FROM DIAGNOSTIC INTENT TO PHYSICAL EXPERIMENTS: GROUNDED EXPERIMENTAL DESIGN WITH LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model agents can interpret diagnostic context and formulate useful questions, while physical systems answer experiments with continuous measurements governed by dynamics, operating conditions, and sensor uncertainty. This creates an interface problem when changing context determines which registered experiments should be considered before numerical design. We introduce PhysGround-BED, a grounded interface in which an LLM constructs a posterior-dependent semantic experiment space by proposing typed intents from the current belief, available capabilities, and operating context. A deterministic compiler maps these intents to executable acquisitions, and a grounded response model evaluates continuous-response expected information gain, selects an acquisition, and updates the posterior. A history-conditional analysis decomposes one-step utility loss into response-model mismatch, compilation omission, and numerical estimation error. We evaluate the interface on distribution-network topology identification, supercomputer telemetry, source localization, damped-oscillator identification, hyperspectral sensing, and battery impedance diagnosis. The results show that posterior-dependent semantic restriction can retain informative acquisitions with fewer numerical candidate evaluations in large catalogs and can translate textual sensor-quality information into grounded acquisition restrictions.

1 INTRODUCTION

Real diagnostic work rarely begins with a fully specified experiment-design problem. An operator may know the observed symptoms, recent maintenance events, unavailable devices, and prohibited operations, yet still lack the evidence needed to identify the system state. Diagnosis must therefore determine both what the available evidence supports and what evidence should be acquired next. As observations accumulate, the useful experiment changes with the remaining uncertainty and the current operating context.

Bayesian experimental design (BED) selects experiments by maximizing expected uncertainty reduction once hypotheses, acquisitions, observations, and constraints have been represented computationally (Lindley, 1956; Chaloner & Verdinelli, 1995; Rainforth et al., 2024). Rule-based and amortized selectors share this structured starting point. The unresolved step precedes the numerical problem: constructing the current acquisition space from operational context. This step becomes important in newly commissioned or rapidly reconfigured systems, where limited state-matched history may prevent training a dedicated selector and maintenance records describe available resources, temporary faults, and diagnostic goals in changing combinations. Distribution-topology methods illustrate the distinction. They infer network structure from registered voltage and power measurements, including noisy and time-varying states (Weng et al., 2017; Yu et al., 2018; 2019). Before such numerical inference or BED can act, the current context must be translated into the registered measurements and interventions relevant to the next design decision.

LLM agents offer a general mechanism for connecting flexible descriptions with structured domain concepts. Scientific agents already extract experimental requirements, organize specialist tools, and invoke validated resources (Qu et al., 2026; Jin et al., 2025; Tripathi et al., 2026). Yet choosing an informative next question requires more than parsing a request. Active Task Disambiguation and BED-LLM show that an agent can reason over competing hypotheses, generate candidate questions, and select among them using information gain (Kobalczyk et al., 2025; Choudhury et al., 2026). Their central insight is that an incomplete task should trigger targeted information acquisition instead

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.

of an immediate final answer. We adopt this insight for physical diagnosis, where the agent proposes an experiment that distinguishes the remaining physical hypotheses.

Physical systems introduce a response channel that changes this formulation. Feeders, batteries, oscillators, hyperspectral sensors, and data centers answer experiments through continuous measurements governed by dynamics, nuisance variation, and sensor uncertainty. The information value of an experiment therefore depends on a calibrated stochastic response model. A useful semantic proposal must also be realized through the actuators, channels, locations, times, wavelengths, or frequencies available in the system. Physical diagnosis consequently requires both a flexible proposal mechanism and a grounded numerical model that connects each executable acquisition to its observations.

We introduce PhysGround-BED, in which an LLM constructs a posterior-dependent semantic experiment space and a physical response model performs numerical design within it. The LLM proposes typed experiment intents from the current posterior, available capabilities, operating context, and summarized history. A deterministic compiler maps these intents to registered acquisitions. A grounded probabilistic backend evaluates the expected information gain of each compiled acquisition under the continuous response model, selects one, and updates the posterior. New evidence thereby changes both the diagnostic belief and the semantic space considered at the next turn.

Our contributions are threefold. First, we formulate a typed experimental interface connecting diagnostic context, semantic experiment families, registered acquisitions, and continuous stochastic responses. Second, we combine belief-conditioned semantic proposal with continuous-observation BED and give a history-conditional decomposition of one-step utility loss into response mismatch, compilation omission, and acquisition-estimation error. Third, we evaluate the resulting composition across six engineered and scientific systems. Distribution topology and real supercomputer telemetry provide the principal applications. Across several large acquisition catalogs, the compiled experiment spaces retain endpoint utility with substantially fewer numerical EIG evaluations. Complementary benchmarks show how textual sensor-quality information redirects acquisition and how qualitative semantic descriptions limit fine-grained experimental resolution.

2 Related Work

Physical and amortized experimental design. Sequential BED selects experiments to reduce latent uncertainty (Lindley, 1956; Chaloner & Verdinelli, 1995; Rainforth et al., 2024). Distribution-topology methods recover structure from statistical dependence, joint parameter estimation, or voltage responses to active and reactive perturbations (Weng et al., 2017; Yu et al., 2018; 2019; Cavraro & Kekatos, 2019; 2018), with related mixed-integer formulations (Taheri et al., 2019). DAD and DiffBED amortize design through offline policy learning (Foster et al., 2021; Saravanan et al., 2026). These approaches operate on numerically represented hypotheses, actions, observations, and objectives. We study the interface through which changing diagnostic context determines the experiments entering that representation.

Grounded LLM agents and formal interfaces. Tool-augmented agents separate semantic reasoning from domain execution. SayCan combines language relevance with learned affordances (Ahn et al., 2022), and LLM+P converts natural-language tasks into classical planning problems (Liu et al., 2023). Scientific and power-system agents similarly organize specialist tools and solvers (Bran et al., 2024; Jin et al., 2025; Tripathi et al., 2026). PhysGround-BED applies this composition to a typed, belief-dependent experimental contrast whose measurement changes the posterior and next proposal.

LLM-driven disambiguation and information gathering. Active Task Disambiguation samples compatible solutions and asks questions that balance their answer partition (Kobalczyk et al., 2025). BED-LLM defines a language joint model, belief update, candidate generator, and EIG estimator (Choudhury et al., 2026). We extend this belief-conditioned proposal pattern to continuous sensor responses represented by a simulator or empirical likelihood. The language distribution covers semantic experiment descriptions, and the physical model supplies the response distribution.

Physics-grounded learning and scientific agents. Physics-informed learning incorporates equations through objectives, architectures, or differentiable solvers (Raissi et al., 2019; Karniadakis et al., 2021). Scientific-agent benchmarks test language models that compose specialized tools (Chen

et al., 2025). PhysGround-BED exposes mechanistic and empirical models through a hypothesis-conditioned response channel shared by continuous EIG and Bayesian updating, covering both interventions and observational queries.

3 Problem Formulation

We formulate physical diagnosis as sequential inference through a hypothesis-conditioned stochastic experiment channel. In domain d, the unknown state is $\theta^\star \in \Theta_d = \{\theta_1, \dots, \theta_K\}$ with prior π_0. The registered catalog $\mathcal{X}_d$ contains acquisitions that can be executed or retrieved. After history $\mathcal{H}_{t-1} = \{(x_\tau, y_\tau)\}_{\tau<t}$, the remaining catalog is $\mathcal{X}_{d,t} = \mathcal{X}_d \setminus \{x_1, \dots, x_{t-1}\}$. At turn t, the agent selects $x_t \in \mathcal{X}_{d,t}$ and receives a potentially multivariate response $y_t \in \mathcal{Y}_d \subseteq \mathbb{R}^{d_y}$. An acquisition may intervene on a controllable input, as in feeder probing, or reveal an unread measurement, as in telemetry, hyperspectral sensing, and impedance diagnosis. Both cases are represented through the dependence of the observation on the latent hypothesis, acquisition, and history. We study finite-hypothesis closed-set identification over a horizon $T \le |\mathcal{X}_d|$, with terminal prediction $\widehat{\theta}_T = \arg\max_{\theta\in\Theta_d} \pi_T(\theta)$.

The environment follows an unknown response law $p_{\mathrm{env},d}(y \mid \theta, x, \mathcal{H}_{t-1})$, while experiment design and inference use a working response model $p_d(y \mid \theta, x, \mathcal{H}_{t-1})$. In a history-independent mechanistic model, a state map $z = F_d(\theta, x, \xi)$ with nuisance variable $\xi \sim p_{\Xi_d}$, measurement operator $\mathcal{M}_d$, and sensor noise ϵ induces

$$p_d(y \mid \theta, x) = \int p_\epsilon(y - \mathcal{M}_d(F_d(\theta, x, \xi)))\, p_{\Xi_d}(\xi)\, \mathrm{d}\xi. \tag{1}$$

When a simulator is unavailable, an empirical distribution calibrated from development measurements provides the same hypothesis-conditioned response channel. After acquiring (x_t, y_t), the working likelihood updates the posterior as

$$\pi_t(\theta) = \frac{\pi_{t-1}(\theta) p_d(y_t \mid \theta, x_t, \mathcal{H}_{t-1})}{\sum_{\theta'\in\Theta_d} \pi_{t-1}(\theta') p_d(y_t \mid \theta', x_t, \mathcal{H}_{t-1})}. \tag{2}$$

The remaining challenge is representational. Diagnostic context may specify a downstream region to test, posterior modes to distinguish, or a reported sensor fault to avoid. Numerical design instead requires a concrete actuator, channel, location, time, wavelength, or frequency. We therefore seek an interface that uses the current belief and operating context to construct executable acquisition candidates while preserving the continuous response model used for design and inference.

4 Method: PhysGround-BED

PhysGround-BED constructs a posterior-dependent semantic experiment space and performs continuous-response Bayesian experimental design within it. At each turn, an LLM proposes typed diagnostic intents from the posterior, semantic resource capabilities, operating context, and summarized history. A deterministic compiler maps valid intents to registered acquisitions. The backend evaluates their physical expected information gain, selects an acquisition, and uses its response to update the posterior. Setting the compiled set equal to the remaining catalog recovers Full Physics BED.

4.1 Typed Belief-Conditioned Semantic Compilation

Definition 1 (Typed physical experimental interface)**.** *A domain interface is* $\mathfrak{I}_d = (\Theta_d, \mathcal{E}_d, \mathcal{X}_d, \ell_d, \Gamma_d, p_d)$. *Here* $\mathcal{E}_d$ *is a typed semantic intent space,* $\mathcal{X}_d$ *is the registered numerical catalog, and* ℓ_d *serializes the diagnostic state and semantic capabilities. Registered action identifiers and numerical controls remain within the compiler. The deterministic mapping* $\Gamma_d : \mathcal{E}_d \times \Delta(\Theta_d) \times 2^{\mathcal{X}_d} \to 2^{\mathcal{X}_d}$ *maps an intent, belief, and remaining catalog to executable candidates. It is sound when* $\Gamma_d(e, \pi, \mathcal{X}') \subseteq \mathcal{X}'$ *for every valid input.*

Compilation quality is measured by the information value omitted from the compiled set, as formalized in Section 4.3.

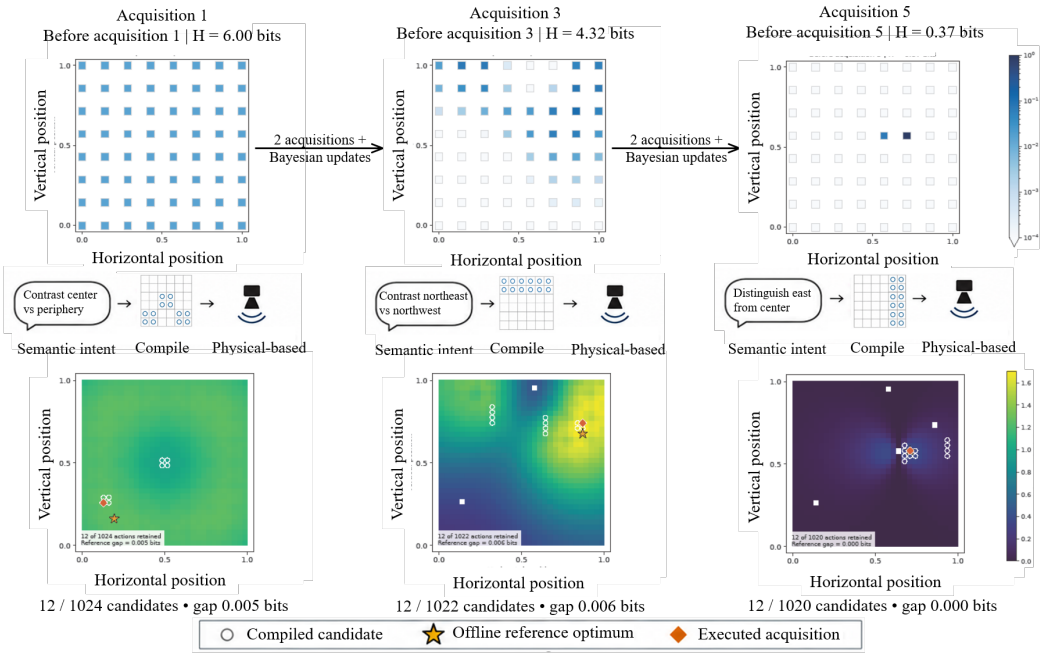


Figure 1: Belief-conditioned semantic compilation in a source-localization episode. Posterior states before acquisitions 1, 3, and 5 condition the semantic intents. The compiler maps each intent to registered candidates, and the physical backend evaluates their continuous-response EIG. Each measurement updates the posterior and reshapes the next contrast. Lower panels show full-catalog EIG in an audit. Circles denote compiled candidates, stars the reference optima, and diamonds the executed acquisitions. Small gaps show that compact candidate sets preserve information value.

Figure 1 shows how posterior concentration changes the hypotheses requiring separation, the proposed semantic family, and the most informative compiled realization. Let q_ϕ be the frozen proposer and $\mathcal{R}_{d,t}$ the semantic capabilities of the remaining resources. The semantic history is $\mathcal{H}^{\text{sem}}_{t-1} = \{(e_\tau, s_d(x_\tau, y_\tau), \pi_\tau)\}_{\tau<t}$, where s_d deterministically summarizes the experiment family, response evidence, and posterior change. The posterior identifies the hypotheses that still require discrimination. The serializer ℓ_d forms the proposal context, after which

$$\widetilde{\mathcal{E}}_t \sim q_\phi\big(\cdot \mid \ell_d(\Theta_d, \mathcal{R}_{d,t}, \pi_{t-1}, \mathcal{H}^{\text{sem}}_{t-1})\big)\,, \tag{3}$$

$$\mathcal{V}_t = \mathsf{Validate}_d(\widetilde{\mathcal{E}}_t, \mathcal{X}_{d,t}), \tag{4}$$

$$\mathcal{C}_t = \bigcup_{e\in\mathcal{V}_t} \Gamma_d(e, \pi_{t-1}, \mathcal{X}_{d,t}). \tag{5}$$

An intent describes a physical contrast through domain-relevant typed fields. In the feeder, resource, actuation mode, direction, and qualitative intensity identify intervention families. In Sol, measurement family, PDU scope, and bank region identify unread channels. Spatial relations specify source sensors, timing and excitation attributes specify oscillator experiments, spectral regions specify AVIRIS bands, and frequency regimes specify battery impedance queries. The LLM composes fields from the current belief and history, and the compiler resolves them against registered resources.

Validation checks schema conformance, resource availability, and compilability before scoring. Invalid proposals receive bounded field-level repair feedback. A predeclared fallback or termination rule handles an empty compilation after the repair budget. Appendix B gives the prompt contract, schemas, compilation rules, and failure handling.

4.2 Continuous Physical Acquisition and Belief Feedback

The response model may arise from a mechanistic simulator, an analytical forward model, or a fitted empirical distribution. Every implementation exposes the likelihood $p_d(y \mid \theta, x, \mathcal{H}_{t-1})$, which supports vector-valued responses, nuisance variation, and history-conditioned observations.

For example, in Sol let o denote channels already observed from the record, a a candidate unread channel, and z_o the standardized observed vector. A joint class-conditional Gaussian gives

$$\mu_{\theta,a|o} = \mu_{\theta,a} + \Sigma_{\theta,ao}\Sigma_{\theta,oo}^{-1}(z_o - \mu_{\theta,o}), \tag{6}$$

$$\sigma^2_{\theta,a|o} = \max\Big(\Sigma_{\theta,aa} - \Sigma_{\theta,ao}\Sigma_{\theta,oo}^{-1}\Sigma_{\theta,oa}, 10^{-4}\Big). \tag{7}$$

Conditioning on z_o accounts for correlation among telemetry channels. Appendix C details the response contract in each domain.

Define the posterior predictive response $p_{t-1}(y \mid x, \mathcal{H}_{t-1}) = \sum_{\theta\in\Theta_d} \pi_{t-1}(\theta) p_d(y \mid \theta, x, \mathcal{H}_{t-1})$. The physical acquisition utility is the continuous-response EIG

$$U_t^{\mathrm{phys}}(x) = I_{p_d}(\Theta; Y \mid x, \mathcal{H}_{t-1}) \tag{8}$$

$$= \mathbb{E}_{\theta\sim\pi_{t-1},\, y\sim p_d}\left[\log_2 \frac{p_d(y \mid \theta, x, \mathcal{H}_{t-1})}{p_{t-1}(y \mid x, \mathcal{H}_{t-1})}\right]. \tag{9}$$

Monte Carlo integration or quadrature evaluates this expectation according to the response model. For a nonempty compiled set, the backend selects $\widehat{x}_t \in \arg\max_{x\in\mathcal{C}_t} \widehat{U}_t^{\mathrm{phys}}(x)$. The observed y_t updates the posterior using Eq. 2, and the resulting belief enters the next proposal. This posterior interface applies to nonlinear simulated interventions and correlated observational measurements. The full-catalog maps in Figure 1 are offline trajectory audits.

4.3 Information Preservation under Semantic Restriction

Assume that $\mathcal{X}_{d,t}$ is finite and that $\mathcal{C}_t$ is a nonempty subset of $\mathcal{X}_{d,t}$. Let $U_t^{\mathrm{env}}(x) = I_{p_{\mathrm{env},d}}(\Theta; Y \mid x, \mathcal{H}_{t-1})$. For a nonempty compiled set, define

$$\varepsilon_t = \sup_{x\in\mathcal{X}_{d,t}} |U_t^{\mathrm{env}}(x) - U_t^{\mathrm{phys}}(x)|, \tag{10}$$

$$\Delta_t^{\mathrm{int}} = \max_{x\in\mathcal{X}_{d,t}} U_t^{\mathrm{phys}}(x) - \max_{x\in\mathcal{C}_t} U_t^{\mathrm{phys}}(x), \tag{11}$$

$$\eta_t = \sup_{x\in\mathcal{C}_t} |\widehat{U}_t^{\mathrm{phys}}(x) - U_t^{\mathrm{phys}}(x)|. \tag{12}$$

Here ε_t measures response-model error, Δ_t^{int} measures information omitted by compilation, and η_t measures numerical acquisition error on retained candidates. A compiled set can preserve utility even when its selected action differs from the full-catalog optimum.

Theorem 1 (One-step interface decomposition)**.** *Assume that $\mathcal{X}_{d,t}$ is finite and $\varnothing \neq \mathcal{C}_t \subseteq \mathcal{X}_{d,t}$. Let $x_t^{\mathrm{env}} \in \arg\max_{x\in\mathcal{X}_{d,t}} U_t^{\mathrm{env}}(x)$, and let $\widehat{x}_t$ maximize $\widehat{U}_t^{\mathrm{phys}}$ over $\mathcal{C}_t$. Then $U_t^{\mathrm{env}}(x_t^{\mathrm{env}}) - U_t^{\mathrm{env}}(\widehat{x}_t) \leq 2\varepsilon_t + \Delta_t^{\mathrm{int}} + 2\eta_t$.*

Proof sketch. Insert the full-catalog physical maximizer and the compiled-set physical maximizer between the two environment utilities. The response-model substitutions contribute $2\varepsilon_t$, semantic restriction contributes Δ_t^{int}, and empirical maximization over the retained candidates contributes $2\eta_t$. The complete proof is given in Appendix A.

Theorem 1 gives a history-conditional perturbation decomposition for an arbitrary compiled set. It complements multiplicative guarantees that rely on submodularity or adaptive submodularity (Golovin & Krause, 2011). Response calibration controls ε_t, candidate retention controls Δ_t^{int}, and numerical accuracy controls η_t. Let $R_T^{\mathrm{loc}} = \sum_{t=1}^T [U_t^{\mathrm{env}}(x_t^{\mathrm{env}}) - U_t^{\mathrm{env}}(\widehat{x}_t)]$. Summing the one-step result along the realized belief trajectory gives $R_T^{\mathrm{loc}} \leq \sum_{t=1}^T \left(2\varepsilon_t + \Delta_t^{\mathrm{int}} + 2\eta_t\right)$. This quantity evaluates each decision at the history reached by the policy. Comparing complete adaptive policies requires a trajectory-level analysis.

The response term can be related to likelihood quality. Define the prior-weighted conditional divergence in nats as $\bar{D}_t(x) = \sum_{\theta\in\Theta_d} \pi_{t-1}(\theta) D_{\mathrm{KL}}(p_{\mathrm{env},d}(\cdot \mid \theta, x, \mathcal{H}_{t-1}) \,\|\, p_d(\cdot \mid \theta, x, \mathcal{H}_{t-1}))$. Suppose $\bar{D}_t(x) \leq \kappa_t(x)$ and let $\delta_t(x) = \sqrt{\kappa_t(x)/2} \leq 1/2$. For finite $|\Theta_d| = K$, the KL chain rule, Pinsker's inequality, and continuity of conditional entropy (Winter, 2016) give

$$|U_t^{\mathrm{env}}(x) - U_t^{\mathrm{phys}}(x)| \leq 2\delta_t(x)\log_2 K + 2h_2(\delta_t(x)), \tag{13}$$

where h_2 is binary entropy in bits. The post-selection full-catalog audits in Figure 1 measure the compilation term Δ_t^{int}.

For any deterministic response summary $B = h(Y)$, the Markov relation $\Theta \to Y \to B$ gives

$$I(\Theta; Y \mid x, \mathcal{H}_{t-1}) = I(\Theta; B \mid x, \mathcal{H}_{t-1}) + I(\Theta; Y \mid B, x, \mathcal{H}_{t-1}) \geq I(\Theta; B \mid x, \mathcal{H}_{t-1}). \quad (14)$$

Equality requires conditional sufficiency of the categorical response. PhysGround-BED retains the continuous measurement for information calculation and Bayesian inference.

5 Benchmarks and Experimental Protocol

The experiments test continuous-response grounding, retention of informative acquisitions under semantic restriction, and transfer across acquisition geometries. Power distribution and data-center telemetry provide the principal application evidence. Four compact systems isolate spatial compilation, continuous-parameter resolution, textual operating context, and response-model transfer.

Power and data center. The power benchmark follows inverter-based topology probing (Cavraro & Kekatos, 2019; 2018). OpenDSS generates multibus voltage responses to registered active- and reactive-power perturbations on the IEEE 13 and IEEE 123 feeders (Dugan & McDermott, 2011; Kersting, 2001). IEEE 13 supports the proposer comparison, and IEEE 123 tests utility retention in a larger catalog. The data-center benchmark diagnoses rack role from PDU telemetry collected on Sol, a research supercomputer operated by Arizona State University (Jennewein et al., 2023). Each acquisition reveals one unread total-power or bank-load channel. Appendix D gives the construction, splits, and budgets.

Complementary systems. Source Location adapts the benchmark used by DAD and DiffBED (Foster et al., 2021; Saravanan et al., 2026) to test spatial shortlists. Damped Oscillator tests the resolution of qualitative experiment descriptions. AVIRIS Indian Pines tests textual sensor-quality context (Baumgardner et al., 2015), and NASA Battery EIS tests cross-cell response-model transfer (Saha & Goebel, 2007). Appendix D gives their construction and splits.

5.1 Comparisons and Measurements

Full Physics BED evaluates EIG over the complete catalog. PhysGround-BED uses the same likelihood and update but scores candidates compiled from belief-conditioned intents. Prompt-Only, Active Task (Kobalczyk et al., 2025), and BED-LLM (Choudhury et al., 2026) serve as language-native response-representation controls. Categorical-response BED and Diagonal-response BED ablate the fidelity of the grounded observation model while retaining numerical acquisition scoring and Bayesian updating. Fixed acquisition, mRMR (Peng et al., 2005), and domain-specific numerical methods provide conventional controls.

We report top-1 accuracy for power, data center, and inverse problems, balanced accuracy for AVIRIS and battery diagnosis, continuous estimation error where appropriate, and the number of numerical candidate scores. Figures show trajectories and mechanism-level behavior. Tables provide endpoint estimates with bootstrap uncertainty. Appendix E defines the metrics and resampling units, and Appendix F reports additional endpoints and audit protocols.

6 Experiments

Our experiments address three questions: how explicit numerical response models affect acquisition quality relative to language-native reasoning, whether belief-conditioned semantic proposals preserve diagnostic utility while reducing numerical scoring, and whether textual operating context can redirect acquisition. We further examine how these effects vary with response-model transfer and semantic resolution.

6.1 Grounded Response Models Determine Information Gain

Table 1 isolates the response distribution used for acquisition and inference. On IEEE 123, the language-native controls remain near the 7.7% uniform reference, whereas Full Physics BED and

Table 1: Endpoint accuracy under response grounding on IEEE 123 (39 episodes) and Sol rack-OOD (48 episodes). Brackets show 95% bootstrap intervals.

Method	Power (%)	Data center (%)
Prompt-Only	5.1 [0.0, 12.8]	41.7 [35.4, 47.9]
Active Task	5.1 [0.0, 12.8]	27.1 [25.0, 31.3]
BED-LLM	7.7 [0.0, 17.9]	37.5 [31.3, 43.8]
Fixed acquisition	92.3 [84.6, 100.0]	75.0 [75.0, 75.0]
Supervised MLP	100.0 [100.0, 100.0]	72.9 [68.8, 75.0]
Categorical-response BED	69.2 [53.8, 82.1]	70.8 [64.6, 75.0]
Diagonal-response BED	100.0 [100.0, 100.0]	72.9 [68.8, 75.0]
Full Physics BED	100.0 [100.0, 100.0]	75.0 [75.0, 75.0]
PhysGround-BED	100.0 [100.0, 100.0]	75.0 [75.0, 75.0]

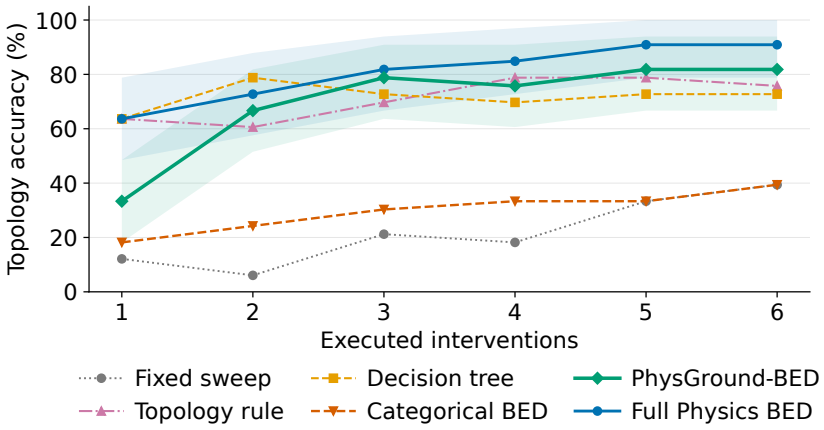


Figure 2: IEEE 13 topology-identification accuracy over six interventions. Categorical-response BED binarizes each voltage response. Shading shows 95% episode-bootstrap intervals for PhysGround-BED and Full Physics BED.

PhysGround-BED reach 100%. Replacing the multivariate response with a binary categorical outcome reduces accuracy to 69.2%, while the continuous-response models reach 100%. Figure 2 shows the corresponding contrast on IEEE 13, where categorical-response BED reaches 39.4%, and Full Physics BED reaches 90.9% after six interventions. Together, these results connect the data-processing relation in Equation 14 to diagnostic performance: categorical compression can remove hypothesis-relevant response structure used by EIG and Bayesian updating.

Sol provides an observation-only counterpart to active feeder probing. At six rack-OOD queries, language-native controls reach 27.1–41.7%, whereas methods with an explicit numerical response model reach 70.8–75.0%. The point estimates increase with response-model expressiveness, although the bootstrap intervals overlap. The remaining error is concentrated: methods reaching 75.0% classify all 12 `condo_compute` episodes as GPU while correctly identifying the other rack roles. Unlike feeder probing, Sol selects correlated measurements from a fixed record, so its errors reflect cross-rack response overlap rather than insufficient intervention design.

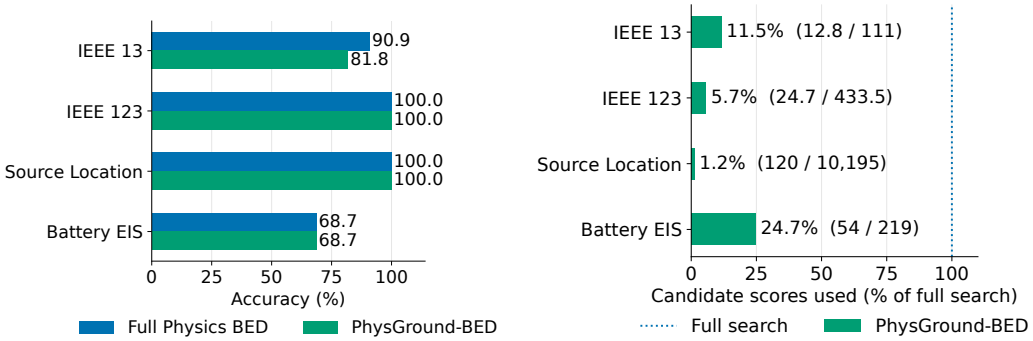
6.2 Semantic Proposals Track Belief and Preserve Useful Experiments

Figure 2 shows the acquisition trajectory. Registered-endpoint accuracy and online numerical scoring are reported in Appendix Table 7. On IEEE 13, PhysGround-BED rises from 33.3% after one intervention to 81.8% at the endpoint. Fixed sweep, topology rule, decision tree, and Full Physics BED end at 39.4%, 75.8%, 72.7%, and 90.9%. Relative to fixed sweep, the paired gain is 42.4 percentage points with a 95% interval of $[24.2, 60.6]$. The differences from the specialized rule and decision tree are smaller than their uncertainty intervals, placing the semantic proposer in the same endpoint range while preserving belief-conditioned adaptation.

The larger IEEE 123 catalog provides a utility audit. PhysGround-BED identifies all targets while scoring 24.7 candidates per episode. A full-search audit over the same histories scores 433.5

Table 2: Fixed-history Source Location proposer audit over 281 informative histories from 50 episodes. Every proposal compiles to 12 candidates. Random typed intents are averaged over 20 schema-matched draws per history. Lower omission and higher retention are better.

Proposer	Δ^{int} (bits)	Retain $\geq 95\%$
Belief-conditioned LLM	0.0504	77.3%
LLM without belief/history	0.1254	37.5%
Deterministic belief heuristic	0.0427	67.2%
Random typed intents	0.1343	47.7%
Static typed intents	0.2662	6.7%



(a) Endpoint diagnostic accuracy.

(b) Selection candidates as a percentage of full-catalog search.

Figure 3: Endpoint performance and numerical acquisition scoring. Panel a compares Full Physics BED and PhysGround-BED. Panel b reports PhysGround-BED scores as a percentage of full search, with mean counts in parentheses. IEEE 123 uses a same-history full-search audit.

candidates, a 94.3% reduction. Exact argmax coverage is 4.5%, yet the mean EIG gap is 0.038 bits. The compiled set therefore retains near-optimal information value even when it omits the exhaustive action index, directly measuring the interface term in Theorem 1.

Controlled proposer audit. To isolate semantic proposal quality from the physical response model, we replay alternative proposers at the same 500 posterior–history states from 50 Source Location episodes while holding the compiler, physical EIG, and 12-candidate budget fixed. On the 281 histories with non-negligible full-catalog EIG, removing belief and history from the LLM increased the mean compilation-omission gap from 0.0504 to 0.1254 bits. The paired increase was 0.0751 bits (95% episode-bootstrap interval $[0.0588, 0.0918]$). A task-specific deterministic belief heuristic attained a numerically smaller mean gap of 0.0427 bits, but its gap difference from the LLM was not resolved ($[-0.0167, 0.0012]$). The LLM nevertheless retained an acquisition with at least 95% of full-catalog utility more frequently than the heuristic (77.3% versus 67.2%), a paired improvement of 10.1 percentage points ($[5.7, 14.6]$). Table 2 further shows that both posterior-aware proposers outperform size-matched random and static intents. Thus, the observed benefit is attributable to belief-conditioned semantic choice rather than candidate-set size alone, while a task-engineered posterior-aware proposer remains a competitive control.

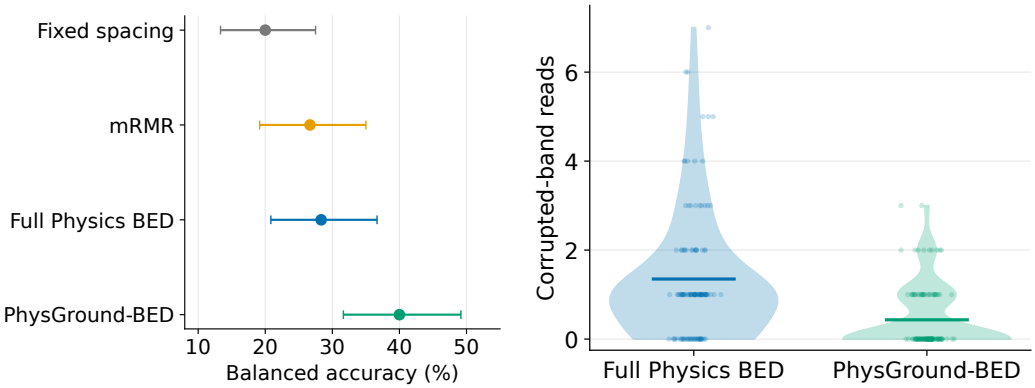
6.3 Semantic Restriction Retains Informative Acquisitions

Figure 3 compares endpoint accuracy and numerical search across the four larger catalogs. The left panel reports retained task performance. The right reports PhysGround-BED candidate scores as a percentage of full search, where lower is better.

Source Location and IEEE 123 match full-search accuracy using 1.2% and 5.7% of full-search candidate scores, with mean audited EIG gaps of 0.028 and 0.038 bits. Battery EIS also matches Full Physics BED at 68.7% balanced accuracy using 24.7% of full-search scoring. Table 3 reports the corresponding Source Location and Battery EIS endpoints. IEEE 13 reaches 81.8% compared with 90.9% using 11.5% of the scores. Sol's 14-channel catalog and broad semantic families use 95.7% of full-search scoring, so its principal role here is to test response grounding rather than search reduction.

Table 3: Cross-system endpoints. Source Location contains 50 paired episodes. Battery EIS contains 120 leave-one-battery-out snapshots. Battery error is SOH MAE.

System	Method	Acc. (%)	Error	Scores
Source Location	Fixed sweep	80.0	0.0273	0
	Full Physics BED	100.0	$< 10^{-10}$	10,195
	PhysGround-BED	100.0	$< 10^{-10}$	120
Battery EIS	Fixed sweep	62.4	0.0717	0
	Full Physics BED	68.7	0.0658	219
	PhysGround-BED	68.7	0.0657	54



(a) Balanced accuracy with 95% bootstrap intervals. (b) Episode-level distribution of corrupted-band reads.

Figure 4: AVIRIS acquisition under textual detector-fault information. Panel a reports diagnostic performance, where farther right is better. Panel b reports fault exposure across 120 episodes, where lower is better. PhysGround-BED converts the operator note into semantic restrictions before physical scoring. Full Physics BED uses the nominal response model alone.

6.4 Textual Operating Context Redirects Grounded Acquisition

AVIRIS tests operating context supplied as an operator note describing unreliable detector regions, information absent from the nominal likelihood. Figure 4 shows the performance and fault-exposure distributions. Exact endpoint values and scoring counts are reported in Appendix Table 8. PhysGround-BED reaches 40.0% balanced accuracy, compared with 28.3% for Full Physics BED, 26.7% for mRMR, and 20.0% for fixed spacing. It averages 0.43 corrupted-band reads and avoids them in 81 of 120 episodes. Full Physics BED averages 1.35 reads and has 34 zero-exposure episodes. This case isolates the translation of textual operating context into grounded acquisition restrictions while retaining the same physical EIG criterion.

The remaining systems delimit the resolution of the semantic interface. PhysGround-BED reaches 78.0% versus 90.0% for Full Physics BED on Damped Oscillator and 45.0% versus 53.3% under nominal AVIRIS sensing. On Sol, broad telemetry intents retain 75.0% rack-OOD accuracy but score 66 of 69 candidates. These results identify the regime anticipated by Δ_t^{int}: semantic families are most effective when they form compact sets that retain high-utility acquisitions.

7 Conclusion

PhysGround-BED casts an LLM as a belief-dependent constructor of semantic experiment spaces and grounds its proposals through deterministic compilation, continuous-response EIG, and Bayesian feedback. Across engineered infrastructure and scientific sensing, the interface retains informative acquisitions with fewer numerical candidate evaluations and converts textual operating context into acquisition restrictions. The analysis identifies response grounding, interface coverage, and acquisition estimation as distinct sources of design quality. These results establish a posterior-adaptive language interface whose experiment proposals acquire numerical meaning through physical simulation and probabilistic inference.

Reproducibility Statement

Appendices B–F specify the proposal contract, typed schemas, deterministic compilation, response models, data splits, acquisition budgets, uncertainty estimates, scoring conventions, and mechanism-audit protocols. The anonymized implementation, benchmark preparation scripts, experiment entry points, and evaluation code are available at `https://anonymous.4open.science/r/physground-bed-18CD/`. Raw Sol telemetry is subject to provider access restrictions and cannot be redistributed, while the repository provides the corresponding parser, preprocessing procedure, and evaluation code for use with an authorized copy. Appendix B documents the proposer configuration and available provider-side decoding controls.

AI Use Statement

Generative AI tools, including OpenAI Codex, were used to identify relevant literature, draft and polish portions of the manuscript, and assist in checking and refining the presentation of mathematical arguments. The research questions, methodology, experimental design, implementation, and scientific conclusions were determined by the authors. The authors manually verified all citations, derivations, claims, and reported results and take responsibility for the final content of the paper.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning*, 2022.

Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band AVIRIS hyperspectral image data set: June 12, 1992 Indian Pine Test Site 3, 2015. Imagery acquired in 1992.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024.

Guido Cavraro and Vassilis Kekatos. Graph algorithms for topology identification using power grid probing. *IEEE Control Systems Letters*, 2(4):689–694, 2018. doi: 10.1109/LCSYS.2018.2846801.

Guido Cavraro and Vassilis Kekatos. Inverter probing for power distribution network topology processing. *IEEE Transactions on Control of Network Systems*, 6(3):980–992, 2019.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. BED-LLM: Intelligent information gathering with LLMs and Bayesian experimental design. In *International Conference on Learning Representations*, 2026.

Roger C. Dugan and Thomas E. McDermott. An open source platform for collaborating on smart grid research. In *IEEE Power and Energy Society General Meeting*, 2011.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395. PMLR, 2021.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011. doi: 10.1613/jair.3278.

Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, William Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobawala, Thirugnanam Jagadeesan, Praful Bhargav Basani, Torey Battelle, Rebecca Belshe, Deb McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski, Dhruvil Shah, Sean M. Dudley, Gil Speyer, and Jason Yalim. The Sol supercomputer at Arizona State University. In *Proceedings of the Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, pp. 296–301. ACM, 2023. doi: 10.1145/3569951.3597573.

Hongwei Jin, Kibaek Kim, and Jonghwan Kwon. GridMind: LLMs-powered agents for power system analysis and operations. *arXiv preprint arXiv:2509.02494*, 2025.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

William H. Kersting. Radial distribution test feeders. In *IEEE Power Engineering Society Winter Meeting*, volume 2, pp. 908–912, 2001.

Katarzyna Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with large language models. In *International Conference on Learning Representations*, 2025.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qian Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. doi: 10.1109/TPAMI.2005.159.

Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C. Cousins, William A. Johnson, Xiaotong Wang, Mihir Shah, Russ B. Altman, Denny Zhou, Mengdi Wang, and Le Cong. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, 10(2):245–258, 2026. doi: 10.1038/s41551-025-01463-z.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Bhaskar Saha and Kai Goebel. Battery data set. NASA Prognostics Data Repository, NASA Ames Research Center, 2007.

Adhithya Saravanan, Rik Knowles, Gavin Kerrigan, and Tom Rainforth. DiffBED: Scaling Bayesian experimental design to high-dimensions. In *International Conference on Learning Representations*, 2026.

Sina Taheri, Vassilis Kekatos, and Guido Cavraro. An MILP approach for distribution grid topology identification using inverter probing. In *IEEE Milan PowerTech*, 2019.

Shivanshu Tripathi, Hamed Mohsenian-Rad, and Maziar Raissi. VeraGrid-Agent: Tool-augmented LLMs for distribution optimal power flow at the grid edge. *arXiv preprint arXiv:2607.25155*, 2026.

Yang Weng, Yizheng Liao, and Ram Rajagopal. Distributed energy resources topology identification via graphical modeling. *IEEE Transactions on Power Systems*, 32(4):2682–2694, 2017. doi: 10.1109/TPWRS.2016.2628876.

Andreas Winter. Tight uniform continuity bounds for quantum entropies: Conditional entropy, relative entropy distance and energy constraints. *Communications in Mathematical Physics*, 347 (1):291–313, 2016. doi: 10.1007/s00220-016-2609-8.

Jiafan Yu, Yang Weng, and Ram Rajagopal. PaToPa: A data-driven parameter and topology joint estimation framework in distribution grids. *IEEE Transactions on Power Systems*, 33(4):4335–4347, 2018. doi: 10.1109/TPWRS.2017.2778194.

Jiafan Yu, Yang Weng, and Ram Rajagopal. PaToPaEM: A data-driven parameter and topology joint estimation framework for time-varying system in distribution grids. *IEEE Transactions on Power Systems*, 34(3):1682–1692, 2019. doi: 10.1109/TPWRS.2018.2888619.

A Proof of the Interface Decomposition

Proof of Theorem 1. Let x_t^{phys} maximize U_t^{phys} on $\mathcal{X}_{d,t}$, and let $x_t^{\mathcal{C}}$ maximize it on $\mathcal{C}_t$. By the definition of ε_t,

$$\begin{aligned} U_t^{\mathrm{env}}(x_t^{\mathrm{env}}) - U_t^{\mathrm{env}}(\widehat{x}_t) &\le U_t^{\mathrm{phys}}(x_t^{\mathrm{env}}) - U_t^{\mathrm{phys}}(\widehat{x}_t) + 2\varepsilon_t \\ &\le U_t^{\mathrm{phys}}(x_t^{\mathrm{phys}}) - U_t^{\mathrm{phys}}(x_t^{\mathcal{C}}) + U_t^{\mathrm{phys}}(x_t^{\mathcal{C}}) - U_t^{\mathrm{phys}}(\widehat{x}_t) + 2\varepsilon_t. \end{aligned}$$

The first difference is Δ_t^{int}. Empirical optimality gives

$$\widehat{U}_t^{\mathrm{phys}}(\widehat{x}_t) \ge \widehat{U}_t^{\mathrm{phys}}(x_t^{\mathcal{C}}).$$

Adding and subtracting these estimated scores and applying the definition of η_t gives

$$U_t^{\mathrm{phys}}(x_t^{\mathcal{C}}) - U_t^{\mathrm{phys}}(\widehat{x}_t) \le 2\eta_t,$$

which proves the result. □

B Language Proposal and Deterministic Compilation

Proposal model and context. We use the provider deployment `gemma4_31b_it` with temperature 0.8. This name is the identifier exposed by the serving endpoint rather than a claim about a separately released checkpoint. At each turn, the model receives a domain description, the current posterior summarized by the leading hypotheses and their probabilities, available resource types, a qualitative history of previous acquisitions and responses, and any operator-supplied text. It does not receive registered action identifiers, numerical control values, response-model parameters, likelihood values, EIG scores, or backend rankings. The history contains the semantic intent issued at each previous turn and a domain-specific qualitative summary of the resulting observation. The numerical observation remains available to the Bayesian backend. We request three typed intents per turn and make at most three model calls, consisting of one initial proposal and up to two repair calls.

Table 4: Language-proposal configuration used in the reported experiments. The endpoint did not expose a controllable decoding seed.

Setting	Value	Setting	Value
Temperature	0.8	Intents per turn	3
Total calls per turn	≤ 3	Output cap	800–1200 tokens
Archived outputs	Yes	Decoding seed	Unavailable

Prompt and output contract. The prompt asks for diagnostic contrasts rather than action commands. Its stable structure is shown below, while domain-specific field names and admissible values are supplied from the corresponding interface.

```
[SYSTEM]
Propose typed diagnostic experiments. Do not invent action IDs,
numerical controls, response values, or likelihoods.
[DOMAIN] hypotheses, observable phenomena, and resource types
[BELIEF] leading hypotheses and posterior probabilities
[HISTORY] previous intents and structured observation summaries
[CONTEXT] optional operator notes or temporary conditions
[OUTPUT] JSON array of typed experiment intents
```

Each object must use declared categorical fields and must include a short diagnostic rationale. The rationale is logged but never enters the numerical acquisition score. Parsing rejects undeclared keys, values outside the schema, numerical commands, action identifiers, and unavailable resources. A repair call returns only the validation error and the allowed schema. The application interfaces do not introduce a heuristic action when no executable candidate is produced. The complementary benchmarks use a logged deterministic public fallback only when all three model calls fail, which preserves paired episode evaluation.

Domain interfaces. Table 5 lists the semantic variables exposed to the proposer. The deterministic compiler first validates resource compatibility and then maps each intent to matching registered actions. Posterior-relative fields such as "between modes" or "outside the dominant region" are resolved from the current belief. Textual exclusions are applied only when they appear in the supplied context. The compiler does not predict measurements or rank candidates by information value.

Table 5: Typed semantic interfaces. Fields describe experiment families rather than catalog entries or numerical commands.

System	Intent fields	Deterministic compilation
Power feeder	resource region, P/Q mode, direction, intensity class, response focus	Retains registered inverter perturbations with matching resource, mode, direction, and qualitative magnitude.
Data center	measurement family, PDU scope, bank region, diagnostic contrast	Retains unread PDU telemetry channels with matching electrical role and location.
Source location	spatial region, posterior relation, placement relation, contrast	Resolves belief-relative regions and returns nearby unqueried sensor positions.
Damped oscillator	excitation regime, observation window, parameter contrast	Maps qualitative frequency and timing regimes to registered excitation-observation pairs.
AVIRIS	spectral region, reliability constraint, class contrast	Maps spectral descriptions to bands and removes regions named unreliable in the operator context.
Battery EIS	frequency regime, impedance component, degradation contrast	Maps electrochemical regimes to available impedance indices on the recorded sweep.

C Response Models and Acquisition Scores

Power-feeder likelihood. OpenDSS generates voltage responses for each topology, operating condition, and registered inverter perturbation. For topology θ and action x, a polynomial-feature ridge model predicts the continuous voltage-difference vector,

$$p_d(y \mid \theta, x) = \mathcal{N}\Big(y; \widehat{\mu}_{\theta,x}, \widehat{\Sigma}_{\theta,x}\Big), \qquad \widehat{\Sigma}_{\theta,x} = (1-\lambda)S_{\theta,x} + \lambda \frac{\mathrm{tr}(S_{\theta,x})}{d_y} I + \epsilon I, \tag{15}$$

where $S_{\theta,x}$ is the residual covariance, λ is selected on development data, and ϵ provides numerical regularization. Training operating conditions and evaluation episodes are disjoint. Means, Cholesky factors, and log determinants are cached by topology and action. EIG is estimated with common-random-number Monte Carlo using 256 mixture samples per candidate.

Data-center likelihood. Telemetry channels are standardized with training statistics. For each rack role, the mean and covariance are estimated from training racks and shrunk toward a diagonal target,

$$\widetilde{\Sigma}_\theta = (1-\rho)S_\theta + \rho\ \mathrm{diag}(S_\theta) + \epsilon I. \tag{16}$$

Previously acquired channels condition the Gaussian distribution of each unread channel through the conditional model in Eq. 7. EIG is evaluated by deterministic Gauss–Hermite quadrature with cached conditional factors. The robust variant inflates the conditional variance using a scale selected on validation racks and then applies the same acquisition and posterior-update equations.

Complementary response models. Source Location uses the benchmark log-normal intensity model directly. The oscillator benchmark uses Gaussian observations around simulated displacement responses. AVIRIS uses class-conditional Gaussian band models estimated from spatially separated training pixels. Battery EIS uses diagonal Gaussian likelihoods for real and imaginary impedance conditioned on degradation stage and training cell. These four systems use deterministic quadrature or cached analytic components for acquisition evaluation. In every domain, the response likelihood used to score an acquisition is also used for the subsequent Bayesian update, while train-test separation and covariance regularization limit reuse of evaluation observations.

Selection and accounting. The backend selects an acquisition by computing physical EIG over the compiled candidate set. Reported candidate scores count the numerical evaluations required for this selection and exclude LLM calls, physical measurements, and full-catalog evaluations used only for post-selection auditing. Full Physics BED requires scores for every remaining acquisition, whereas PhysGround-BED selects from the compiled subset and fixed schedules use no EIG scoring. Candidate counts measure numerical search exposure rather than elapsed-time acceleration.

D Benchmark Construction and Splits

Table 6: Evaluation protocol. Budgets denote sequential acquisitions per episode.

System	Data source	Evaluation split	Budgets
Power feeder	OpenDSS IEEE 13 and IEEE 123 feeders	operating conditions with paired topology episodes	1–6 and 1–10
Data center	ASU Sol PDU telemetry	chronological and rack-OOD evaluation	$1, 2, 4, 6$
Source location	synthetic benchmark response law	paired test-source episodes	1–10
Damped oscillator	simulated displacement trajectories	paired test-parameter episodes	1–6
AVIRIS	Indian Pines hyperspectral cube	disjoint spatial training and test tiles	1–8
Battery EIS	NASA Ames battery archive	leave-one-battery-out evaluation	1–6

Application systems. The power benchmark forms alternative radial feeder configurations by applying valid branch and tie-switch edits to the standard IEEE feeders. Registered acquisitions are positive or negative active- and reactive-power perturbations at available inverter sites, and the response is the vector of measured voltage changes. Catalog construction removes configurations or actions that fail radiality, power-flow convergence, or declared operating limits. The data-center benchmark contains 10,094 one-minute PDU records collected on the Sol research supercomputer at Arizona State University. Chronological evaluation tests later timestamps, while rack-OOD evaluation tests complete unseen racks. Each acquisition reveals one previously unread telemetry channel.

Mechanism benchmarks. Source Location follows the response law used by DAD and related BED studies and evaluates paired test source positions. Damped Oscillator evaluates identification of damping and stiffness from sequentially selected excitation-observation pairs. AVIRIS uses spatially disjoint tiles to reduce pixel leakage. Its fault condition corrupts selected spectral regions and provides the language-assisted policy with an operator note naming unreliable regions, isolating translation of textual context into acquisition restrictions. Battery EIS uses 120 archived impedance snapshots from four cells with leave-one-battery-out evaluation. Because pointwise frequencies are absent from the archive, semantic frequency regions are assigned by logarithmic interpolation between the documented sweep endpoints and are used only by the compiler.

E Evaluation Metrics and Uncertainty

Comparison interfaces. Full Physics BED and PhysGround-BED receive the same posterior and continuous response likelihood, but only Full Physics BED scores the complete remaining catalog. Fixed acquisition, topology rules, decision trees, and mRMR select registered acquisitions directly and do not invoke the compiler or physical EIG scorer. Supervised MLP trains a discriminative hypothesis classifier on development responses augmented with acquisition identity. At each turn, it estimates the expected entropy reduction of every remaining acquisition using action-specific confusion matrices estimated from development or calibration data, then selects the acquisition with the largest value. After observing its response, the classifier probabilities provide evidence for updating the current posterior. Prompt-Only predicts the latent label from the serialized language context. Active Task and BED-LLM generate language-native queries and update their language-level beliefs without access to the calibrated continuous likelihood used by the grounded methods.

Outcome metrics. Discrete tasks report top-1 or balanced accuracy. Source Location reports Euclidean RMSE, Damped Oscillator reports parameter RMSE and normalized RMSE, and Battery EIS reports balanced degradation-stage accuracy and state-of-health MAE. Search exposure is reported as the number of numerical candidate scores per episode and as the reduction relative to Full Physics BED. Post-selection audits report the EIG difference between the executed acquisition and an offline full-catalog reference evaluated at the same history.

Uncertainty estimates. Curves use 95% percentile bootstrap intervals with 5,000 resamples. Resampling preserves the experimental unit: complete episodes for the feeder and inverse problems, paired timestamp blocks for Sol, spatial tiles for AVIRIS, and batteries for EIS. Each Sol block contains one episode from every rack role. The rack-OOD split fixes one held-out rack per role, so its interval measures variation across sampled timestamp blocks rather than variation across possible held-out racks. Methods are paired on the same episode or block whenever their archived trajectories permit paired evaluation. Tables report the corresponding endpoint estimate with its bootstrap interval when space permits.

F Additional Experimental Results and Mechanism Audits

Power endpoint comparisons. Table 7 reports the registered-endpoint comparisons summarized alongside the acquisition trajectory in Section 6.

Table 7: Power endpoint comparisons. IEEE 13 uses six interventions and IEEE 123 uses ten. Scores are mean online numerical acquisition evaluations per episode.

	IEEE 13		IEEE 123	
Method	Acc. (%)	Scores	Acc. (%)	Scores
Fixed sweep	39.4	0.0	92.3	0.0
Topology rule	75.8	0.0	100.0	0.0
Decision tree	72.7	0.0	100.0	0.0
Categorical-response BED	39.4	111.0	69.2	445.0
Full Physics BED	90.9	111.0	100.0	445.0
PhysGround-BED	81.8	12.8	100.0	24.7

AVIRIS sensor-fault endpoints. Table 8 gives the exact endpoint values and numerical scoring counts corresponding to Figure 4.

Table 8: AVIRIS sensor-fault endpoint over 120 spatial test pixels and eight acquisitions. Fault reads count selected bands in corrupted detector regions.

Method	Balanced acc. (%)	Fault reads	Scores
Fixed spectral sweep	20.0	1.75	0
mRMR	26.7	2.00	0
Full Physics BED	28.3	1.35	1,572
PhysGround-BED	40.0	0.43	96

State-dependent interface audit. Figure 1 replays one Source Location test episode at acquisitions 1, 3, and 5. The upper panels show the posterior before acquisition. The middle row records the semantic intent and compiled candidate geometry. The lower panels evaluate an offline full-grid EIG surface only for diagnosis of the interface, overlaying compiled candidates, the offline reference optimum, and the executed acquisition. The displayed gap is the reference EIG minus the EIG of the executed acquisition at the same posterior. This audit visualizes state-dependent coverage and is not presented as a full regret certificate.

Fixed-history proposer audit. We reconstruct the posterior and remaining catalog immediately before each of the 500 acquisitions on the 50 archived Source Location trajectories. Each proposer emits three typed intents, and the same deterministic compiler retains at most four registered acquisitions per intent. We evaluate the exact physical EIG of the resulting candidate set and compute $\Delta_t^{\mathrm{int}} = \max_{x \in \mathcal{X}_t} U_t^{\mathrm{phys}}(x) - \max_{x \in \mathcal{C}_t} U_t^{\mathrm{phys}}(x)$. The no-belief condition retains the same

model, domain description, semantic capabilities, output schema, decoding settings, compiler, and 12-candidate budget as the full condition, removing only the posterior-belief and interaction-history fields. All 500 calls in each LLM condition produced valid, nonempty compilations without fallback. The deterministic heuristic uses posterior geometry, the static baseline uses three predeclared typed intents, and the random baseline uses the same 12-candidate budget. Its reported metrics average over 20 independent schema-matched proposals per history rather than pooling their candidates. Informative histories have full-catalog EIG above 10^{-6} bits. Confidence intervals resample complete episodes and preserve the pairing between proposers. Because this audit changes neither visited histories nor physical observations, it isolates candidate retention rather than end-to-end policy performance.