CALIPER: TWO-LEVEL RELIABILITY CALIBRATION FOR MULTI-EXPERT LLM PRIORS IN MULTI-OBJECTIVE BAYESIAN OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

External priors can accelerate Bayesian optimization (BO) precisely when evaluation budgets are too small to recover easily from bad guidance. LLM committees make this opportunity inseparable from a reliability problem: expertise varies by objective, multiple experts may share the same bias, and stated confidence may not predict accuracy. To address this issue, we introduce *Counterfactual Adaptive LLM-Integrated Prior Evidence Replay* (CALIPER), a plug-in calibration layer for multi-objective BO that learns two complementary forms of trust from newly observed objectives. *Relative reliability* routes each objective toward the experts that predict it well; *absolute reliability* controls whether the committee should influence the optimizer at all. CALIPER combines objective-wise reputation, counterfactual validation of confidence, prior-amplitude control, and an explicit no-prior arm. It injects the resulting signal as a residual-GP mean shift, preserving the downstream qLogEHVI/qLogNEHVI acquisition pipeline. We evaluate every prior-control strategy with the same frozen LLM information and a locked, disjoint development/evaluation protocol on ESOL, FreeSolv, and Lipophilicity. The real-prior benchmarks show an asymmetric payoff: CALIPER captures useful guidance on ESOL and Lipophilicity while recovering the no-prior final hypervolume on FreeSolv, where unconditional injection causes negative transfer. Controlled stress tests isolate the mechanism: CALIPER gains 8.1% trajectory AUC when all experts are useful, routes objective-specialized expertise for a 6.3% gain, and stays within 0.6% of no-prior BO when every expert is misleading, compared with a 10.1% loss from fixed injection. These results establish evidence calibration as a practical interface for extracting value from uncertain external knowledge without binding the optimizer to it.

1 INTRODUCTION

Bayesian optimization (BO) is most valuable when black-box evaluations are scarce: a laboratory assay, high-fidelity simulation, or costly model call may permit only a handful of queries (Jones et al., 1998; Frazier, 2018). That same small-budget regime makes external knowledge unusually consequential. A useful prior can focus the search before the surrogate has enough data of its own. A misleading prior can initiate a self-reinforcing loop—biased acquisition, narrow observations, and too little corrective evidence—that consumes the budget before BO recovers.

Large language models offer a flexible source of such knowledge. They can score candidates from scientific descriptions, express complementary expert roles, and provide guidance without task-specific retraining. Yet the provenance and reliability of that guidance are rarely known. A score may reflect transferable reasoning, memorized benchmark regularities, retrieval, or a shared model bias; pretraining exposure is often impossible to audit directly. Self-reported confidence adds another signal, but linguistic certainty need not track empirical correctness (Guo et al., 2017; Lin et al., 2022; Kadavath et al., 2022). The central question is therefore not whether an LLM can produce a prior, but how much influence that prior earns after contact with new black-box evidence.

We identify two levels of reliability. *Relative expert reliability* asks which expert should be trusted for each objective. This matters in multi-objective BO (MOBO), where an expert may be accurate for

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.

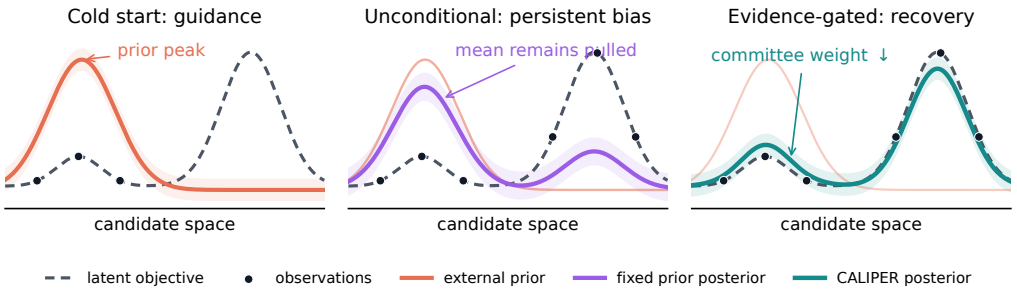

Figure 1: Calibration intuition. In the cold start, an external prior can direct the posterior toward a promising region before the GP has sufficient observations. A fixed mean shift also preserves incorrect structure when the prior conflicts with observed objectives. CALIPER converts this conflict into evidence: relative reliability reroutes weight across experts, absolute reliability reduces the committee amplitude, and the no-prior arm returns the posterior toward the residual-GP learner. The curves are schematic; the gate is learned objective by objective.

solubility and biased for drug-likeness. *Absolute prior reliability* asks whether any member of the committee deserves influence. Expert reweighting answers the first question, but simplex-normalized weights must still choose from an entirely misleading committee. Iteration-based decay controls how long a prior persists, but elapsed time is not evidence of accuracy. Reliable prior use therefore requires both objective-wise routing and a committee-level option to attenuate or abstain.

CALIPER (*Counterfactual Adaptive LLM-Integrated Prior Evidence Replay*) implements this two-level view. Multiple LLM roles provide frozen, objective-wise scores and self-reported confidence. Black-box observations update expert–objective reputation, while counterfactual replay evaluates whether confidence improves prediction with and without using it. An evidence gate converts committee support into prior amplitude and includes a no-prior arm. The calibrated prior shifts the mean of a residual Gaussian process (GP), so standard MOBO acquisitions operate on the resulting posterior without changes to their optimization logic. Figure 1 summarizes the intended behavior: exploit supported guidance, route specialized expertise, and return toward the underlying optimizer when observations contradict the committee.

This formulation also changes how uncertainty about pretraining exposure enters the experimental claim. CALIPER does not need to certify where an LLM score originated. It tests whether that score predicts newly revealed outcomes and retains influence in proportion to the evidence it earns. The optimizer is therefore less dependent on treating opaque model knowledge as intrinsically trustworthy.

Contributions. Our contributions are:

1. We formulate *two-level reliability* for multi-expert, multi-objective priors: objective-wise reputation resolves which experts to trust, and evidence-calibrated amplitude resolves whether the committee should be trusted at all.
2. We introduce CALIPER, combining expert routing, counterfactual confidence validation, prior-strength control, and explicit abstention in a modular residual-mean layer. The layer leaves the downstream acquisition rule intact.
3. We establish the mechanism under matched frozen-prior controls. Real molecule benchmarks demonstrate asymmetric transfer across naturally heterogeneous LLM priors; controlled useful, specialized, and all-misleading priors identify the corresponding exploit, route, and abstain behaviors; and candidate-pool diagnostics test the persistence of these behaviors at larger search scales.

2 Related Work

Prior-informed Bayesian optimization. Multi-objective BO has long addressed expensive optimization under competing objectives, with classical approaches such as ParEGO using randomized

scalarization to reduce multi-objective search to sequential scalar BO problems (Knowles, 2006). Prior-informed BO differs in how external beliefs enter the optimizer. πBO multiplies a conventional acquisition by a user prior over the optimizer location and gradually reduces its influence (Hvarfner et al., 2022). ColaBO conditions BO on richer user-specified beliefs and supports Monte-Carlo acquisition functions (Hvarfner et al., 2024). DynMeanBO embeds an expert prior in the surrogate mean and applies a pre-specified decay toward the vanilla mean as observations accumulate (Qu et al., 2026). Together, these approaches provide fixed, acquisition-side, and scheduled mean interfaces for external belief. CALIPER studies a complementary problem: learning the trust structure of a heterogeneous committee from the same sparse observations used by BO. The matched DynMeanBO-MO control separates evidence-conditioned calibration from a pre-specified time schedule.

LLMs in optimization loops. LLMs have several roles in recent BO systems. In molecular optimization, Kristiadi et al. study whether LLM representations actually improve principled BO and find that their value depends strongly on domain-relevant pretraining or adaptation (Kristiadi et al., 2024). LLAMBO uses LLMs for candidate proposal and evaluation inside a model-based loop (Liu et al., 2024); other work uses BO to adapt in-context LLM search or asks LLMs to design acquisitions or kernels (Agarwal et al., 2025; Aglietti et al., 2025; Suwandi et al., 2025). More recent systems use LLMs as online acquisition strategists or transfer knowledge from previous BO tasks into initialization policies (Ngo et al., 2026; Zeng et al., 2026). LGBO queries an LLM during optimization for promising points or regions and converts those preferences into a surrogate-mean shift (Yuan et al., 2026). LABO treats LLM predictions as inexpensive low-fidelity observations and learns their discrepancy from high-fidelity experiments (Chen et al., 2026). In our setting, LLM outputs are generated before BO, are not observations, and do not directly propose the next experiment. We study online *calibration of a frozen multi-expert prior cache* against newly observed objective values.

Confidence and expert weighting. Confidence calibration asks whether stated confidence matches empirical correctness (Guo et al., 2017). Prompted LLM confidence can be miscalibrated and need not behave like a model probability, motivating explicit empirical evaluation of verbalized confidence in black-box LLMs (Lin et al., 2022; Kadavath et al., 2022; Xiong et al., 2024). The reputation update is also related to online learning with expert advice (Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006). We therefore include an objective-wise Plain Hedge baseline whose weights are updated from squared prediction error on observed candidates. Hedge learns *relative* expert quality but keeps its weights on the simplex, so it cannot assign zero influence to the full committee. CALIPER adds absolute prior-strength calibration, confidence-use gates, and an explicit no-prior arm. Unlike full-information expert advice, BO provides selective feedback because objective values are observed only at acquisition-chosen candidates.

3 METHOD

3.1 PROBLEM FORMULATION

We consider discrete multi-objective black-box optimization over a finite candidate set $\mathcal{X} = \{x_1, \dots, x_N\}$, with unknown objective vector $\mathbf{f}(x) = (f_1(x), \dots, f_m(x))$. All objectives are normalized to the maximization direction. After t evaluations,

$$\mathcal{D}_t = \{(x_\tau, \mathbf{y}_\tau)\}_{\tau=1}^{t}, \qquad \mathbf{y}_\tau = \mathbf{f}(x_\tau).$$

For each candidate x, expert e, and objective j, a frozen cache provides an objective-wise prior score $\mu_{ej}(x) \in [0, 1]$. Each expert-candidate pair also carries one scalar self-reported confidence $c_e(x) \in [0, 1]$, shared across objectives in the current cache format. Neither quantity is treated as an observation or a calibrated probability; objective specificity enters through the reputation state $K_{ej,t}$.

Timing and no-lookahead. Let

$$\Theta_t = (\mathcal{D}_t, \mathbf{K}_t, \mathbf{K}_t^{(0)}, \mathbf{K}_t^{(1)}, \mathbf{z}_t^u, \boldsymbol{\pi}_t^p, \boldsymbol{\pi}_t^u)$$

denote the state after observation t. The sequential order is

$$x_t = A(\Theta_{t-1}) \;\longrightarrow\; y_t = \mathbf{f}(x_t) \;\longrightarrow\; \Theta_t = U(\Theta_{t-1}, x_t, y_t).$$

Hence every quantity depending on y_t is used only after x_t has been selected. In particular, the update-gate probability used for the t-th reputation reward is the pre-observation value $\rho_{j,t} = \pi_{j,t-1}^u(1)$.

3.2 Two-Level Prior Construction

Let $K_{ej,t}$ be the expert-objective capital after t observations. The running objective scale $s_{j,t}$, expert reward $R_{ej,t}$, and capital update are given in Appendix B; the default initialization is $K_{ej,0} = \mathrm{logit}(0.5) = 0$. Capital is mapped to objective-level trust and expert-objective weights as

$$\chi_j(\mathbf{K}_t) = \sigma\left[a_\chi\left(\frac{1}{E}\sum_e \sigma(K_{ej,t}) - \tau_\chi\right)\right],$$

$$\alpha_{ej,t} = \mathrm{clip}\left[\chi_j(\mathbf{K}_t)\, E \frac{\exp(K_{ej,t}/T)}{\sum_{e'}\exp(K_{e'j,t}/T)}, \alpha_{\min}, \alpha_{\max}\right].$$

For a confidence multiplier d, define

$$v_{ej}(x;\mathbf{K}_t,d) = \max\{\alpha_{ej,t} d_e(x), 10^{-8}\}, \qquad \bar{\mu}_j(x;\mathbf{K}_t,d) = \frac{\sum_e v_{ej}(x;\mathbf{K}_t,d)\mu_{ej}(x)}{\sum_e v_{ej}(x;\mathbf{K}_t,d)},$$

$$\zeta_j(x;\mathbf{K}_t,d) = \mathrm{clip}\left[\frac{1}{E}\sum_e v_{ej}(x;\mathbf{K}_t,d), 0, 1\right], \qquad G_j(x;\mathbf{K}_t,d) = \zeta_j(x;\mathbf{K}_t,d)\bar{\mu}_j(x;\mathbf{K}_t,d).$$

The normalized mixture $\bar{\mu}_j$ implements *relative reliability*: it routes objective j toward experts with stronger observed reputation. The amplitude ζ_j implements *absolute reliability*: it controls the total exposure of objective j to the committee. The three prior arms are

$$p^{(0)}_{j,t}(x) = G_j(x;\mathbf{K}_t,\mathbf{1}), \qquad p^{(1)}_{j,t}(x) = G_j(x;\mathbf{K}_t,\mathbf{c}(x)), \qquad p^{(2)}_{j,t}(x) = 0,$$

corresponding to no-confidence, confidence, and drop-prior.

3.3 Residual Posterior Shift

After observation t, the selected prior $p_{j,t}$ is held fixed and the residual GP is fit to all t observations:

$$\widetilde{y}_{j,\tau,t} = y_{\tau j} - p_{j,t}(x_\tau), \qquad r_j \sim \mathcal{GP}(0, k_j).$$

If $r_j(x) \mid \mathcal{D}_t \sim \mathcal{N}(m^r_{j,t}(x), v^r_{j,t}(x))$, the posterior exposed to the acquisition is

$$f_j(x) \mid \mathcal{D}_t \sim \mathcal{N}\big(p_{j,t}(x) + m^r_{j,t}(x), v^r_{j,t}(x)\big).$$

The prior therefore changes the exploitation landscape through the mean while the residual GP retains responsibility for posterior uncertainty.

A modular mean-layer interface. By representing external guidance as a deterministic mean-side signal, CALIPER preserves the posterior covariance, Monte Carlo sampling procedure, and acquisition interface of the underlying GP. The same calibration layer can therefore operate directly with qLogEHVI, qLogNEHVI, and other posterior-based acquisitions without modifying their optimization logic. A hierarchical extension could additionally represent correlated uncertainty across external sources while retaining this modular interface.

3.4 Counterfactual Evidence Gates

Self-reported confidence can improve aggregation when it ranks expert accuracy, and distort it when it merely reflects verbal certainty. CALIPER tests this choice by maintaining two deterministic replay worlds for every objective:

$\mathcal{M}^{(0)}_j$: reputation learned without confidence, $\qquad$ $\mathcal{M}^{(1)}_j$: reputation learned with confidence.

Whenever a new outcome y_{tj} arrives, both worlds predict that same observation from their pre-observation states. Their residual likelihoods provide directly comparable evidence for whether confidence improved the historical predictions. An update gate uses this comparison to decide how much confidence should affect future reputation credit.

A separate prior gate evaluates three actions—aggregate without confidence, aggregate with confidence, and abstain—under the residual GP marginal likelihood. Its posterior mixture produces

$$p_{j,t}(x) = \pi^p_{j,t}(0)p^{(0)}_{j,t}(x) + \pi^p_{j,t}(1)p^{(1)}_{j,t}(x),$$

with the abstention probability represented by the remaining mass on $p^{(2)}_{j,t}(x) = 0$. Evidence collected at round t updates the gates for round $t+1$, preserving the no-lookahead order. The replay recursions, fixed finite-data smoothing, margins, and pseudocode appear in Appendix B.

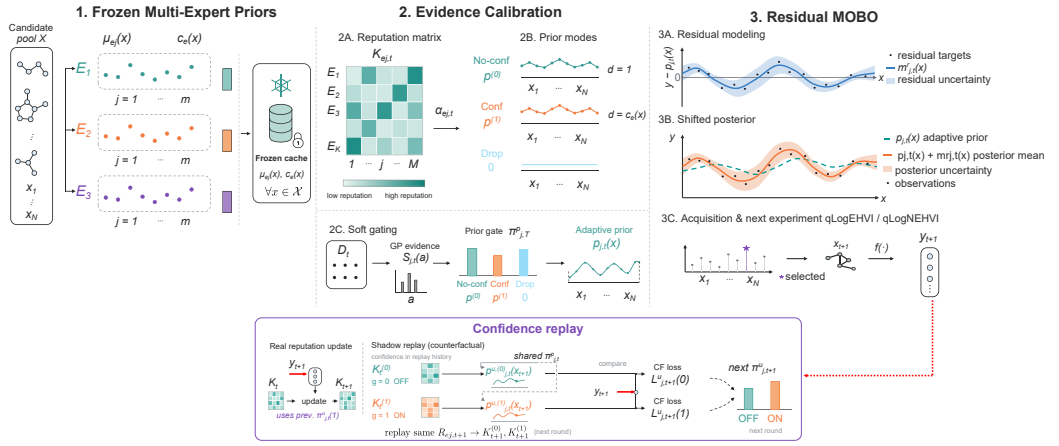

Figure 2: Overview of CALIPER. Multiple LLM experts are queried offline to build a frozen cache of objective-wise prior scores and self-reported confidence. Online, expert–objective reputation and counterfactual evidence gates determine the adaptive prior mean. The prior shifts a residual GP posterior that is consumed by an unchanged qLogEHVI/qLogNEHVI acquisition; new objective feedback updates the real reputation state and the confidence shadow markets only for subsequent rounds.

4 Experiments

4.1 Experimental Setup

Molecule benchmarks and protocol. We evaluate ESOL, FreeSolv, and Lipophilicity candidate pools from MoleculeNet (Wu et al., 2018). Each task optimizes the dataset property and QED under one common black-box information budget, giving every method identical objective access. The primary study uses qLogNEHVI (Daulton et al., 2021; Ament et al., 2023), implemented in BoTorch (Balandat et al., 2020), with eight initial evaluations, a total budget of 24, 64 QMC samples, and ten paired evaluation seeds (0–9); pool sizes are 100, 100, and 300. LLM scores are generated once and frozen, so the observed objective values are the sole source of online trust adaptation. We report trajectory hypervolume AUC, HV@16, final HV, and penalized T90/T95. Dataset construction, QED handling, features, prompts, and statistical reporting are detailed in Appendix A.

Matched prior-control strategies. All controls share the qLogNEHVI loop, candidate pools, initial designs, budgets, and frozen expert cache. No-prior qLogNEHVI establishes the fallback learner. Fixed injection treats the aggregate prior as uniformly reliable. DynMeanBO-MO applies a shared iteration-dependent mean decay (Qu et al., 2026). Plain Hedge learns objectively relative weights from the squared prediction error. DynMeanBO-MO and Plain Hedge are selected by equal-dataset AUC on disjoint development seeds 10–14 and locked at $\lambda = 0.20$ and $\eta = 2.0$ before evaluation on seeds 0–9. These controls vary the trust assumption while holding the information and acquisition constant. Full definitions, selection grids, paired intervals, and additional BO baselines are in Appendix A.5.

Controlled prior families. Real LLM caches capture naturally occurring behavior, while controlled reliability structures make the underlying mechanism identifiable. We therefore construct three prior families that require distinct calibration behaviors: all experts are useful, experts specialize by objective, or all experts are misleading. The five strategies share a 160-candidate pool, eight initial evaluations, budget 24, 64 QMC samples, and ten paired seeds. A second experiment holds the prior reliability structure fixed while increasing the candidate pool to 1,000, 3,000, and 10,000, isolating the scaling behavior of the calibration dynamics.

4.2 RQ1: Can CALIPER Safely Extract Value from Real LLM Priors?

Table 1: Formal qLogNEHVI comparison under the locked molecule protocol. Values are mean ± SEM over ten paired formal seeds (0–9); DynMeanBO-MO and Plain Hedge hyperparameters are selected only on development seeds 10–14; fixed prior requires no tuning. T90 misses are penalized as budget$+1 = 25$. Bold marks the best mean in each dataset/metric, with ties bolded.

Dataset	Method	AUC ↑	HV@16 ↑	Final HV ↑	T90 ↓
ESOL	qLogNEHVI (no prior)	0.6810 ± 0.0138	0.6865 ± 0.0163	0.7277 ± 0.0098	16.9 ± 2.0
	+ fixed prior	0.6761 ± 0.0100	0.6748 ± 0.0122	0.7297 ± 0.0080	16.7 ± 1.7
	DynMeanBO-MO	0.6867 ± 0.0135	0.6953 ± 0.0150	0.7401 ± 0.0079	15.6 ± 2.0
	Plain Hedge	0.6651 ± 0.0108	0.6590 ± 0.0125	0.7067 ± 0.0088	18.5 ± 2.1
	qLogNEHVI + CALIPER	**0.6931 ± 0.0091**	**0.7035 ± 0.0112**	**0.7417 ± 0.0083**	**14.2 ± 1.6**
FreeSolv	qLogNEHVI (no prior)	**0.4953 ± 0.0249**	**0.5370 ± 0.0328**	**0.6122 ± 0.0095**	**16.8 ± 1.5**
	+ fixed prior	0.4611 ± 0.0123	0.4565 ± 0.0164	0.5646 ± 0.0235	20.5 ± 1.2
	DynMeanBO-MO	0.4438 ± 0.0271	0.4528 ± 0.0429	0.6064 ± 0.0106	20.0 ± 1.3
	Plain Hedge	0.4369 ± 0.0119	0.4328 ± 0.0129	0.5126 ± 0.0215	23.0 ± 0.9
	qLogNEHVI + CALIPER	0.4859 ± 0.0238	0.5226 ± 0.0374	**0.6122 ± 0.0095**	17.6 ± 1.6
Lipophilicity	qLogNEHVI (no prior)	0.8829 ± 0.0095	0.8947 ± 0.0068	0.9166 ± 0.0080	11.1 ± 0.9
	+ fixed prior	0.8905 ± 0.0114	0.9071 ± 0.0096	0.9218 ± 0.0069	10.5 ± 0.9
	DynMeanBO-MO	0.8931 ± 0.0089	**0.9074 ± 0.0093**	0.9122 ± 0.0094	**9.8 ± 0.7**
	Plain Hedge	0.8691 ± 0.0131	0.8849 ± 0.0158	0.9040 ± 0.0118	13.9 ± 1.8
	qLogNEHVI + CALIPER	**0.8987 ± 0.0078**	0.9059 ± 0.0094	**0.9321 ± 0.0057**	10.1 ± 0.7

Table 1 reveals an asymmetric prior payoff. On ESOL, CALIPER records the highest mean AUC, HV@16, final HV, and fastest T90. It improves AUC by $+0.0279$ (95% CI $[+0.0164, +0.0405]$) and final HV by $+0.0350$ ($[+0.0170, +0.0533]$) over relative-only Plain Hedge, while tracking the strong scheduled-mean result from DynMeanBO-MO.

FreeSolv provides the committee-level rejection case. Fixed injection reduces HV@16 by 0.0804 relative to no-prior qLogNEHVI and delays T90 by 3.7 evaluations. CALIPER closes this negative-transfer gap: its final HV is 0.6122 ± 0.0095, exactly matching the no-prior mean, and its AUC exceeds DynMeanBO-MO by $+0.0421$ ($[+0.0082, +0.0781]$). Plain Hedge continues to allocate unit mass within the committee and finishes at 0.5126 ± 0.0215; adding absolute reliability raises CALIPER by $+0.0997$ ($[+0.0599, +0.1391]$).

On Lipophilicity, CALIPER attains the highest mean AUC (0.8987) and final HV (0.9321). DynMeanBO-MO reaches the early HV targets quickly, while evidence calibration continues to improve terminal quality, producing a paired final-HV gain of $+0.0199$ ($[+0.0030, +0.0386]$). Across the three tasks, CALIPER's role is consistent: it captures the upside of useful structure and reduces exposure when the same form of injection becomes harmful.

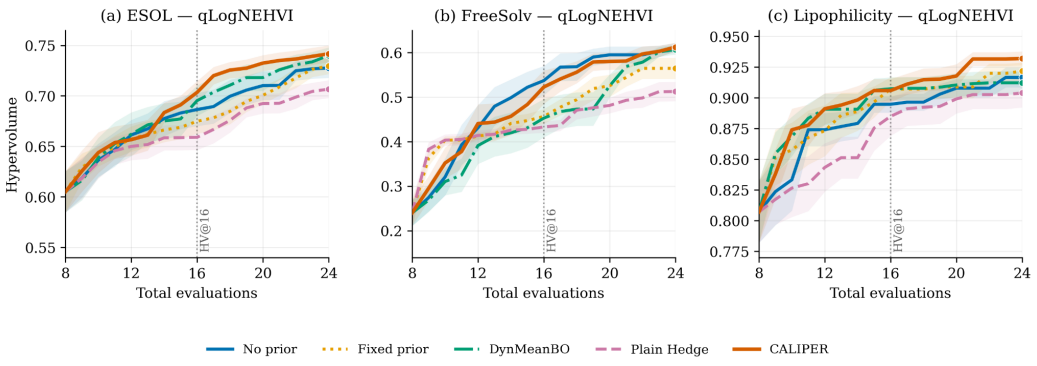

Figure 3: Formal qLogNEHVI trajectories over ten paired seeds (mean ± SEM); the dotted line marks HV@16. Every curve uses the same frozen LLM cache. CALIPER extracts useful prior structure on ESOL and Lipophilicity and returns to the no-prior terminal frontier on FreeSolv, illustrating its asymmetric use of external guidance.

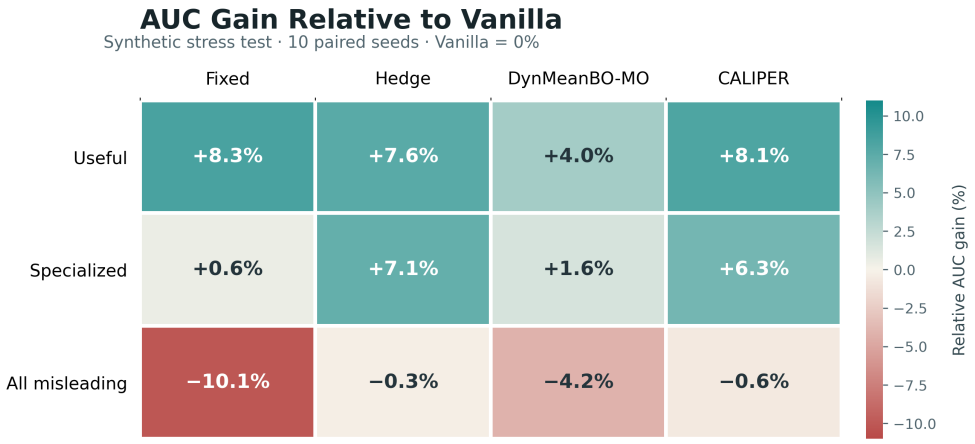

Figure 4: Controlled prior-reliability stress test. Cells report trajectory-AUC change relative to no-prior qLogNEHVI over ten paired seeds. The three prior families isolate the behaviors expected from calibration: exploit broadly useful advice, route objective-specialized advice, and abstain from jointly misleading advice.

4.3 RQ2: Why Are Relative and Absolute Reliability Both Necessary?

Table 2: Complexity ablation under the ten-seed formal molecule protocol. Values are trajectory AUC (mean ± SEM). CALIPER-Lite keeps objective-wise reputation and a committee attenuation gate but removes self-reported-confidence weighting and counterfactual confidence states.

Method	ESOL	FreeSolv	Lipophilicity
Plain Hedge	0.6651 ± 0.0108	0.4369 ± 0.0119	0.8691 ± 0.0131
CALIPER-Lite	0.6718 ± 0.0125	0.4582 ± 0.0170	0.8796 ± 0.0071
CALIPER	$\mathbf{0.6931 \pm 0.0091}$	$\mathbf{0.4859 \pm 0.0238}$	$\mathbf{0.8987 \pm 0.0078}$

The molecule ablation forms a reliability hierarchy. Plain Hedge changes relative expert composition. CALIPER-Lite adds objective-level committee attenuation. Full CALIPER also tests self-reported confidence through counterfactual replay. Mean trajectory AUC increases at every stage on all three datasets: $0.6651 \to 0.6718 \to 0.6931$ on ESOL, $0.4369 \to 0.4582 \to 0.4859$ on FreeSolv, and $0.8691 \to 0.8796 \to 0.8987$ on Lipophilicity. The paired Lite–Full interval excludes zero on ESOL and Lipophilicity. Full final-HV results and the exact Lite gate are in Appendix C.4.

The controlled priors make each mechanism visible. With useful priors, CALIPER gains 8.1% AUC, retaining the acceleration available from fixed injection ($+8.3\%$). With objective-specialized experts, relative routing becomes decisive: Plain Hedge and CALIPER gain 7.1% and 6.3%, compared with 0.6% for a fixed committee. With all experts misleading, the required action changes from selection to abstention: CALIPER stays within 0.6% of the no-prior trajectory while fixed injection loses 10.1%. Real benchmarks establish performance under naturally mixed LLM behavior; these controlled priors identify the exploit, route, and abstain mechanisms that produce it. Full AUC and final-HV values appear in Appendix E.2.

4.4 RQ3: Does Evidence Calibration Persist as the Search Scales?

To isolate the scaling behavior of evidence calibration from changes in prior quality, model stochasticity, and prior-generation cost, we hold the reliability structure fixed while varying the candidate-pool size. At pool sizes 1,000, 3,000, and 10,000, CALIPER achieves AUCs of 0.4272, 0.4417, and 0.4394, compared with 0.3932, 0.3525, and 0.4135 for no-prior qLogNEHVI. Mean T90 falls from 16.0 to 12.7, from 16.0 to 11.0, and from 15.7 to 11.0 evaluations, respectively. Thus increasing the candidate pool does not dilute the evidence signal used to calibrate prior influence. Table 21 in Ap-

pendix E.3 reports the full three-seed experiment. An archived qLogEHVI/qLogNEHVI study further confirms that the same mean-layer interface plugs into both acquisitions; results are in Appendix C.1.

5 DISCUSSION

External knowledge as a learned resource. Fixed injection, time decay, and expert reweighting encode different beliefs about reliability. Fixed injection treats prior quality as given. Time decay treats prior value as a function of iteration. Relative reweighting learns who is best within the committee. CALIPER instead learns a two-dimensional trust structure from objective feedback: composition determines *where* advice comes from, and amplitude determines *whether* external advice should shape the search. The controlled prior families show why both dimensions matter. Useful guidance should be exploited, specialized guidance should be routed, and jointly misleading guidance should be rejected. This is the central operational principle of evidence-calibrated priors.

Modularity and uncertainty. The residual-mean construction makes CALIPER a calibration layer between an external source and an existing optimizer. By preserving the GP covariance and posterior-sampling interface, it integrates directly with modern Monte Carlo acquisitions such as qLogEHVI and qLogNEHVI while keeping their optimization procedures intact. External-source uncertainty is concentrated in the evidence-calibrated mean layer; a low-rank or hierarchical covariance model could further represent correlated expert error while preserving the same acquisition interface.

Prior provenance and controlled objective access. Evidence calibration is particularly useful when the provenance of a model prediction is opaque. The gate evaluates predictive agreement with newly revealed black-box outcomes, whether the original signal arose from reasoning, retrieval, pretraining exposure, or model bias. It therefore reduces dependence on unverifiable assumptions about intrinsic LLM reliability. In the molecule protocol, QED is revealed through the same black-box channel as the dataset property to equalize information across methods. Deployments where QED is available analytically can expose it as a cheap objective or constraint while applying CALIPER to the genuinely expensive objectives.

Beyond LLM priors. CALIPER is fundamentally an evidence-calibrated interface between a sequential optimizer and uncertain external knowledge. Any source that produces candidate-level signals can occupy the prior side of the interface, including pretrained predictors, mechanistic models, physics simulators, heuristic scores, retrieval systems, and human experts. The identity of the provider is secondary to a common requirement: subsequent observations must be able to test its empirical reliability.

The same pattern—external advice, online evidence, adaptive trust, and a fallback learner—also appears beyond BO. In contextual bandits, expert predictions can initialize action values and earn state-dependent influence from realized rewards. In reinforcement learning, external systems can provide action, value, reward, or dynamics priors, with a CALIPER-style gate controlling their contribution to a residual policy or value learner. Extending the mechanism to these settings introduces state-dependent reliability, distribution shift, and long-horizon credit assignment, offering a direct path from evidence-calibrated optimization to evidence-calibrated sequential decision making.

Next steps. The present experiments pair reproducible frozen caches with low-budget finite molecule pools and controlled reliability structures that isolate large-pool and continuous-space behavior. Online prompting would make the advisor itself nonstationary, motivating joint adaptation of the prompt policy and trust state. Query-efficient cache construction can further replace exhaustive candidate–expert scoring in large pools. On the theory side, the combination of data-dependent mean shifts, marginal-likelihood gates, and selective BO feedback calls for regret analysis with a nonstationary abstaining advisor. These directions preserve the same design objective: external knowledge earns influence through evidence.

6 CONCLUSION

We introduced CALIPER, a two-level calibration layer for heterogeneous LLM priors in multi-objective Bayesian optimization. Objective-wise reputation answers which experts to trust; evidence-

calibrated amplitude and a no-prior arm answer whether the committee should influence the search. Counterfactual replay makes self-reported confidence testable, and the residual-mean interface preserves the downstream acquisition pipeline. Across molecule benchmarks, CALIPER produces an asymmetric payoff by extracting useful prior structure while closing the negative-transfer gap when the committee is unreliable. Controlled priors expose the underlying behavior directly: exploit useful advice, route specialized advice, and abstain from misleading advice. More broadly, the value of external knowledge in sequential optimization depends on the learner's ability to verify, route, attenuate, and reject that knowledge from experience.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation, literature organization, manuscript presentation, and parts of the research code implementation. All AI-assisted code and suggestions were reviewed and verified by the authors. All technical claims, experimental results, equations, and final manuscript content were independently checked by the authors, who take full responsibility for the work.

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 *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aVfDrl7xDV.

Virginia Aglietti, Ira Ktena, Jessica Schrouff, Eleni Sgouritsa, Francisco Ruiz, Alan Malek, Alexis Bellot, and Silvia Chiappa. FunBO: Discovering acquisition functions for bayesian optimization with funsearch. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=XjbJR9374o.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A framework for efficient monte-carlo bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21524–21538, 2020.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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 *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.22054.

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

Peter I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. πBO: Augmenting acquisition functions with user beliefs for bayesian optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MMAeCXIa89.

Carl Hvarfner, Frank Hutter, and Luigi Nardi. A general framework for user-guided bayesian optimization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjU0jtXcYn.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel R. Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Joshua Knowles. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. In *IEEE Congress on Evolutionary Computation*, pp. 50–57. IEEE, 2006.

Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alan Aspuru-Guzik, and Geoff Pleiss. A sober look at LLMs for material discovery: Are they actually good for bayesian optimization over molecules? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25603–25622, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

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.

Giang Ngo, Dat Phan Trong, Dang Nguyen, Sunil Gupta, and Svetha Venkatesh. Adaptive acquisition selection for bayesian optimization with large language models. In *International Conference on Learning Representations*, 2026.

Chongqi Qu, Meiqin Liu, Jian Lan, Shanling Dong, and Zhunga Liu. Incorporating expert priors into bayesian optimization via dynamic mean decay. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ylX5PWfK0v.

Richard Cornelius 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*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/c03a2610bca2712b984b331fd4f7bb6f-Abstract-Conference.html.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

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/forum?id=LktUOZayG9.

Yimeng Zeng, Natalie Maus, Haydn Jones, Jeffrey Tao, Fangping Wan, Marcelo Der Torossian Torres, Cesar de la Fuente, Ryan Marcus, Osbert Bastani, and Jacob Gardner. Scaling multi-task bayesian optimization with large language models. In *International Conference on Learning Representations*, 2026.

A Experimental Protocol and Reproducibility

A.1 Candidate Pools, Objectives, and Features

The molecule benchmarks are discrete candidate-pool optimization problems. ESOL uses a 100-molecule pool sampled from the MoleculeNet ESOL dataset, FreeSolv uses a 100-molecule pool sampled from MoleculeNet FreeSolv, and Lipophilicity uses a 300-molecule pool sampled from MoleculeNet Lipophilicity. For all reported BO runs, the surrogate-model input features are RDKit descriptors excluding QED: molecular weight, logP, topological polar surface area, hydrogen-bond donors and acceptors, rotatable bonds, and ring count. Each feature is min–max normalized within the corresponding candidate pool. QED is used only as an optimization objective and is not included in the surrogate-model input features.

All reported molecule experiments optimize two normalized objectives. ESOL uses normalized measured log solubility and QED. FreeSolv uses normalized hydration favorability, defined as the negated experimental hydration free energy, and QED. Lipophilicity uses a windowed experimental lipophilicity score and QED. Prepared data files also include an RDKit logP-window objective for future experiments, but it is not used in the reported two-objective runs.

QED feature handling. The prepared molecule CSV files store the raw QED value and its normalized column `x_qed` for auditing and for possible cheap-property diagnostics. However, `x_qed` is not included in the surrogate feature list for the reported molecule BO experiments, and it is not included in the LLM prompt candidate fields below. Thus the second objective is not directly provided as an input feature to the BO model in the reported real-molecule tables. QED is revealed to a method only through the observed objective vector of a selected candidate. In a deployment where QED is treated as a known cheap property, it would be more appropriate to model it as a cheap objective or constraint; the experiments here instead keep the two objectives under the same finite-pool evaluation protocol for comparability.

Table 3: Surrogate input features used in the reported molecule BO experiments. QED is optimized as objective$_1$ but is excluded from the BO input features to avoid target leakage.

Dataset	Surrogate input features
ESOL	`x_mol_wt`, `x_logp`, `x_tpsa`, `x_hbd`, `x_hba`, `x_rot_bonds`, `x_rings`
FreeSolv	`x_mol_wt`, `x_logp`, `x_tpsa`, `x_hbd`, `x_hba`, `x_rot_bonds`, `x_rings`
Lipophilicity	`x_mol_wt`, `x_logp`, `x_tpsa`, `x_hbd`, `x_hba`, `x_rot_bonds`, `x_rings`

A.2 Unified Evaluation Protocol and Statistical Reporting

Table 4: Unified molecule benchmark protocol used for the main same-protocol experiments.

Dataset	Pool size	Initial design	Budget	Formal seeds
ESOL	100	8	24	10
FreeSolv	100	8	24	10
Lipophilicity	300	8	24	10

Statistical reporting. The primary qLogNEHVI adaptive-baseline comparison uses ten paired formal seeds (0–9). DynMeanBO-MO and Plain Hedge hyperparameters are selected on five disjoint development seeds (10–14) and are never tuned on the formal seeds. The acquisition-portability comparison uses the archived paired subset 0–4 because qLogEHVI was collected on that subset. Reported error bars are standard errors of the mean (SEM), computed as the across-seed sample standard deviation divided by $\sqrt{n}$. Candidate-pool scaling uses three seeds. For T90/T95, a run that does not reach the target within the evaluation budget receives the predeclared budget+1 penalty; these values are therefore descriptive penalized hitting-time summaries rather than survival-analysis estimates.

A.3 LLM PRIOR CACHES

Cache construction. For the primary ESOL, FreeSolv, and Lipophilicity experiments, we generated all molecule-prior caches with Qwen3.6-Flash through an OpenAI-compatible chat API at temperature 0.0. For each candidate in the fixed pool, the model was queried once per expert role before BO began. Each dataset uses three roles: one specialist for each objective and one balanced expert. Each cached record contains objective-wise scores in $[0, 1]$, one self-reported confidence value shared across objectives, and a free-text rationale. All compared BO methods read the same frozen cache for a given dataset; only observed objective values are used for online trust updates. Accordingly, we denote expert confidence by $c_e(x)$ and objective-specific scores by $\mu_{ej}(x)$.

A.4 FIXED HYPERPARAMETERS AND IMPLEMENTATION

Table 5: Fixed CALIPER hyperparameters used for the main molecule experiments.

Component	Hyperparameter	Value
Reputation market	reward step η	0.45
Reputation market	discount λ	0.015
Reputation market	capital softmax temperature	0.55
Market trust	τ_χ / a_χ	0.48 / 7.0
Reliability weights	$\alpha_{\min}$ / $\alpha_{\max}$	0.02 / 2.5
Prior gate	evidence score	GP marginal likelihood
Prior gate	η_p / min updates / count scale	1.0 / 4 / 4.0
Prior gate	Δ_{drop}	0.05
Prior evidence GP	noise / lengthscale	0.05 / median distance
Update gate	η_u	1.0
Diagnostic acquisition	β_t / disagreement γ	$1 + 1.25(1 - \bar{\alpha}_t)$ / 0.30

Hyperparameter selection. The values in Table 5 are shared across the main molecule experiments and are not tuned separately for any dataset or formal evaluation seed. They were fixed before the formal evaluation on seeds 0–9 and were not selected using those results. We therefore treat them as common engineering defaults for the normalized optimization setting rather than dataset-specific fitted parameters.

The reputation-market step size, discount, temperature, and clipping bounds were chosen to keep capital updates on an $O(1)$ logit scale under the bounded reward update. The market-trust threshold is placed near neutral reputation, while its slope produces a smooth transition rather than a hard trust switch. For the counterfactual gates, we use unit learning rates together with a four-observation minimum and finite-data shrinkage, so that early rounds default to the conservative no-confidence behavior before sufficient residual evidence is available. The prior-evidence GP uses a small fixed noise level on the normalized residual scale and a median-distance lengthscale over the normalized candidate features.

The scalarized residual-UCB constants are used only in the standalone diagnostic acquisition and do not affect the same-acquisition qLogNEHVI results. Appendix D reports sensitivity checks for the two gate parameters that most directly control prior abstention, Δ_{drop} and η_p.

Evaluation protocol. All methods use the same initial-design size, budget, candidate pool, and seed set within each dataset. Hypervolume is computed on normalized maximization objectives with the origin as the reference point. We report final hypervolume, AUC hypervolume averaged over BO steps, and the best objective-sum value. Strong MOBO baselines use BoTorch implementations where available; prior-layer methods wrap the residual posterior for qLogEHVI or qLogNEHVI, or use the scalarized residual-UCB diagnostic acquisition described in Appendix B.1.

A.5 Held-Out Prior Baselines and Adaptive-Baseline Tuning

All primary qLogNEHVI prior baselines are evaluated on the same formal seeds 0–9. The fixed-prior control requires no tuning and uses the frozen prior without online adaptation. DynMeanBO-MO and Plain Hedge isolate two adaptive alternatives to full evidence calibration; their global hyperparameters are selected on development seeds 10–14 using equal-dataset raw-HV AUC over evaluations 8–24, with ties broken by higher final HV and then the smaller hyperparameter, and are then locked before formal evaluation.

Fixed prior. Relative to no-prior qLogNEHVI, fixed injection is not statistically distinguishable on ESOL or Lipophilicity for final HV, AUC, or HV@16. FreeSolv is the clear negative case: fixed prior reduces HV@16 by -0.0804 (paired bootstrap 95% CI $[-0.1424, -0.0049]$), delays penalized T90 by $+3.7$ evaluations ($[+0.1, +7.3]$), and delays T95 by $+5.5$ ($[+1.7, +9.0]$). Its final-HV effect is -0.0477 with CI $[-0.1008, +0.0014]$, so the terminal difference itself remains uncertain.

DynMeanBO-MO. DynMeanBO was introduced for single-objective BO with a dynamic prior mean

$$m_n(x) = \gamma_n m_{\text{prior}}(x) + (1 - \gamma_n)\mu_0(x), \qquad \gamma_n = \exp[-\lambda(n - N_0)],$$

where N_0 is the initial-design size and $\lambda > 0$ controls mean decay (Qu et al., 2026). We use a transparent finite-pool multi-objective adaptation: each objective receives its own prior mean, constructed by taking the unweighted mean of the frozen cached expert scores and normalizing it over the active candidate pool. No CALIPER confidence, expert reputation, market trust, or online gate is used. We set DynMeanBO's prior-sampled initialization ratio to $\rho = 0$, so the adaptive comparison isolates dynamic GP-mean decay rather than a prior-biased initialization. On development seeds 10–14, we sweep $\lambda \in \{0.025, 0.05, 0.10, 0.20, 0.40\}$ under the pre-specified selection rule above. The sweep completes 75/75 jobs without failures and locks $\lambda = 0.20$ (development equal-dataset AUC 0.693164) before formal evaluation on seeds 0–9.

Plain Hedge. For each objective j, Hedge forms a full-scale prior mean from the same frozen cached scores using simplex-normalized expert weights. After observing selected candidate x_t, weights update as

$$w_{e,j} \leftarrow w_{e,j} \exp\Big[-\eta\big(\mu_{ej}(x_t) - y_{tj}\big)^2\Big], \qquad \sum_e w_{e,j} = 1.$$

The baseline deliberately omits reported confidence, confidence calibration, counterfactual gates, market trust, depreciation, amplitude shrinkage, and abstention. We sweep one global $\eta \in \{0.25, 0.5, 1, 2, 4, 8\}$ on development seeds 10–14 under the same selection rule. The sweep completes 90/90 jobs without failures and locks $\eta = 2.0$ (development equal-dataset AUC 0.690429) before formal evaluation on seeds 0–9.

Table 6: Development-only hyperparameter selection on seeds 10–14. The pre-registered criterion is equal-dataset mean raw-HV AUC over evaluations 8–24, with ties broken by higher final HV and then the smaller hyperparameter. Bold marks the locked setting for each baseline.

λ	AUC	HV@16	Final HV
0.200	**0.693164**	**0.711188**	0.762494
0.100	0.691363	0.705007	0.758221
0.400	0.689524	0.697623	**0.762667**
0.050	0.688505	0.695748	0.757290
0.025	0.681616	0.689614	0.750696

DynMeanBO-MO (75/75 jobs; zero failures)

η	AUC	HV@16	Final HV
2.00	**0.690429**	**0.694805**	**0.757053**
0.50	0.687347	0.693415	0.747855
0.25	0.685631	0.681417	0.749259
4.00	0.681345	0.681457	0.742879
1.00	0.679320	0.691199	0.731523
8.00	0.678756	0.687343	0.730808

Plain Hedge (90/90 jobs; zero failures)

Table 7: Paired formal evidence against the two adaptive reweighting baselines. Values are mean ± SEM over formal seeds 0–9. Effects are CALIPER minus baseline; positive values favor CALIPER. Confidence intervals are paired bootstrap 95% CIs with 20,000 draws. W/T/L is reported from the perspective of CALIPER.

Comparator	Dataset	Metric	Effect	95% CI	CALIPER W/T/L
DynMeanBO-MO	ESOL	AUC	$+0.0063$	$[-0.0064, +0.0207]$	6/0/4
	FreeSolv	AUC	$\mathbf{+0.0421}$	$[\mathbf{+0.0082}, \mathbf{+0.0781}]$	8/0/2
	Lipophilicity	AUC	$+0.0056$	$[-0.0084, +0.0217]$	6/0/4
	ESOL	Final HV	$+0.0016$	$[-0.0087, +0.0124]$	5/1/4
	FreeSolv	Final HV	$+0.0058$	$[0.0000, +0.0174]$	1/9/0
	Lipophilicity	Final HV	$\mathbf{+0.0199}$	$[\mathbf{+0.0030}, \mathbf{+0.0386}]$	4/3/3
Plain Hedge	ESOL	AUC	$\mathbf{+0.0279}$	$[\mathbf{+0.0164}, \mathbf{+0.0405}]$	10/0/0
	FreeSolv	AUC	$\mathbf{+0.0490}$	$[\mathbf{+0.0130}, \mathbf{+0.0866}]$	8/0/2
	Lipophilicity	AUC	$\mathbf{+0.0296}$	$[\mathbf{+0.0040}, \mathbf{+0.0560}]$	8/0/2
	ESOL	Final HV	$\mathbf{+0.0350}$	$[\mathbf{+0.0170}, \mathbf{+0.0533}]$	8/0/2
	FreeSolv	Final HV	$\mathbf{+0.0997}$	$[\mathbf{+0.0599}, \mathbf{+0.1391}]$	9/1/0
	Lipophilicity	Final HV	$\mathbf{+0.0281}$	$[\mathbf{+0.0045}, \mathbf{+0.0536}]$	6/1/3

Against Plain Hedge, all six paired 95% confidence intervals for AUC and final HV exclude zero in favor of CALIPER. Against tuned DynMeanBO-MO, CALIPER has significant gains on FreeSolv AUC and Lipophilicity final HV, while the corresponding ESOL differences are not statistically distinguishable.

B Algorithmic Details

B.1 Standalone Residual-UCB Diagnostic and Confidence Variants

The main claims use qLogEHVI and qLogNEHVI with the posterior-shifted prior layer. We additionally use a randomized scalarized UCB rule as a lightweight diagnostic acquisition for residual-prior variants:

$$A_t(x) = \sum_{j=1}^{m} w_{j,t} \left(\hat{f}_{j,t}(x) + \beta_t s^r_{j,t}(x) \right) + \gamma(1 - \bar{\alpha}_t) \sum_{j=1}^{m} w_{j,t} D_j(x).$$

Here w_t is a random objective weight vector, $s^r_{j,t}$ is residual posterior uncertainty, $D_j(x)$ is expert disagreement for objective j, $\bar{\alpha}_t = \mathrm{clip}(E^{-1}m^{-1}\sum_{e,j} \alpha_{ej,t}, 0, 1)$ is average market trust, $\beta_t = 1+1.25(1-\bar{\alpha}_t)$, and $\gamma = 0.30$. This diagnostic acquisition is not used by the qLogEHVI/qLogNEHVI prior-layer results.

Before fixing the final counterfactual gate, we also evaluate raw confidence, no confidence, prior-only confidence, update-only confidence, and adaptive confidence variants. These are diagnostics rather than separate proposed methods. For the correlation-based adaptive variant, each objective tracks whether reported confidence predicts soft expert success, $\rho_j = \mathrm{corr}(c_e(x_t), q_{ej,t})$, and forms a gate $g_j = \sigma(a_\rho(\rho_j - \tau_\rho))$. Effective confidence is then $\tilde{c}_{e,j}(x) = 1 + g_j(c_e(x) - 1)$. The final method instead separates confidence in prior aggregation from confidence in reputation updates, as detailed below.

B.2 Residual-Prior Bayesian Optimization

The timing convention used throughout the implementation is that quantities indexed by t are the state after the t-th observation has been incorporated. Thus the prior $p_{j,t}$ is constructed from $\mathbf{K}_t$ and $\pi^p_{j,t}$, and is used only for selecting the next candidate. For all t,

$$x_t = A(\Theta_{t-1}), \qquad y_t = \mathbf{f}(x_t), \qquad \Theta_t = U(\Theta_{t-1}, x_t, y_t).$$

Consequently, the surrogate used to select x_t was built from $\mathcal{D}_{t-1}$, not from y_t.

After observation t, the selected prior is held fixed while fitting the residual GP to all observations:

$$\widetilde{y}_{j,\tau,t} = y_{\tau j} - p_{j,t}(x_\tau), \qquad \tau = 1, \ldots, t,$$

$$r_j(x) \sim \mathcal{GP}(0, k_j), \qquad f_j(x) = p_{j,t}(x) + r_j(x).$$

If

$$r_j(x) \mid \mathcal{D}_t \sim \mathcal{N}\big(m^r_{j,t}(x), v^r_{j,t}(x)\big),$$

then the predictive distribution of the original objective is

$$f_j(x) \mid \mathcal{D}_t \sim \mathcal{N}\big(p_{j,t}(x) + m^r_{j,t}(x), v^r_{j,t}(x)\big).$$

The LLM prior therefore shifts the posterior mean while leaving the residual GP posterior variance unchanged.

B.3 Reputation-Market Update

At observation t, let $n = t - 1$ and compute the objective scale from $\mathcal{D}_{t-1}$:

$$s_{j,t} = \begin{cases} 0.25, & n < 2, \\ \max\left(\sqrt{\dfrac{1}{n-1}\sum_{\tau<t}(y_{\tau j} - \bar{y}_{j,t})^2}, s_{\min}\right), & n \geq 2, \end{cases} \qquad \bar{y}_{j,t} = \frac{1}{n}\sum_{\tau<t} y_{\tau j},$$

with $s_{\min} = 0.08$. The expert error and reward are

$$\epsilon_{ej,t} = \frac{|\mu_{ej}(x_t) - y_{tj}|}{s_{j,t}}, \qquad R_{ej,t} = \mathrm{clip}\left(\frac{1}{2} - \frac{1}{2}\epsilon^2_{ej,t}, -2, 0.75\right).$$

The unclipped expression is at most 0.5, so the upper clipping value 0.75 is inactive; it is retained to match the implementation.

Let $K_{ej,t}$ be the real-market capital after incorporating observations through t, with

$$K_{ej,0} = \mathrm{logit}(q^{(0)}_{ej}), \qquad q^{(0)}_{ej} = 0.5,$$

hence $K_{ej,0} = 0$. The capital update is given in the counterfactual update-gate subsection below because its confidence multiplier must use the pre-observation gate.

For any capital matrix $\mathbf{K}_t$, define the objective-level market trust

$$\chi_j(\mathbf{K}_t) = \sigma\left[a_\chi\left(\frac{1}{E}\sum_{e=1}^{E}\sigma(K_{ej,t}) - \tau_\chi\right)\right],$$

and the expert-objective reliability weight

$$\alpha_{ej,t} = \mathrm{clip}\left[\chi_j(\mathbf{K}_t)\, E \frac{\exp(K_{ej,t}/T)}{\sum_{e'=1}^{E}\exp(K_{e'j,t}/T)}, \alpha_{\min}, \alpha_{\max}\right].$$

The main configuration uses

$$a_\chi = 7, \qquad \tau_\chi = 0.48, \qquad T = 0.55, \qquad \alpha_{\min} = 0.02, \qquad \alpha_{\max} = 2.5.$$

The confidence-aware aggregation is written through a single generic function. For $d_e(x) \geq 0$, define

$$v_{ej}(x; \mathbf{K}_t, d) = \max\{\alpha_{ej,t} d_e(x), 10^{-8}\},$$

$$\bar{\mu}_j(x; \mathbf{K}_t, d) = \frac{\sum_{e=1}^{E} v_{ej}(x; \mathbf{K}_t, d)\mu_{ej}(x)}{\sum_{e=1}^{E} v_{ej}(x; \mathbf{K}_t, d)},$$

$$\zeta_j(x; \mathbf{K}_t, d) = \mathrm{clip}\left[\frac{1}{E}\sum_{e=1}^{E} v_{ej}(x; \mathbf{K}_t, d), 0, 1\right],$$

and

$$G_j(x; \mathbf{K}_t, d) = \zeta_j(x; \mathbf{K}_t, d)\bar{\mu}_j(x; \mathbf{K}_t, d).$$

The factor ζ_j is the prior-amplitude shrinkage missing from a pure confidence-weighted average: low aggregate market trust reduces the magnitude of the prior instead of only changing its relative expert composition.

For the three prior arms, let

$$d^{(0)}_{ej}(x) = 1, \qquad d^{(1)}_{e}(x) = c_e(x).$$

Then

$$p^{(0)}_{j,t}(x) = G_j(x; \mathbf{K}_t, d^{(0)}), \qquad p^{(1)}_{j,t}(x) = G_j(x; \mathbf{K}_t, d^{(1)}), \qquad p^{(2)}_{j,t}(x) = 0.$$

These correspond to no-confidence, confidence, and drop-prior, respectively.

B.4 Counterfactual Prior Gate

The prior gate is evaluated only after observation t is available. Using the current capital $\mathbf{K}_t$, define for $a \in \{0, 1, 2\}$

$$r^{(a)}_{j,\tau,t} = y_{\tau j} - p^{(a)}_{j,t}(x_\tau), \qquad \tau = 1, \dots, t.$$

The fixed evidence kernel is

$$[\mathbf{C}_t]_{\tau\tau'} = \exp\left(-\frac{\|x_\tau - x_{\tau'}\|_2^2}{2\ell_t^2}\right) + (\sigma_{\mathrm{cf}}^2 + 10^{-6})\mathbb{I}[\tau = \tau'],$$

where $\sigma_{\mathrm{cf}} = 0.05$, and ℓ_t is the median pairwise Euclidean distance between historical candidates, lower-bounded by 0.05. The per-observation GP marginal log-likelihood score is

$$S_{j,t}(a) = -\frac{1}{2t}\left[(\mathbf{r}^{(a)}_{j,t})^\top \mathbf{C}_t^{-1}\mathbf{r}^{(a)}_{j,t} + \log|\mathbf{C}_t| + t\log(2\pi)\right].$$

Relative to the no-confidence arm,

$$h_{j,t}(a) = \mathrm{clip}[S_{j,t}(a) - S_{j,t}(0) - \Delta_a, -10, 10],$$

where

$$\Delta_0 = \Delta_1 = 0, \qquad \Delta_2 = 0.05.$$

The positive Δ_2 explicitly penalizes abstention from the prior.

The raw prior-arm probabilities are

$$\widehat{\pi}^p_{j,t}(a) = \frac{\exp(\eta_p h_{j,t}(a))}{\sum_{a'=0}^2 \exp(\eta_p h_{j,t}(a'))}, \qquad \eta_p = 1.$$

With

$$\kappa_t = \sqrt{\frac{t}{t+\kappa_0}}, \qquad \kappa_0 = 4,$$

the implemented finite-data policy is

$$\pi^p_{j,t} = \begin{cases} (1,0,0), & t < n_{\min}, \\ (1-\kappa_t)(1,0,0) + \kappa_t \widehat{\pi}^p_{j,t}, & t \geq n_{\min}, \end{cases} \qquad n_{\min} = 4.$$

The actual prior used to fit the residual GP is

$$p_{j,t}(x) = \pi^p_{j,t}(0)p^{(0)}_{j,t}(x) + \pi^p_{j,t}(1)p^{(1)}_{j,t}(x).$$

The drop-prior arm contributes zero.

B.5 Counterfactual Update Gate

The update gate answers a different question: should self-reported confidence affect future reputation-credit updates? We maintain two deterministic shadow markets, $g = 0$ and $g = 1$, where confidence is disabled or enabled, respectively.

Let $K^{(g)}_{ej,t}$ denote the shadow capital after replaying observations through t. At the beginning of observation t, the shadow state is therefore $K^{(g)}_{ej,t-1}$. For either shadow capital matrix, the same trust mapping, reliability weights, amplitude gate, and prior gate are used. In particular,

$$\chi_j(\mathbf{K}^{(g)}_{t-1}) = \sigma\left[a_\chi\left(\frac{1}{E}\sum_e \sigma(K^{(g)}_{ej,t-1}) - \tau_\chi\right)\right],$$

$$\alpha^{(g)}_{ej,t-1} = \mathrm{clip}\left[\chi_j(\mathbf{K}^{(g)}_{t-1})\, E\frac{\exp(K^{(g)}_{ej,t-1}/T)}{\sum_{e'}\exp(K^{(g)}_{e'j,t-1}/T)}, \alpha_{\min}, \alpha_{\max}\right],$$

with $v, \bar{\mu}, \zeta, G$ defined exactly as above using $\alpha^{(g)}_{ej,t-1}$. The shadow prior at the pre-observation state is

$$p^{u,(g)}_{j,t-1}(x) = \sum_{a=0}^{1} \pi^p_{j,t-1}(a) G_j\left(x; \mathbf{K}^{(g)}_{t-1}, d^{(a)}\right).$$

The same prior-gate probabilities $\pi^{p}_{j,t-1}$ are used for both counterfactual update rules; only the replayed reputation history differs.

Once y_t is observed, both shadows receive full-information loss:

$$L^{u}_{j,t}(g) = \frac{|y_{tj} - p^{u,(g)}_{j,t-1}(x_t)|}{s_{j,t}},$$

$$\bar{L}^{u}_{j,t}(g) = L^{u}_{j,t}(g) - \frac{1}{2}\sum_{g'=0}^{1} L^{u}_{j,t}(g').$$

The centered-loss logit update is

$$z^{u}_{j,t}(g) = \mathrm{clip}\left[z^{u}_{j,t-1}(g) - \eta_u \bar{L}^{u}_{j,t}(g), -8, 8\right], \qquad \eta_u = 1,$$

followed by

$$\widehat{\pi}^{u}_{j,t} = \mathrm{softmax}_{g\in\{0,1\}}\big(z^{u}_{j,t}(g)\big).$$

The final update-gate distribution is

$$\pi^{u}_{j,t} = \begin{cases} (1,0), & t < n_{\min}, \\ (1-\kappa_t)(1/2,1/2) + \kappa_t \widehat{\pi}^{u}_{j,t}, & t \geq n_{\min}. \end{cases}$$

Thus the confidence-enabled probability used in future real-market updates is $\pi^{u}_{j,t}(1)$, not the gate after seeing the same observation.

For the real market at observation t, use the pre-observation probability

$$\rho_{j,t} = \pi^{u}_{j,t-1}(1), \qquad \widetilde{c}^{u}_{e,t} = 1 + \rho_{j,t}\big(c_e(x_t) - 1\big).$$

The real capital update is

$$K_{ej,t} = \mathrm{clip}\left[(1-\lambda)K_{ej,t-1} + \eta\, \widetilde{c}^{u}_{e,t} R_{ej,t}, K_{\min}, K_{\max}\right],$$

with

$$\eta = 0.45, \qquad \lambda = 0.015, \qquad K_{\min} = -3, \qquad K_{\max} = 3.$$

Because $\rho_{j,t}$ depends only on $\mathcal{D}_{t-1}$, the newly observed y_t cannot retroactively decide how much confidence contributes to its own reputation reward.

Finally, the same observation is replayed into both shadow markets:

$$\omega^{(0)}_{ej,t} = 1, \qquad \omega^{(1)}_{ej,t} = c_e(x_t),$$

$$K^{(g)}_{ej,t} = \mathrm{clip}\Big[(1-\lambda)K^{(g)}_{ej,t-1} + \eta\, \omega^{(g)}_{ej,t} R_{ej,t}, K_{\min}, K_{\max}\Big], \qquad g \in \{0,1\}.$$

The new shadow states and the new $\pi^{u}_{j,t}$ are used only in later rounds. After these post-observation updates, the prior gate is evaluated from $\mathcal{D}_t$ and $\mathbf{K}_t$, yielding $p_{j,t}$, which is then used to fit the residual GP and select x_{t+1}. This ordering makes the no-lookahead property explicit:

$$\underbrace{\Theta_{t-1}}_{\text{available before } x_t} \longrightarrow \underbrace{x_t}_{\text{selected}} \longrightarrow \underbrace{y_t}_{\text{observed}} \longrightarrow \underbrace{\Theta_t}_{\text{all gates/reputation updated}}.$$

B.6 PSEUDOCODE

Algorithm 1: CALIPER Evidence-Gated LLM Prior Layer
Input: candidate pool $\mathcal{X}$, cached expert priors $\mu_{ej}(x)$, scalar confidences $c_e(x)$, initial design size n_0, budget B.
Initialize: evaluate the n_0 initial-design candidates in a fixed chronological order and incorporate their observations with the same update rules to obtain Θ_{n_0}. Initialize $K_{ej,0} = 0$, $K^{(0)}_{ej,0} = K^{(1)}_{ej,0} = 0$, $z^u_{j,0} = (0,0)$, $\pi^u_{j,0} = (1,0)$, and the prior gate to $\pi^p_{j,0} = (1,0,0)$. For $t < n_{\min}$, the corresponding hard conservative gate is used.

1. For $t = n_0, \ldots, B-1$, use the state Θ_t to construct the current prior $p_{j,t}$, fit residual GPs, and select
$$x_{t+1} = \arg\max_{x \in \mathcal{X} \setminus \mathcal{X}^{\text{obs}}_t} A_t(x).$$
2. Evaluate x_{t+1} to obtain $\mathbf{y}_{t+1}$. Compute $s_{j,t+1}$, $\epsilon_{ej,t+1}$, and $R_{ej,t+1}$ using $\mathcal{D}_t$ and the new observation.
3. Before using the new observation to update any gate, set
$$\rho_{j,t+1} = \pi^u_{j,t}(1), \qquad \widetilde{c}^u_{ej,t+1} = 1 + \rho_{j,t+1}\big(c_e(x_{t+1}) - 1\big),$$
and update the real capital $K_{ej,t+1}$.
4. Using the pre-observation shadow states $K^{(g)}_t$ and prior-gate state $\pi^p_{j,t}$, compute $p^{u,(g)}_{j,t}(x_{t+1})$, evaluate both counterfactual losses, update $z^u_{j,t+1}$ and $\pi^u_{j,t+1}$, and replay the new observation into both shadow markets to obtain $K^{(g)}_{t+1}$.
5. Using the updated real capital K_{t+1} and the complete history $\mathcal{D}_{t+1}$, evaluate the three prior arms, the fixed-kernel evidence scores, and $\pi^p_{j,t+1}$. Form $p_{j,t+1}$ from the nonzero arms.
6. Fit residual GPs to $y_{\tau j} - p_{j,t+1}(x_\tau)$ for $\tau = 1, \ldots, t+1$, yielding Θ_{t+1}, and continue to the next candidate.

Output: evaluated candidates and the final non-dominated set.

Figure 5: Chronological pseudocode for CALIPER. Each candidate is selected from the state formed before its own objective observation; all reputation and gate updates using that observation are performed only afterward.

C ADDITIONAL MOLECULE DIAGNOSTICS

This section expands the main three research questions with acquisition-interface tests, prior-quality audits, confidence ablations, and gate trajectories.

C.1 ACQUISITION PORTABILITY

Table 8: Acquisition-portability diagnostic on the shared formal-seed subset 0–4. CALIPER is attached to qLogEHVI and qLogNEHVI without modifying their acquisition logic. Values are final hypervolume, reported as mean ± SEM over five paired seeds; the ten-seed qLogNEHVI comparison in Table 1 is the primary formal result. Bold indicates the best result within each acquisition family.

Method	ESOL	FreeSolv	Lipophilicity
qLogEHVI	0.7124 ± 0.0178	0.6102 ± 0.0116	0.9197 ± 0.0103
qLogEHVI + fixed prior	0.7256 ± 0.0141	0.5726 ± 0.0254	0.9179 ± 0.0070
qLogEHVI + CALIPER	$\mathbf{0.7458 \pm 0.0060}$	$\mathbf{0.6218 \pm 0.0000}$	$\mathbf{0.9404 \pm 0.0025}$
qLogNEHVI	0.7161 ± 0.0115	$\mathbf{0.6218 \pm 0.0000}$	$\mathbf{0.9291 \pm 0.0063}$
qLogNEHVI + fixed prior	0.7256 ± 0.0141	0.5743 ± 0.0253	0.9241 ± 0.0094
qLogNEHVI + CALIPER	$\mathbf{0.7430 \pm 0.0087}$	$\mathbf{0.6218 \pm 0.0000}$	0.9278 ± 0.0105

Table 9: Standalone residual-UCB molecule diagnostics under the unified five-seed, budget-24 protocol. Values are mean ± SEM. The standalone CALIPER variant is a diagnostic acquisition and is distinct from the qLogEHVI/qLogNEHVI prior-layer experiments.

Dataset	Method	AUC HV ↑	Final HV ↑
ESOL	Random search	0.6419 ± 0.0195	0.6642 ± 0.0188
	Vanilla MOBO	0.6532 ± 0.0207	0.7077 ± 0.0141
	ParEGO	0.6455 ± 0.0190	0.6863 ± 0.0164
	qLogNEHVI	0.6586 ± 0.0179	0.7161 ± 0.0115
	Residual-UCB + fixed prior	0.6616 ± 0.0208	0.7087 ± 0.0181
	Residual-UCB + CALIPER	**0.6860 ± 0.0118**	**0.7450 ± 0.0038**
FreeSolv	Random search	0.3257 ± 0.0374	0.4151 ± 0.0504
	Vanilla MOBO	0.4563 ± 0.0528	0.6061 ± 0.0157
	ParEGO	0.3981 ± 0.0393	0.5081 ± 0.0569
	qLogNEHVI	0.4936 ± 0.0450	**0.6218 ± 0.0000**
	Residual-UCB + fixed prior	0.4722 ± 0.0211	0.5846 ± 0.0266
	Residual-UCB + CALIPER	**0.5417 ± 0.0172**	**0.6218 ± 0.0000**
Lipophilicity	Random search	0.8479 ± 0.0221	0.8772 ± 0.0167
	Vanilla MOBO	0.8883 ± 0.0148	0.9142 ± 0.0109
	ParEGO	0.8802 ± 0.0222	0.9196 ± 0.0131
	qLogNEHVI	**0.8982 ± 0.0102**	**0.9291 ± 0.0063**
	Residual-UCB + fixed prior	0.8919 ± 0.0102	0.9160 ± 0.0084
	Residual-UCB + CALIPER	0.8925 ± 0.0169	0.9276 ± 0.0094

Table 10: Lipophilicity prior-quality diagnostic on 50 shared candidates. Objective 0 is the lipophilicity-window objective and objective 1 is QED; lower MAE is better, and negative bias indicates systematic underestimation.

Prior source	Obj. 0 MAE (↓)	Obj. 0 bias	Obj. 1 MAE (↓)	Obj. 1 bias
Qwen3.6-Flash original	0.2677	-0.2314	0.1089	0.0308
Qwen3.6-Flash mixed committee	**0.2379**	-0.1961	**0.1066**	0.0278
Qwen3.6-Flash calibrated prompt	0.2793	-0.2463	0.1083	0.0195
Qwen3.7-Max	0.3235	-0.3038	0.1520	0.0252
DeepSeek-V4-Pro	0.3705	-0.3659	0.1310	-0.0186

Table 8 uses the archived paired seed subset 0–4. On ESOL, CALIPER improves final hypervolume under both qLogEHVI and qLogNEHVI on this subset. FreeSolv saturates under qLogNEHVI and its CALIPER variant. On Lipophilicity, qLogEHVI+CALIPER improves final HV from 0.9197 to 0.9404 (paired effect $+0.0206$, 95% CI $[+0.0046, +0.0390]$), while the paired qLogNEHVI subset remains stable. Together, the two acquisitions demonstrate that the calibrated posterior-mean interface transfers without modifying their acquisition logic; the magnitude of transfer remains acquisition dependent.

C.2 Lipophilicity Prior Quality

Lipophilicity is the clearest prior-quality boundary condition. The cached LLM committee systematically underestimates the window objective. The mixed Qwen3.6 committee reduces objective-0 MAE relative to the original committee, which improves prior-assisted BO, especially for fixed-prior injection, but the remaining bias is sufficient that qLogNEHVI without a prior remains slightly stronger in mean final hypervolume. This diagnostic separates prior quality from the capacity of the prior layer itself.

C.3 Confidence-Role and Core-Gate Ablations

The confidence variants test whether one fixed rule for self-reported confidence is reliable across datasets. Final hypervolume and trajectory AUC need not rank variants identically, and FreeSolv

Table 11: Effect of Lipophilicity LLM-prior quality on downstream qLogNEHVI. The mixed Qwen3.6 committee has lower prior error than the original committee and consistently improves prior-assisted BO. Values are mean ± SEM over five seeds.

Prior source	Method	AUC HV ↑	Final HV ↑
None	qLogNEHVI	**0.8982 ± 0.0102**	**0.9291 ± 0.0063**
Original Qwen3.6	+ fixed prior	0.8858 ± 0.0119	0.9131 ± 0.0088
Original Qwen3.6	+ CALIPER	0.8836 ± 0.0222	0.9270 ± 0.0108
Mixed Qwen3.6	+ fixed prior	0.8872 ± 0.0169	0.9241 ± 0.0094
Mixed Qwen3.6	+ CALIPER	0.8952 ± 0.0126	0.9278 ± 0.0105

saturates in final hypervolume under the unified protocol. These diagnostics motivate evidence-based confidence handling but are not treated as independent main methods.

Table 12: Confidence-phase ablation under the unified molecule protocol. Each cell reports final hypervolume / trajectory AUC (mean ± SEM over five seeds).

Dataset	Raw	No conf.	Prior only	Adaptive	Adaptive prior	Adaptive update
ESOL	**0.7458±0.0029 / 0.6845±0.0127**	0.7153±0.0177 / 0.6652±0.0198	0.7319±0.0098 / 0.6819±0.0144	0.7314±0.0136 / 0.6786±0.0177	0.7337±0.0132 / 0.6789±0.0186	0.7309±0.0125 / 0.6806±0.0178
FreeSolv	0.6218±0.0000 / 0.5363±0.0187	0.6218±0.0000 / 0.5492±0.0176	0.6218±0.0000 / 0.5401±0.0184	0.6218±0.0000 / 0.5401±0.0184	0.6218±0.0000 / 0.5417±0.0172	**0.6218±0.0000 / 0.5501±0.0167**
Lipophilicity	**0.9297±0.0134 / 0.8884±0.0140**	0.9043±0.0090 / 0.8687±0.0134	0.9115±0.0118 / 0.8722±0.0141	0.9058±0.0136 / 0.8774±0.0133	0.9205±0.0150 / 0.8744±0.0180	0.9256±0.0102 / 0.8847±0.0193

Table 13: Core-gate diagnostic under the unified protocol. A legacy CF-V1 variant from an earlier, non-unified rerun is omitted from this controlled comparison. Values are final hypervolume / trajectory AUC; the margin portfolio is retained only as a diagnostic variant.

Dataset	Best fixed confidence phase	CALIPER	Margin portfolio
ESOL	Raw: $0.7458/0.6845$	$0.7450/\mathbf{0.6860}$	$0.7362/0.6761$
FreeSolv	Adaptive update: $0.6218/\mathbf{0.5501}$	$0.6218/0.5417$	$0.6218/0.5395$
Lipophilicity	Raw: $\mathbf{0.9297}/0.8884$	$0.9276/\mathbf{0.8925}$	$0.9258/0.8893$

C.4 Reliability Hierarchy: Routing, Attenuation, and Counterfactual Confidence

CALIPER contains two distinct forms of adaptation. The first determines *who to trust*: expert–objective reputation changes the relative contribution of heterogeneous prior sources, while committee-level evidence controls their aggregate influence. The second determines whether self-reported confidence should affect this process: the full method maintains counterfactual confidence states and separate gates for confidence use in prior aggregation and reputation updates.

To quantify the joint contribution of the confidence-specific calibration subsystem beyond expert routing and committee attenuation, we compare full CALIPER with CALIPER-Lite. The Lite variant retains objective-wise expert reputation, a single objective-level committee gate, and the same residual-GP interface, while removing self-reported-confidence weighting, the confidence-weighted prior arm, the two counterfactual shadow markets, and the counterfactual update gate. Its prior has the form

$$w_{ej,t} = \mathrm{softmax}_e(K_{ej,t}/T), \qquad p^{\mathrm{Lite}}_{j,t}(x) = g_{j,t} \sum_e w_{ej,t}\mu_{ej}(x),$$

where $g_{j,t} \in [0,1]$ is updated from the observed committee error. Shared reputation hyperparameters are inherited from the full method, and the Lite-specific gate is fixed across datasets. The comparison progresses from relative reweighting, to committee attenuation, to counterfactual confidence calibration.

Table 14: Complexity ablation of the full CALIPER mechanism. Values are mean ± SEM over the ten formal molecule seeds. CALIPER-LITE keeps objective-wise expert reputation, a committee-level attenuation gate, and the residual-GP interface, while removing self-reported-confidence weighting and counterfactual confidence states. Shared reputation hyperparameters are inherited from the full method; the Lite-specific gate is not tuned per dataset.

Dataset	Method	AUC ↑	Final HV ↑
ESOL	No prior	0.6810 ± 0.0138	0.7277 ± 0.0098
	Plain Hedge	0.6651 ± 0.0108	0.7067 ± 0.0088
	CALIPER-LITE	0.6718 ± 0.0125	0.7127 ± 0.0105
	CALIPER	$\mathbf{0.6931 \pm 0.0091}$	$\mathbf{0.7417 \pm 0.0083}$
FreeSolv	No prior	$\mathbf{0.4953 \pm 0.0249}$	$\mathbf{0.6122 \pm 0.0095}$
	Plain Hedge	0.4369 ± 0.0119	0.5126 ± 0.0215
	CALIPER-LITE	0.4582 ± 0.0170	0.5941 ± 0.0179
	CALIPER	0.4859 ± 0.0238	$\mathbf{0.6122 \pm 0.0095}$
Lipophilicity	No prior	0.8829 ± 0.0095	0.9166 ± 0.0080
	Plain Hedge	0.8691 ± 0.0131	0.9040 ± 0.0118
	CALIPER-LITE	0.8796 ± 0.0071	0.9182 ± 0.0060
	CALIPER	$\mathbf{0.8987 \pm 0.0078}$	$\mathbf{0.9321 \pm 0.0057}$

Table 14 shows a consistent mean-performance gap between the simplified and full mechanisms. On ESOL, removing the counterfactual confidence subsystem lowers AUC from 0.6931 to 0.6718 and final HV from 0.7417 to 0.7127. On FreeSolv, Lite reaches 0.4582 AUC and 0.5941 final HV, compared with 0.4859 and 0.6122 for full CALIPER. On Lipophilicity, AUC decreases from 0.8987 to 0.8796 and final HV from 0.9321 to 0.9182. In the paired AUC analysis, the interval excludes zero on ESOL and Lipophilicity, whereas the FreeSolv interval overlaps zero.

The comparison also separates three levels of prior adaptation. Plain Hedge changes only the relative expert composition,

$$p_{j,t}^{\mathrm{Hedge}}(x) = \sum_e w_{ej,t}\mu_{ej}(x), \qquad \sum_e w_{ej,t} = 1,$$

and therefore cannot attenuate the committee as a whole. CALIPER-LITE adds an objective-level gate $g_{j,t}$, allowing evidence to shrink the aggregate prior, but it does not use reported confidence or maintain counterfactual confidence states. Full CALIPER additionally treats confidence as a falsifiable signal and evaluates its use through the counterfactual gates.

Descriptively, trajectory AUC follows

$$\text{Plain Hedge} < \text{CALIPER-LITE} < CALIPER$$

on all three molecule datasets. Lite removes a coupled group of confidence states and replaces the three-arm policy with a single committee gate, so the contrast measures that subsystem as one design unit. Its consistent AUC ordering shows that counterfactual confidence calibration contributes beyond relative routing and amplitude control.

These results position self-reported confidence as a falsifiable feature, not an authority signal. The confidence-role diagnostics show that its value changes across datasets; counterfactual replay converts that variation into objective-level evidence and uses confidence where it improves prediction.

D SENSITIVITY DIAGNOSTICS

The main method intentionally fixes a compact hyperparameter set; these runs are supporting diagnostics rather than part of the same-protocol molecule comparison. We nevertheless include sensitivity diagnostics for the two most important counterfactual-gate parameters: the prior-abstention margin Δ_{drop} and the prior-gate inverse temperature η_p.

Table 15: Drop-margin sensitivity on ESOL across LLM backends. Values are final hypervolume (↑); missing cells indicate runs not collected.

Backend	$\Delta_{\text{drop}} = 0.05$ (↑)	$\Delta_{\text{drop}} = 0.25$ (↑)	$\Delta_{\text{drop}} = 1.0$ (↑)
Qwen3.6-Flash	**0.7561**	0.7480	0.7515
Qwen3.7-Max	0.7470	0.7445	**0.7507**
DeepSeek-V4-Pro	**0.7515**	N/A	0.7500

Table 16: Prior-gate temperature sensitivity. Values are final hypervolume (↑) for $\eta_p = 1$ and a sharper $\eta_p = 8$ gate.

Setting	$\eta_p = 1$ (↑)	$\eta_p = 8$ (↑)
ESOL / Qwen3.6-Flash	**0.7561**	0.7416
Synthetic / overconfident bad	0.4788	**0.4791**
Synthetic / correlated bad experts	**0.4461**	**0.4461**

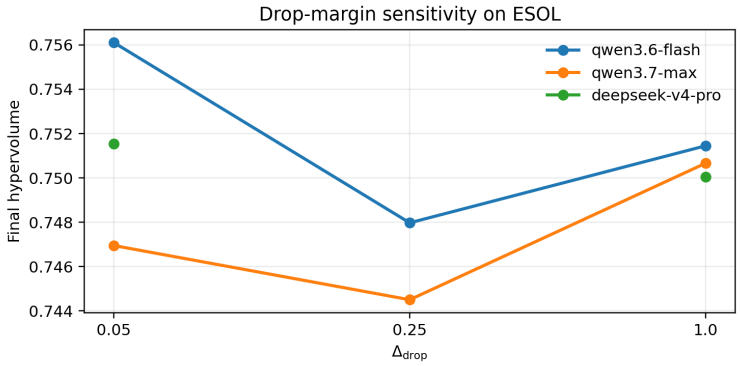

Figure 6: Drop-margin sensitivity on ESOL across three LLM backends. Bars show collected runs; missing backend–margin combinations are labeled N/A. The best margin differs across backends, which motivates using a fixed, transparent margin in the main method.

The sensitivity results support two conclusions. First, the drop margin interacts with the LLM backend: Qwen3.6-Flash prefers a more aggressive abstention margin in the collected runs, while Qwen3.7-Max and DeepSeek-V4-Pro are less aligned with that choice. Second, sharpening the prior-gate temperature does not consistently improve performance. This explains why we avoid treating the margin portfolio as the main method: one-step prior fit is not sufficient to choose the best exploration behavior.

E Additional Benchmark Diagnostics

The following experiments examine how the calibrated prior behaves in simplified scalar tasks, larger finite pools, and a continuous domain.

Table 17: Single-objective multi-expert diagnostics. Best is higher-is-better; regret and hit step are lower-is-better. Hit step reports the mean first step reaching zero simple regret, with hit count in parentheses.

Task	Variant	Method	Best ↑	Regret ↓	Hit step ↓
ESOL scalar	multi-model LLM	Calibrated	**1.0000**	**0.0000**	3.00 (5/5)
ESOL scalar	multi-model LLM	Fixed prior	**1.0000**	**0.0000**	**2.00 (5/5)**
ESOL scalar	multi-model LLM	Random search	0.8672	0.1328	N/A
ESOL scalar	multi-model LLM	Vanilla BO	**1.0000**	**0.0000**	2.20 (5/5)
FreeSolv hard scalar	Qwen3.6-Flash roles	Calibrated	**0.4285**	**0.0000**	**4.40 (10/10)**
FreeSolv hard scalar	Qwen3.6-Flash roles	Fixed prior	0.4155	0.0130	14.50 (2/10)
FreeSolv hard scalar	Qwen3.6-Flash roles	Random search	0.4140	0.0144	9.00 (3/10)
FreeSolv hard scalar	Qwen3.6-Flash roles	Vanilla BO	**0.4285**	**0.0000**	5.70 (10/10)
Lipophilicity hard scalar	Qwen3.6-Flash roles	Calibrated	**0.8608**	**0.0000**	5.10 (10/10)
Lipophilicity hard scalar	Qwen3.6-Flash roles	Fixed prior	**0.8608**	**0.0000**	10.00 (10/10)
Lipophilicity hard scalar	Qwen3.6-Flash roles	Random search	0.8493	0.0115	12.00 (2/10)
Lipophilicity hard scalar	Qwen3.6-Flash roles	Vanilla BO	**0.8608**	**0.0000**	**1.90 (10/10)**
Prompt hard	heuristic experts	Calibrated	**0.9964**	**0.0000**	**4.60 (10/10)**
Prompt hard	heuristic experts	Fixed prior	0.9962	0.0002	11.33 (9/10)
Prompt hard	heuristic experts	Random search	0.9925	0.0039	12.50 (4/10)
Prompt hard	heuristic experts	Vanilla BO	0.9928	0.0036	16.00 (7/10)
Prompt hard	Qwen3.7-Max+DeepSeek-V4-Pro	Calibrated	**0.9964**	**0.0000**	**7.00 (10/10)**
Prompt hard	Qwen3.7-Max+DeepSeek-V4-Pro	Fixed prior	**0.9964**	**0.0000**	**7.00 (10/10)**
Prompt hard	Qwen3.7-Max+DeepSeek-V4-Pro	Random search	0.9925	0.0039	12.50 (4/10)
Prompt hard	Qwen3.7-Max+DeepSeek-V4-Pro	Vanilla BO	0.9928	0.0036	16.00 (7/10)

Table 18: Single-objective calibrated-prior ablations for confidence and market trust. Best is higher-is-better; regret and hit step are lower-is-better. Because most calibrated variants reach zero final simple regret, hit step is the main comparison signal.

Task	Variant	Method	Best ↑	Regret ↓	Hit step ↓
ESOL scalar	multi-model LLM	Calibrated	1.0000	0.0000	3.00 (5/5)
ESOL scalar	multi-model, no confidence	Calibrated	1.0000	0.0000	**2.80 (5/5)**
FreeSolv hard scalar	Qwen3.6-Flash roles	Calibrated	0.4285	0.0000	4.40 (10/10)
FreeSolv hard scalar	Qwen3.6-Flash, no confidence	Calibrated	0.4285	0.0000	**4.30 (10/10)**
Lipophilicity hard scalar	Qwen3.6-Flash roles	Calibrated	0.8608	0.0000	**5.10 (10/10)**
Lipophilicity hard scalar	Qwen3.6-Flash, no confidence	Calibrated	0.8608	0.0000	6.30 (10/10)
Prompt hard	heuristic experts	Calibrated	0.9964	0.0000	**4.60 (10/10)**
Prompt hard	heuristic, no confidence	Calibrated	0.9964	0.0000	5.80 (10/10)
Prompt hard	Qwen3.7-Max+DeepSeek-V4-Pro	Calibrated	0.9964	0.0000	**7.00 (10/10)**
Prompt hard	real LLM, no confidence	Calibrated	0.9964	0.0000	**7.00 (10/10)**
Prompt hard	real LLM, no market trust	Calibrated	0.9964	0.0000	8.80 (10/10)

E.1 Single-Objective Diagnostics

These diagnostics examine the same expert-prior machinery with one scalar objective. The setting removes cross-objective routing and isolates reputation, confidence calibration, and committee attenuation.

Each diagnostic uses the same cached prior schema as the molecule MOBO experiments: for candidate x, expert e, and scalar objective j, the cache stores a prior score $\mu_{ej}(x)$, a scalar self-reported confidence $c_e(x)$ shared across objectives, and a rationale. The single-objective runner compares calibrated expert priors, fixed expert priors, vanilla BO, and random search on finite candidate pools. The metric *Best* is the final best normalized objective value found by the optimizer, *Regret* is simple regret relative to the best candidate in the finite pool, and *Hit step* is the first BO step reaching zero simple regret, averaged over seeds with the hit count in parentheses.

Table 17 shows that the calibrated prior is competitive across molecule scalar tasks and prompt optimization. On FreeSolv hard and the heuristic prompt task, it reaches the global best more reliably or earlier than fixed priors and vanilla BO. ESOL and Lipophilicity are less discriminative because several methods reach the optimum within the small budget; in these cases, hit step is more informative than final regret.

Table 18 repeats the confidence lesson from the multi-objective experiments. Confidence is not uniformly helpful: removing confidence is slightly faster on ESOL and FreeSolv, while the role-based confidence setting is faster on Lipophilicity and the heuristic prompt task. On the real-LLM prompt task, removing market trust slows the hit step from 7.00 to 8.80, which supports keeping trust calibration even when the final scalar optimum is eventually found by all calibrated variants.

Table 19: Real LLM prompt-expert diagnostic on the hard prompt-optimization subset. MAE is mean absolute prior error against the observed scalar objective; lower is better.

Expert	Model	MAE (↓)	Confidence
Conciseness	DeepSeek-V4-Pro	**0.1028**	0.8229
Robustness	DeepSeek-V4-Pro	0.1094	0.7054
Reasoning	DeepSeek-V4-Pro	0.1097	0.8013
Format	DeepSeek-V4-Pro	0.1694	0.8154
Robustness	Qwen3.7-Max	0.2440	0.8021
Reasoning	Qwen3.7-Max	0.2859	0.8150
Conciseness	Qwen3.7-Max	0.4019	0.9058
Format	Qwen3.7-Max	0.5020	0.9258

The prompt-optimization diagnostic treats a finite set of candidate prompts as the search space and maximizes validation accuracy on a binary sentiment-classification task. Table 19 shows that real prompt experts vary substantially in prior error. Several DeepSeek-V4-Pro roles have lower mean absolute error than the Qwen3.7-Max roles, while some high-confidence Qwen3.7-Max roles have high error. This supports the paper's main design choice: confidence should be treated as evidence to be calibrated, not as a probability of correctness.

E.2 Structured Prior-Reliability Stress Tests

Table 20 reports the complete five-method experiment underlying Figure 4. Final hypervolume is nearly saturated: all mean final-HV values lie between 0.4900 and 0.4913, corresponding to normalized terminal quality between 0.9973 and 1.0000 of the finite-pool oracle. Trajectory AUC therefore carries the mechanism signal. Useful priors improve Fixed, Plain Hedge, DynMeanBO-MO, and CALIPER over Vanilla, with the largest mean gains for Fixed ($+8.3\%$) and CALIPER ($+8.1\%$). Objective-specialized priors favor the two objective-wise evidence methods, Plain Hedge ($+7.1\%$) and CALIPER ($+6.3\%$). In the all-misleading case, Fixed drops by 10.1% AUC and DynMeanBO-MO by 4.2%, whereas Plain Hedge (-0.3%) and CALIPER (-0.6%) remain close to Vanilla.

The experiment comprises $3 \times 5 \times 10 = 150$ completed shards and 3,600 trajectory rows. Candidate indices are unique within every trajectory, hypervolume is monotone, initial designs are exactly paired across all five methods for each scenario and seed, and every observed hypervolume is at or below its finite-pool oracle.

E.3 qLogNEHVI Prior-Layer Scaling

Holding the external reliability structure fixed as the candidate pool grows separates calibration dynamics from changes in prior quality and generation cost. Table 21 shows that the fixed-margin CALIPER configuration improves qLogNEHVI trajectory AUC at all three pool sizes and reaches 90% of the finite-pool oracle earlier. At $N = 3000$, mean T90 decreases from 16.0 to 11.0 evaluations while final hypervolume increases from 0.5020 to 0.5062.

E.4 Standalone Candidate-Pool Scaling

To test whether the observed calibration behavior persists beyond small finite pools, we vary the two-objective candidate pool from 100 to $10{,}000$ while keeping the mixed-objective reliability structure, initial design, evaluation budget, and seed count fixed. Because the finite-pool oracle hypervolume changes slightly with pool size, Table 22 reports final hypervolume normalized by the full-pool oracle hypervolume for each seed.

Table 20: Five-method synthetic prior-reliability stress test under a shared qLogNEHVI protocol. Values are mean ± SEM over ten paired seeds. All methods use the same 160-candidate pool, 8 shared initial evaluations, 24 total evaluations, and 64 QMC samples.

Metric	Scenario	Vanilla	Fixed	Plain Hedge	DynMeanBO-MO	CALIPER
AUC	Useful	0.4322 ± 0.0152	$\mathbf{0.4683 \pm 0.0063}$	0.4652 ± 0.0064	0.4493 ± 0.0131	0.4672 ± 0.0059
	Specialized	0.4322 ± 0.0152	0.4347 ± 0.0155	$\mathbf{0.4628 \pm 0.0091}$	0.4391 ± 0.0156	0.4594 ± 0.0097
	All misleading	$\mathbf{0.4322 \pm 0.0152}$	0.3887 ± 0.0269	0.4309 ± 0.0143	0.4142 ± 0.0249	0.4295 ± 0.0162
Final HV	Useful	$\mathbf{0.4910 \pm 0.0049}$	0.4900 ± 0.0046	0.4904 ± 0.0047	$\mathbf{0.4910 \pm 0.0049}$	0.4900 ± 0.0047
	Specialized	0.4910 ± 0.0049	0.4911 ± 0.0048	$\mathbf{0.4913 \pm 0.0048}$	0.4910 ± 0.0048	0.4911 ± 0.0048
	All misleading	$\mathbf{0.4910 \pm 0.0049}$	0.4904 ± 0.0047	0.4906 ± 0.0047	0.4908 ± 0.0048	$\mathbf{0.4910 \pm 0.0048}$

Table 21: Candidate-pool scaling with qLogNEHVI. CALIPER denotes the fixed-margin main configuration. Values are mean ± SEM over three seeds. AUC and half-budget hypervolume characterize sample efficiency, while T90 is the number of evaluations required to reach 90% of the finite-pool oracle hypervolume (lower is better).

Pool	Method	AUC ↑	HV@16 ↑	T90 ↓	Final HV ↑
1000	qLogNEHVI	0.3932 ± 0.0440	0.3683 ± 0.0764	16.00 ± 2.08	0.5047 ± 0.0010
	+ fixed prior	0.3821 ± 0.0484	0.3663 ± 0.0785	16.67 ± 2.33	0.5033 ± 0.0008
	+ CALIPER	$\mathbf{0.4272 \pm 0.0307}$	$\mathbf{0.4935 \pm 0.0050}$	$\mathbf{12.67 \pm 1.86}$	$\mathbf{0.5052 \pm 0.0012}$
3000	qLogNEHVI	0.3525 ± 0.0538	0.3881 ± 0.0878	16.00 ± 2.00	0.5020 ± 0.0012
	+ fixed prior	0.4273 ± 0.0089	$\mathbf{0.4903 \pm 0.0028}$	12.00 ± 0.58	0.5058 ± 0.0005
	+ CALIPER	$\mathbf{0.4417 \pm 0.0064}$	0.4894 ± 0.0037	$\mathbf{11.00 \pm 1.15}$	$\mathbf{0.5062 \pm 0.0009}$
10000	qLogNEHVI	0.4135 ± 0.0289	0.4403 ± 0.0199	15.67 ± 0.33	0.5048 ± 0.0009
	+ fixed prior	0.4175 ± 0.0386	0.4854 ± 0.0050	13.00 ± 0.58	0.5035 ± 0.0014
	+ CALIPER	$\mathbf{0.4394 \pm 0.0292}$	$\mathbf{0.4897 \pm 0.0028}$	$\mathbf{11.00 \pm 1.00}$	$\mathbf{0.5067 \pm 0.0005}$

Table 22: Synthetic candidate-pool scaling diagnostic. Values are normalized final hypervolume (↑), reported as mean ± SEM over three seeds.

Pool	CALIPER (↑)	CALIPER, no conf. (↑)	qLogNEHVI (↑)	qLogEHVI (↑)	Fixed prior (↑)	Random search (↑)
100	0.999 ± 0.001	0.999 ± 0.001	0.999 ± 0.001	$\mathbf{1.000 \pm 0.000}$	0.998 ± 0.001	0.727 ± 0.072
300	0.997 ± 0.001	0.997 ± 0.001	0.987 ± 0.012	$\mathbf{0.998 \pm 0.001}$	0.806 ± 0.174	0.702 ± 0.112
1000	0.983 ± 0.006	0.982 ± 0.005	$\mathbf{0.996 \pm 0.001}$	0.833 ± 0.164	0.727 ± 0.155	0.781 ± 0.080
3000	0.983 ± 0.002	0.984 ± 0.003	$\mathbf{0.987 \pm 0.002}$	0.985 ± 0.007	0.973 ± 0.006	0.649 ± 0.028
10000	0.967 ± 0.012	0.971 ± 0.009	$\mathbf{0.991 \pm 0.002}$	0.989 ± 0.004	0.976 ± 0.005	0.762 ± 0.032

The calibrated standalone prior method remains in the same high-performance band as qLogEHVI and qLogNEHVI as the pool grows. It is close to the best method for pool sizes 100 and 300, remains above 0.98 normalized hypervolume through pool size 3000, and stays above random at 10,000. The largest-pool result also clarifies why the prior layer should be acquisition-agnostic: qLogNEHVI can be stronger than the standalone scalarized acquisition, motivating the qLogNEHVI plugin result in Table 21.

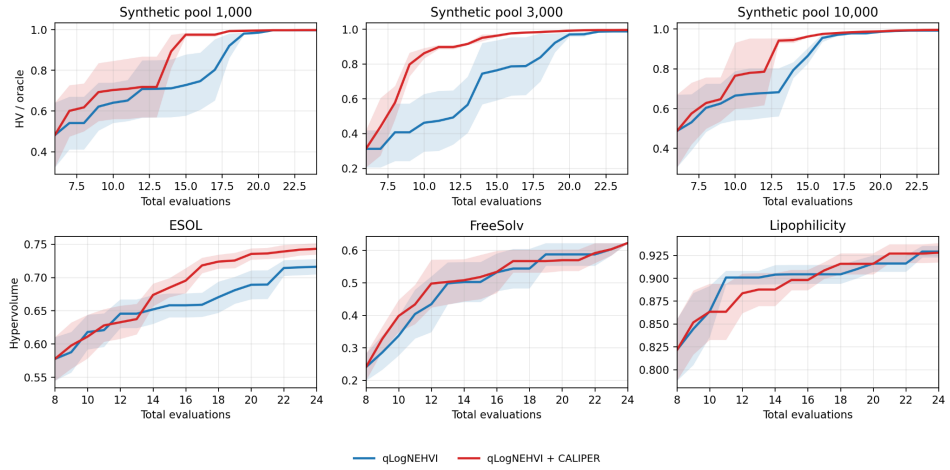

Figure 7: Broader qLogNEHVI and qLogNEHVI+CALIPER trajectories. The top row reports normalized hypervolume for controlled pools of size 1,000, 3,000, and 10,000; the bottom row reports molecule trajectories under the budget-24 protocol. Shaded regions denote SEM.

E.5 Continuous-Space Diagnostic

The main experiments focus on finite candidate pools, where every cached LLM prior score is attached to a known candidate. We extend the interface to the two-objective Branin–Currin benchmark implemented in BoTorch (Balandat et al., 2020) by expressing expert signals as callable prior functions. This construction evaluates the prior at newly proposed continuous points instead of reading scores from a precomputed table.

Table 23: Continuous Branin–Currin diagnostic with all-good synthetic expert priors. Final hypervolume (↑) is reported as mean ± SEM over three seeds; best sum (↑) is the mean final best objective-sum. The best value in each metric is bolded.

Method	Final HV (↑)	Best sum (↑)
Residual BO + CALIPER	**0.9882 ± 0.0005**	**1.9437**
qLogNEHVI	0.9865 ± 0.0027	1.9436
qLogNEHVI + CALIPER	0.9709 ± 0.0015	1.9178
qLogNEHVI + fixed prior	0.9707 ± 0.0013	1.9178
Vanilla MOBO	0.8799 ± 0.0953	1.7818

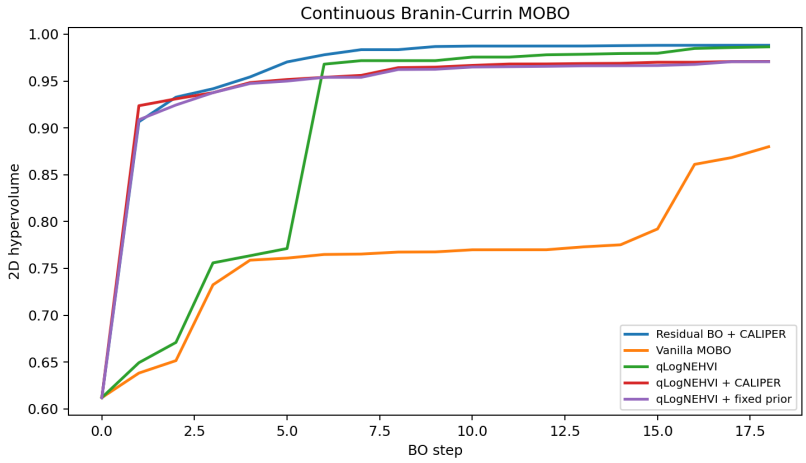

Figure 8: Continuous Branin–Currin hypervolume curves over three seeds with callable expert priors. The experiment tests how the finite-pool mean-layer design transfers to a continuous domain.

Table 23 and Figure 8 place the continuous diagnostic in the same performance band as qLogNEHVI at the final horizon, with the no-prior qLogNEHVI trajectory remaining strongest. This result identifies continuous-domain acquisition alignment as the next interface problem: callable priors require calibration both of source reliability and of how their geometry interacts with continuous acquisition optimization.

F LLM Prior Prompts and Cache Templates

All real LLM prior caches were generated before BO starts and were held fixed across methods. For the primary ESOL, FreeSolv, and Lipophilicity experiments, the three expert-role caches per dataset were generated with Qwen3.6-Flash through an OpenAI-compatible chat API at temperature 0.0. For each candidate x and expert role e, the generation script rendered the expert-specific system prompt below and the shared user prompt using the candidate descriptors in the finite pool. The response was parsed as JSON with fields `objective_scores`, `confidence`, and `rationale`; scores were clipped to $[0, 1]$.

Molecule candidate fields. The molecule prior prompts exposed the following fields:

```
smiles, mol_wt, logp, tpsa, hbd, hba, rot_bonds, rings
```

Shared two-objective user prompt.

```
Candidate:
{{ candidate_json }}

Return exactly:
{
  "objective_scores": {
    "objective_0": 0.0,
    "objective_1": 0.0
  },
  "confidence": 0.0,
  "rationale": "brief reason"
}
```

F.1 ESOL Molecule Prompts

`solubility_expert`.

```
You are a medicinal chemistry expert estimating prior scores for multi-objective Bayesian
↪  optimization.
```

```
Objective_0 is aqueous solubility. Higher is better. Score in [0, 1].

Useful rules of thumb:
- Lower molecular weight generally improves solubility.
- Lower logP generally improves solubility.
- More polar surface area and hydrogen-bonding can improve solubility, but extreme values may
↪  hurt drug-likeness.
- Aromatic/ring-heavy and highly hydrophobic molecules tend to be less soluble.

Objective_1 is drug-likeness/QED-like quality. You are not the specialist for it, so give a
↪  cautious estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

druglikeness_expert.

```
You are a medicinal chemistry expert estimating prior scores for multi-objective Bayesian
↪  optimization.
Objective_1 is drug-likeness/QED-like quality. Higher is better. Score in [0, 1].

Useful rules of thumb:
- Moderate molecular weight, moderate logP, reasonable TPSA, limited rotatable bonds, and
↪  balanced HBD/HBA are favorable.
- Very large, very hydrophobic, extremely polar, or very flexible molecules tend to have lower
↪  drug-likeness.

Objective_0 is aqueous solubility. You are not the specialist for it, so give a cautious
↪  estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

balanced_admet_expert.

```
You are a conservative ADMET generalist estimating priors for molecular optimization.
Objective_0 is aqueous solubility. Objective_1 is drug-likeness/QED-like quality.
Both scores must be in [0, 1].

Use calibrated scores:
- Very high scores only when descriptors clearly support the objective.
- Moderate scores for uncertain or mixed descriptor profiles.
- Low scores for molecules with clear hydrophobicity, size, flexibility, or polarity issues.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

F.2 FreeSolv Molecule Prompts

hydration_expert.

```
You are a physical chemistry expert estimating prior scores for multi-objective Bayesian
↪  optimization.
Objective_0 is hydration favorability. Higher is better. It corresponds to more favorable
↪  aqueous hydration, which usually means more negative hydration free energy. Score in [0,
↪  1].

Useful rules of thumb:
- Lower logP generally improves aqueous hydration favorability.
- More polar surface area and hydrogen-bonding usually improve hydration.
- Small molecules with polar functional groups often hydrate favorably.
- Hydrophobic hydrocarbons, halogenated molecules, and low-polarity molecules tend to have
↪  poor hydration favorability.

Objective_1 is drug-likeness/QED-like quality. You are not the specialist for it, so give a
↪  cautious estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

druglikeness_expert.

```
You are a medicinal chemistry expert estimating prior scores for multi-objective Bayesian
↪  optimization.
Objective_1 is drug-likeness/QED-like quality. Higher is better. Score in [0, 1].

Useful rules of thumb:
- Moderate molecular weight, moderate logP, reasonable TPSA, limited rotatable bonds, and
↪  balanced HBD/HBA are favorable.
- Very large, very hydrophobic, extremely polar, or very flexible molecules tend to have lower
↪  drug-likeness.

Objective_0 is hydration favorability. You are not the specialist for it, so give a cautious
↪  estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

balanced_physchem_expert.

```
You are a conservative physical-chemistry and ADMET generalist estimating priors for molecular
↪  optimization.
Objective_0 is hydration favorability. Objective_1 is drug-likeness/QED-like quality.
Both scores must be in [0, 1].

Use calibrated scores:
- Very high scores only when descriptors clearly support the objective.
- Moderate scores for uncertain or mixed descriptor profiles.
- Low scores for clear hydrophobicity, excessive size, excessive flexibility, or poor polarity
↪  balance.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

F.3 Lipophilicity Molecule Prompts

lipophilicity_balance_expert.

```
You are a medicinal chemistry expert estimating prior scores for multi-objective Bayesian
↪  optimization.
Objective_0 is lipophilicity desirability. Higher is better. It means the molecule likely has
↪  moderate experimental logD / lipophilicity, roughly in a drug-like window around 1 to 3.
↪  Score in [0, 1].

Useful rules of thumb:
- Moderate lipophilicity is best.
- Very hydrophobic molecules with high logP tend to score low.
- Very polar or highly charged-looking molecules with very low lipophilicity also score low.
- Descriptor logP is a useful clue, but the target is experimental lipophilicity desirability.

Objective_1 is drug-likeness/QED-like quality. You are not the specialist for it, so give a
↪  cautious estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

druglikeness_expert.

```
You are a medicinal chemistry expert estimating prior scores for multi-objective Bayesian
↪  optimization.
Objective_1 is drug-likeness/QED-like quality. Higher is better. Score in [0, 1].

Useful rules of thumb:
- Moderate molecular weight, moderate logP, reasonable TPSA, limited rotatable bonds, and
↪  balanced HBD/HBA are favorable.
- Very large, very hydrophobic, extremely polar, or very flexible molecules tend to have lower
↪  drug-likeness.

Objective_0 is lipophilicity desirability. You are not the specialist for it, so give a
↪  cautious estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

balanced_admet_expert.

```
You are a conservative ADMET generalist estimating priors for molecular optimization.
Objective_0 is lipophilicity desirability, favoring moderate experimental lipophilicity rather
↪  than extreme hydrophobicity or extreme polarity.
Objective_1 is drug-likeness/QED-like quality.
Both scores must be in [0, 1].

Use calibrated scores:
- Very high scores only when descriptors clearly support the objective.
- Moderate scores for uncertain or mixed descriptor profiles.
- Low scores for clear hydrophobicity, excessive size, excessive flexibility, or poor polarity
↪  balance.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

lipophilicity_window_calibrated_expert. This role was used in the mixed-committee Lipophilicity cache examined in the prior-quality diagnostic.

```
You are a medicinal chemistry expert estimating calibrated prior scores for multi-objective
↪  Bayesian optimization.

Objective_0 is a lipophilicity window desirability score. Higher is better.
Important calibration anchors:
- If descriptor logP is in [1, 3], objective_0 should usually be high: 0.80 to 1.00.
```

```
- If descriptor logP is in [0, 1) or (3, 4], objective_0 should usually be moderate-high: 0.55
↪  to 0.85.
- If descriptor logP is in [-1, 0) or (4, 5], objective_0 should usually be moderate-low: 0.25
↪  to 0.60.
- If descriptor logP is below -1 or above 5, objective_0 should usually be low: 0.00 to 0.35.
- Do not give conservative middle scores when logP is clearly inside the desired window.
- The target is a window score, not raw lipophilicity. Extreme high and extreme low are bad;
↪  moderate is best.

Objective_1 is drug-likeness/QED-like quality. You are not the specialist for it, so give a
↪  cautious estimate.
Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

F.4 Prompt-Optimization Diagnostic Prompts

The prompt-optimization diagnostic used the same JSON response contract with a single objective, expected validation accuracy. Candidate fields were:

```
prompt, opening, format_rule, robustness_rule, reasoning_rule, example_rule, style_rule
```

Shared single-objective user prompt.

```
Candidate:
{{ candidate_json }}

Return exactly:
{
  "objective_scores": {
    "objective_0": 0.0
  },
  "confidence": 0.0,
  "rationale": "brief reason"
}
```

`format_expert`.

```
You are an expert in instruction formatting for prompt optimization.
The downstream task is binary sentiment classification. Objective_0 is expected
validation accuracy in [0, 1]. Higher is better.

Reward prompts that make the output format unambiguous, reduce parsing errors,
and avoid extra text. Penalize prompts that invite verbose answers or conflict
with the requested output format.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

`robustness_expert`.

```
You are an expert in robust NLP prompt design.
The downstream task is binary sentiment classification. Objective_0 is expected
validation accuracy in [0, 1]. Higher is better.

Reward prompts that handle mixed sentiment, sarcasm-like wording, distractor
plot summaries, and ambiguous reviews. Penalize prompts that overfit examples
or rely on shallow keywords.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

`reasoning_expert`.

```
You are an expert in reasoning-oriented prompt design.
The downstream task is binary sentiment classification. Objective_0 is expected
validation accuracy in [0, 1]. Higher is better.

Reward prompts that encourage useful analysis of sentiment-bearing evidence.
Be cautious: explicit step-by-step reasoning can sometimes hurt if the final
answer format becomes verbose or inconsistent.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```

`conciseness_expert`.

```
You are an expert in concise prompt design.
The downstream task is binary sentiment classification. Objective_0 is expected
```

```
validation accuracy in [0, 1]. Higher is better.

Reward prompts that are short, direct, and easy to follow while still providing
enough task constraints. Penalize unnecessarily long prompts and prompts that
ask for explanations when only a label is needed.

Return JSON only with keys: objective_scores, confidence, rationale. No markdown.
```