learned predictive components to identify promising subspaces before fine-grained optimization. While these approaches demonstrate that LLMs can provide useful structural guidance, their guidance is primarily derived from pretrained knowledge, learned representations, or domain-specific information rather than statistically estimated evidence from the target system itself. This can be limiting when prior knowledge is weak, incomplete, or poorly aligned with the experimental setting. Our framework addresses this gap by grounding collective LLM reasoning in system-specific statistical evidence obtained from the target optimization problem.

2 METHOD

We optimize $f : \mathcal{X} \to \mathbb{R}$ on

$$\mathcal{X} = \left(\prod_{j=1}^{d} \mathcal{L}_j \right) \cap \mathcal{C}$$

under total evaluation budget T, where $\mathcal{C}$ denotes feasibility constraints and

$$\mathcal{F} = \{(j, \mathrm{type}_j, \mathcal{L}_j, \mathrm{desc}_j)\}_{j=1}^{d}$$

contains the factor metadata. The history $\mathcal{D}$ stores all measured configuration–response pairs, including replicates.

A budget-aware router selects either DOE-FIRST (Path A) or DIRECT-SEARCH (Path B). Path A evaluates a prespecified experimental design, extracts design-matched statistical evidence $\mathcal{E}$, and uses collective LLM reasoning to refine the feasible search region. Path B bypasses the DoE and statistical stages and directly initializes the search using aggregated LLM proposals. Both paths then enter the same probabilistic LLM–GP–random sequential optimization procedure.

2.1 PRE-SEARCH ROUTING AND DOE SELECTION

We select the DoE family *before observing responses*. The default rule in Table 2 follows standard factorial and response-surface design practice (Montgomery, 2019; Box & Wilson, 1951), space-filling design principles (McKay et al., 1979; Johnson et al., 1990; Morris & Mitchell, 1995), and optimal-design methods for constrained design spaces (Atkinson et al., 2007; Montgomery et al., 2002). An LLM may optionally recommend among eligible design families using only factor metadata, dimension, constraints, and prespecified qualitative prior structure. The user may override the default before outcomes are observed; the final choice is recorded.

Let n_{DoE} denote the smallest design judged useful under the selected rule. For model-based designs, the required regression basis must be estimable; for space-filling designs, the required run count and space-filling criterion are prespecified. The routing rule is

$$\boxed{\text{PATH A} \iff \text{a useful DoE exists with } n_{\mathrm{DoE}} \le \rho_{\mathrm{DoE}} T}$$

and otherwise PATH B is used. We use $\rho_{\mathrm{DoE}} = 0.30$ as a default budget allocation, while allowing a different prespecified value when domain knowledge or evaluation cost makes a larger or smaller initial design preferable.

Algorithm 1 Pre-Search Router

Require: $\mathcal{F}, \mathcal{C}, T, \rho_{\mathrm{DoE}}$
1: Select a prespecified design–analysis pair using Table 2
2: Determine the smallest useful design size n_{DoE}
3: **if** $n_{\mathrm{DoE}} \leq \rho_{\mathrm{DoE}} T$ **then**
4: **return** DoE-First
5: **else**
6: **return** Direct-Search
7: **end if**

Algorithm 2 Design-Adaptive Statistical Agent

Require: $\mathcal{D}_{\mathrm{DoE}}, \mathcal{F}$, selected design–analysis rule
1: **if** factorial or response-surface design **then**
2: Fit the prespecified regression basis
3: Report estimable effects, interactions, curvature, ANOVA statistics, and diagnostics
4: **else**
5: Fit the prespecified RF/ExtraTrees model
6: Report cross-validation, feature importance, local support, and top surrogate candidates
7: **end if**
8: Flag invalid or unavailable quantities
9: **return** statistical evidence $\mathcal{E}$

2.2 Design-Adaptive Statistical Agent

The Statistical Agent does not select its primary model after observing the responses. Instead, the analysis is matched to the prespecified DoE in Appendix Table 2. Factorial designs use contrast-coded regression and ANOVA; ordered multi-level factors may additionally use orthogonal polynomial contrasts; CCD-type designs use second-order response-surface models. For space-filling designs targeting nonlinear, multimodal, or higher-dimensional responses, we use Random Forest or ExtraTrees models (Breiman, 2001; Geurts et al., 2006), together with cross-validation, feature importance, local support around strong observations, and surrogate-ranked candidates.

Multiple compatible diagnostics may be reported, but the LLM interprets the resulting evidence rather than choosing which statistical method to trust.

2.3 Collective LLM Reasoning and General Region Aggregation

We represent continuous regions, observed anchors, discrete subsets, and mixed-domain proposals through a common feasible-region representation. Agent A_m receives

$$\mathcal{I}_{\mathrm{LLM}} = \mathcal{F} \oplus \mathcal{C} \oplus \mathcal{E} \oplus \mathcal{D}_{\mathrm{summary}}, \qquad q^{(m)} \in \mathcal{Q},$$

where the problem-dependent descriptor $q^{(m)}$ may specify a continuous subregion, observed DoE anchors, a feasible discrete subset, or a mixed-domain proposal. A prespecified deterministic operator

$$\Gamma : \mathcal{Q} \times (\mathcal{D}_{\mathrm{DoE}}, \mathcal{E}, \mathcal{F}, \mathcal{C}) \to 2^{\mathcal{X}}, \qquad \mathcal{R}^{(m)} = \Gamma(q^{(m)}; \mathcal{D}_{\mathrm{DoE}}, \mathcal{E}, \mathcal{F}, \mathcal{C}) \subseteq \mathcal{X}$$

maps each descriptor to a canonical feasible region. For example,

$$\Gamma(q^{(m)}) = \begin{cases} (\bigcup_k R_k^{(m)}) \cap \mathcal{X}, & \text{continuous-region proposal,} \\ (\bigcup_{i \in S^{(m)}} N_i) \cap \mathcal{X}, & \text{observed-anchor proposal,} \\ S^{(m)} \cap \mathcal{X}, & \text{discrete-subset proposal.} \end{cases}$$

Here N_i is a prespecified neighborhood of observed anchor x_i; its geometry is algorithmic rather than LLM-selected. Mixed-domain proposals are handled analogously using native factor domains.

Let ν be counting measure for finite domains, Lebesgue measure for bounded continuous domains, and their product for mixed domains. We define

$$J_\nu(\mathcal{R}, \mathcal{R}') = \frac{\nu(\mathcal{R} \cap \mathcal{R}')}{\nu(\mathcal{R} \cup \mathcal{R}')}, \qquad d_\nu(\mathcal{R}, \mathcal{R}') = 1 - J_\nu(\mathcal{R}, \mathcal{R}').$$

Regions differing only on a ν-null boundary are identified. More generally, the aggregation below may use any prespecified metric d on the canonical region space.

Algorithm 3 Canonical Region Medoid Aggregation

Require: descriptors $q^{(1:M)}$
1: **for** $m = 1, \dots, M$ **do**
2: $\quad \mathcal{R}^{(m)} \leftarrow \Gamma(q^{(m)}; \mathcal{D}_{\mathrm{DoE}}, \mathcal{E}, \mathcal{F}, \mathcal{C})$
3: **end for**
4: $m^\star \leftarrow \arg\min_m \sum_k d(\mathcal{R}^{(m)}, \mathcal{R}^{(k)})$
5: **return** $\mathcal{R}^{(m^\star)}$ and agreement A

The aggregate is the medoid

$$m^\star = \arg\min_m \sum_{k=1}^{M} d(\mathcal{R}^{(m)}, \mathcal{R}^{(k)}), \qquad \mathcal{R}^\star = \mathcal{R}^{(m^\star)}, \qquad A = \frac{1}{M} \sum_{m=1}^{M} J_\nu(\mathcal{R}^{(m)}, \mathcal{R}^\star),$$

where A measures agreement among proposals rather than correctness.

Theorem 1 (Majority-robust region medoid)**.** *Let $(\mathfrak{R}, d)$ be the metric space of admissible canonical regions and $\mathcal{R}^\dagger \in \mathfrak{R}$. If $g > M/2$ proposals satisfy $d(\mathcal{R}^{(m)}, \mathcal{R}^\dagger) \le r$, then*

$$d(\mathcal{R}^\star, \mathcal{R}^\dagger) \le \frac{M + 2g - 2}{2g - M}\, r.$$

If these events are independent with probability at least $p > 1/2$, then for any $\alpha \in (1/2, p)$,

$$\Pr\left[d(\mathcal{R}^\star, \mathcal{R}^\dagger) \le \frac{1 + 2\alpha}{2\alpha - 1} r \right] \ge 1 - e^{-2M(p-\alpha)^2}.$$

The proof is in Appendix A.1; independence is required only for the probability statement and is not implied by using different LLM agents.

2.4 Region Refinement and Probabilistic LLM–GP Search

The canonical aggregate $\mathcal{R}^\star$ defines the initial local search region. Its subsequent treatment depends on the descriptor type. Direct continuous or discrete proposals may remain fixed until the next refinement stage. For anchor-based proposals, each selected anchor initializes an algorithmic neighborhood or trust region. For normalized numerical coordinates ϕ_j,

$$\mathcal{T}_{i,s} = \left\{ x \in \mathcal{X} : |\phi_j(x_j) - \phi_j(x_{ij})| \le \tfrac{1}{2} L_{i,s} w_{j,s},\ \forall j \right\} \cap \mathcal{C}.$$

Dimension weights $w_{j,s} > 0$ are obtained through a prespecified mapping from $\mathcal{E}$ and normalized to geometric mean one; if unavailable, $w_{j,s} = 1$. Following trust-region Bayesian optimization (Eriksson et al., 2019), repeated local improvements expand $L_{i,s}$ and repeated failures shrink it. The LLM selects promising descriptors or anchors, but does not select adaptive trust-region widths.

During sequential optimization, branch probabilities follow Algorithm 4. Motivated by the early-LLM/late-GP principle of LLINBO (Chang et al., 2025), we use a polynomial decay of LLM influence with default parameters

$$\rho_0 = 0.10, \qquad \lambda_0 = 0.80, \qquad \gamma = 4.$$

Here ρ_0 preserves a fixed global-random exploration component, while $\gamma = 4$ shifts probability relatively quickly from LLM-guided search toward GP-guided search as target-system evidence accumulates. A larger ρ_0 may be prespecified for strongly multimodal or poorly characterized objectives. These values are recorded tuning choices rather than parameters inherited from LLINBO. The LLM and GP branches search the current local canonical region, whereas the random branch samples from the full feasible domain $\mathcal{X}$.

For the LLM branch, each agent proposes $x^{(m)} \in \mathcal{A}_s^{\mathrm{loc}}$. Under a fixed normalized numerical/one-hot embedding ψ,

$$\bar{z} = \frac{1}{M} \sum_{m=1}^{M} \psi(x^{(m)}), \qquad x^{\mathrm{LLM}} \in \arg\min_{x \in \mathcal{A}_s^{\mathrm{loc}}} \|\psi(x) - \bar{z}\|_2.$$

Projection enforces feasibility and the current local region.

Algorithm 4 Probabilistic LLM–GP Search

Require: $\mathcal{D}, \mathcal{X}, \mathcal{R}_s, s, T_{\mathrm{BO}}, \lambda_0, \gamma, \rho_0$
Fit GP to $\mathcal{D}$
$u_s \leftarrow s/\max\{1, T_{\mathrm{BO}} - 1\}$
$\rho_{\mathrm{random}} \leftarrow \rho_0$ ▷ default $\rho_0 = 0.10$
$\rho_{\mathrm{LLM}} \leftarrow (1 - \rho_{\mathrm{random}})\lambda_0(1 - u_s)^{\gamma}$ ▷ default $\gamma = 4$
$\rho_{\mathrm{GP}} \leftarrow 1 - \rho_{\mathrm{random}} - \rho_{\mathrm{LLM}}$
Draw $b_s \sim \mathrm{Categorical}(\rho_{\mathrm{LLM}}, \rho_{\mathrm{GP}}, \rho_{\mathrm{random}})$
if $b_s = \mathrm{GP}$ **then**
    return GP acquisition maximizer in $\mathcal{R}_s$
else if $b_s = \mathrm{random}$ **then**
    return a feasible random point from the full domain $\mathcal{X}$
else
    Obtain feasible LLM candidates in $\mathcal{R}_s$
    return their projected-mean aggregate
end if

Algorithm 5 Evidence-Grounded Bayesian Optimization with Collective LLM Reasoning

Require: $\mathcal{F}, \mathcal{C}, T, K_0, \{A_m\}_{m=1}^{M}$, GP, shared tuning parameters
$\mathcal{D} \leftarrow \emptyset, t \leftarrow 0$
$b \leftarrow$ Algorithm 1
if $b = $ DoE-First **then**
    Evaluate selected DoE and obtain $\mathcal{D}_{\mathrm{DoE}}$
    $\mathcal{E} \leftarrow$ Algorithm 2$(\mathcal{D}_{\mathrm{DoE}})$
    Obtain refinement descriptors $q^{(1:M)}$
    $\mathcal{R}^{\star} \leftarrow$ Algorithm 3$(q^{(1:M)})$
else
    $\mathcal{E} \leftarrow \emptyset$
    Evaluate aggregated LLM initialization of size K_0
    $\mathcal{R}^{\star} \leftarrow \mathcal{X}$
end if
while $t < T$ **do**
    Update the current local region $\mathcal{R}_s$ when its canonicalization rule is adaptive
    $x \leftarrow$ Algorithm 4 $(\mathcal{D}, \mathcal{X}, \mathcal{R}_s, \ldots)$
    Evaluate$(\{x\}; \mathcal{D}, t)$
end while
return the configuration with the best observed response

Theorem 2 (Projected mean aggregation)**.** *Let $\mathcal{A}_s^{\mathrm{loc}}$ be nonempty and compact and let $\psi(\mathcal{A}_s^{\mathrm{loc}})$ be compact. For feasible proposals $z_m = \psi(x^{(m)})$ and any $z^{\star} = \psi(x^{\star})$,*

$$\|\bar{z} - z^{\star}\|_2^2 = \frac{1}{M}\sum_m \|z_m - z^{\star}\|_2^2 - \frac{1}{M}\sum_m \|z_m - \bar{z}\|_2^2,$$

and exact projection satisfies

$$\|\psi(x^{\mathrm{LLM}}) - z^{\star}\|_2 \le 2\|\bar{z} - z^{\star}\|_2.$$

The result requires neither independence nor unbiasedness and bounds representation-space distance rather than objective improvement.

2.5 Overall Algorithm

Evaluate$(\mathcal{S}; \mathcal{D}, t)$ appends all measured responses and increments t by the charged evaluations.

2.6 Sequential Refinement and Budget Accounting

Initialization, DoE evaluations, sequential queries, and replicates all count toward T. An expert-approved later stage may allocate a new local DoE, reapply its prespecified statistical analysis, and regenerate refinement descriptors provided the remaining budget is sufficient. All refinements remain inside the original feasible domain, and all user interventions, budget reallocations, and parameter changes are recorded.

3 NUMERICAL STUDY

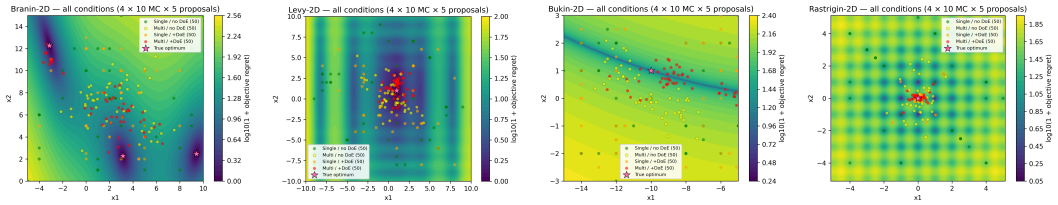


Figure 2: Initialization proposal ablation visualization across the four two-dimensional benchmark functions over 10 Monte Carlo replications. **Red** denotes multi-agent reasoning with DoE evidence, **orange** single-agent reasoning with DoE evidence, **yellow** multi-agent reasoning without DoE evidence, and **green** single-agent reasoning without DoE evidence. The **pink star** denotes the true global optimum.

Effect of Collective LLM Reasoning and DOE Evidence. We isolate the effect of collective LLM reasoning and DoE evidence on initialization quality. The seven-agent ensemble consists of Claude-Sonnet-4.6, GPT-4o-Mini, Kimi-K2.6, MiniMax-M2.5, Gemini-3.5-Flash-Lite, Qwen-Flash, and DeepSeek-V4-Pro, while the single-agent baseline uses GPT-4o-Mini. All agents receive identical information following the prompt structure in Appendix A.4 and the synthetic description cards in Table 3 or Table 4.

The first experiment is an initialization ablation that separates the contributions of multi-agent aggregation and DoE evidence. For each optimization task, the single agent or multi-agent ensemble proposes five initial points, either with or without DoE evidence. Across 10 Monte Carlo replications, Figure 2 shows the distribution of these proposals over the two-dimensional domains. The multi-agent proposals with DoE evidence are visibly more concentrated around the global optimum. Table 1 reports the corresponding quantitative results.

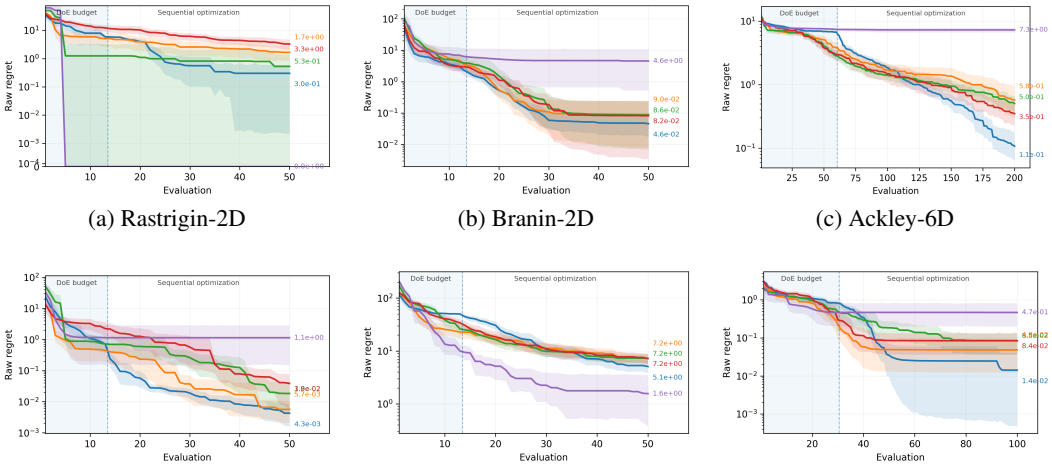
As shown in Table 1, collective reasoning and DoE evidence each improve initialization quality relative to the single-agent/no-DoE baseline. Their combination performs best, reducing the average normalized distance to the optimum to 0.1308 and average objective regret to 19.91. This suggests that statistical evidence and cross-agent aggregation provide complementary gains in initialization quality.

Condition	Avg Dist.	Regret
Single / no DoE	.3511	56.11
Multi / no DoE	.1951	35.91
Single / +DoE	.2510	46.09
Multi / +DoE	**.1308**	**19.91**

Table 1: Initialization ablation results.

In the second experiment, we further compare single-agent and collective LLM proposals under richer prior information, with a different set of LLMs, corresponding to Path B of our framework. Each agent receives the same function description provided in Table 4 and is asked to propose a single promising configuration. The collective proposal is constructed by averaging individual proposals and projecting it back onto the feasible domain. As shown in Appendix Figure 5, across $N_{\mathrm{MC}} = 100$ Monte Carlo replications, projected-mean aggregation consistently reduces the expected normalized distance to the global optimum relative to individual-agent proposals across benchmarks. This result is consistent with the broader "wisdom of crowds" principle, whereby aggregating heterogeneous judgments can improve collective accuracy and problem solving (Galton, 1907; Hong & Page, 2004).

Black-box Function Optimization. We compare our method with GP- and LLM-assisted BO baselines, including LLAMBO (Liu et al., 2024), LLINBO-transient (Chang et al., 2025), standard GP-BO (Srinivas et al., 2010), and an LLM warm-start GP baseline, on six noisy continuous benchmark functions—Levy-$2D$, Rastrigin-$2D$, Branin-$2D$, Bukin-$2D$, Hartmann-$4D$, and Ackley-$6D$. Within each benchmark, all methods use the same search domain, reference global optimum, and total evaluation budget, with additive observation noise and initialization evaluations counted toward the budget. Our method uses a 13-point hybrid DoE for all $2D$ functions, consisting of 9 CCD points and 4 maximin points, except for Rastrigin-$2D$, where we use 8 CCD points and 5 maximin points because the global optimum lies at the origin. For Hartmann-$4D$, we use a 30-point maximin Latin hypercube, while for Ackley-$6D$, we use a 60-point maximin Latin hypercube. Results are aggregated over $N_{\mathrm{MC}} = 20$ Monte Carlo runs. Figure 3 reports evolution of mean raw global simple regret.



(a) Rastrigin-2D (b) Branin-2D (c) Ackley-6D

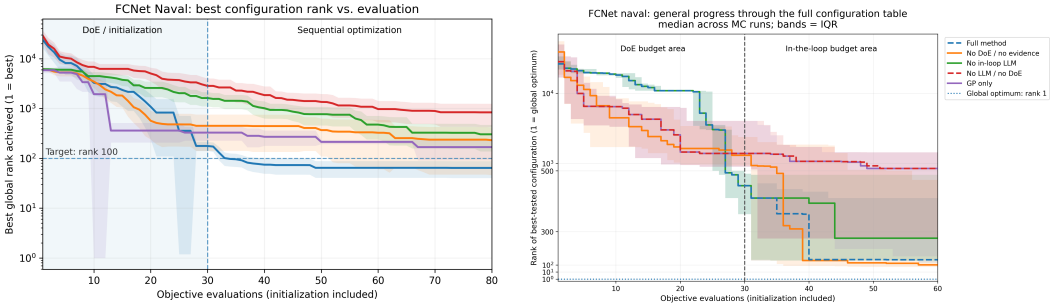
(d) Levy-2D (e) Bukin-2D (f) Hartmann-4D

Figure 3: Mean raw global simple regret across Monte Carlo runs for the six benchmark functions. The shaded region denotes the initial DoE budget for the proposed method. **Blue** denotes our method, **orange** LLINBO-Transient, **green** warm-start GP, **red** random-start GP, and **purple** LLAMBO.

Overall, our method achieves the lowest final regret on Rastrigin-$2D$, Levy-$2D$, Ackley-$6D$, and Hartmann-$4D$, while remaining competitive on Branin-$2D$ and Bukin-$2D$. The gains are especially pronounced on Ackley-$6D$ and Hartmann-$4D$, where evidence-grounded initialization followed by sequential optimization continues to reduce regret after competing methods begin to plateau.

HPO Benchmark and Component Ablation. We also evaluate our framework on FCNet Naval, a standard hyperparameter-optimization benchmark (Klein & Hutter, 2019), in two experiments reported in Figure 3. First, we compare our method against LLINBO-Transient, LLM warm-start GP, random-start GP, and LLAMBO. Our method allocates 30 evaluations to an outcome-independent D-optimal DoE before sequential optimization. Figure 3(a) shows that it reaches the Top-100 region earlier than most baselines and attains the best final configuration rank, suggesting that statistical screening provides a useful starting point when semantic prior information is limited.

Second, we perform a component ablation using a self-designed, practical 30-run block-factorial DoE: 3^2 designs for hidden-layer widths, 2^3 for activation functions and learning-rate schedule, 2^2 for batch size and initial learning rate, and 3^2 for dropout factors. The 2^k blocks provide ANOVA evidence, while the 3^2 blocks additionally support quadratic RSM and orthogonal-polynomial analysis. Figure 3(b) compares the full framework with variants that remove DoE evidence, in-loop LLM guidance, both LLM and DoE components, or rely on GP alone. The full method reaches strong configurations earlier, entering the top 0.5% after roughly 38 iterations, and continues improving beyond the DoE phase. In contrast, the ablated variants mostly plateau at worse ranks, suggesting that statistical evidence and LLM-guided refinement contribute at distinct stages of the search.



(a) Method comparison on FCNet Naval. (b) Ablation of statistical guidance and LLM components.

4 Application to Software Configuration Optimization

A typical application of our framework is software configuration optimization. We evaluate it on the x264 configurable video encoder using 1,152 measured configurations across 17 input videos (Alves Pereira et al., 2020). To create a low-prior setting, configuration names are anonymized, and the LLM receives only factor types, admissible values, and observed evaluations. Encoding time is treated as the black-box objective. All methods use the same feasible domain and total evaluation budget: 8 evaluations for warm-starting or DoE construction and 42 subsequent evaluations for sequential optimization. For this low-budget setting, we ask ChatGPT-5.5 to propose suitable DoE structures; the two variants shown are the complete set of structures it recommends.

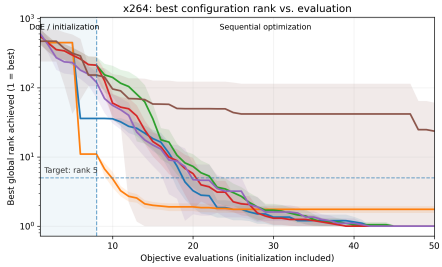


Figure 4: X264 benchmark results. The curves correspond to our 6 D-opt + 2 maximin variant (**blue**), our 5 D-opt + 3 maximin variant (**orange**), LLINBO-Transient (**green**), LLM warm-start GP (**red**), random-start GP (**purple**), and LLAMBO (**brown**).

Across 20 Monte Carlo replications, our proposed variants reach the Top-5 configuration set earlier than the main baselines. The 6 D-opt + 2 maximin variant has a mean hitting time of 17.0 evaluations, compared with 18.45 for LLINBO-Transient, 18.60 for LLM warm-start GP, and 17.50 for random-start GP. The 5 D-opt + 3 maximin variant performs substantially better, with a mean hitting time of 9.50 evaluations. LLAMBO reaches the target in only 3 of 5 runs, with a mean hitting time of 19.0 conditional on success.

A critical observation is that performance remains sensitive to the initial DoE structure. In particular, the **orange** 5 D-opt + 3 maximin variant reaches the Top-1% region very quickly but does not consistently attain the globally best configuration within the available budget. This suggests that an aggressively structured initial design may improve early localization while also introducing search bias that limits later exploration.

5 Discussion

We introduce an evidence-grounded Bayesian optimization framework combining statistical design, collective LLM reasoning, and GP–LLM sequential search. Across challenging benchmarks with limited prior knowledge, the method improves sample efficiency and robustness while preserving feasibility, highlighting the value of grounding LLM guidance in target-system evidence rather than pretrained priors alone.

A key advantage is rapid improvement after the DoE phase. When the initial design captures meaningful global and local structure, LLM agents can identify promising subregions and reduce subsequent search effort, producing a pronounced advantage shortly after the DoE budget is exhausted. The framework is also cost-efficient: with seven agents, each LLM-selected iteration requires seven HTTP requests, yielding less than 40 expected requests over a 25-step sequential stage including initial region refinement, versus roughly 550 for LLAMBO in our implementation.

Limitations Our framework is sensitive to the initial DoE structure. As shown in Section 3, CCD performs well on smooth or structured two-dimensional surfaces, where response surface modeling can recover useful information from few evaluations, but this advantage weakens on irregular or higher-dimensional landscapes. The performance is therefore strong on Branin and Levy but weaker on Bukin, whose narrow, non-smooth valley is poorly captured by a small CCD and low-order approximation. A second limitation is potential benchmark contamination in pretrained LLMs. Standard benchmark functions may have appeared in model training data or prior optimization studies, allowing an LLM to infer the objective from minimal descriptions. This may partly explain why LLAMBO repeatedly identifies the Rastrigin-2D optimum at $(0, 0)$ during warm starting.

Future Work We are interested in fine-tuning LLMs for experimental design, enabling them to recommend suitable DoE structures given budget, dimensionality, factor types, and expected response geometry. A further direction is an adaptive design-selection mechanism that chooses among factorial, fractional factorial, CCD, and other space-filling designs from early observations, improving robustness on irregular and higher-dimensional landscapes.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing, literature retrieval and discovery, research ideation and experimental implementation, generation of synthetic benchmark data and experimental artifacts, and development and verification of mathematical arguments and proofs. AI tools also assisted with notation, pseudocode, debugging, methodological alternatives, and figure and flowchart preparation. They were not used as a substitute for author judgment: the authors independently reviewed the methodology, experiments, mathematical claims, references, results, and final manuscript and retain full responsibility for the accuracy and integrity of the work.

ETHICS STATEMENT

This work studies optimization methodology using synthetic benchmarks and public benchmark datasets and does not involve human subjects or sensitive personal data. We are not aware of material privacy, fairness, or safety risks beyond the general risks associated with automated optimization systems. The authors adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide algorithmic details, benchmark settings, evaluation budgets, prompts, and implementation conventions in the main text and appendix. Complete proofs of the theoretical results are included in the appendix. Experimental configurations, Monte Carlo protocols, and baseline definitions are documented to support independent reproduction of the reported results.

REFERENCES

Juliana Alves Pereira, Mathieu Acher, Hugo Martin, and Jean-Marc Jézéquel. Sampling effect on performance prediction of configurable systems: A case study. In *Proceedings of the 2020 ACM/SPEC International Conference on Performance Engineering*, pp. 277–288. Association for Computing Machinery, 2020. doi: 10.1145/3358960.3379137.

Anthony C. Atkinson, Alexander N. Donev, and Randall D. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, 2007. doi: 10.1093/oso/9780199296590.001.0001.

David Eric Austin, Anton Korikov, Armin Toroghi, and Scott Sanner. Bayesian optimization with LLM-based acquisition functions for natural language preference elicitation, 2024. URL https://arxiv.org/abs/2405.00981.

George E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B*, 13(1):1–45, 1951. doi: 10.1111/j.2517-6161.1951.tb00067.x.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

Chih-Yu Chang, Milad Azvar, Chinedum Okwudire, and Raed Al Kontar. LLINBO: Trustworthy LLM-in-the-loop bayesian optimization. *arXiv preprint arXiv:2505.14756*, 2025. URL https://arxiv.org/abs/2505.14756.

David Eriksson, Michael Pearce, Jacob R. 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.

Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Initializing bayesian hyperparameter optimization via meta-learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1128–1135, 2015.

Francis Galton. Vox populi. *Nature*, 75(1949):450–451, 1907. doi: 10.1038/075450a0.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Rushil Gupta, Jason Hartford, and Bang Liu. LLMs for bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15482–15510, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.838. URL https://aclanthology.org/2025.findings-emnlp.838/.

Dong Han, Zhehong Ai, Pengxiang Cai, Shanya Lu, Jianpeng Chen, Zihao Ye, Shuzhou Sun, Ben Gao, Lingli Ge, Weida Wang, Xiangxin Zhou, Xihui Liu, Mao Su, Wanli Ouyang, Lei Bai, Dongzhan Zhou, Tao Xu, Yuqiang Li, and Shufei Zhang. ChemBOMAS: Accelerated BO in chemistry with LLM-enhanced multi-agent system. *arXiv preprint arXiv:2509.08736*, 2025. URL https://arxiv.org/abs/2509.08736.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004. doi: 10.1073/pnas.0403723101.

Carl Hvarfner, David Eriksson, Eytan Bakshy, and Max Balandat. Informed initialization for bayesian optimization and active learning. 2025. URL https://arxiv.org/abs/2510.23681.

Mark E. Johnson, Leslie M. Moore, and Donald Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148, 1990. doi: 10.1016/0378-3758(90)90122-B.

Aaron Klein and Frank Hutter. Tabular benchmarks for joint architecture and hyperparameter optimization. *arXiv preprint arXiv:1905.04970*, 2019. doi: 10.48550/arXiv.1905.04970. URL https://arxiv.org/abs/1905.04970.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OOxotBmGol.

Michael D. McKay, Richard J. Beckman, and William J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979. doi: 10.1080/00401706.1979.10489755.

Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 9 edition, 2017.

Douglas C. Montgomery. *Design and Analysis of Experiments*. Wiley, 10 edition, 2019.

Douglas C. Montgomery, Elvira N. Loredo, Duangporn Jearkpaporn, and Murat Caner Testik. Experimental designs for constrained regions. *Quality Engineering*, 14(4):587–601, 2002. doi: 10.1081/QEN-120003561.

Max D. Morris and Toby J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381–402, 1995. doi: 10.1016/0378-3758(94)00035-T.

Klara Neumayr, Pronoy Dutta, and Christian Prehal. Data-driven investigation of aqueous zinc anode degradation using design of experiments and bayesian optimization. *Batteries & Supercaps*, 9(6):e70347, 2026. doi: 10.1002/batt.70347.

Bojana Ranković, Ryan-Rhys Griffiths, and Philippe Schwaller. Large language models as uncertainty-calibrated optimizers for experimental discovery. *Nature Machine Intelligence*, 8:1466–1477, 2026. doi: 10.1038/s42256-026-01283-z. URL https://doi.org/10.1038/s42256-026-01283-z.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022, 2010.

Niki van Stein and Thomas Bäck. LLaMEA: A large language model evolutionary algorithm for automatically generating metaheuristics. *IEEE Transactions on Evolutionary Computation*, 29 (2):331–345, 2025. Preprint first posted in 2024 as arXiv:2405.20132.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html.

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 *International Conference on Learning Representations*, 2026. URL https://openreview.net/.

Yimeng Zeng, Natalie Maus, Haydn Thomas Jones, Jeffrey Tao, Fangping Wan, Marcelo Der Torossian Torres, Cesar de la Fuente-Nunez, Ryan Marcus, Osbert Bastani, and Jacob R. Gardner. BOLT: Large scale multi-task bayesian optimization with large language models, 2025. URL https://arxiv.org/abs/2503.08131.

A Appendix

A.1 Proof of Theorem 1

Let G denote the g proposals satisfying $d(\mathcal{R}^{(m)}, \mathcal{R}^{\dagger}) \leq r$, choose any $q \in G$, and let

$$D = d(\mathcal{R}^{\star}, \mathcal{R}^{\dagger}).$$

For every $k \in G$, the reverse triangle inequality implies

$$d(\mathcal{R}^{\star}, \mathcal{R}^{(k)}) \geq D - r.$$

By medoid optimality,

$$\sum_{k} d(\mathcal{R}^{\star}, \mathcal{R}^{(k)}) \leq \sum_{k} d(\mathcal{R}^{(q)}, \mathcal{R}^{(k)}).$$

For $k \in G$, $d(\mathcal{R}^{(q)}, \mathcal{R}^{(k)}) \leq 2r$, with the q term equal to zero. For $k \notin G$, the triangle inequality gives

$$d(\mathcal{R}^{(q)}, \mathcal{R}^{(k)}) \leq r + D + d(\mathcal{R}^{\star}, \mathcal{R}^{(k)}).$$

After cancelling the terms for $k \notin G$,

$$g(D - r) \leq 2r(g - 1) + (M - g)(D + r),$$

and therefore

$$(2g - M)D \leq (M + 2g - 2)r.$$

Since $g > M/2$, division by $2g - M$ gives the deterministic result.

For the probabilistic statement, let G_m indicate whether proposal m lies within radius r of $\mathcal{R}^{\dagger}$. Under independence with $\Pr(G_m = 1) \geq p$, Hoeffding's inequality gives, for $\alpha \in (1/2, p)$,

$$\Pr(g < \alpha M) \leq e^{-2M(p-\alpha)^2}.$$

On $g \geq \alpha M$, the deterministic coefficient satisfies

$$\frac{M + 2g - 2}{2g - M} \leq \frac{1 + 2\alpha}{2\alpha - 1},$$

which proves the stated probability bound.

A.2 Proof of Theorem 2

Proof. Since $\bar{z} = M^{-1} \sum_{m=1}^{M} z_m$, we have

$$\sum_{m=1}^{M} (z_m - \bar{z}) = 0.$$

For each m, decompose $z_m - z^\star = (z_m - \bar{z}) + (\bar{z} - z^\star)$. Expanding the squared Euclidean norm gives

$$\begin{aligned} \|z_m - z^\star\|_2^2 = \|z_m - \bar{z}\|_2^2 + \|\bar{z} - z^\star\|_2^2 \\ + 2(z_m - \bar{z})^\top (\bar{z} - z^\star). \end{aligned}$$

Averaging over m eliminates the cross term because $\sum_m (z_m - \bar{z}) = 0$. Consequently,

$$\frac{1}{M} \sum_{m=1}^{M} \|z_m - z^\star\|_2^2 = \frac{1}{M} \sum_{m=1}^{M} \|z_m - \bar{z}\|_2^2 + \|\bar{z} - z^\star\|_2^2.$$

Rearranging proves the first identity.

For the projection bound, compactness of $\psi(\mathcal{A}_r)$ ensures that an exact nearest feasible projection exists. Since $x^\star \in \mathcal{A}_r$, optimality of the projection implies

$$\|\psi(x^{\mathrm{LLM}}) - \bar{z}\|_2 \le \|\psi(x^\star) - \bar{z}\|_2 = \|z^\star - \bar{z}\|_2.$$

Applying the triangle inequality,

$$\begin{aligned} \|\psi(x^{\mathrm{LLM}}) - z^\star\|_2 &\le \|\psi(x^{\mathrm{LLM}}) - \bar{z}\|_2 + \|\bar{z} - z^\star\|_2 \\ &\le 2\|\bar{z} - z^\star\|_2. \end{aligned}$$

This proves the second inequality. Any exact minimizer may be used when the projection has ties. □

A.3 DOE Selection

Problem structure	Default DoE	Primary analysis
Finite/discrete factors; interaction screening	Full/fractional factorial	Contrast regression and ANOVA (Montgomery, 2019)
Low-d, smooth quantitative response	CCD; optional maximin augmentation	Quadratic RSM, interactions, and curvature (Box & Wilson, 1951; Montgomery, 2019)
Multimodal/oscillatory or weak shape prior	Maximin/LHS; low-d may use CCD-shell + maximin	ExtraTrees/RF with cross-validation (McKay et al., 1979; Johnson et al., 1990; Morris & Mitchell, 1995)
Moderate/high-d continuous problem	Maximin Latin hypercube	ExtraTrees/RF with cross-validation (McKay et al., 1979; Morris & Mitchell, 1995)
Irregular, mixed, or constrained domain	Constrained D-optimal design	Design-matched regression or nonlinear ensemble (Atkinson et al., 2007; Montgomery et al., 2002)

Table 2: Prespecified DoE and statistical-analysis rule. Hybrid CCD–maximin designs are an augmentation strategy combining structured response-surface points with additional space-filling points.

A.4 LLM Prompt Construction and Design Philosophy

Prompt architecture. We use three stage-specific prompt families rather than a single generic optimization prompt. Let $\oplus$ denote text concatenation. For factor j, let $\mathrm{Factor}_j(\cdot)$ encode its description and the domain information available at the current stage. The prompts have the forms

$$\mathrm{Prompt}_{\mathrm{init}} = \mathrm{Sys}_{\mathrm{init}} \oplus \mathrm{Obj} \oplus \bigoplus_{j=1}^{d} \mathrm{Factor}_j(\mathcal{L}_j) \oplus \mathcal{D}_{\mathrm{selected}} \oplus \mathrm{Out}_{\mathrm{point}},$$

$$\mathrm{Prompt}_{\mathrm{region}} = \mathrm{Sys}_{\mathrm{region}} \oplus \mathrm{Obj} \oplus \bigoplus_{j=1}^{d} \mathrm{Factor}_j(\widetilde{\mathcal{L}}_{j,r-1}) \oplus \mathcal{E}_r \oplus \mathcal{K}_r \oplus \mathcal{G}_r \oplus \mathcal{D}_{\mathrm{DoE}}^{\mathrm{best}} \oplus \mathrm{Out}_{\mathrm{region}},$$

$$\mathrm{Prompt}_{\mathrm{loop},t} = \mathrm{Sys}_{\mathrm{loop}} \oplus \mathrm{Obj} \oplus t \oplus \bigoplus_{j=1}^{d} \mathrm{Factor}_j(\widetilde{\mathcal{L}}_{j,r}) \oplus \mathcal{E}_r \oplus \mathcal{K}_r \oplus \mathcal{G}_r \oplus \mathcal{D}_t^{\mathrm{best}} \oplus \mathcal{D}_t^{\mathrm{recent}} \oplus \mathrm{Out}_{\mathrm{point}}.$$

Here $\mathcal{E}_r$ is the design-matched statistical evidence, $\mathcal{K}_r$ contains the supported statistical terms, $\mathcal{G}_r$ contains model diagnostics, and $\widetilde{\mathcal{L}}_{j,r}$ is factor j's current refined admissible domain. In the implementation, $\mathcal{D}_{\mathrm{DoE}}^{\mathrm{best}}$ contains at most eight DoE observations, while $\mathcal{D}_t^{\mathrm{best}}$ and $\mathcal{D}_t^{\mathrm{recent}}$ contain at most ten and eight observations, respectively.

Design philosophy. The prompts follow five principles. First, information is stage-specific: Path B initialization receives no DoE-derived statistical evidence, post-DoE reasoning receives design-matched evidence, and in-loop reasoning additionally receives sequential outcomes. Second, statistical evidence is supplied explicitly rather than asking the LLM to infer structure from raw observations alone. Third, interactions, trends, and curvature are interpreted jointly with main effects, so a weak main effect does not by itself remove a factor involved in another supported term. Fourth, the LLM is an advisory proposal mechanism: outputs are validated against the current admissible domain and projected to feasibility when necessary before evaluation. Fifth, prompts explicitly prohibit access to hidden optima, unqueried outcomes, and external experimental databases. Thus the LLM does not generate pseudo-observations or modify the GP posterior.

All agents at a given decision point receive the same task information and output schema. Model diversity produces multiple proposals, which are aggregated only after feasibility validation. The API layer additionally requires a single syntactically valid JSON object, simplifying automatic parsing and preventing free-form text from entering the optimization logic.

A.4.1 DIRECT-PATH INITIALIZATION PROMPT

This prompt is used only when the router selects DIRECT-SEARCH, so no DoE-derived statistical evidence is available.

System prompt.

```
You are one reasoning agent helping initialize an expensive
experimental optimization problem.

There is no DoE-derived statistical evidence yet.

Use only:
- the factor descriptions,
- the original admissible domains,
- and the initial points already selected.

Propose ONE promising experimental configuration.

Return JSON only:

{
  "point": {
    "factor_name": value
  },
  "reason": "brief explanation"
}

Rules:
1. Include every factor exactly once.
2. Every value must lie inside its supplied admissible domain.
3. Do not invent experimental outcomes.
4. Do not assume access to a hidden optimum.
5. Avoid simply repeating an already-selected configuration.
```

User-prompt template.

```
Objective:
```

```
{objective}

Factor information:
{factor_information}

Initial configurations already selected:
{selected_records}

Propose one useful next initialization configuration.

The optimization code will validate your proposal against the
current feasible domain and project it to feasibility when necessary.
```

The placeholders are replaced at runtime by the factor descriptions and current bounds, and by previously selected direct-path initialization points. Each agent's proposal is validated against the original feasible domain. The feasible proposals are aggregated in the common embedding and projected onto the admissible next-query domain:

$$x^{(1:M)} \longrightarrow \bar{z} \longrightarrow \Pi_{\mathcal{A}_1}(\bar{z}).$$

A.4.2 POST-DOE REGION-REASONING PROMPT

When Path A is selected, the design–model pair is fixed before the current DoE responses are observed. The LLM agents are invoked only after those responses have been analyzed by the Statistical Agent. Their role is to translate the resulting design-matched statistical evidence into candidate search regions.

System prompt.

```
You are one reasoning agent in a multi-agent experimental
optimization system.

Use ONLY:
1. supplied factor descriptions,
2. current admissible factor domains,
3. observed DoE results,
4. statistical evidence on effects, trends, interactions, or curvature,
5. model-diagnostic information.

Your task is to propose ONE promising admissible subdomain
for every factor. For continuous factors, return a non-degenerate
interval. For integer, categorical, or other finite-level factors,
return a nonempty subset of the supplied admissible values.

Do NOT assume access to:
- the hidden optimum,
- unqueried outcomes,
- external experimental databases.

Return valid JSON only:

{
  "domain": {
    "factor_name": [admissible values or interval endpoints]
  },
  "reason": "brief data-grounded explanation"
}

Rules:
1. Include every factor exactly once.
```

```
2. Every proposed subdomain must be contained in the supplied domain.
3. Continuous intervals must satisfy lower < upper; finite-level
   proposals may contain only supplied admissible values.
4. Do not invent experimental observations.
5. Interpret coefficients using the supplied design and model coding.
6. Consider interactions, trends, and curvature jointly with main effects.
7. If evidence is uncertain, prefer a wider region rather than
   unjustified narrowing.
```

User-prompt template.

```
Objective:
{objective}

Factor information:
{factor_information}

Design-matched statistical evidence:
{statistical_records}

Terms meeting the optimization-allocation thresholds:
{supported_records}

Model diagnostics:
{diagnostic_records}

Best observed DoE experiments:
{observed_records}

Important interpretation:

- Interpret coefficients using the supplied design and model coding.
- Supported-term thresholds are allocation heuristics, not proof that
  unsupported terms are irrelevant.
- Treat the fitted statistical model as local evidence rather than an
  exact global response model.
- Consider interactions, trends, and curvature jointly with main effects.

Propose ONE promising admissible factor-domain tuple for the
next optimization stage.
```

A.4.3 In-Loop Candidate Prompt

During sequential optimization, this prompt is generated only when the probabilistic branch selector chooses the LLM branch. Consequently, an LLM call is not required at every BO iteration.

System prompt.

```
You are one reasoning agent in the sequential stage of an
expensive experimental optimization problem.

Use ONLY:
1. factor descriptions,
2. the current refined admissible domain,
3. experimentally observed outcomes,
4. design-matched statistical evidence,
5. supported-term information,
6. model-diagnostic information.
```

```
Propose ONE next experimental configuration.

Do NOT assume access to:
- the hidden benchmark optimum,
- unqueried outcomes,
- external experimental databases.

Return valid JSON only:

{
  "point": {
    "factor_name": value
  },
  "reason": "brief data-grounded explanation"
}

Rules:
1. Include every factor exactly once.
2. Every proposed value must lie inside the current refined admissible domain.
3. Do not invent objective values.
4. Consider interactions, trends, and curvature jointly with main effects.
5. A weak main effect does not make a factor irrelevant if another
   supported term involving that factor is informative.
6. Avoid simply repeating the strongest observed point unless
   the evidence strongly supports local exploitation.
7. The optimization code will validate the proposal and, when necessary,
   project it onto the current admissible next-query domain.
```

User-prompt template.

```
Objective:
{objective}

Total experimental evaluations already used:
{t}

Current factor information:
{factor_information}

Design-matched statistical evidence:
{statistical_records}

Supported terms used for refinement:
{supported_records}

Model diagnostics:
{diagnostic_records}

Best observed experimental configurations:
{best_records}

Most recent experimental observations:
{recent_records}

Important:
- Treat the fitted statistical model as local evidence.
- Consider interactions, trends, and curvature jointly.
- Do not use hidden benchmark information.
- Propose one next experimental configuration.
```

For factor j, `factor_information` contains its description and current refined admissible domain. The implementation supplies the ten best observations and eight most recent observations when available. For M valid agent proposals, the multi-agent point is obtained by projection onto the current admissible next-query domain:

$$z_m = \psi(x^{(m)}), \qquad \bar{z} = \frac{1}{M}\sum_{m=1}^{M} z_m, \qquad x^{\mathrm{LLM}} \in \arg\min_{x \in \mathcal{A}_r} \|\psi(x) - \bar{z}\|_2.$$

A.4.4 Output Enforcement

The API wrapper appends an additional JSON-only instruction to each system prompt. In equivalent form,

```
CRITICAL OUTPUT FORMAT:
Return ONLY one syntactically valid JSON object.
Do not output analysis, reasoning, Markdown, code fences,
introductions, or explanations outside the JSON object.
```

and appends to the user message:

```
Return only the requested JSON object.
```

Provider-specific wording differs slightly, but the constraint is the same. Outputs that fail parsing or feasibility validation are retried up to the prespecified retry limit; remaining failures are replaced by a feasible fallback according to the implementation protocol.

A.5 Description Card for Synthetic Initialization

Simulation functions	**Description Card `[Function Patterns]`**
Levy-$2D$	Range: $x_j \in [-10, 10]$, $j = 1, 2$. A smooth multimodal function with several local variations and nontrivial interactions across the search space.
Rastrigin-$2D$	Range: $x_j \in [-5.12, 5.12]$, $j = 1, 2$. A smooth, non-convex, and highly periodic function containing many regularly distributed local optima.
Branin-$2D$	Range: $x_1 \in [-5, 10]$, $x_2 \in [0, 15]$. A smooth non-convex function with multiple separated basins and a curved response surface.
Bukin-$2D$	Range: $x_1 \in [-15, -5]$, $x_2 \in [-3, 3]$. A challenging non-smooth function characterized by a narrow curved valley and strong changes in function value across nearby points.
Hartmann-$4D$	Range: $x_j \in [0, 1]$, $j = 1, \ldots, 4$. A smooth, non-convex, and non-separable function with multiple localized basins and interactions among dimensions.
Ackley-$6D$	Range: $x_j \in [-5, 5]$, $j = 1, \ldots, 6$. A multimodal function with a relatively flat global structure combined with many local oscillations, making exploration in higher dimensions challenging.

Table 3: Description cards for the benchmark functions used in the initial-proposal experiment. The cards provide only general structural information about each optimization landscape and do not reveal optimizer locations or detailed analytical properties.

Simulation functions	Description Card `[Function Patterns]`
Levy-$2D$	Range: $x_j \in [-10, 10]$, $j = 1, 2$. Smooth and highly multimodal after the transformation $w_j = 1+(x_j-1)/4$. Rough Lipschitz scale: $L \approx 65$. Favorable regions tend to lie near the central transformed basin, while far positive or negative regions create oscillatory traps.
Rastrigin-$2D$	Range: $x_j \in [-5.12, 5.12]$, $j = 1, 2$. Smooth, non-convex, and strongly periodic. Since $\partial f/\partial x_j = 2x_j + 20\pi \sin(2\pi x_j)$, a conservative scale is $L \lesssim 104$. The central region is generally more favorable, but many evenly spaced local traps remain.
Branin-$2D$	Range: $x_1 \in [-5, 10]$, $x_2 \in [0, 15]$. Smooth non-convex surface with several separated favorable basins connected by a curved valley. Rough Lipschitz scale: $L \approx 115$. Good regions may occur in different parts of the box, so over-exploiting one basin can be misleading.
Bukin-$2D$	Range: $x_1 \in [-15, -5]$, $x_2 \in [-3, 3]$. Steep, narrow, and non-smooth because of absolute-value and square-root terms. Away from the sharp valley, e.g. $\lvert x_2 - 0.01x_1^2\rvert \geq 0.01$, a rough local scale is $L \lesssim 523$. Promising points lie near a curved valley in the negative-x_1, low-positive-x_2 region.
Hartmann-$4D$	Range: $x_j \in [0, 1]$, $j = 1, \ldots, 4$. Smooth, non-convex, and non-separable. It is composed of weighted anisotropic Gaussian-like bumps. Practical numerical Lipschitz scale: $L \approx 15$. Favorable regions are localized inside the unit cube, typically involving small-to-intermediate coordinate values rather than boundary solutions.
Ackley-$6D$	Range: $x_j \in [-5, 5]$, $j = 1, \ldots, 6$. Lipschitz on the box, with rough scale $L \lesssim 8.6$. Smooth away from the central radial kink-like region. The function has a nearly flat outer region and many local oscillations; favorable regions are concentrated near the center, but sparse proposals can miss the main basin.

Table 4: Descriptive card for benchmark functions used in the initial-proposal experiment. The descriptions provide structural information suitable for Bayesian optimization prompts without revealing exact optimizer coordinates. The reported L values are rough Lipschitz scales, not tight theoretical constants.

A.6 Initialization Ablation Second Experiment

A.7 Implementation Conventions

Model and design. Fix candidate design families, statistical models and bases, term hierarchy, replication plan, and design-search seed before observing DoE responses. Candidate designs may include full/fractional factorial, central composite, and constrained D-optimal designs as factor types and inferential goals permit. All design–model pairs must pass admissibility, hierarchy, estimability, and budget checks. Numerical D-optimal routines use the determinant criterion when applicable; global optimality is not assumed unless certified. Replicates provide pure-error information; unavailable diagnostics and nonestimable terms are flagged. Record the selected design, statistical model, and actual run count in the router.

Feasible-domain refinement. At each stage, store the original admissible domains separately from the stage-specific refined domains. Every refined domain must satisfy $\widetilde{\mathcal{L}}_{j,r} \subseteq \mathcal{L}_j$. Continuous factors retain continuous intervals; integer, categorical, and other finite-level factors retain only admissible native values. Apply $\mathcal{C}$ after combining the factor domains and require $\mathcal{X}_r \neq \emptyset$. If refinement produces an empty feasible domain, retain the previous domain or broaden the proposed subdomains before any evaluation. Approved later rounds may restore previously excluded admissible values. All observations from all stages are retained for GP fitting.

Search and review. The embedding $\psi : \mathcal{X} \to \mathbb{R}^{d_\psi}$ uses fixed numerical scalings and category encodings across stages. Projection is performed onto the current admissible next-query domain $\mathcal{A}_r$. GP acquisition optimization respects the native factor domains and original feasibility constraints. GP fitting uses measured outcomes only. The branch is sampled before generating its proposal, so unselected branches need not be computed. Record LLM versions, prompts, sampling settings, retry limits, seeds, refinement thresholds, K_0, stage budgets, and human review decisions. Allocate a stage's budget before its first query; stop when its budget is exhausted or $\mathcal{A}_r$ is empty. Any DoE replicates and direct-path initialization runs consume the same total budget T.

B Notation

Problem and Budgets

Table 5: Effect of the decay exponent k on proposal-selection probabilities. We define $p_{\mathrm{LLM}}(u) = 0.72(1-u)^k$, $p_{\mathrm{GP}}(u) = 0.90 - p_{\mathrm{LLM}}(u)$, and fix $p_{\mathrm{rand}} = 0.10$. The main experiments use $k = 4$.

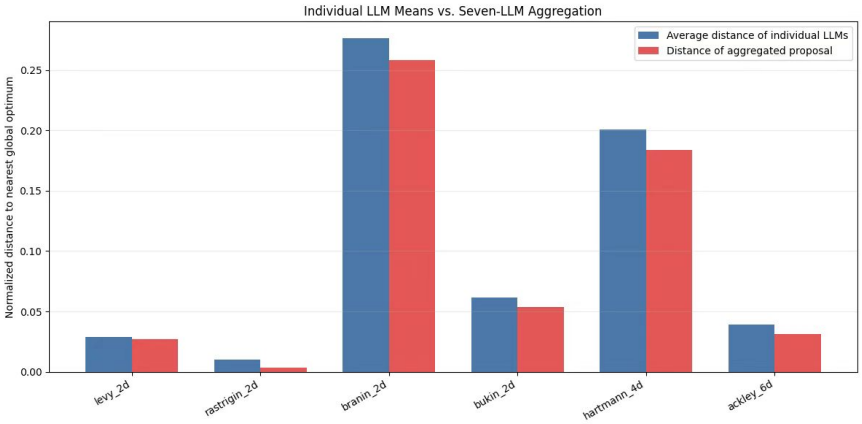
k	Progress u	p_{LLM}	p_{GP}	p_{rand}
1.0	0.00	0.720	0.180	0.100
	0.25	0.540	0.360	0.100
	0.50	0.360	0.540	0.100
	0.75	0.180	0.720	0.100
	1.00	0.000	0.900	0.100
1.5	0.00	0.720	0.180	0.100
	0.25	0.468	0.432	0.100
	0.50	0.255	0.645	0.100
	0.75	0.090	0.810	0.100
	1.00	0.000	0.900	0.100
2.0	0.00	0.720	0.180	0.100
	0.25	0.405	0.495	0.100
	0.50	0.180	0.720	0.100
	0.75	0.045	0.855	0.100
	1.00	0.000	0.900	0.100
2.5	0.00	0.720	0.180	0.100
	0.25	0.351	0.549	0.100
	0.50	0.127	0.773	0.100
	0.75	0.023	0.878	0.100
	1.00	0.000	0.900	0.100
3.0	0.00	0.720	0.180	0.100
	0.25	0.304	0.596	0.100
	0.50	0.090	0.810	0.100
	0.75	0.011	0.889	0.100
	1.00	0.000	0.900	0.100
3.5	0.00	0.720	0.180	0.100
	0.25	0.263	0.637	0.100
	0.50	0.064	0.836	0.100
	0.75	0.006	0.894	0.100
	1.00	0.000	0.900	0.100
4.0	0.00	**0.720**	**0.180**	0.100
	0.25	**0.228**	**0.672**	0.100
	0.50	**0.045**	**0.855**	0.100
	0.75	**0.003**	**0.897**	0.100
	1.00	**0.000**	**0.900**	0.100

$f,\ x$	Objective function and experimental configuration
$d,\ x_j$	Number of factors and value of factor j
$\mathcal{L}_j$	Original admissible domain of factor j
$\mathcal{C}, \mathcal{X}$	Feasibility constraints and original feasible domain
$\mathcal{F}$	Factor metadata: type, admissible domain, and description
$T,\ t$	Total evaluation budget and number of evaluations used
n_{DoE}	Smallest useful DoE size under the selected design rule
ρ_{DoE}	Maximum budget fraction allocated to the initial DoE
K_0	Direct-search initialization size under Path B
$s,\ T_{\mathrm{BO}}$	Sequential-search iteration and sequential optimization budget
$\mathcal{D}$	History of measured configuration–response pairs
$\mathcal{D}_{\mathrm{DoE}}$	Observations collected from the initial designed experiment
$\mathcal{D}_{\mathrm{summary}}$	Observed-data summary supplied to LLM agents

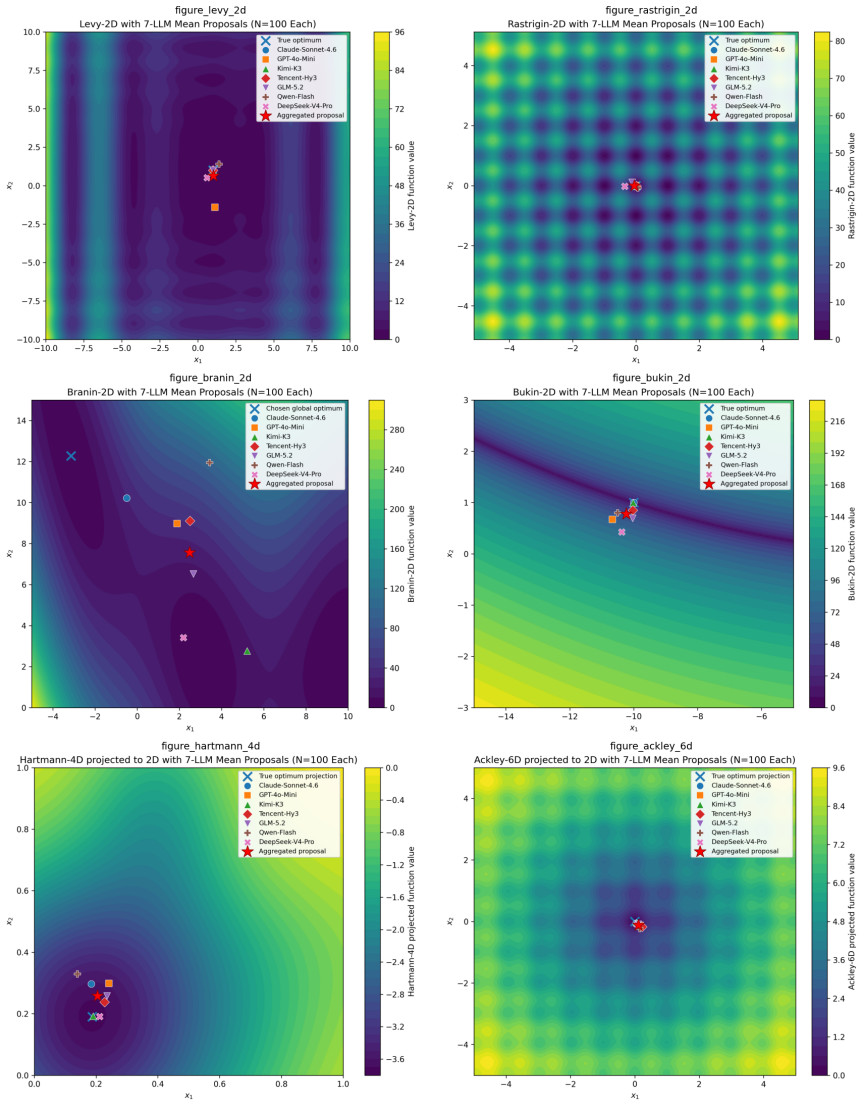
Statistical Evidence and Prompt State

$\mathcal{E}$	Design-matched statistical evidence extracted from the DoE
r	Refinement-stage index
$\mathcal{E}_r$	Statistical evidence available at refinement stage r
$\mathcal{K}_r$	Supported statistical terms supplied to the LLM at stage r
$\mathcal{G}_r$	Statistical-model diagnostics supplied at stage r
$\widetilde{\mathcal{L}}_{j,r}$	Refined admissible domain of factor j at stage r
$\mathcal{X}_r$	Constraint-feasible refined domain at stage r
$\mathcal{A}_r$	Current admissible next-query domain at stage r
$\oplus$	Text concatenation operator used in prompt construction
	Collective LLM Reasoning and Region Aggregation
$M,\ A_m$	Number of LLM agents and reasoning agent m
$\mathcal{I}_{\mathrm{LLM}}$	Combined information supplied to each reasoning agent
$\mathcal{Q},\ q^{(m)}$	Descriptor space and refinement descriptor proposed by agent m
Γ	Deterministic operator mapping a descriptor to a feasible canonical region
$R_k^{(m)}$	Continuous subregion component proposed by agent m
$S^{(m)}$	Set of observed anchors or feasible discrete configurations proposed by agent m
$N_i,\ x_i$	Prespecified neighborhood and corresponding observed anchor
$\mathcal{R}^{(m)}$	Canonical feasible search region induced by agent m
ν	Counting, Lebesgue, or corresponding product measure on the domain
$J_\nu(\mathcal{R},\mathcal{R}')$	Measure-based Jaccard similarity between canonical regions
$d_\nu(\mathcal{R},\mathcal{R}')$	Jaccard distance $1 - J_\nu(\mathcal{R},\mathcal{R}')$
d	Prespecified metric used for canonical-region aggregation
$m^\star$	Medoid agent index
$\mathcal{R}^\star$	Aggregate canonical region selected by medoid aggregation
A	Mean agreement of agent regions with the medoid region
	Region Refinement
$\mathcal{R}_s$	Current local canonical search region at iteration s
ϕ_j	Normalized numerical coordinate for factor j
$\mathcal{T}_{i,s}$	Adaptive trust region centered at observed anchor x_i
$L_{i,s}$	Scalar size of trust region $\mathcal{T}_{i,s}$
$w_{j,s}$	Evidence-derived dimension weight for factor j
$\mathcal{A}_s^{\mathrm{loc}}$	Current feasible local next-query domain
	Probabilistic Search and Point Aggregation

u_s	Normalized progress through the sequential-search budget
ρ_0	Fixed probability assigned to global random exploration
λ_0, γ	Initial LLM-search weight and polynomial decay exponent
ρ_{LLM}	Probability of selecting the LLM-guided branch
ρ_{GP}	Probability of selecting the GP-guided branch
ρ_{random}	Probability of selecting global random exploration
b_s	Search branch selected at sequential iteration s
$x^{(m)}$	Feasible point proposed by LLM agent m
$\psi,\ d_\psi$	Fixed numerical/one-hot embedding and its dimension
z_m	Embedded proposal $\psi(x^{(m)})$ of agent m
$\bar{z}$	Mean of the embedded LLM proposals
x^{LLM}	Feasible projected-mean LLM proposal
$x^\star,\ z^\star$	Reference feasible point and its embedding in the projected-mean theorem
	Medoid-Theorem Quantities
$\mathfrak{R}$	Metric space of admissible canonical regions
$\mathcal{R}^\dagger$	Reference region in the medoid robustness theorem
r (theorem)	Distance tolerance in the medoid theorem, not the refinement-stage index
g	Number of proposals within distance r of $\mathcal{R}^\dagger$
p	Lower bound on individual proposal success probability
α	Target fraction of proposals satisfying the proximity condition



(a) Aggregation quality.



(b) Initialization landscape.

Figure 5: Collective LLM reasoning and initialization quality.