Cost-Aware Preference-Based Bayesian Optimization for Structured Prompt Configuration of Large Language Models

Muhammad Amir Saeed*[1], Sadia Saba†[1], and Antonio Candelieri‡[1]

[1]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

Abstract

Prompt configuration strongly affects the accuracy, reliability, and cost of large language models (LLMs), but jointly tuning instructions, output contracts, demonstrations, reasoning cues, and decoding settings creates a large mixed discrete black-box optimization problem. We propose **MNL-PCO**, a cost-aware, preference-based Bayesian optimization framework that uses a ridge-regularized multinomial-logit (MNL) discrete-choice surrogate over structured prompt components. The surrogate is learned from pairwise preferences induced by observed configuration scores and provides directly interpretable component utilities without requiring learned prompt embeddings or kernel design. Evaluations are performed at three validation-subset fidelities (15%, 40%, and 100%); expected improvement is weighted by a fidelity-information factor, normalized by a pre-evaluation token-cost proxy, and combined with periodic full-fidelity promotion.

We evaluate MNL-PCO in two stages. A controlled study covers four simulation regimes, 30 paired seeds per regime, a 23,328-configuration space, and 11 methods/ablations. MNL-PCO obtains the highest mean selected-prompt accuracy among the principal methods in all four regimes (0.907–0.960), significantly outperforming Hyperband and random-search baselines after Holm correction. On 100 unseen simulated configurations, its mean held-out Spearman correlation is 0.807 versus 0.667 for direct-score ridge. We then evaluate two local open-weight models—Llama 3.2 1B and Qwen2.5 1.5B—on IMDb, SST-2, and AG News with five paired seeds per setting and actual token/runtime accounting. MNL-PCO has the best mean test accuracy in four of six model–task settings and the best average accuracy rank (1.83), but it does not dominate IMDb, and the direct-score ridge surrogate ranks unseen live configurations better. These results support MNL-PCO as a competitive, interpretable structured-search method while also identifying the limits of an additive MNL utility model on real LLM behavior.

Keywords: Bayesian optimization; automatic prompt optimization; multinomial logit; preference learning; multi-fidelity optimization; structured prompts; large language models.

1 Introduction

Large language models (LLMs) are unusually sensitive to the way a task is presented. A change in instruction wording, output format, demonstration count, demonstration order, reasoning cue, temperature, or generation length can alter both predictive quality and computational cost, even

*Corresponding author. Email: m.saeed@campus.unimib.it; ORCID: 0000-0003-1650-8194
†Email: s.saba5@campus.unimib.it; ORCID: 0009-0006-3166-8953
‡Email: antonio.candelieri@unimib.it; ORCID: 0000-0003-1431-576X

when the underlying task and model are unchanged. This sensitivity has made prompt design a practical optimization problem rather than a purely linguistic one. Manual prompt engineering remains useful, but it becomes difficult to reproduce and expensive to scale when many prompt components must be tuned jointly.

Automatic prompt optimization (APO) has therefore developed along several lines. LLM-as-optimizer methods generate and refine natural-language instructions (Zhou et al., 2023; Pryzant et al., 2023; Yang et al., 2024); evolutionary and reflective methods search open-ended prompt text (Guo et al., 2024; Fernando et al., 2023; Agrawal et al., 2025); program-level systems jointly optimize instructions and demonstrations (Khattab et al., 2024; Opsahl-Ong et al., 2024); and recent work increasingly emphasizes explicit structure, latency, interpretability, and cost (Kim et al., 2026; Zhu et al., 2026; Liu et al., 2026; Xu et al., 2025). These approaches are effective when the search target is free-form text, but many deployed systems expose a different decision problem: a practitioner has a finite menu of instruction templates, output contracts, reasoning policies, example-selection choices, and decoding settings, and wishes to identify a high-performing configuration under a query or token budget.

We study this latter problem, which we call *structured prompt configuration optimization*. It is a mixed discrete black-box optimization problem. The objective is observed only after querying an LLM on an evaluation subset; categorical components dominate the search space; evaluations are noisy because the validation subset changes and model outputs can fail to parse; and the cost of a trial depends on both the number of examples and the prompt/generation length. Bayesian optimization (BO) is attractive because it is designed for expensive black-box objectives (Snoek et al., 2012; Frazier, 2018). However, standard Gaussian-process (GP) BO is most natural in continuous spaces and typically needs kernels, relaxations, or learned representations to exploit prompt structure (Sabbatella et al., 2024; Schneider et al., 2024; Tomar et al., 2025).

This paper investigates a different surrogate. We propose **MNL-PCO** (Multinomial-Logit Prompt Configuration Optimization), an extension of MNL-BO (Saeed and Candelieri, 2026) to cost-aware, multi-fidelity prompt configuration. The surrogate is a ridge-regularized multinomial-logit model trained from pairwise preferences induced by observed configuration scores. For paired alternatives, this is the familiar binary-logit/Bradley–Terry form. Categorical prompt components are represented with interpretable indicator contrasts, while ordinal decoding variables are standardized. The model therefore avoids learned prompt embeddings and kernel design, and its coefficients remain directly tied to controllable configuration components.

MNL-PCO also addresses evaluation cost. Three fidelities evaluate a candidate on 15%, 40%, or 100% of the validation set. Expected improvement is calculated on latent utility and weighted by a fidelity-information term; candidate–fidelity pairs are then normalized by a deterministic pre-evaluation token-cost proxy. Periodic promotion re-evaluates promising low-fidelity configurations at full fidelity. Actual prompt tokens, generated tokens, and inference time are recorded separately in the live experiments, allowing the planned search budget to be distinguished from realized resource consumption.

The empirical study has two layers. First, a controlled simulation uses the complete 23,328-configuration space under additive, interaction, nonlinear, and fidelity-biased response surfaces. Each regime is repeated for 30 paired random seeds and includes 11 methods and ablations. This layer isolates algorithmic behavior and permits evaluation against a known optimum. Second, we run local open-weight LLMs through Ollama on IMDb, SST-2, and AG News using Llama 3.2 1B Instruct and Qwen2.5 1.5B Instruct. The live study uses five paired seeds per model–task combination, six principal methods, actual token/runtime accounting, and a representative two-component ablation on IMDb with Llama 3.2 1B.

The results support a deliberately qualified conclusion. In simulation, MNL-PCO is the highest-accuracy principal method in all four regimes and significantly exceeds Hyperband and both random-search baselines on selected-prompt accuracy after Holm correction. Its held-out ranking model is also markedly stronger than direct-score ridge: mean Spearman correlation across regimes is 0.807 versus 0.667, and pairwise ranking accuracy is 0.811 versus 0.742. On live LLMs, MNL-PCO has the best mean selected-prompt accuracy in four of six model–task combinations and the best average accuracy rank (1.83), but five seeds per setting are insufficient for strong per-setting significance claims. Moreover, the live held-out surrogate diagnostic reverses the simulation ordering: direct-score ridge has higher mean held-out rank correlation in all six settings. We treat this discrepancy as evidence about where the proposed surrogate helps and where its additive utility model is too restrictive, rather than hiding it behind an aggregate score.

The paper makes four contributions:

1. We formulate structured LLM prompting as a preference-based mixed-variable BO problem and adapt an MNL discrete-choice surrogate so that the learned parameters correspond to prompt components rather than latent embeddings.
2. We develop a cost-normalized, fidelity-weighted acquisition and promotion procedure for subset-based prompt evaluation, while separating planned budget units from realized token and runtime measurements.
3. We provide a broad controlled study with four simulation regimes, 30 paired seeds, 11 methods/ablations, held-out surrogate validation, simple regret, cost-normalized AUC, target-reaching analysis, and multiplicity-corrected paired tests.
4. We provide live validation across two local open-weight models and three classification tasks, with actual resource accounting and a representative real-model ablation. The live results are reported with the same emphasis on negative and null findings as on wins.

The contribution is thus not "prompt optimization by BO" itself, nor multi-fidelity evaluation itself. Both are established. The distinctive question is whether an interpretable preference-based discrete-choice surrogate can be a useful alternative inside a structured, cost-aware prompt-configuration optimizer.

2 Related Work

2.1 Automatic prompt optimization

Early APO systems showed that prompt design can be automated without updating the target model. APE generates and scores candidate instructions (Zhou et al., 2023); ProTeGi uses natural-language critiques as textual gradients (Pryzant et al., 2023); and OPRO places optimization trajectories in the context of an LLM and asks it to propose improved solutions (Yang et al., 2024). EvoPrompt and PromptBreeder use evolutionary search to mutate and recombine natural-language prompts (Guo et al., 2024; Fernando et al., 2023). DSPy and MIPROv2 broaden this view to language-model programs in which instructions and demonstrations are optimized jointly (Khattab et al., 2024; Opsahl-Ong et al., 2024), while TextGrad represents feedback and optimization steps in natural language (Yuksekgonul et al., 2024). GEPA adds reflective evolution and a Pareto mechanism for retaining complementary prompt improvements (Agrawal et al., 2025).

The 2026 literature increasingly questions monolithic prompt rewriting. LLPO predicts structured prompt fields with a lightweight classifier to reduce optimization latency (Kim et al., 2026); MePO

explicitly organizes optimization around model-agnostic prompt merits (Zhu et al., 2026); and aPSF discovers and intervenes on semantic prompt factors rather than editing the whole prompt at once (Liu et al., 2026). LCP learns from contrasts between stronger and weaker prompts (Li, Aggarwal, et al., 2026). These developments make structured and interpretable optimization especially relevant. MNL-PCO differs by assuming that the controllable prompt grammar is specified in advance and by learning a statistical utility model over that grammar rather than generating new text.

2.2 Bayesian optimization for prompts

BO is a natural framework when prompt evaluations are expensive and black-box. Sabbatella et al. (2024) formulate hard prompt tuning as BO over a continuous relaxation of discrete token choices. InstructZero optimizes a soft prefix that is converted into an instruction for a black-box model (Chen, Chen, et al., 2023). BODE-GEN performs GP-based BO in an embedding space and uses an auxiliary LLM to bridge between discrete prompts and continuous representations (Tomar et al., 2025). HbBoPs is particularly close to our setting: it uses a structure-aware deep-kernel GP and Hyperband to allocate the number of validation examples used to evaluate prompt components (Schneider et al., 2024).

MNL-PCO does not claim that BO for prompts is new. Its difference lies in the surrogate and in the unit of interpretability. Instead of learning smoothness in a continuous or embedded prompt space, it estimates component-level utility contrasts from preferences between evaluated configurations. The search remains finite and structured, and no embedding model is required.

2.3 Preference-based optimization

Preference learning is well established in BO (Brochu et al., 2010; González et al., 2017) and in discrete-choice theory (Train, 2009). Recent prompt optimizers also use preferences, but often at a different level. PDO formulates label-free prompt selection as a dueling-bandit problem using pairwise feedback from an LLM judge (Wu, Verma, et al., 2026). PrefPO similarly uses an LLM discriminator to compare outputs and guide a prompt optimizer (Singhal et al., 2026). In contrast, MNL-PCO constructs preferences between *configurations* from task scores and fits an explicit parametric utility function. The preference formulation is therefore a surrogate-learning device rather than a replacement for task labels.

2.4 Cost-aware and multi-fidelity optimization

Multi-fidelity BO allocates cheap, noisy evaluations and expensive, accurate evaluations adaptively (Kandasamy et al., 2017; Wu, Toscano-Palmerin, et al., 2020). Hyperband provides a non-Bayesian resource-allocation baseline based on successive halving (Li, Jamieson, et al., 2018). In prompt selection, HbBoPs uses the number of validation instances as fidelity (Schneider et al., 2024). EcoTune uses token-based fidelity, a token-aware expected-improvement criterion, and dynamic fidelity scheduling for inference hyperparameter optimization (Xu et al., 2025). Efficient fidelity allocation is also appearing in broader LLM hyperparameter optimization (Chen, Xiao, et al., 2026).

Our subset fidelities and cost normalization should therefore be viewed as an adaptation of established ideas. The methodological question is whether they work effectively when acquisition is driven by a preference-based MNL utility model. This is why the experiments include MNL-Full, MNL-NoCost, MNL-NoPromotion, Random-MF, Random-Full, Hyperband, and GP-Full.

2.5 Categorical and mixed-variable BO

Mixed-variable BO has been approached through tree-based surrogates, specialized kernels, and hybrid categorical/continuous models (Hutter et al., 2011; Ru et al., 2020; Deshwal et al., 2022). MNL-BO introduced a discrete-choice surrogate for categorical and mixed-variable optimization (Saeed and Candelieri, 2026). MNL-PCO inherits that motivation but changes the application and the optimization loop substantially: evaluations are multi-fidelity, the response is a prompt-performance statistic with parsing failures, resource use is explicit, and the experiments include live LLM inference.

Positioning. Table 1 summarizes the distinction most relevant to this work. We avoid ranking prior methods by a generic "interpretability" label; instead, we distinguish whether the optimization model exposes parameters directly associated with controllable prompt components.

Table 1: Positioning relative to representative prompt optimizers.

Method	Search object	Optimization engine	Multi-fidelity	Component-level statistical utility
APE / OPRO	Free-form prompt text	LLM proposal	No	No
EvoPrompt / GEPA	Free-form prompt text	Evolution / reflection	No	No
BODE-GEN	Prompt text via embeddings	GP BO	No	No
HbBoPs	Structured prompt components	Deep-kernel GP + Hyperband	Yes	No
PDO / PrefPO	Prompt text	Preference-driven LLM/bandit	No	No
LLPO / aPSF	Structured fields/factors	Classifier / interventional updates	Task-dependent	Not an MNL utility model
MNL-PCO	**Structured configuration**	**Preference MNL + BO**	**Yes**	**Yes**

3 Methodology

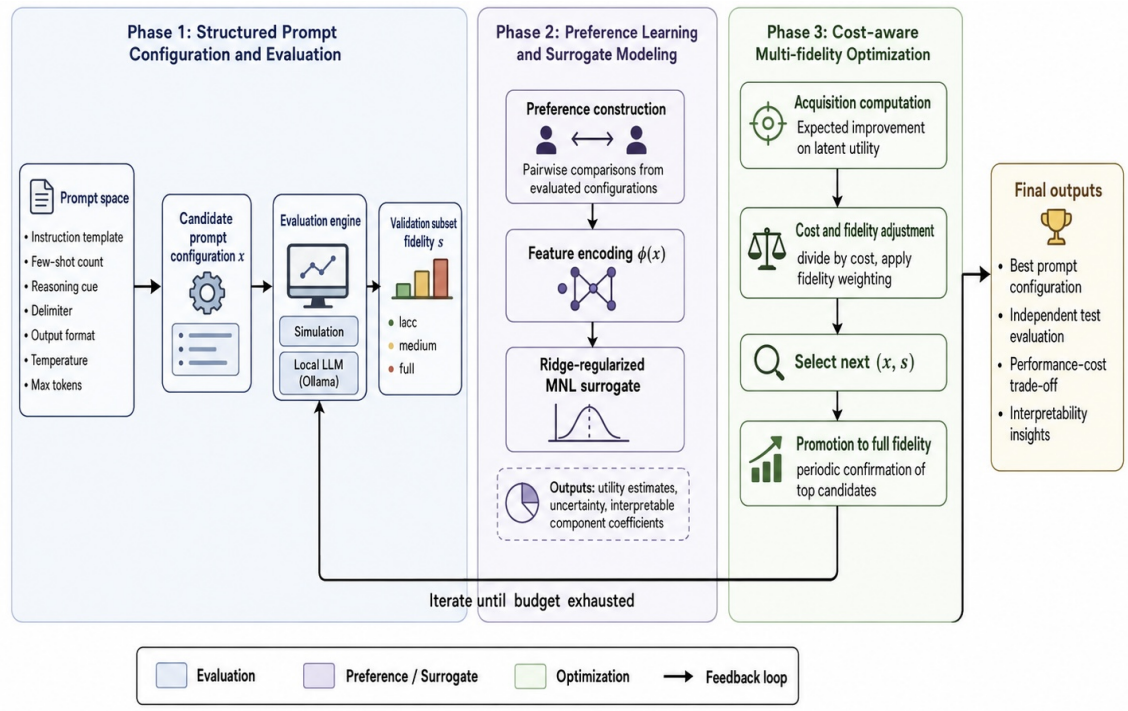

Figure 1: Architecture of MNL-PCO for structured prompt optimization. Phase 1 defines the structured prompt-configuration space and evaluates candidate configurations at different validation fidelities using either simulation or a local LLM. Phase 2 converts evaluation history into pairwise preferences, encodes prompt configurations, and fits a ridge-regularized multinomial logit (MNL) surrogate that provides utility estimates, uncertainty information, and interpretable component effects. Phase 3 uses these quantities in a cost-aware multi-fidelity acquisition rule to select the next configuration–fidelity pair and periodically promotes strong candidates to full-fidelity confirmation. The process iterates until the evaluation budget is exhausted, after which the best configuration is tested independently and used for performance-cost and interpretability analysis.

Figure 1 provides an overview of the complete MNL-PCO pipeline. The method begins from a structured prompt space whose dimensions correspond to interpretable prompt components such as the instruction template, few-shot count, reasoning cue, delimiter, output format, temperature, and maximum token budget. Candidate prompt configurations are then evaluated on a validation subset at a selected fidelity level, producing observed performance and resource-consumption signals.

The second phase transforms these observations into pairwise preference comparisons and fits a ridge-regularized MNL surrogate over the encoded prompt components. This surrogate serves two purposes: it guides the search through latent-utility estimates and uncertainty-aware acquisition, and it retains interpretability through component-level coefficients. In the final phase, MNL-PCO applies a cost-aware multi-fidelity acquisition rule to select the next configuration and fidelity, while periodically promoting strong candidates to full-fidelity evaluation. This loop continues until the budget is exhausted, after which the best discovered configuration is evaluated on an independent test set.

The remainder of this section explains each part of the pipeline in detail, including the prompt-space construction, preference generation, MNL surrogate estimation, and cost-aware multi-fidelity acquisition.

Table 2: Structured prompt-configuration space. The Cartesian product contains 23,328 configurations.

Component	Type	Domain
Instruction template	categorical	4 variants
Output format	categorical	4 variants
Reasoning cue	categorical	3 variants
Example ordering	categorical	fixed / positive-first / negative-first
Delimiter	categorical	backticks / quotes / plain
Chain-of-thought cue	binary	0 / 1
Few-shot count	integer	0 / 1 / 2
Temperature	ordinal numeric	0.0 / 0.2 / 0.5
Maximum generated tokens	ordinal integer	8 / 16 / 32

3.1 Problem formulation

Let $\mathcal{X}$ be a finite structured prompt-configuration space. A configuration $\mathbf{x} \in \mathcal{X}$ specifies prompt and decoding choices, and an evaluation at fidelity s returns an observed score

$$\tilde{f}(\mathbf{x}, s) = f(\mathbf{x}) + \epsilon(\mathbf{x}, s), \tag{1}$$

where $f(\mathbf{x})$ is the unknown full-evaluation performance and ϵ represents sampling noise, model stochasticity, and parsing effects. The goal is to identify a high-performing full-fidelity configuration while respecting an evaluation budget B.

The experimental grammar contains nine components:

Categorical levels are encoded as indicator contrasts and the two decoding variables are standardized. This is an encoding, but unlike prompt-embedding approaches it does not learn a latent semantic representation: each coefficient remains attached to an explicit component or level.

3.2 Preference-based MNL surrogate

The surrogate assigns a latent utility

$$u(\mathbf{x}) = \phi(\mathbf{x})^\top \beta, \tag{2}$$

where $\phi(\mathbf{x})$ is the structured feature vector. If configuration a is preferred to configuration b, then

$$P(\mathbf{x}_a \succ \mathbf{x}_b) = \sigma\Big[(\phi(\mathbf{x}_a) - \phi(\mathbf{x}_b))^\top \beta\Big], \tag{3}$$

with $\sigma(t) = 1/(1 + \exp(-t))$. The parameters are estimated with a ridge-penalized pairwise log-likelihood,

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] - \frac{\lambda}{2} \|\beta\|_2^2. \tag{4}$$

The implementation solves Eq. (4) by penalized iteratively reweighted least squares (IRLS), which is stable under the separation and small-sample conditions that occur early in the search.

Comparisons are constructed from evaluated configurations. Scores from different fidelities are adjusted using the observed sample counts and a small fidelity pseudocount before cross-fidelity comparisons are formed; near-ties are discarded. The simulation study includes a same-fidelity-only ablation to test whether these adjusted cross-fidelity comparisons are actually necessary.

3.3 Uncertainty approximation and acquisition

At the fitted solution, the regularized information matrix gives a Laplace-style covariance approximation

$$\Sigma_\beta \approx (H + \lambda I)^{-1}. \tag{5}$$

For a candidate $\mathbf{x}$,

$$\mu(\mathbf{x}) = \phi(\mathbf{x})^\top \hat{\beta}, \tag{6}$$
$$\sigma^2(\mathbf{x}) = \phi(\mathbf{x})^\top \Sigma_\beta \phi(\mathbf{x}). \tag{7}$$

The default acquisition is expected improvement (EI) on the utility scale,

$$\alpha_{\mathrm{EI}}(\mathbf{x}) = (\mu - u^+ - \xi)\Phi(z) + \sigma\varphi(z), \qquad z = \frac{\mu - u^+ - \xi}{\sigma}. \tag{8}$$

The implementation also supports UCB, but the reported experiments use the configured EI rule.

3.4 Cost-aware multi-fidelity selection

We use three subset fidelities,

$$\rho_s \in \{0.15, 0.40, 1.00\}, \tag{9}$$

corresponding to low, medium, and full validation subsets. A base acquisition independent of fidelity would make a simple "EI divided by cost" rule favor the cheapest fidelity too aggressively. MNL-PCO therefore applies the information factor

$$I_s(\mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{\sigma^2(\mathbf{x}) + \tau_s^2}, \qquad \tau_s^2 = \eta^2 \left(\frac{1}{\rho_s} - 1\right) \bar{\sigma}^2, \tag{10}$$

where $\bar{\sigma}^2$ is a robust reference scale based on the candidate-pool predictive variance. The candidate–fidelity pair is selected according to

$$(\mathbf{x}^\star, s^\star) = \arg\max_{\mathbf{x}, s} \frac{\alpha(\mathbf{x}) I_s(\mathbf{x})}{\widehat{c}(\mathbf{x}, s)}. \tag{11}$$

The planned cost $\widehat{c}$ is computed before inference from the number of evaluation examples, an approximate prompt-token count, and the configured maximum output length. This deterministic quantity is used to match budgets fairly across methods. For live Ollama runs we additionally record the actual prompt token count, generated token count, total duration, prompt-evaluation duration, and generation duration returned by the inference API. Consequently, "assigned budget" and "realized tokens/time" are separate quantities in the results.

3.5 Promotion

Every fixed number of BO iterations, the best candidate observed only at lower fidelity is considered for promotion to full fidelity. Promotion is executed only if it fits inside the remaining budget. This step has two purposes: it prevents the search from ending with a candidate supported only by a small validation subset, and it supplies high-fidelity comparisons that anchor later MNL fits. The no-promotion ablation isolates its contribution.

3.6 Reference methods and ablations

The experiments use the following methods under a shared budget:

MNL-PCO Full method: preference MNL, three fidelities, information weighting, cost normalization, and promotion.

Direct-Ridge-MF Same multi-fidelity search architecture but with ridge regression on observed scores rather than preferences.

MNL-Full MNL surrogate evaluated only at full fidelity.

MNL-NoPromotion MNL-PCO without periodic full-fidelity promotion.

MNL-NoInfo Removes the fidelity-information factor in Eq. (10).

MNL-NoCost Removes candidate-specific cost normalization.

MNL-SameFidelity Builds MNL comparisons only within the same fidelity.

Random-MF Random configurations and fidelities subject to the shared budget.

Random-Full Random search at full fidelity only (simulation study).

Hyperband Successive-halving style allocation over the same subset fidelities.

GP-Full RBF-kernel Gaussian-process BO with EI at full fidelity.

The GP and Hyperband baselines are reference implementations inside the common R harness; they are not the original HbBoPs implementation. This distinction matters when interpreting external comparisons.

4 Experimental Design

4.1 Controlled simulation study

The simulation study uses the same 23,328 structured configurations as the live experiments but replaces LLM inference with a deterministic ground-truth probability model followed by reproducible stochastic observations. Four regimes are used.

Additive. Prompt components contribute independent main effects. This favors an additive utility surrogate and tests whether MNL-PCO can recover a well-specified ranking.

Interaction. The objective adds component interactions, including few-shot–temperature, chain-of-thought–generation-length, and output-format–generation-length effects. This introduces misspecification for the additive MNL.

Nonlinear. Additional nonlinear temperature and generation-length effects and component interactions create a more strongly misspecified response surface.

Fidelity bias. The nonlinear structure is retained, and low/medium fidelities receive configuration-dependent bias so that cheap evaluations are not merely higher-variance versions of the full objective.

Each regime uses 30 paired seeds (1701–1730). MNL-PCO starts with 20 initial evaluations and performs 40 BO iterations. Validation, independent test, and held-out surrogate sets contain 500, 2,000, and 500 observations, respectively; 100 unseen prompt configurations are sampled for held-out surrogate validation. Candidate pools contain 120 configurations per BO step. All 11 methods and ablations are run under the budget consumed by the MNL-PCO reference in the corresponding seed. The simulation target accuracy is 0.85.

Because the full simulation response surface is known, we report simple regret in addition to test accuracy, cost-normalized area under the incumbent full-fidelity curve (AUC), target-reaching probability, cost to target, and budget utilization.

4.2 Live LLM study

The live study evaluates the six principal methods on two local instruction-tuned models: Llama 3.2 1B Instruct (Meta, 2024) and Qwen2.5 1.5B Instruct (Qwen Team, 2024). Inference is performed locally through Ollama (Ollama, 2026). The use of relatively small models is intentional: it permits repeated matched-budget optimization on commodity Apple-silicon hardware, but it also limits the scope of generalization.

We use three classification datasets:

- **IMDb** binary movie-review sentiment (Maas et al., 2011);
- **SST-2** binary sentence-level sentiment from the Stanford Sentiment Treebank / GLUE distribution (Socher et al., 2013; Wang et al., 2018);
- **AG News** four-class topic classification (Zhang et al., 2015).

For each model–task combination, five paired seeds are run. The live profile uses 8 initial evaluations and 16 BO iterations for MNL-PCO, 100 validation examples, 300 independent test examples, 80 held-out surrogate-evaluation examples, 20 unseen held-out configurations, a candidate pool of 50, six GP initialization points, and nine initial Hyperband configurations. Target accuracies are 0.80 for IMDb, 0.75 for SST-2, and 0.55 for AG News.

The test set and held-out surrogate set are fixed within each model–task combination, while the BO validation subset changes with the replication seed. Strict parsing is used: an unparseable output counts as an incorrect prediction. This prevents a lenient parser from turning formatting failures into artificial accuracy gains.

4.3 Representative live ablation

The two computationally most important MNL components—promotion and cost normalization—are additionally ablated on IMDb with Llama 3.2 1B for five seeds. This study contains MNL-NoPromotion and MNL-NoCost alongside the six core methods. The broader MNL-NoInfo and MNL-SameFidelity ablations are evaluated in the 30-seed simulations, where statistical power is substantially higher.

Table 3: Simulation selected-prompt test accuracy over 30 paired seeds. Entries are mean ± 95% t confidence interval. Bold denotes the highest mean among the principal methods in each regime.

Method	Additive	Interaction	Nonlinear	Fidelity bias
MNL-PCO	**0.907 ± 0.005**	**0.954 ± 0.004**	**0.959 ± 0.002**	**0.960 ± 0.001**
Direct-Ridge-MF	0.893 ± 0.006	0.936 ± 0.009	0.956 ± 0.003	0.956 ± 0.005
MNL-Full	0.879 ± 0.010	0.927 ± 0.012	0.948 ± 0.008	0.957 ± 0.003
GP-Full	0.870 ± 0.009	0.927 ± 0.009	0.954 ± 0.005	0.953 ± 0.005
Hyperband	0.854 ± 0.012	0.897 ± 0.012	0.917 ± 0.013	0.933 ± 0.010
Random-MF	0.821 ± 0.014	0.836 ± 0.015	0.868 ± 0.017	0.872 ± 0.016
Random-Full	0.820 ± 0.012	0.873 ± 0.016	0.915 ± 0.015	0.891 ± 0.015

4.4 Evaluation metrics and statistical analysis

The primary endpoint is the held-out test accuracy of the configuration selected at the end of each search. We also report macro-F1, best full-fidelity validation accuracy, strict parse rate, normalized AUC of the incumbent full-fidelity validation score against spent budget, target-reaching rate, penalized cost to target, evaluation counts, actual tokens, and wall-clock search time.

For $n \geq 5$, uncertainty around means is reported with 95% Student-t intervals. Pairwise comparisons use the Wilcoxon signed-rank test on paired seeds, a paired bootstrap confidence interval for the mean difference, and a rank-biserial effect size. Within each metric, p-values are corrected with Holm's procedure. Target-reaching rates are accompanied by Wilson intervals in the output package. The live study has only five seeds per setting; we therefore treat its per-setting inferential tests as low-powered and emphasize effect directions, confidence intervals, and cross-setting ranks rather than binary significance labels.

4.5 Reproducibility

The complete optimizer and baselines are implemented in R. All randomization that affects candidate pools, data subsets, simulated outcomes, and model generation seeds is derived from stable replication-specific seeds. Live prediction responses are cached, and method-level checkpoints allow interrupted runs to resume. The result archives supplied with this manuscript contain the aggregate tables and publication figures used below.

5 Results I: Controlled Simulation

5.1 Optimization quality

Table 3 shows a consistent pattern. MNL-PCO has the highest mean selected-prompt test accuracy among the principal methods in all four regimes: 0.907 in the additive regime, 0.954 in the interaction regime, 0.959 in the nonlinear regime, and 0.960 in the fidelity-biased regime. Relative to Random-MF, the corresponding gains are 8.58, 11.77, 9.02, and 8.82 percentage points. Relative to Hyperband, the gains are 5.30, 5.75, 4.12, and 2.72 percentage points.

The comparison with stronger model-based baselines is more nuanced. Against Direct-Ridge-MF, MNL-PCO improves selected test accuracy by 1.39 percentage points in the additive regime and 1.77 points in the interaction regime; both differences remain significant after Holm correction. In the nonlinear and fidelity-biased regimes, the mean advantages shrink to 0.30 and 0.38 points and are not Holm-significant. Against GP-Full, MNL-PCO is significantly better in the additive, interaction, and fidelity-biased simulations, while the nonlinear selected-test difference is small and non-significant.

These results are noteworthy because the simulation progressively violates the additive MNL specification. MNL-PCO does not collapse when interactions, nonlinear effects, or fidelity-dependent bias are introduced. At the same time, the shrinking advantage over Direct-Ridge-MF in the more difficult regimes indicates that the preference-based linear utility should not be interpreted as uniformly superior to direct regression.

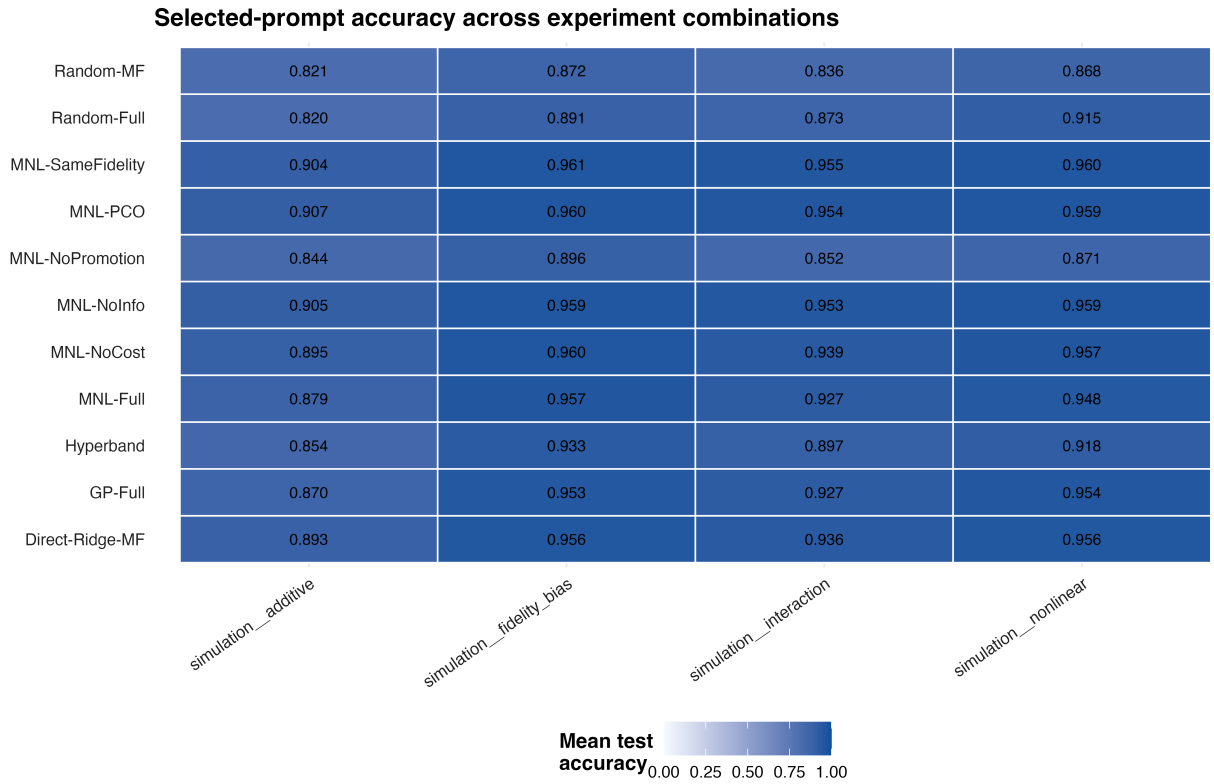

Figure 2: Selected-prompt test accuracy across the four simulation regimes and all methods. Cell values are mean accuracies over 30 paired seeds. MNL-PCO is the highest-mean principal method in each regime, while the complete ablation set reveals which components are responsible for the overall behavior.

Figure 2 provides the broadest visual summary of the controlled study. The advantage of MNL-PCO is especially clear relative to Random-MF, Random-Full, Hyperband, and the no-promotion ablation. The figure also shows that the differences among the strongest model-based variants are much smaller. In particular, MNL-NoInfo and MNL-SameFidelity remain close to the complete method in several regimes. Thus, the heatmap supports a component-level interpretation rather than a claim that every element of the full acquisition rule is independently essential.

5.2 Cost-normalized convergence and regret

MNL-PCO's mean cost-normalized AUC is 0.826, 0.855, 0.873, and 0.879 in the additive, interaction, nonlinear, and fidelity-biased regimes, respectively. It reaches the 0.85 target in all 120 MNL-PCO replications. Its simple regret is 0.0063 in the additive regime, 0.0060 in the interaction regime, 6.9×10^{-4} in the nonlinear regime, and 3.6×10^{-4} under fidelity bias. These values show that the configurations returned by MNL-PCO lie close to the known simulated optimum even when the response surface departs from the MNL specification.

The cost story is not one of universal dominance. GP-Full, for example, obtains slightly higher AUC in some of the more complex regimes, and full-fidelity methods can occasionally reach the target quickly when they encounter a strong candidate early. The more defensible conclusion is that MNL-PCO combines high final accuracy with consistently strong cost-normalized convergence under a common evaluation budget.

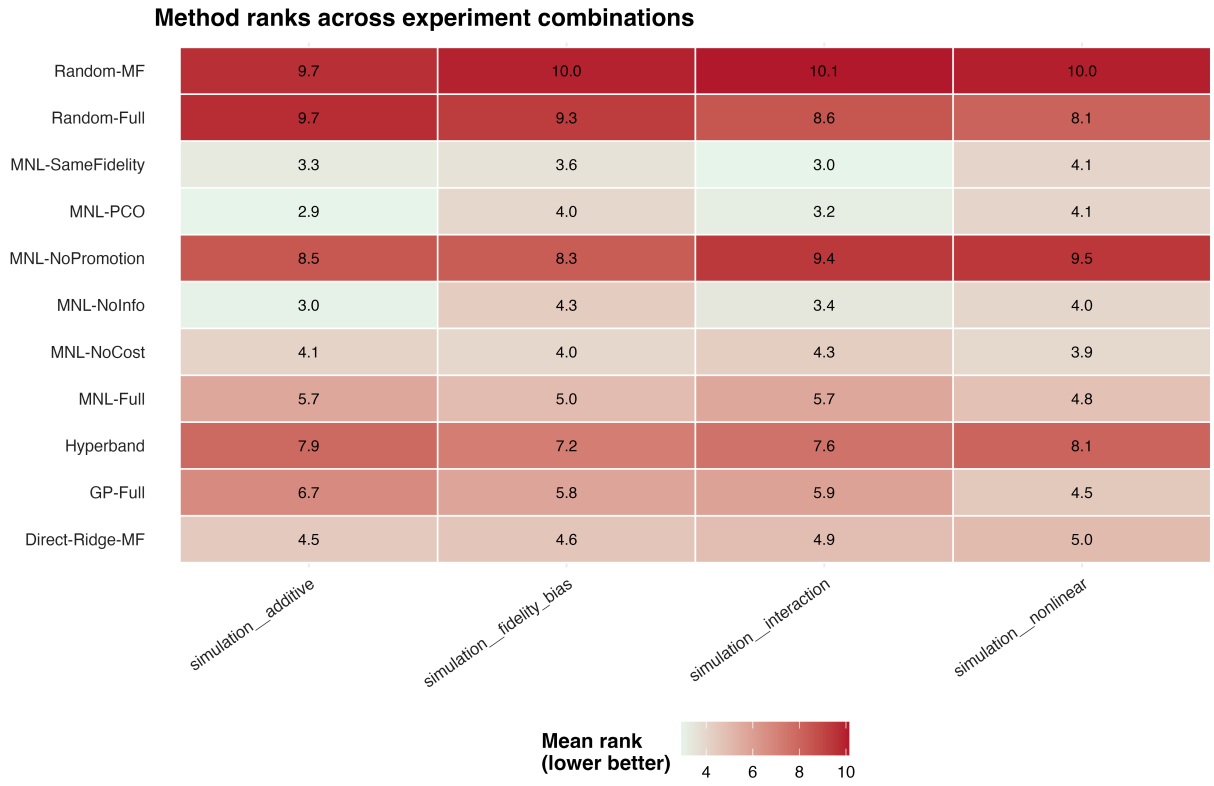

Figure 3: Mean method ranks across the primary outcomes of the four simulation regimes. Lower values indicate better performance. The ranking view complements the raw accuracy results by summarizing performance across multiple optimization criteria rather than selected-prompt accuracy alone.

Figure 3 shows that the favorable performance of MNL-PCO is not confined to a single accuracy statistic. The method remains among the leading approaches across the broader set of reported outcomes, whereas the random-search variants and the no-promotion ablation tend to occupy the lower ranks. The figure also reinforces an important qualification: rankings vary across outcomes and regimes, so MNL-PCO should be described as consistently competitive rather than uniformly optimal on every cost or trajectory metric.

5.3 Surrogate validity

Held-out surrogate validation directly tests the proposed modeling claim. The MNL surrogate is fitted using search history and then evaluated on 100 unseen configurations at full fidelity. Across all four regimes, its rank correlation and pairwise ranking accuracy exceed those of the direct-score ridge surrogate. Mean Spearman correlation ranges from 0.793 to 0.837 for MNL and from 0.604 to 0.736 for direct ridge. Pairwise accuracy ranges from 0.801 to 0.827 for MNL, compared with 0.719 to 0.770 for direct ridge.

Table 4: Held-out surrogate validation in simulation (100 unseen configurations; 30 seeds). MNL consistently improves rank correlation and pairwise ranking accuracy over direct-score ridge.

Regime	MNL ρ	Ridge ρ	MNL pair acc.	Ridge pair acc.
Additive	0.793	0.604	0.806	0.719
Interaction	0.804	0.644	0.809	0.735
Nonlinear	0.837	0.736	0.827	0.770
Fidelity bias	0.794	0.685	0.801	0.744

Table 5: Simulation ablation results: selected-prompt test accuracy over 30 paired seeds. The no-promotion variant is consistently worse; removing the information factor or restricting comparisons to the same fidelity changes mean accuracy only slightly.

Method	Additive	Interaction	Nonlinear	Fidelity bias
MNL-PCO	0.907	0.954	0.959	0.960
MNL-NoPromotion	0.844	0.852	0.871	0.896
MNL-NoCost	0.895	0.939	0.957	0.960
MNL-NoInfo	0.905	0.953	0.959	0.959
MNL-SameFidelity	0.904	0.955	0.960	0.961
MNL-Full	0.879	0.927	0.948	0.957

The improvement is not limited to the correctly specified additive regime; it persists under interaction, nonlinear, and fidelity-biased simulation. This result supports the idea that pairwise training can be robust to noise in absolute performance scores when the underlying ordering remains learnable. A representative visualization of this comparison in the nonlinear regime is provided in Appendix Figure 9.

5.4 Ablations

The clearest ablation result concerns promotion. Removing promotion reduces selected test accuracy by 6.25 percentage points in the additive regime, 10.26 points in the interaction regime, 8.78 points in the nonlinear regime, and 6.43 points under fidelity bias. All four differences are strongly significant after Holm correction. Promotion is therefore not merely an implementation detail: it anchors inexpensive multi-fidelity screening to reliable full-fidelity evidence.

Cost normalization is beneficial in the additive and interaction regimes, where MNL-PCO significantly exceeds MNL-NoCost, but its incremental effect is small in the nonlinear and fidelity-biased regimes. The information factor is less decisive: MNL-NoInfo is statistically indistinguishable from MNL-PCO on selected test accuracy in every regime. Likewise, MNL-SameFidelity is statistically tied with the adjusted cross-fidelity method. These null ablations are informative because they identify promotion, rather than every component of the acquisition rule, as the most stable contributor to the method's performance.

Appendix Figure 7 presents the complete paired effect-size view for the simulation study, while Appendix Figure 8 illustrates the evolution of simple regret under a representative interaction regime. These supplementary diagnostics support the conclusions above without duplicating the primary accuracy and rank summaries in the main text.

Table 6: Live Ollama selected-prompt test accuracy over five paired seeds. Each test set contains 300 examples. Bold denotes the highest mean in each model–task setting.

Method	IMDb/L3.2	SST-2/L3.2	AG/L3.2	IMDb/Q2.5	SST-2/Q2.5	AG/Q2.5
MNL-PCO	0.798	**0.799**	**0.487**	0.870	**0.907**	**0.730**
Direct-Ridge-MF	**0.833**	0.739	0.429	0.873	0.901	0.698
MNL-Full	0.773	0.760	0.422	**0.879**	0.901	0.717
GP-Full	0.785	0.775	0.395	0.874	0.901	0.729
Hyperband	0.778	0.772	0.447	0.877	0.891	0.707
Random-MF	0.771	0.739	0.403	0.856	0.894	0.723

L3.2 = Llama 3.2 1B; Q2.5 = Qwen2.5 1.5B. Per-setting Holm-adjusted pairwise tests have low power at $n = 5$; no universal dominance claim is made.

6 Results II: Live LLM Experiments

6.1 Selected-prompt accuracy across models and tasks

The live results are less uniform than the simulation results, which is both expected and informative. MNL-PCO has the highest mean selected-prompt test accuracy in four of the six model–task combinations: AG News with Llama 3.2 1B (0.487), AG News with Qwen2.5 1.5B (0.730), SST-2 with Llama 3.2 1B (0.799), and SST-2 with Qwen2.5 1.5B (0.907). It is not the highest-mean method on IMDb. Direct-Ridge-MF reaches 0.833 versus 0.798 for MNL-PCO with Llama, while MNL-Full reaches 0.879 versus 0.870 with Qwen.

Across the six settings, MNL-PCO has the best descriptive average test rank, 1.83. Its unweighted mean test accuracy is 0.765, compared with 0.745 for Direct-Ridge-MF, 0.743 for GP-Full, 0.742 for MNL-Full, 0.745 for Hyperband, and 0.731 for Random-MF. The unweighted average should not be interpreted as a pooled benchmark score because the tasks differ in difficulty and class structure; the average within-setting rank is the safer cross-setting summary.

Table 7: Cross-setting descriptive summary over the six live model–task combinations. Mean accuracy is an unweighted descriptive average across heterogeneous tasks; mean rank is the more appropriate cross-setting comparison (lower is better).

Method	Mean test acc.	Mean test rank	Mean AUC rank
MNL-PCO	**0.765**	**1.83**	3.00
GP-Full	0.743	3.25	3.83
MNL-Full	0.742	3.58	2.83
Direct-Ridge-MF	0.745	3.67	2.83
Hyperband	0.745	3.67	4.83
Random-MF	0.731	5.00	3.67

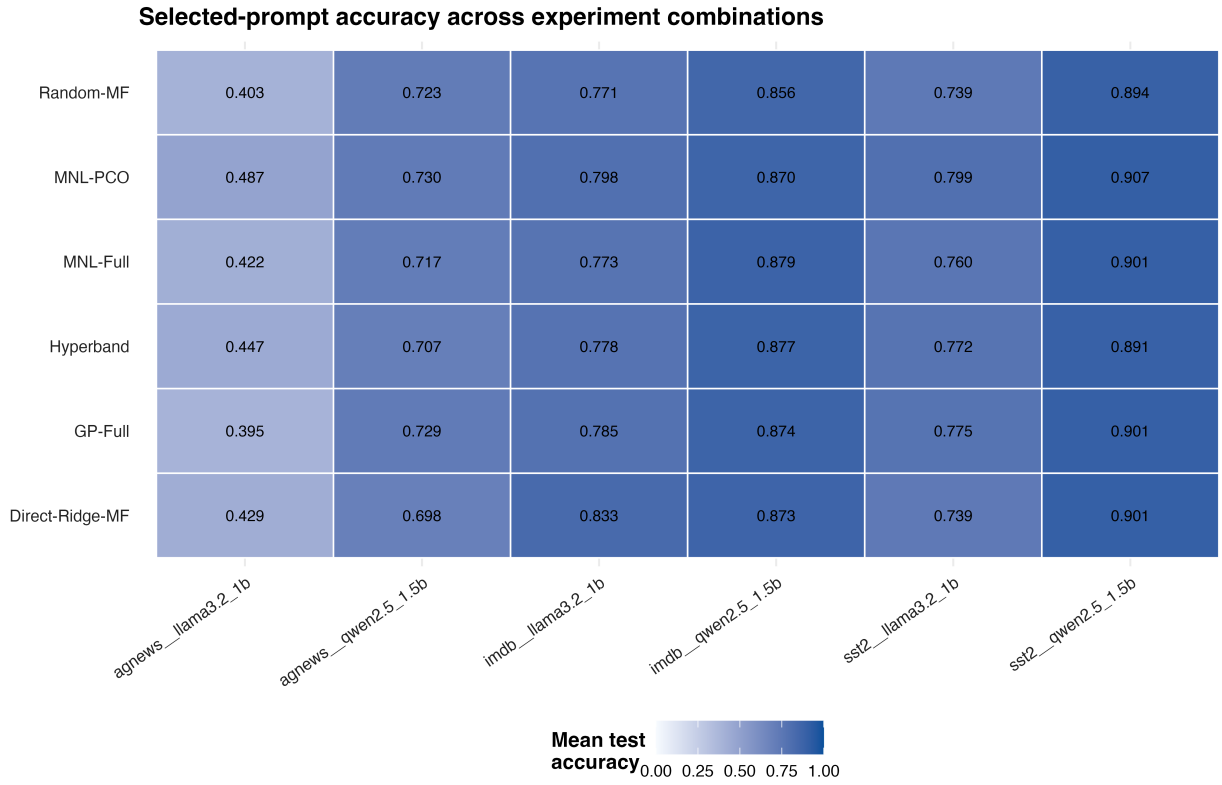

Figure 4: Selected-prompt test accuracy across the six live model–task combinations. Cell values are five-seed means. MNL-PCO achieves the highest mean accuracy in four of the six settings, while the two IMDb experiments show that the advantage does not hold uniformly across tasks.

Figure 4 makes the heterogeneity of the real-model results immediately visible. MNL-PCO is strongest on both SST-2 settings and both AG News settings, including the difficult AG News/Llama combination where all methods have relatively low absolute accuracy. The IMDb columns tell a different story: Direct-Ridge-MF is strongest for Llama and MNL-Full is marginally strongest for Qwen. This pattern supports the intended conclusion that MNL-PCO transfers beyond simulation, while also ruling out a claim of universal dominance.

With only five seeds per setting, none of the per-setting MNL-PCO-versus-baseline selected-test comparisons survives Holm correction. The paired bootstrap intervals nevertheless reveal meaningful effect directions. On AG News with Llama, MNL-PCO is 5.8 percentage points above Direct-Ridge-MF and 9.3 points above GP-Full on average. On SST-2 with Llama it is 6.0 points above

Direct-Ridge-MF. Conversely, on IMDb with Llama, Direct-Ridge-MF is 3.47 points higher than MNL-PCO. These heterogeneous outcomes are important because they show where the method helps and where a flexible direct-score surrogate remains competitive.

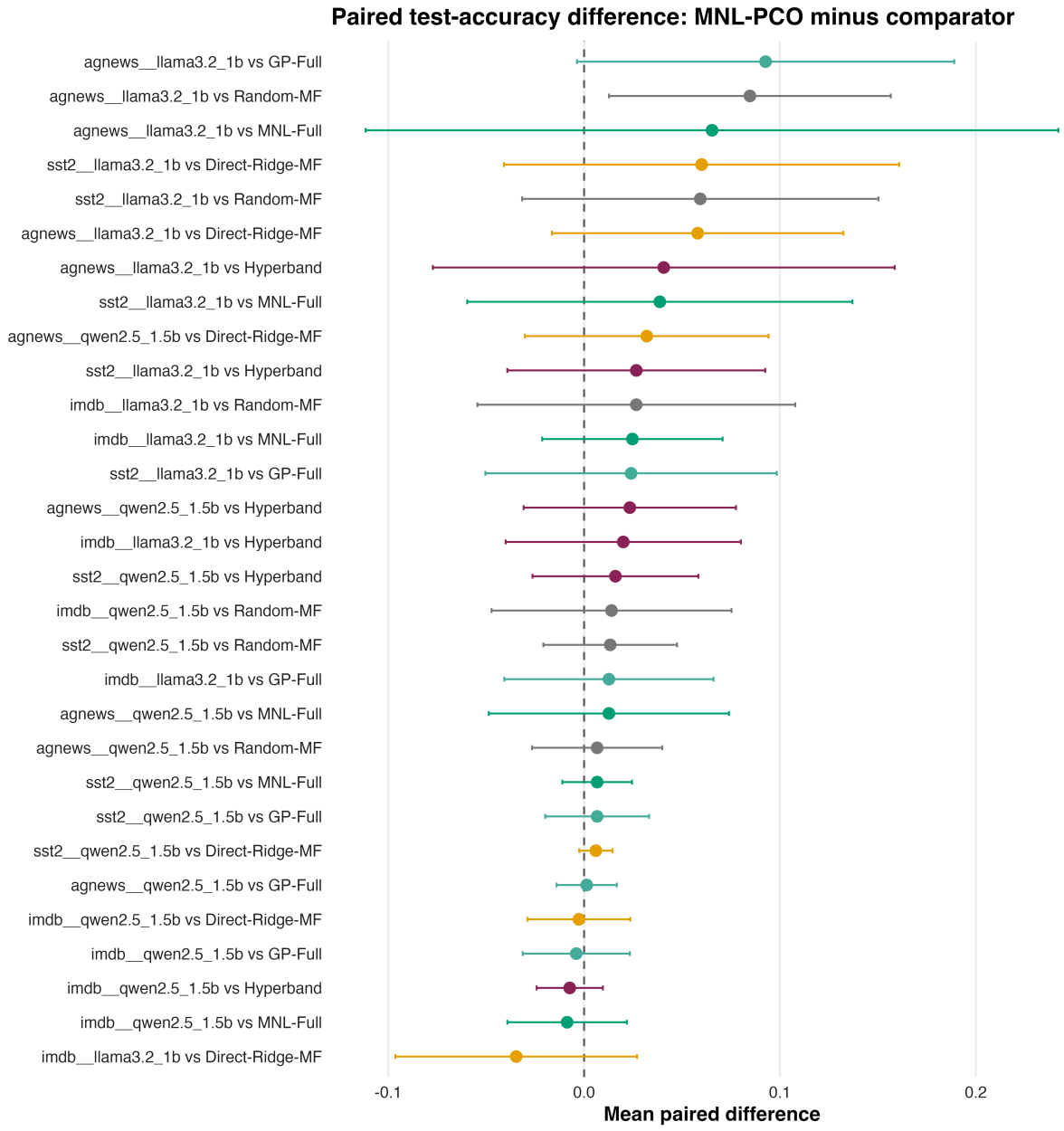

Figure 5: Paired selected-prompt test-accuracy differences for the live experiments, reported as MNL-PCO minus each comparator. Positive values favor MNL-PCO and negative values favor the comparator; horizontal intervals show the uncertainty from the five paired seeds.

Figure 5 complements the mean-accuracy heatmap by showing the magnitude and uncertainty of the paired differences. Most point estimates lie to the right of zero, particularly for several AG News and SST-2 comparisons, but many intervals overlap zero because only five paired replications are available. The negative IMDb/Llama comparison against Direct-Ridge-MF is also clearly visible. The figure therefore supports a cross-setting competitiveness claim while visually explaining why the individual live comparisons are not presented as definitive significance results.

Table 8: MNL-PCO mean live-search resource use per replication (five paired seeds). Token and wall-clock measurements come from Ollama response metadata; the search budget itself is assigned with the deterministic pre-evaluation token proxy described in Section 3.4.

Setting	Actual search tokens	Search time (s)	Evaluations	Parse rate	Target reached
IMDb / Llama 3.2 1B	304,454	1,451	24.0	0.971	1.00
SST-2 / Llama 3.2 1B	130,706	1,017	24.0	0.966	1.00
AG News / Llama 3.2 1B	194,744	787	24.0	0.990	0.20
IMDb / Qwen2.5 1.5B	307,255	1,165	24.0	1.000	1.00
SST-2 / Qwen2.5 1.5B	144,049	681	24.0	0.997	1.00
AG News / Qwen2.5 1.5B	225,935	979	24.0	0.951	1.00

Table 9: Held-out surrogate validation on live LLMs (20 unseen configurations, 80 evaluation examples, five seeds). Unlike simulation, direct-score ridge has higher mean rank correlation and pairwise accuracy in every setting; differences are not Holm-significant at $n = 5$.

Setting	MNL ρ	Ridge ρ	MNL pair acc.	Ridge pair acc.
IMDb / Llama 3.2 1B	0.095	0.178	0.534	0.571
SST-2 / Llama 3.2 1B	0.276	0.395	0.607	0.650
AG News / Llama 3.2 1B	0.314	0.370	0.618	0.634
IMDb / Qwen2.5 1.5B	0.280	0.307	0.599	0.621
SST-2 / Qwen2.5 1.5B	0.280	0.319	0.596	0.615
AG News / Qwen2.5 1.5B	0.234	0.363	0.581	0.636

6.2 Target reaching, tokens, and runtime

The target accuracy is task-specific (0.80 for IMDb, 0.75 for SST-2, and 0.55 for AG News). MNL-PCO reaches the target in all seeds for five of the six combinations. The exception is AG News with Llama 3.2 1B, where the mean selected accuracy is below the target and only one of five search runs reaches it. This is also the weakest absolute-performing setting for every method, suggesting a model–task limitation rather than a failure unique to MNL-PCO.

Actual search-token consumption for MNL-PCO ranges from approximately 131k tokens on SST-2/Llama to 307k on IMDb/Qwen. Mean recorded search time ranges from 681 to 1,451 seconds of Ollama-reported inference duration. These resource totals are similar in magnitude to those of the competing methods under the matched planned budget; the accuracy performance is therefore not explained by MNL-PCO receiving substantially more inference.

MNL-PCO's mean cost-normalized AUC across the six settings is 0.679. It does not have the best AUC rank overall: MNL-Full and Direct-Ridge-MF each average an AUC rank of 2.83, compared with 3.00 for MNL-PCO. This distinguishes *final selection quality* from *trajectory efficiency*. MNL-PCO has the strongest average final-accuracy rank, but is not uniformly the fastest method to accumulate a strong incumbent.

The complementary rank heatmap is retained in Appendix Figure 10. It provides a useful compact view of within-setting ranks but is secondary to the primary selected-prompt accuracy endpoint reported in Figure 4.

6.3 Held-out surrogate diagnostics

The live held-out analysis is the main counterweight to the simulation narrative. MNL rank correlation is modest, ranging from 0.095 to 0.314, and direct-score ridge has a higher mean Spearman correlation in every model–task combination. The same directional pattern holds for pairwise ranking accuracy. Averaged descriptively across the six settings, MNL obtains $\rho = 0.246$ and pairwise accuracy 0.589, whereas direct ridge obtains $\rho = 0.322$ and pairwise accuracy 0.621.

Table 10: Representative live ablation on IMDb with Llama 3.2 1B (five seeds). Differences are directional rather than statistically conclusive at this sample size.

Method	Test acc.	AUC	Target rate
MNL-PCO	0.798	0.746	1.00
MNL-NoPromotion	0.748	0.718	0.80
MNL-NoCost	0.781	0.730	1.00
MNL-Full	0.773	0.733	1.00
Direct-Ridge-MF	0.833	0.746	1.00
GP-Full	0.785	0.717	0.60
Hyperband	0.778	0.705	0.60
Random-MF	0.771	0.720	0.60

At five seeds, these differences do not survive Holm correction. Nevertheless, their consistency means that the MNL surrogate should not be described as a superior *global* ranking model on live LLM data. Instead, the optimization and held-out results together support a narrower interpretation: pairwise MNL can guide an effective sequential search even when its global out-of-sample ranking is weaker than direct regression. Promotion and repeated full-fidelity confirmation may reduce the consequences of imperfect global ranking, although establishing that mechanism causally would require a larger targeted ablation.

A representative held-out-surrogate visualization is provided in Appendix Figure 13; the complete model–task-specific held-out and calibration figures remain in the supplementary results archive.

6.4 Representative real-model ablation

On IMDb/Llama, removing promotion reduces mean test accuracy from 0.798 to 0.748, a five-percentage-point drop; removing cost normalization reduces it to 0.781. Neither difference is significant after multiplicity correction with only five seeds, and the no-promotion interval is relatively wide. The direction is nevertheless consistent with the much stronger 30-seed simulation ablation, where promotion is the dominant component.

Direct-Ridge-MF is the highest-mean method in this representative live setting (0.833), reinforcing the need to interpret the experiment as a component analysis rather than as evidence that MNL-PCO universally dominates direct regression.

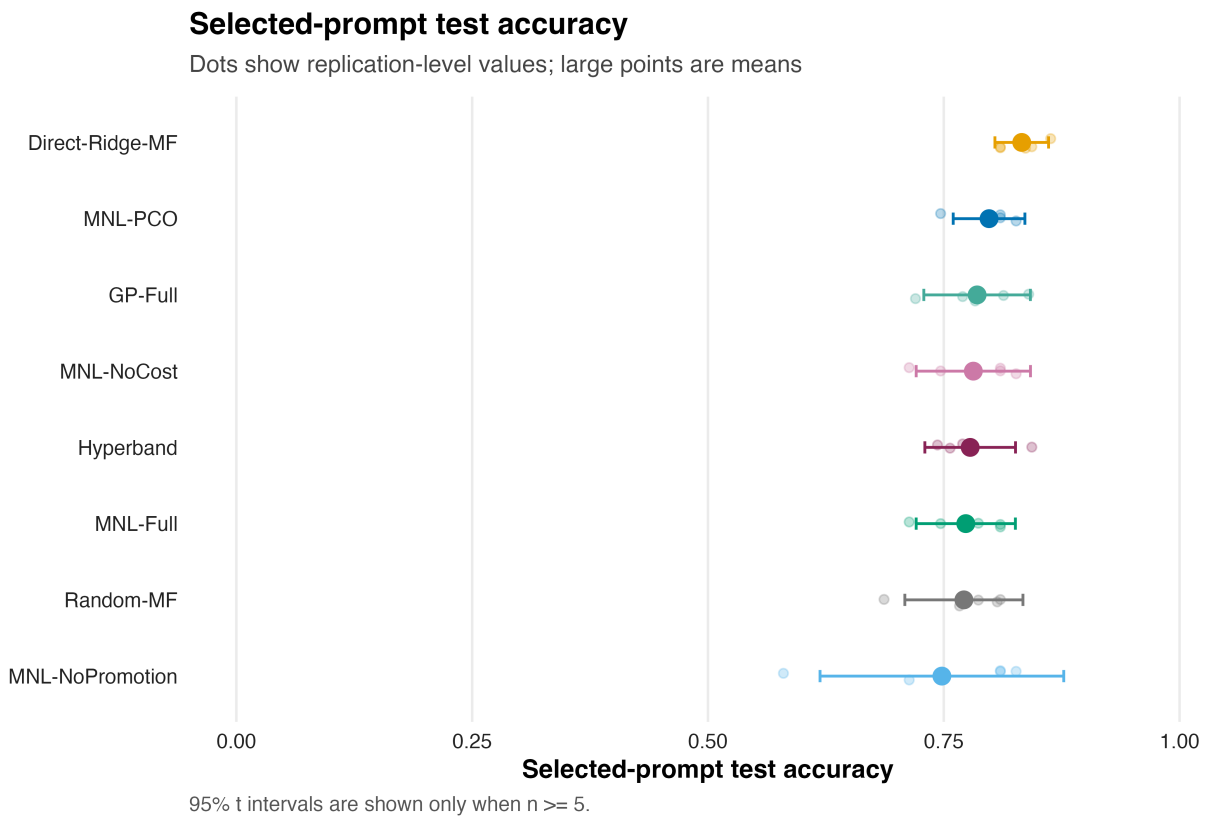

Figure 6: Representative live ablation on IMDb with Llama 3.2 1B. Small points show replication-level selected-prompt test accuracies, large points show method means, and intervals show 95% t intervals where available. Removing promotion produces the largest deterioration among the two MNL-PCO components ablated in the live study.

Figure 6 provides a useful visual link between the large simulation ablation and the smaller real-model study. The MNL-NoPromotion distribution is shifted downward relative to MNL-PCO, whereas the no-cost variant remains much closer to the complete method. The figure also makes clear that Direct-Ridge-MF performs strongly on this particular task. Consequently, the live ablation supports promotion as an important mechanism without being used to claim overall MNL superiority.

Additional diagnostics for this experiment are reported in Appendix Figures 11–14. They show how the methods allocate fidelities, how their cost-normalized trajectories compare, how MNL and direct ridge behave on unseen configurations, and how stable the fitted MNL coefficients are across replications.

7 Discussion

7.1 Where MNL-PCO works

The most convincing evidence for MNL-PCO comes from the controlled study. Across four response surfaces and 120 paired simulation replications, the full method is consistently near the top, reaches the target in every replication, and has very small simple regret. Its advantage over random and Hyperband-style allocation is large and robust. The pairwise surrogate also has substantially better held-out ranking quality than direct-score ridge in simulation.

The live study shows that this advantage can transfer, but not mechanically. MNL-PCO leads four of six model–task settings and has the best average final-accuracy rank, including the largest advantage on the difficult AG News/Llama setting. This indicates that the method is not merely overfit to the synthetic construction.

7.2 Where direct regression is stronger

The live held-out analysis is equally important. Direct-Ridge-MF is a better global ranking surrogate in all six live settings by mean Spearman correlation and pairwise ranking accuracy, and it wins the IMDb/Llama final-accuracy comparison. This is plausible because the true relationship between prompt components and LLM behavior can contain interactions, discontinuities, and task-specific formatting effects that an additive logit utility cannot express.

The implication is not that preference learning fails. Rather, the result separates two possible claims that are often conflated. MNL-PCO can be a useful *optimizer* even when MNL is not the best *global predictor*. The optimizer only needs sufficiently informative local ordering, uncertainty, exploration, and confirmation to discover strong configurations. Future work should test this explanation directly by comparing additive MNL, interaction MNL, mixed logit, and nonlinear preference models under the same acquisition loop.

7.3 Interpretability

MNL coefficients provide a form of structural interpretability that latent prompt embeddings do not: they indicate which explicit configuration contrasts are associated with higher or lower learned utility in a given task/model/seed. The live coefficient tables also show why this interpretability should be treated as *conditional*, not universal. For example, output-format contrasts and chain-of-thought cues can have different signs across Llama and Qwen settings. A coefficient is therefore evidence about the local optimization environment, not a general prompt-engineering law.

In simulation, where ground-truth effects are known, coefficient directions and held-out ranking provide a meaningful validation of this interpretability. In live LLMs, coefficient plots should be read as diagnostic summaries and hypothesis generators. A larger number of seeds and a richer preference model would be needed for stable causal claims about individual prompt components.

7.4 The role of multi-fidelity promotion

The strongest mechanistic result is the promotion ablation. In simulation, removing promotion produces a large and significant accuracy loss in all four regimes. In the representative live experiment, the mean loss is again substantial (five points) but uncertain at $n = 5$. Promotion appears to solve a practical problem that pure cost-normalized acquisition cannot: cheap subsets are useful for screening, but the optimizer needs recurring full-fidelity anchors to avoid selecting a candidate whose apparent advantage is an artifact of subset noise.

7.5 Practical interpretation

For practitioners with a finite prompt grammar and a strict evaluation budget, MNL-PCO offers three advantages. First, it can search heterogeneous categorical and decoding choices without a separate embedding model. Second, it exposes task/model-specific component utilities. Third, it naturally combines cheap subset evaluations with full-fidelity confirmation. The cost of these advantages is a

stronger structural assumption than a flexible nonlinear surrogate. The empirical results suggest that this trade-off is favorable in many settings, but not all.

8 Limitations and Future Work

Small live models. The real experiments use Llama 3.2 1B and Qwen2.5 1.5B because the full paired study was run locally on an 8 GB Apple-silicon machine. This makes the evaluation reproducible and affordable, but it does not establish behavior on 7B–70B models or proprietary frontier systems. Extending the same protocol to larger open-weight models and at least one API model is the most important external-validity step.

Five live seeds. The simulation study has 30 paired seeds per regime, but the live study has five per setting because each optimization run requires many local generations. The resulting live confidence intervals are wide, and Holm-corrected per-setting tests have low power. We therefore treat the live study as cross-model/task validation rather than definitive significance evidence. Ten or more paired seeds would strengthen the paper if compute becomes available.

Additive utility specification. The MNL surrogate is additive in the encoded component effects. Although the simulation includes interaction and nonlinear regimes, the live held-out results show that direct-score ridge can rank unseen real configurations better. Mixed logit, explicit low-order interaction terms, hierarchical logit, or nonlinear preference models are natural extensions. The challenge is to increase expressiveness without losing the coefficient-level interpretability that motivates the method.

IIA and preference construction. Standard MNL inherits the independence-of-irrelevant-alternatives assumption. In addition, preferences are induced from task scores rather than elicited directly from users or judges. A future study should compare score-induced preferences, human preferences, and LLM-judge preferences within the same structured surrogate.

Reference implementations. GP-Full and Hyperband are implemented in the common R harness to ensure identical budgets and evaluation code. They are not direct executions of HbBoPs, and the live study does not include the original PDO code. Reproducing the experiment with authors' reference implementations would improve external validity.

Finite prompt grammar. MNL-PCO optimizes a fixed configuration grammar; it does not invent new instructions. This is a strength for auditability and reproducibility but a limitation when the best prompt lies outside the predefined alternatives. A promising hybrid is to let a generative optimizer propose new levels while MNL-PCO selects and allocates evaluation resources among the growing structured alternatives.

Cost model. Budget assignment uses a deterministic pre-query token proxy so that affordability can be checked before inference. Actual Ollama tokens and time are recorded after each query, but they are not known when a candidate is selected. Future versions could learn a calibrated runtime/token predictor online and optimize expected monetary or energy cost directly.

Datasets and tasks. The live tasks are short-form classification benchmarks. Structured prompt choices may behave differently for generation, retrieval-augmented QA, coding, tool use, and long-context reasoning. Recent systems such as aPSF, LCP, and GEPA target more open-ended reasoning settings (Liu et al., 2026; Li, Aggarwal, et al., 2026; Agrawal et al., 2025); evaluating MNL-PCO there would test whether configuration-level BO remains useful when output scoring becomes more complex.

Future research. The next methodological priorities are: (i) interaction-augmented and mixed-logit surrogates; (ii) adaptive fidelity definitions based on learned token/runtime models; (iii) larger models and more seeds; (iv) original-code comparisons with structured prompt-selection baselines; (v) generative expansion of the configuration grammar; and (vi) hierarchical analyses that pool information across tasks while retaining task-specific coefficients.

9 Conclusion

We introduced MNL-PCO, a cost-aware preference-based Bayesian optimization framework for structured LLM prompt configuration. The method replaces the usual GP or embedding surrogate with a ridge-regularized multinomial-logit utility model trained from pairwise configuration preferences, and combines that surrogate with subset-based fidelities, cost normalization, and periodic full-fidelity promotion.

The controlled evidence is strong. Across four simulation regimes, 30 paired seeds, and 11 methods/ablations, MNL-PCO is the best principal method on mean selected-prompt accuracy and significantly exceeds Hyperband and random-search baselines in every regime. Its held-out simulated ranking is also stronger than direct-score ridge. Promotion is the clearest indispensable component.

The live evidence is encouraging but more heterogeneous. Across Llama 3.2 1B and Qwen2.5 1.5B on IMDb, SST-2, and AG News, MNL-PCO has the best mean test accuracy in four of six settings and the best average accuracy rank. It does not dominate IMDb, and its live held-out ranking is weaker than direct-score ridge. These findings sharpen rather than weaken the contribution: MNL-PCO is a competitive and interpretable search strategy, not a universal replacement for flexible regression surrogates.

The main lesson is therefore conditional. When prompt design can be represented as a finite structured grammar and evaluations are expensive, a preference-based discrete-choice surrogate can provide an effective optimization signal and directly interpretable component utilities. When prompt interactions dominate, the same structure becomes a modeling constraint. Developing richer preference surrogates while preserving this interpretability is the central path forward.

Data and code availability

The data and source code supporting the findings of this study, including the experimental configurations, aggregate result tables, statistical analyses, and publication figures, are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Muhammad Amir Saeed: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing—original draft, Visualization.

Sadia Saba: Methodology, Software, Validation, Visualization, Writing—review & editing.

Antonio Candelieri: Supervision, Methodology, Validation, Visualization, Writing—review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author statement

All authors contributed to the development of this work and the preparation of the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgements

Muhammad Amir Saeed and Sadia Saba acknowledge the University of Milano-Bicocca for doctoral scholarship support. The authors also acknowledge the University of Milano-Bicocca for providing the computing facilities used in this research.

A Additional Empirical Material

The main manuscript deliberately retains five figures that directly support the central empirical argument: two summarize the controlled simulation study, two summarize cross-model/task live performance, and one isolates the principal real-model ablation. The experiment code generates a much larger set of diagnostics. These additional figures are retained here and in the accompanying reproducibility archive because they provide useful evidence about convergence, uncertainty, fidelity use, surrogate behavior, and coefficient stability, but reproducing all of them in the main text would obscure the primary results.

The complete publication-figure archive is organized under:

- `figures/supplement/simulation/`,
- `figures/supplement/ollama_core/`, and
- `figures/supplement/ollama_ablation/`.

For each experiment combination, the archive includes the available test-accuracy, cost-normalized AUC, shared-budget convergence, target-reaching, penalized cost-to-target, fidelity-allocation, budget-utilization, parse-rate, token-cost-sensitivity, accuracy–resource Pareto, evaluation-count, method-rank, held-out-surrogate, calibration, and MNL coefficient diagnostics. Simulation combinations

additionally include simple-regret analyses. The corresponding aggregate CSV tables, paired tests, held-out summaries, and coefficient-stability tables are stored in `data_tables/`.

A.1 Additional simulation diagnostics

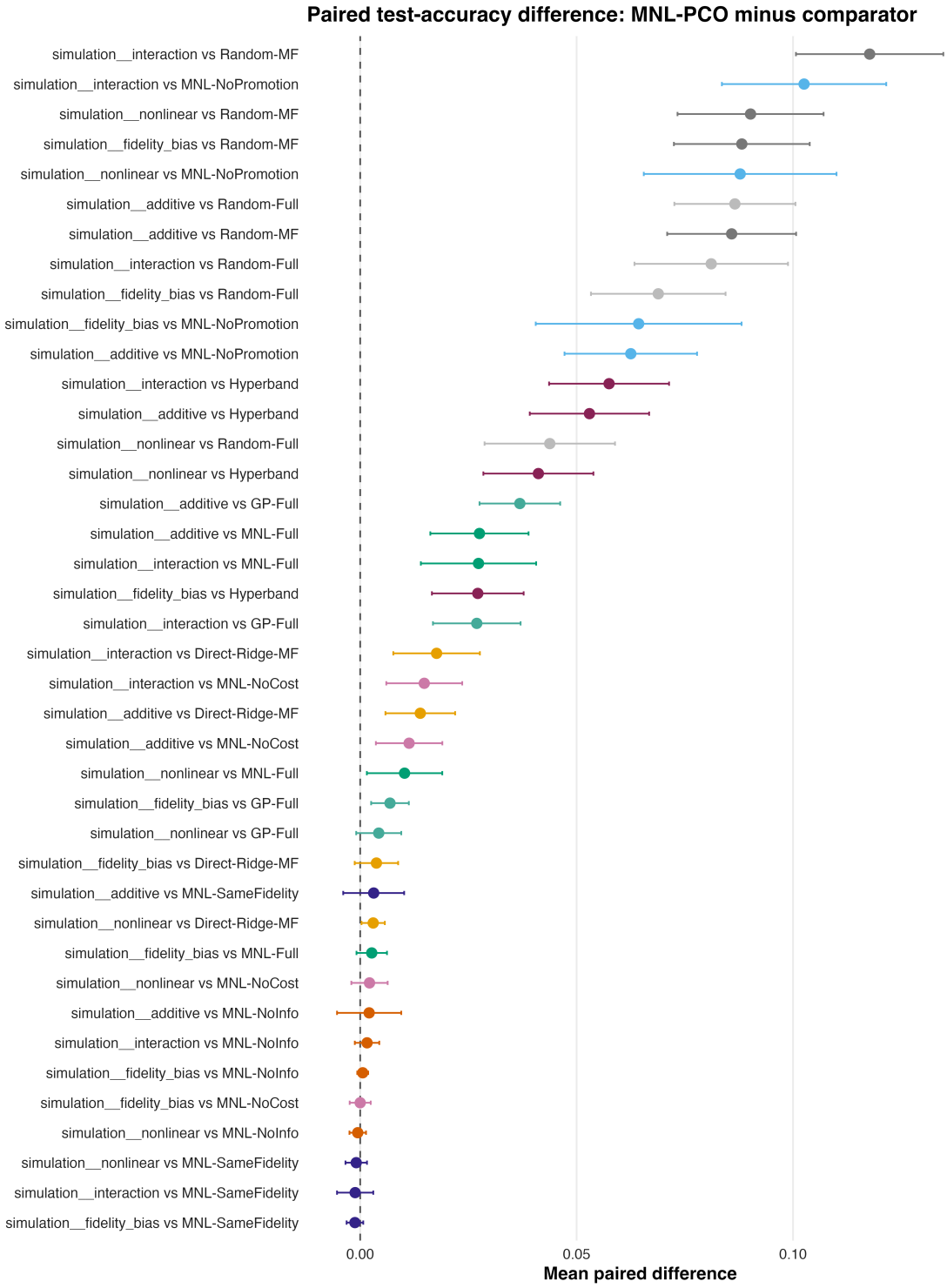

Figure 7: Paired MNL-PCO selected-test-accuracy differences versus competing methods across the simulation regimes. Positive differences favor MNL-PCO. The plot complements the aggregate heatmap in the main text by displaying the magnitude and uncertainty of the paired method contrasts.

Figure 7 expands the simulation comparison beyond the mean values shown in Figure 2. The largest positive effects occur against random search and the no-promotion variant, with substantial advantages also visible against Hyperband in several regimes. By contrast, the differences against MNL-NoInfo and MNL-SameFidelity lie close to zero. This graphical pattern is consistent with the ablation analysis in the main text: promotion is strongly supported, whereas the information factor and cross-fidelity comparison rule provide smaller and more context-dependent incremental effects.

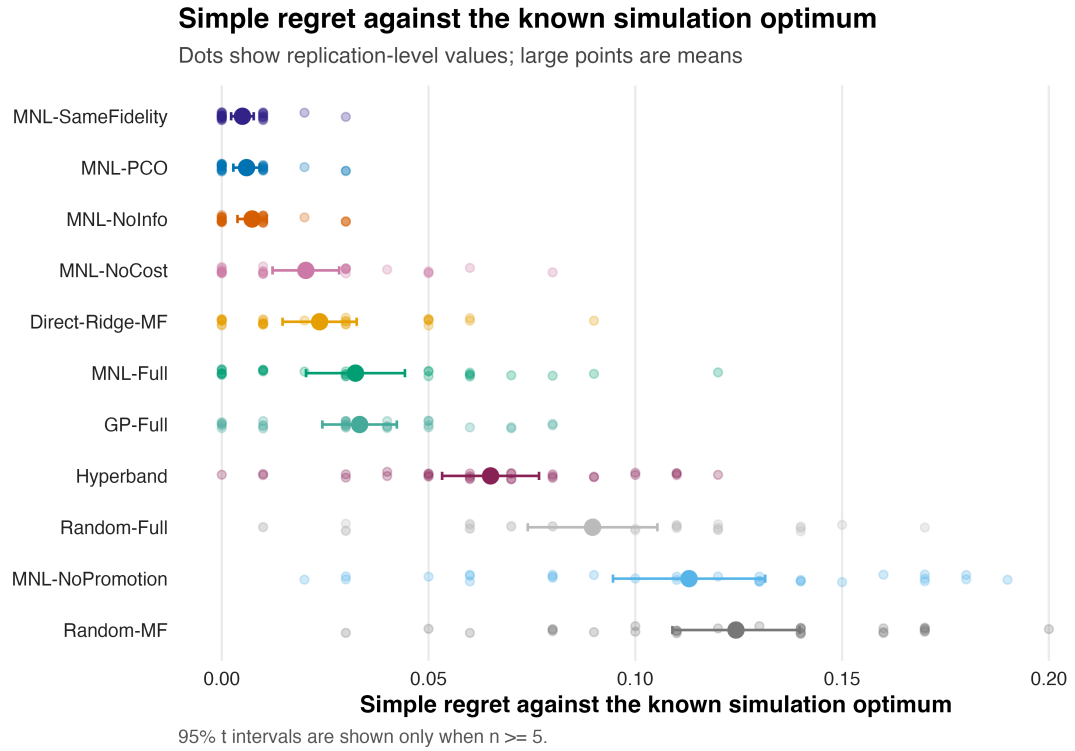

Figure 8: Simple-regret trajectories in the interaction simulation. Because the simulated response surface is known, regret can be measured relative to the true optimum; the figure illustrates how the terminal regret values reported in the main text emerge over the shared evaluation budget.

Figure 8 provides a trajectory-level view that is not available in the live experiments, where the true global optimum is unknown. The interaction regime is shown as a representative case because it deliberately violates the additive MNL specification. The corresponding regret plots for the additive, nonlinear, and fidelity-biased regimes are included in the supplementary archive.

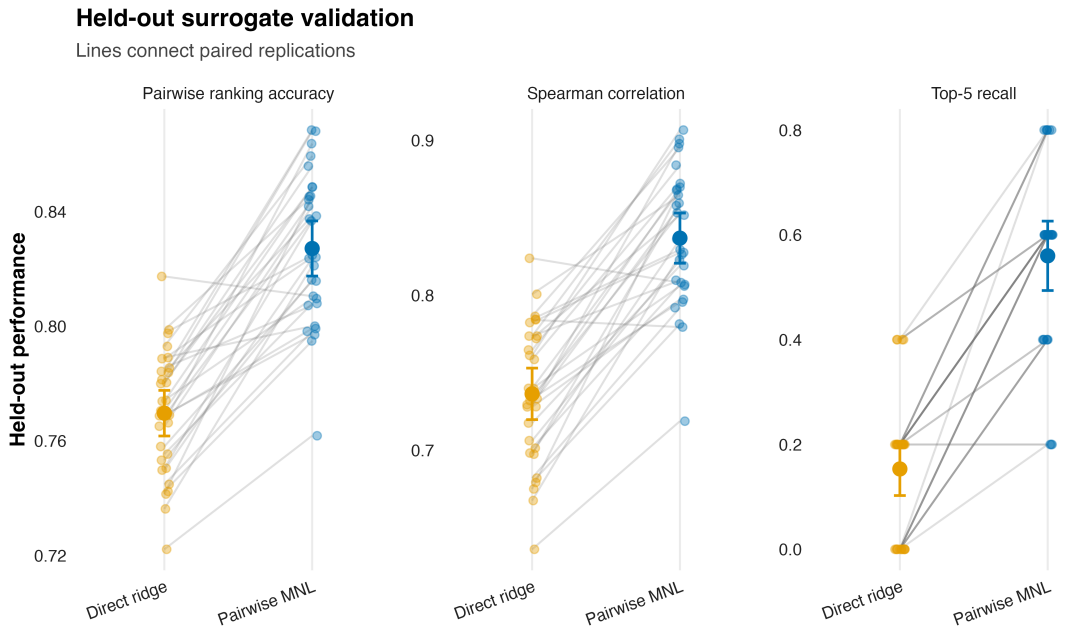

Figure 9: Held-out MNL-versus-direct-ridge surrogate comparison in the nonlinear simulation regime. This representative misspecified regime illustrates the ranking advantage summarized across all four regimes in Table 4.

Figure 9 illustrates why the simulation held-out-surrogate result is important. Even in a nonlinear regime for which an additive MNL utility is misspecified, the preference surrogate retains strong ranking performance on unseen configurations. The main manuscript reports the full four-regime summary rather than relying on this single representative plot.

A.2 Additional live-LLM diagnostics

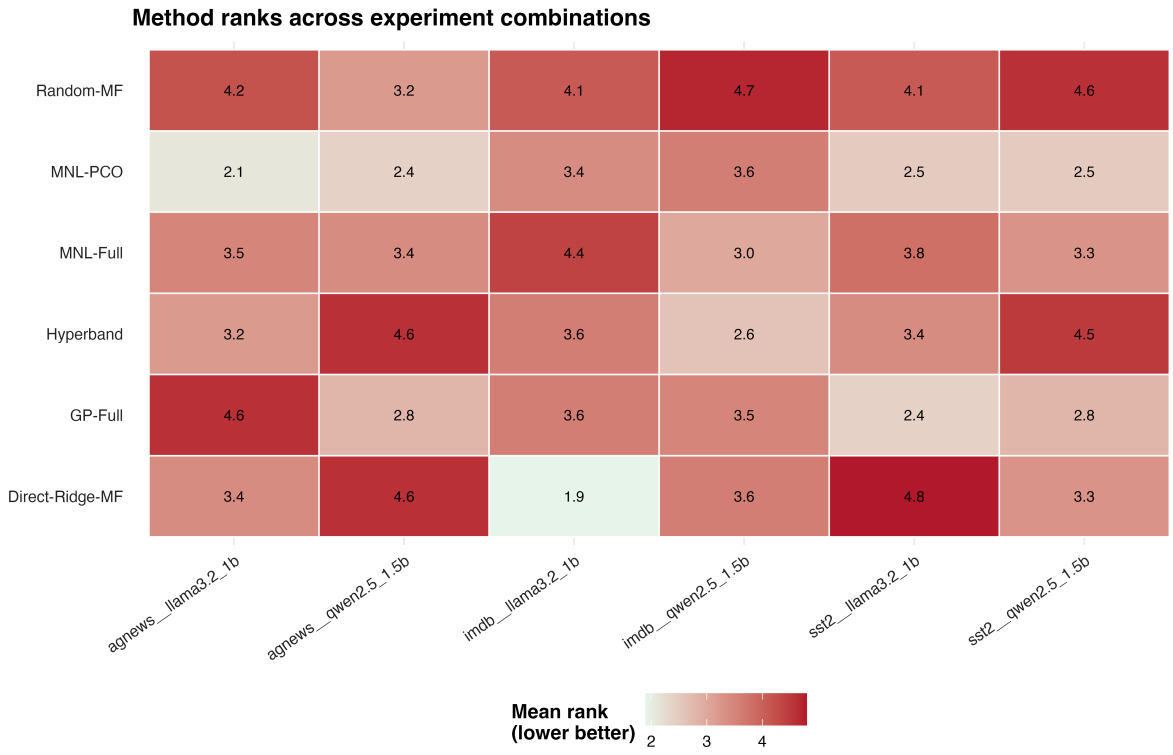

Figure 10: Mean method ranks across the six live model–task combinations. Lower values indicate better performance. The figure complements the selected-accuracy heatmap in the main text by showing how within-setting method ordering varies across models and tasks.

Figure 10 shows that MNL-PCO generally occupies a strong rank position, particularly in the AG News/Llama and SST-2 settings, but that the ordering is task dependent. Direct-Ridge-MF performs especially well on IMDb/Llama, while other model-based methods remain competitive in the Qwen settings. This reinforces the main-text conclusion that average cross-setting rank is favorable to MNL-PCO without implying universal dominance.

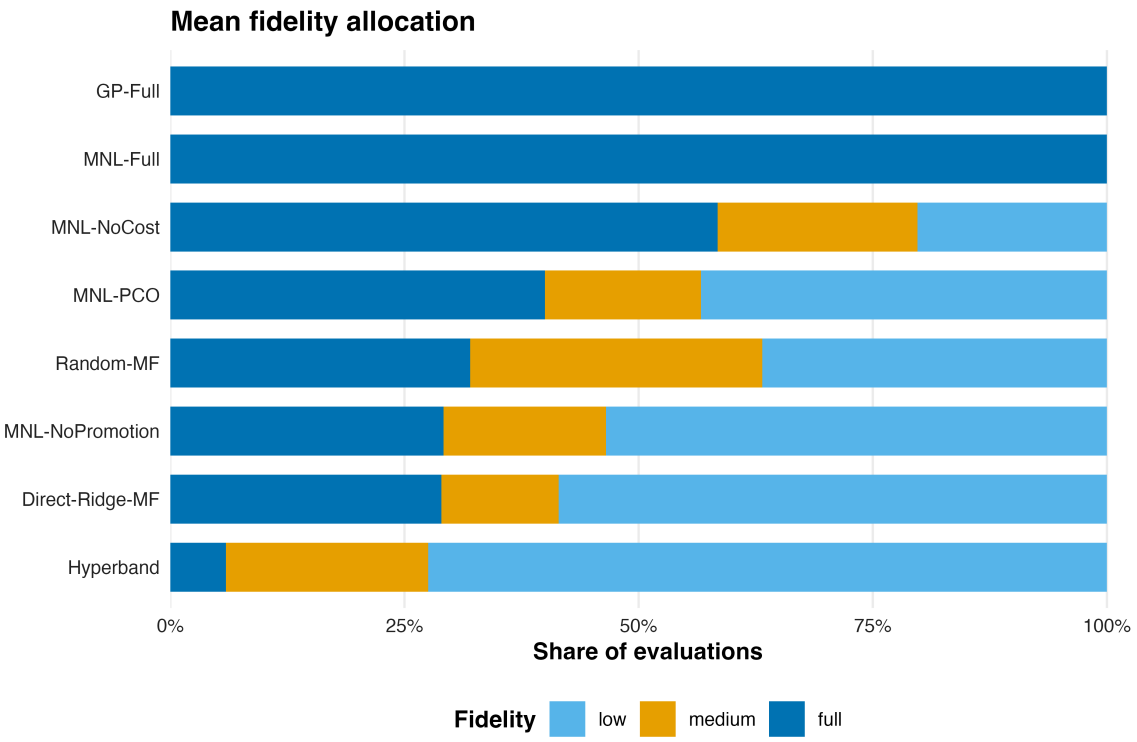

Figure 11: Mean fidelity allocation in the representative IMDb/Llama live ablation. Full-fidelity-only methods allocate all evaluations to the full subset, while the multi-fidelity methods distribute evaluations across low, medium, and full fidelities.

Figure 11 makes the resource-allocation mechanism visible. MNL-Full and GP-Full, by construction, evaluate only at full fidelity. MNL-PCO instead mixes inexpensive low- and medium-fidelity screening with full-fidelity evaluations, while Hyperband places a larger share of its evaluations at the inexpensive levels. The figure therefore helps explain why promotion is needed: a multi-fidelity method must exploit cheap screening without allowing low-fidelity evidence to determine the final choice uncorrected.

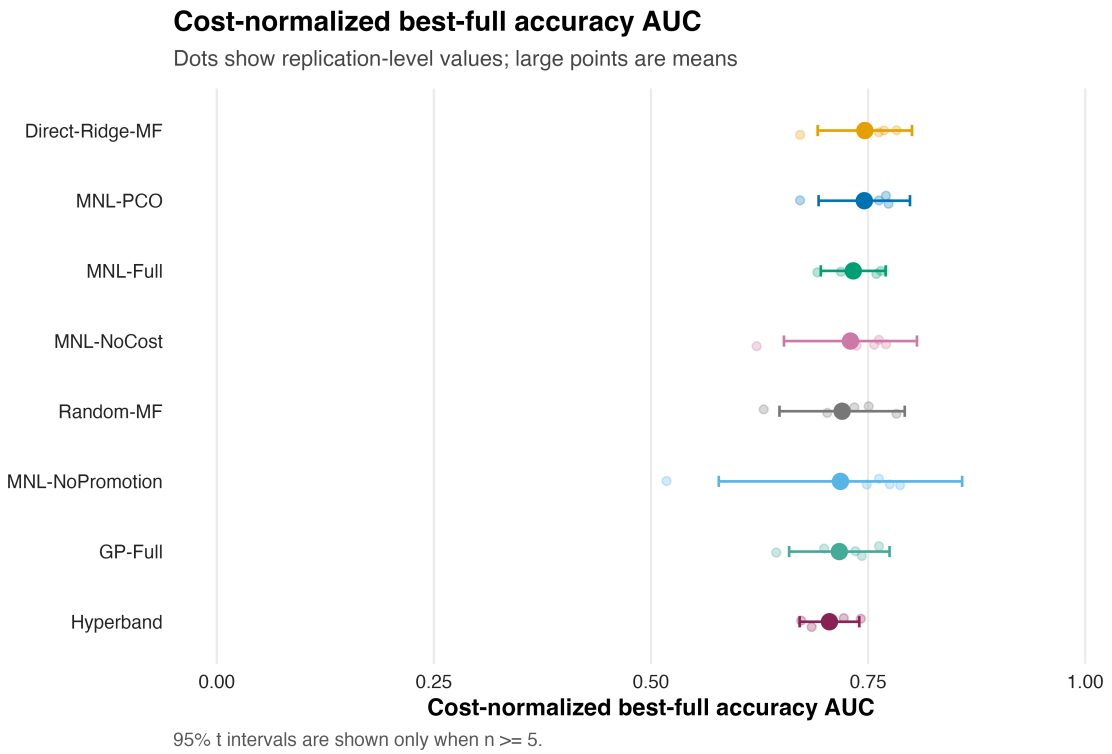

Figure 12: Cost-normalized best-full-accuracy AUC in the representative IMDb/Llama ablation. Small points denote replication-level values, large points denote means, and intervals show 95% t intervals where available.

Figure 12 complements the final selected-prompt accuracy reported in Figure 6. MNL-PCO and Direct-Ridge-MF have similar mean AUC in this representative setting, while several alternatives are lower. The no-promotion variant also exhibits substantial replication-to-replication variability. This illustrates why final accuracy and trajectory efficiency are treated as distinct outcomes in the paper.

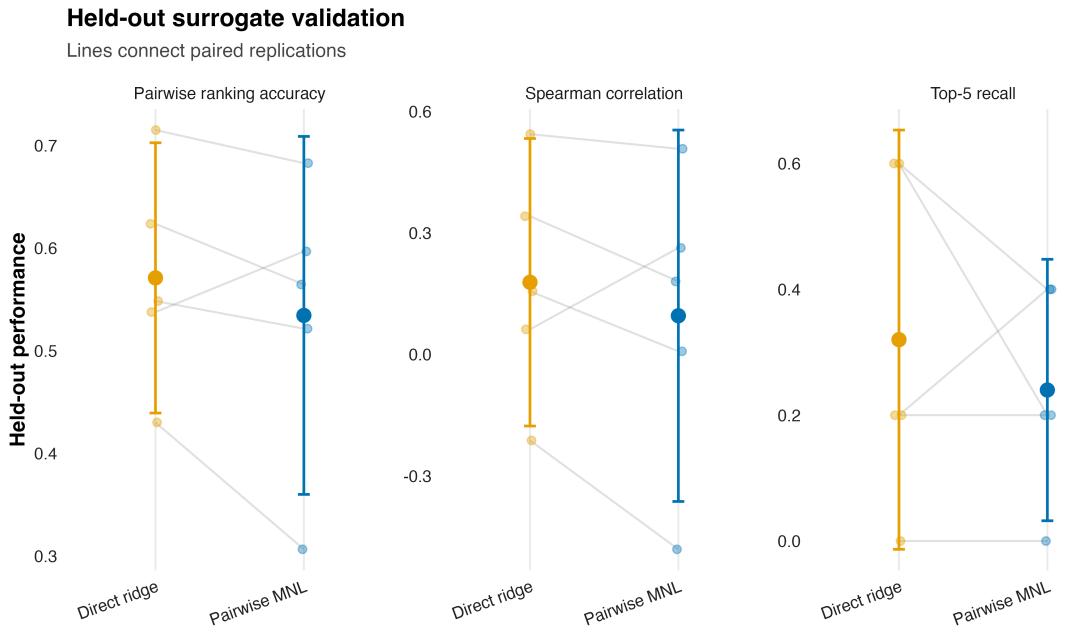

Figure 13: Held-out surrogate comparison in the representative IMDb/Llama ablation. The figure illustrates the real-data reversal relative to the simulation study: direct-score ridge is at least as competitive as, and in the aggregate live study ranks unseen configurations better than, the additive MNL surrogate.

Figure 13 should be read together with the six-setting held-out statistics reported in the main text. The representative plot is not used to establish a universal difference by itself; rather, it illustrates the broader empirical finding that direct ridge obtains higher mean held-out rank correlation and pairwise ranking accuracy across all six live settings. This negative result motivates richer interaction-aware or mixed-logit preference surrogates.

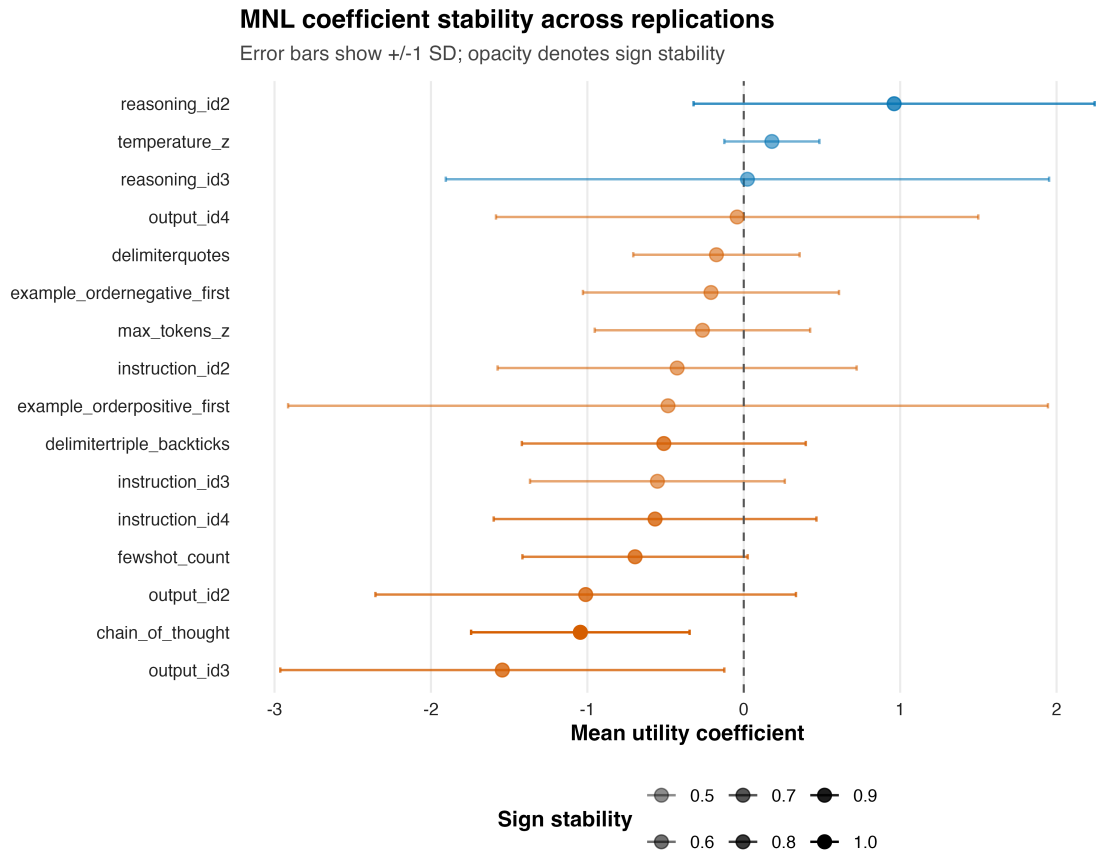

Figure 14: Stability of fitted MNL component coefficients across the five replications of the representative IMDb/Llama experiment. The figure is included as a diagnostic of structural interpretability rather than as a causal analysis of prompt components.

Figure 14 illustrates the interpretability advantage and its limitation. MNL coefficients remain attached to explicit, controllable prompt components, making their direction and stability easy to inspect. At the same time, several effects vary across replications, and the live held-out results show that coefficient interpretability should not be confused with globally superior predictive accuracy. We therefore treat these coefficients as task- and model-specific diagnostic summaries rather than general prompt-engineering rules.

References

Agrawal, Lakshya A., Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab (2025). *GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning*. arXiv: 2507.19457.

Brochu, Eric, Vlad M. Cora, and Nando de Freitas (2010). *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. arXiv: 1012.2599.

Chen, Lichang, Jiuhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou (2023). *InstructZero: Efficient Instruction Optimization for Black-Box Large Language Models*. arXiv: 2306.03082.

Chen, Minping, Bowen Xiao, Du Liang, Chuxuan Zeng, and Zeyi Wen (2026). "Efficient Hyperparameter Optimization for LLM Reinforcement Learning". In: *Proceedings of the 64th Annual Meeting*

of the Association for Computational Linguistics, pp. 27540–27552. DOI: 10.18653/v1/2026.acl-long.1271.

Deshwal, Aryan, Syrine Belakaria, and Janardhan Rao Doppa (2022). "Bayesian Optimization over Hybrid Spaces". In: *Proceedings of ICML*.

Fernando, Chrisantha, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel (2023). *Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution*. arXiv: 2309.16797.

Frazier, Peter I. (2018). "A Tutorial on Bayesian Optimization". In: *arXiv preprint arXiv:1807.02811*.

González, Javier, Zhenwen Dai, Andreas Damianou, and Neil D. Lawrence (2017). "Preferential Bayesian Optimization". In: *Proceedings of ICML*.

Guo, Qingyan, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang (2024). "Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers". In: *International Conference on Learning Representations*.

Hutter, Frank, Holger H. Hoos, and Kevin Leyton-Brown (2011). "Sequential Model-Based Optimization for General Algorithm Configuration". In: *Learning and Intelligent Optimization*.

Kandasamy, Kirthevasan, Gautam Dasarathy, Junier B. Oliva, Jeff Schneider, and Barnabás Póczos (2017). "Multi-fidelity Bayesian Optimisation with Continuous Approximations". In: *Proceedings of ICML*.

Khattab, Omar, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Haq, Ashutosh Sharma, Martin Josifoski, Carlos Guestrin, and Matei Zaharia (2024). "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines". In: *International Conference on Learning Representations*.

Kim, Hee-Soo, Jun-Young Kim, Jeong-Hwan Lee, Seong-Jin Park, and Kang-Min Kim (2026). "Don't Generate, Classify! Low-Latency Prompt Optimization with Structured Complementary Prompt". In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 4364–4383. DOI: 10.18653/v1/2026.eacl-long.204.

Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar (2018). "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization". In: *Journal of Machine Learning Research* 18.185, pp. 1–52.

Li, Mingqi, Karan Aggarwal, Yong Xie, Aitzaz Ahmad, and Stephen Lau (2026). "Learning from Contrastive Prompts: An Automated Prompt Optimization Framework". In: *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 163–188. DOI: 10.18653/v1/2026.findings-acl.9.

Liu, Haoyue, Zhichao Wang, Yongxin Guo, Haoran Shou, and Xiaoying Tang (2026). "Adaptive Prompt Structure Factorization: A Framework for Self-Discovering and Optimizing Compositional Prompt Programs". In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 11690–11714. DOI: 10.18653/v1/2026.acl-long.537.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts (2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 142–150.

Meta (2024). *Llama 3.2 1B Instruct Model Card*. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct. Model released 25 September 2024.

Ollama (2026). *Ollama API Documentation*. https://docs.ollama.com/api/generate. Accessed 9 August 2026.

Opsahl-Ong, Krista, Michael J. Ryan, Josh Purtell, Matei Zaharia, and Omar Khattab (2024). "Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs". In: *Proceedings of EMNLP*.

Pryzant, Reid, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng (2023). "Automatic Prompt Optimization with "Gradient Descent" and Beam Search". In: *Proceedings of EMNLP*.

Qwen Team (2024). *Qwen2.5 Technical Report*. arXiv: 2412.15115.

Ru, Binxin, Ahsan Alvi, Vu Nguyen, Michael A. Osborne, and Stephen J. Roberts (2020). "Bayesian Optimisation over Multiple Continuous and Categorical Inputs". In: *Proceedings of ICML*.

Sabbatella, Antonio, Andrea Ponti, Ilaria Giordani, Antonio Candelieri, and Francesco Archetti (2024). "Prompt Optimization in Large Language Models". In: *Mathematics* 12.6, p. 929. DOI: 10.3390/math12060929.

Saeed, Muhammad Amir and Antonio Candelieri (2026). "Bayesian Optimization for Categorical and Mixed Variables Using a Multinomial Logit Surrogate". In: *Algorithms* 19.5, p. 361. DOI: 10.3390/a19050361.

Schneider, Lennart, Martin Wistuba, Aaron Klein, Jacek Golebiowski, Giovanni Zappella, and Felice Antonio Merra (2024). "Hyperband-based Bayesian Optimization for Black-box Prompt Selection". In: *arXiv preprint arXiv:2412.07820*.

Singhal, Rahul, Pradyumna Tambwekar, and Karime Maamari (2026). *PrefPO: Pairwise Preference Prompt Optimization*. arXiv: 2603.19311.

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts (2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of EMNLP*, pp. 1631–1642.

Tomar, Shlok, Aryan Deshwal, Ethan Villalovoz, Mattia Fazzini, Haipeng Cai, and Janardhan Rao Doppa (2025). *An Exploratory Study of Bayesian Prompt Optimization for Test-Driven Code Generation with Large Language Models*. arXiv: 2512.15076.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge University Press.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*.

Wu, Jian, Saul Toscano-Palmerin, Peter I. Frazier, and Andrew Gordon Wilson (2020). "Practical Multi-fidelity Bayesian Optimization for Hyperparameter Tuning". In: *Proceedings of UAI*.

Wu, Yuanchen, Saurabh Verma, Justin Lee, Fangzhou Xiong, Poppy Zhang, Amel Awadelkarim, Xu Chen, Yubai Yuan, and Shawndra Hill (2026). "LLM Prompt Duel Optimizer: Efficient Label-Free Prompt Optimization". In: *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10066–10089. DOI: 10.18653/v1/2026.findings-acl.490.

Xu, Yuebin, Zhiyi Chen, and Zeyi Wen (2025). "EcoTune: Token-Efficient Multi-Fidelity Hyperparameter Optimization for Large Language Model Inference". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7735–7745. DOI: 10.18653/v1/2025.emnlp-main.394.

Yang, Chengrun, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen (2024). "Large Language Models as Optimizers". In: *International Conference on Learning Representations*.

Yuksekgonul, Mert, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou (2024). *TextGrad: Automatic "Differentiation" via Text*. arXiv: 2406.07496.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). "Character-level Convolutional Networks for Text Classification". In: *Advances in Neural Information Processing Systems*.

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba (2023). "Large Language Models Are Human-Level Prompt Engineers". In: *International Conference on Learning Representations*.

Zhu, Zixiao, Hanzhang Zhou, Zijian Feng, Tianjiao Li, Chua Jia Jim Deryl, Lee Onn Mak, Gee Wah Ng, and Kezhi Mao (2026). "Rethinking Prompt Optimizers: From Prompt Merits to

Optimization". In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 864–892. DOI: 10.18653/v1/2026.eacl-long.38.