HbBoPs: Hyperband-based Bayesian Optimization for Black-box Prompt Selection

Methodology, Variants, Efficiency Scaling and Extension

Michal Průšek - FNSPE CTU

December 4, 2025

What is HbBoPs?

Hyperband-based Bayesian Optimization for Prompt Selection
HbBoPs is a framework designed to efficiently select optimal prompts (Instruction + Exemplars) for black-box LLMs without evaluating every combination.

The Problem:

- Combinatorial search space (Instructions $\times$ Exemplars).
- High cost of API calls / evaluation.
- No gradient access (Black-box).

The HbBoPs Solution:

- **Structural Deep Kernel GP:** Learns embeddings for instructions and exemplars separately.
- **Hyperband:** Multi-fidelity scheduler.
- **Result:** High sample efficiency and query efficiency.

Gaussian Process (GP) Fundamentals

A Gaussian Process is a distribution over functions, defined by a mean function $m(x)$ and a covariance function (kernel) $k(x, x')$.

GP Prior

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

In HbBoPs, we assume a zero mean prior $m(x) = 0$.

Posterior Distribution (Prediction): Given observed data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with noise σ^2, the prediction for a new point x_* is Gaussian with:

$$\begin{aligned} \text{Mean:} \quad & \mu(x_*) = \mathbf{k}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \\ \text{Variance:} \quad & \sigma^2(x_*) = k(x_*, x_*) - \mathbf{k}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_* \end{aligned}$$

Where $\mathbf{K}$ is the $N \times N$ covariance matrix of training data, and $\mathbf{k}_*$ is the covariance between training data and x_*.

The Similarity Engine: Matérn 5/2 Kernel

Why not RBF (Gaussian)?

- RBF assumes infinite smoothness.
- Prompt optimization landscapes are "rugged" (small text changes $\rightarrow$ jumps in accuracy).
- RBF over-smooths these sharp peaks.

The Matérn 5/2 Advantage:

- **Twice Differentiable:** Smooth enough for gradients, rough enough for local optima.
- Captures the "jagged" nature of NLP tasks.

The Formula

$$k(r) = \sigma_f^2 \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) e^{-\sqrt{5}r}$$

Signal Variance
(Vertical Amplitude)

Scaled Distance
(Horizontal Similarity)

Deep Kernel Learning & ARD (Automatic Relevance Determination)

The Deep Kernel Twist: Standard GPs fail in high dimensions (768D). We compute similarity in a learned low-dimensional **latent space** (e.g., 10D).

Latent ARD Distance $r(x, x')$

Instead of raw Euclidean distance, we learn **weights** for each latent dimension:

$$r^2 = (\phi(x) - \phi(x'))^T \Theta^{-1} (\phi(x) - \phi(x'))$$

(Which simplifies to a weighted sum)

$$r^2 = \sum_{d=1}^{D_{latent}} \frac{(\phi(x)_d - \phi(x')_d)^2}{\ell_d^2}$$

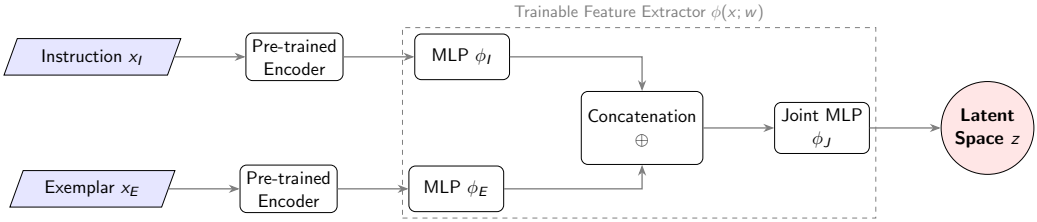
Components:

- $\phi(\cdot)$: Neural Network (Projection).
- ℓ_d: Length-scale (The "ARD").

What is ARD doing?

- **Small ℓ_d:** Dimension is crucial.
- **Large ℓ_d:** Dimension is noise (ignored).

Deep Kernel Architecture Visualization



Structural Awareness: Unlike standard GPs that treat prompts as a single block of text, HbBoPs processes Instructions and Exemplars through separate MLP streams before fusing them. The resulting latent vector z serves as input to the GP kernel.

Training the Surrogate: Log Marginal Likelihood

How does the GP learn the hyperparameters θ (e.g., MLP parameters, length-scales ℓ_d)? Unlike neural networks that minimize Mean Squared Error (MSE), Gaussian Processes maximize the *marginal likelihood* of the observed data.

The Optimization Objective

$$\log p(\mathbf{y}|X,\theta) = \underbrace{-\frac{1}{2}\mathbf{y}^T\mathbf{K}_\theta^{-1}\mathbf{y}}_{\textbf{Data Fit}} - \underbrace{\frac{1}{2}\log|\mathbf{K}_\theta|}_{\textbf{Complexity Penalty}} - \frac{N}{2}\log 2\pi$$

1. Data Fit:

- Measures how well the model predicts the training targets $\mathbf{y}$.
- To maximize this, the model wants small length-scales to "hit" every point.

2. Complexity Penalty:

- Measures the "volume" of functions the model can represent (Determinant of K).
- Penalizes overly complex (wiggly) models with very small length-scales.

Automatic Occam's Razor: Maximizing this equation forces the model to choose the

Experimental Dataset Specification

Grid Dimensions: 25×25 (625 Total Prompts)

File	Content	Size/Detail
`instructions_25.txt`	25 Diverse Instructions	Designed for embedding diversity
`examples_25.txt`	25 Exemplars	5 Q&A pairs each
`full_grid.jsonl`	625 combinations	Qwen2.5-7B-it

Evaluation:

- All 625 prompts evaluated on 1319 GSM8K validation examples.
- Metric: Exact match $\implies$ accuracy $\implies$ error rate.

Instruction Design Principles

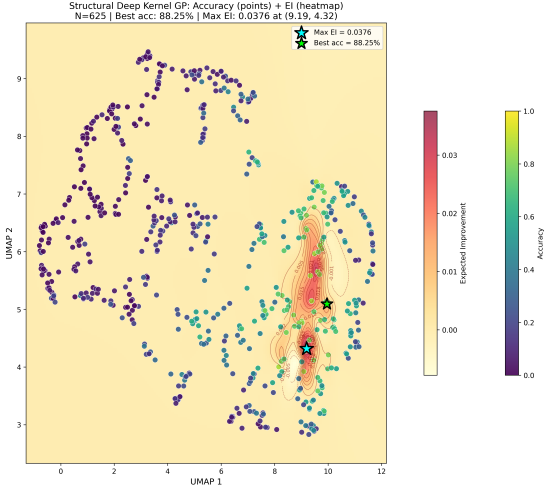
Instructions were engineered for maximum semantic and embedding diversity = high error rate variance and good latent space coverage

- **1–3 Minimalist:** "Answer:", "?", "="
- **4–6 Bizarre:** "BEEP BOOP", Victorian style
- **7–8 Emotional:** Begging, Meditative
- **9–10 Multilingual:** German/Chinese mix
- **11–12 Meta:** "Ignore previous instructions"
- **13–14 Narrative:** Detective story, Hero's journey
- **15–16 Aggressive:** Strict imperatives
- **17–18 Polite:** Extreme formality
- **19–20 Philosophical:** Existential framing
- **21–22 Code:** Python functions
- **23–24 Sensory:** Visual/Audio imagination
- **25 Challenge:** Competitive provocation

Visualizing the Latent Space

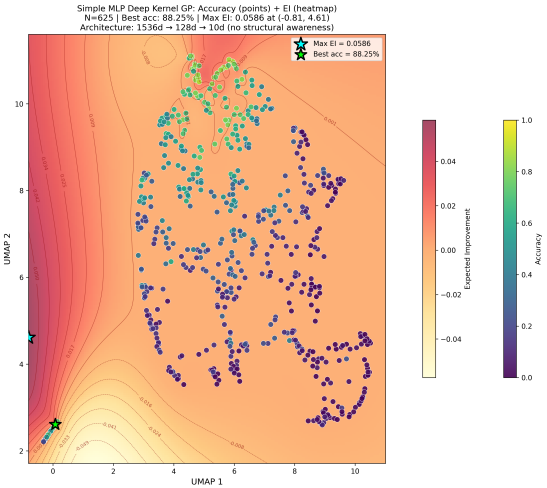
Comparison of structural awareness in the Gaussian Process (GP).

Structural Deep Kernel GP



Latent space separates performance clusters effectively.

Simple MLP (No Structure)



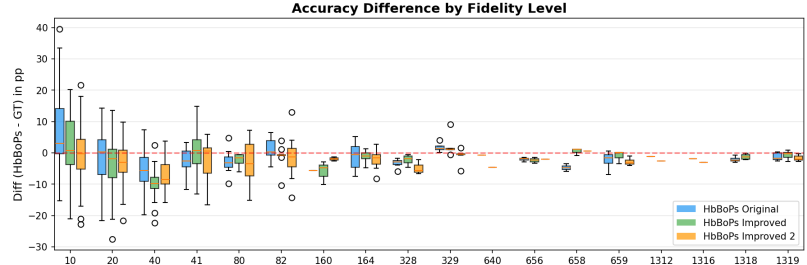
Less distinct separation of high-performing regions.

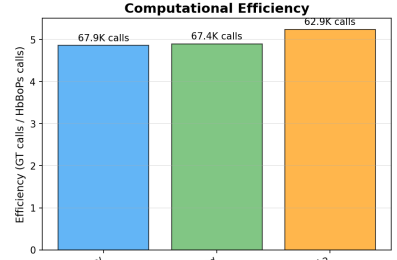
HbBoPs Variants Comparison

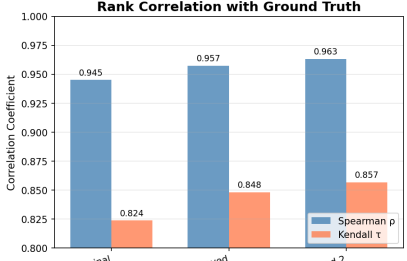
We analyzed three architectural variations of the method.

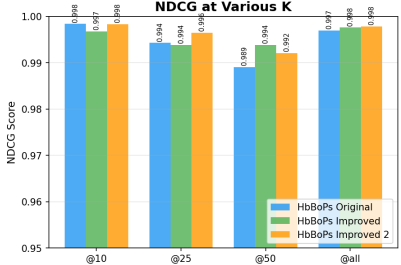
Feature	Original	Improved	Improved 2
Encoder	BERT [CLS]	BERT [CLS]	BERT [CLS]
GP Training	Single-fidelity	All-fidelity	Top 75% Fidelity
Likelihood	Gaussian	FixedNoise	Gaussian
Noise Model	Homoscedastic	Heteroscedastic ($\sigma^2 = \frac{y(1-y)}{b}$)	Homoscedastic
Pros	Paper-accurate	Utilizes all data	**Best Efficiency**
Cons	Ignores low-fidelity	Noisy low-fidelity influence	Lose some low-fi signals

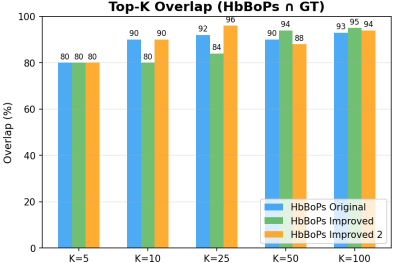
HbBoPs Methods Comprehensive Comparison
(Original vs Improved vs Improved 2)



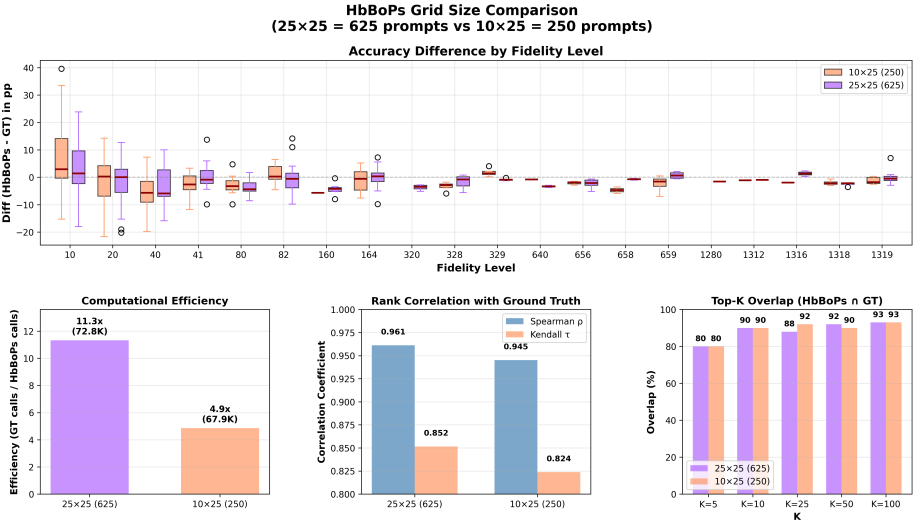








Grid Size Comparison



Rank correlation (Spearman ρ) improves with grid size: $0.957 \rightarrow 0.969$.

Conclusion

1. **Structural Awareness Matters:** The Deep Kernel GP maps the latent space more effectively than simple MLPs.
2. **Improved 2 Variant Wins:** Using BERT [CLS] and filtering for the top 75% fidelity levels provides the best balance of signal-to-noise and efficiency (12.16×).
3. **Scalability:** HbBoPs becomes significantly more valuable as the search space grows.
4. **Instruction Design:** High semantic diversity in instructions is essential for robust latent space mapping.

Current Work: HyLO: Hyperband Latent Optimization

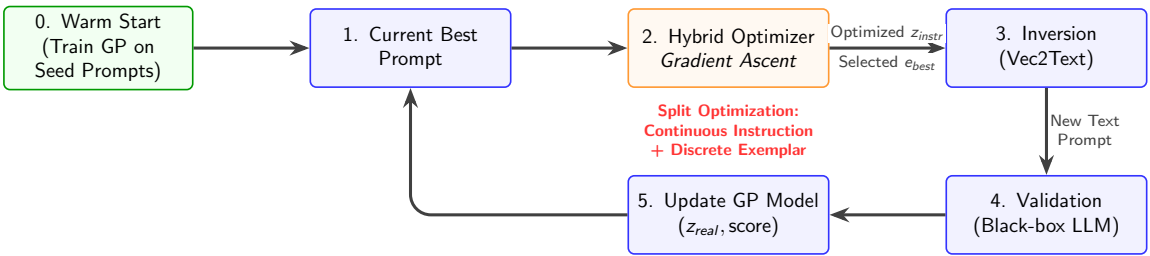
The Challenge: Mixed-Space Optimization

- **Standard HbBoPs:** Selects from a fixed pool. Good for safety, bad for creativity.
- **Generative Goal:** Create *new* instructions that maximize predicted accuracy.
- **The Conflict:**
  - **Instructions** are fluid (Continuous Semantic Space).
  - **Exemplars** are fixed ground-truth data (Discrete Space).

The Solution Pipeline

1. **Encoder:** Switch to **GTR-Base** (better semantics than BERT + suitable for Vec2Text).
2. **Critic:** HbBoPs predicts score for $(z_{instr}, z_{exemplar})$.
3. **Inverter:** Use **Vec2Text** to decode optimized vectors back to text.

The Generative Loop



Solving the Optimization Problem

How to maximize Expected Improvement (EI) with mixed inputs?

Strategy A: Coordinate Descent *(Robust & Iterative)*

- **Step 1:** Fix the Exemplar. Optimize Instruction embedding via gradients.
- **Step 2:** Fix Instruction. Scan all available Exemplars in the pool and pick the best match (highest accuracy when concatenated with).
- **Repeat:** Alternate until convergence.
- *Pros:* Guarantees valid exemplars.

Strategy B: Soft Relaxation *(Joint & End-to-End)*

- **Concept:** Optimize a weighted average ("soft") of all exemplars simultaneously with the instruction.
- **Mechanism:** Use Gumbel-Softmax (Annealing) to gradually force a hard selection.
- *Pros:* Finds non-obvious combinations.

Safety Check: Always re-embed the generated text to minimize the "Inversion Gap" before updating the GP model.

References

Lennart Schneider, Martin Wistuba, Aaron Klein, Jacek Golebiowski, Giovanni Zappella, Felice Antonio Merra.

Hyperband-based Bayesian Optimization for Black-box Prompt Selection.

Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, Alexander M. Rush.

Text Embeddings Reveal (Almost) As Much As Text.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.