TextBO: Bayesian Optimization in Language Space for Eval-Efficient Self-Improving AI

Enoch Hyunwook Kang*
University of Washington

Hema Yoganarasimhan
University of Washington

Abstract

Large Language Models (LLMs) have enabled self-improving AI systems that iteratively generate, evaluate, and refine their outcomes. Recent studies show that prompt-optimization-based self-improvement can outperform state-of-the-art reinforcement-learning fine-tuning of LLMs, but performance is typically measured by *generation* efficiency. However, in many applications, the constraint is *evaluation* efficiency: obtaining reliable feedback is far more costly than generating candidates. In this paper, we propose TextBO, a self-improving algorithm that achieves evaluation-efficiency by provably emulating gradient-based UCB-BO in language space. We empirically validate TextBO on automated ad-alignment tasks agentic AI tasks, demonstrating superior performance per evaluation compared to GEPA. We also evaluate TextBO's Best-of-N multi-step textual-gradient mechanism on agentic AI benchmarks by augmenting GEPA with it and show that it significantly outperforms standard GEPA. For the full paper access, refer to https://arxiv.org/abs/2511.12063.

1 Introduction

1.1 Self-improving AI for Data-driven Decision-making

Data-centric decision making has long relied on an iterative, human-driven cycle of proposing new solutions, testing them, and learning from the results (left panel of Figure 1). Here, the bottleneck has been the time required to propose new solutions, which took from days to weeks.

Recent advances in generative AI have begun to transform this paradigm. Candidate generation is now fast and scalable: for example, AI models can produce high-quality ads in minutes (Jansen et al. 2024; Hartmann et al. 2025). In parallel, LLM-based systems can autonomously conduct analyses by summarizing performance patterns and hypothesizing drivers of success (Ferber et al. 2024; Guo et al. 2024b; Ghosh et al. 2024; Wiedemer et al. 2025). These advances enabled an automated, continuously updating data-centric decision making systems what we refer to as *self-improving* or *self-evolving* AI (Mantia et al. 2025; Chen et al. 2025; Silver & Sutton 2025) (right panel of Figure 1). Here, the time required to propose new solutions is no longer the bottleneck; the time required to evaluate solutions (e.g., using A/B testing or bandit experiments) is the new bottleneck.

Figure 1: Comparison of human-driven and AI-driven self-improvement cycle.

*Please ask all correspondance to: ehwkang@uw.edu and hemay@uw.edu

For example, AI models like AlphaFold can propose protein structures quickly (Jumper et al. 2021), but their functional properties must still be tested through slow and noisy wet-lab experiments; similarly, AlphaChip can generate chip layouts in silico (Goldie et al. 2024), but performance ultimately depends on fabrication and benchmarking in real-world hardware systems, which often takes weeks.

In sum, the primary goal for self-improving AI systems in scientific and societal settings should be *evaluation efficiency*, i.e., reaching high-performing solutions using as few costly evaluations as possible. However, the central line of progress in self-improving AI, which is based on *iterative prompt optimization*[1], has largely focused on *generation efficiency*—the number of LLM calls (or candidate solution generations) needed to reach a target performance level (Nguyen et al. 2016; Ding et al. 2023; Hu et al. 2024; Liu et al. 2024b; Zhang et al. 2025; Wang et al. 2024; Agrawal et al. 2025).

In this paper, we propose TextBO, an *evaluation-efficient*, theoretically grounded, and empirically effective framework for prompt-optimization-based self-improving AI. Our key and novel insight is that iterative prompt-optimization-based self-improving AI can be provably cast as *Bayesian Optimization (BO)* in the language space. This recasting then allows us to inherit the key properties of BO, in particular, its optimal evaluation efficiency (Wang et al. 2023; Whitehouse et al. 2023).

In addition to theoretically proving that TextBO emulates BO in language space, we empirically demonstrate the performance of TextBO through two experiments. In the ad optimization experiments, we show that TextBO significantly outperforms Best-of-N (Snell et al. 2024) and GEPA (Agrawal et al. 2025) across all eight scenarios. Our numerical experiments show that TextBO can reach good performance both in settings where there is significant heterogeneity in the preferences of the target population as well as in settings with homogeneous preferences. We also conduct a series of agentic AI benchmark experiments to assess whether the same Best-of-N multi-step update principle improves evaluation efficiency. Specifically, we consider the experiments used to evaluate GEPA in Agrawal et al. (2025) – HotpotQA, HoVer, and PUPA. We then compare the performance of TextBO-GEPA (GEPA augmented with Best-of-N gradient sampling and multiple gradient steps) to GEPA. Across all experiments, we show that TextBO-GEPA consistently outperforms GEPA.

2 Problem Definition

Let Φ denote a generative AI model (or more generally, an AI system) that, given a prompt $\pi \in \Pi$, produces a candidate solution $\Phi(\pi)$.[2] This candidate solution $\Phi(\pi)$, in conjunction with some environmental variables $x \in \mathcal{X}$ produces an output $y(\Phi(\pi), x)$, where y is drawn from a distribution over an output space $\mathcal{Y}$.

Throughout, we allow for the distribution over $\mathcal{X}$ to depend on the prompt π. This is important for practical applications where such dependencies are common. To capture such dependencies, we denote the input distribution over $\mathcal{X}$ by $\mathcal{D}_{\mathcal{X}}(\pi)$. In a fully randomized test, the distribution over $\mathcal{X}$ is independent of π, and we can denote the distribution of $\mathcal{X}$ as $\mathcal{D}_{\mathcal{X}}$.

To evaluate the output, we use a scoring rule: a function $r : \mathcal{Y} \to [0, 1]$ that assigns a scalar effectiveness score to each $y \in \mathcal{Y}$. In the digital marketing example, r is a marketer-defined function that maps observed user responses (clicks, conversions, spend) to a normalized scalar effectiveness score. Then, our optimization objective is to discover the prompt π^* that maximizes the objective function:

$$J(\pi) := \mathbb{E}[r(y(\Phi(\pi), x))]. \tag{1}$$

Self-improving AI algorithm: Given a scoring rule r, a *self-improving AI algorithm* $\mathcal{A}$ optimizes $J(\pi) := \mathbb{E}[r(y(\Phi(\pi), x))]$ by iteratively analyzing the previous history, proposing a new prompt, and observing its noisy score. Formally, starting from the initial prompt π_0, a self-improving AI algorithm proceeds in discrete optimization iterations, indexed by $t = 1, 2, \ldots$; at the beginning of iteration t, the algorithm has access to the history of past prompts and their observed scores, then selects a new prompt π_t, deploys the corresponding candidate solution $\Phi(\pi_t)$, and receives a score feedback on its

[1] Recent papers show that such approaches often significantly outperform reinforcement-learning-based fine-tuning (Agrawal et al. 2025)

[2] LLM-based AI systems can also be modified by tuning the LLM's internal parameters using LLM fine-tuning methods such as GRPO (Liu et al. 2024a). In this paper, we focus on prompt optimization, which fixes those internal parameters throughout.

performance. We define the score s_t of the prompt π_t as the empirical average reward at t, i.e.,

$$s_t := \frac{1}{L}\sum_{l=1}^{L} r\left(y\left(\Phi\left(\pi_t\right), x_t^{(l)}\right)\right), \quad x_t^{(l)} \sim \mathcal{D}_{\mathcal{X}}\left(\pi_t\right), \tag{2}$$

where L denotes the number of input samples $x_t^{(l)}$ we evaluate for each t. Note that we compare the performance of self-improving algorithms in terms of the progress of s_t across iterations t.

Definition 1 (Self-improving AI algorithm)**.** *Fix a generative model Φ and a scoring rule r. For each iteration $t \geq 1$, let the history of past evaluations be $H_{t-1} := \{(\pi_\tau, \Phi(\pi_\tau), s_\tau)\}_{\tau=0}^{t-1}$, where s_τ is the empirically evaluated score of π_τ, defined as*

$$s_t = \frac{1}{L}\sum_{l=1}^{L} r\big(y(\Phi(\pi_t), x_t^{(l)})\big), \quad x_t^{(l)} \sim \mathcal{D}_{\mathcal{X}}(\pi_t). \tag{3}$$

Then, a self-improving AI algorithm is defined as function $\mathcal{A}$ that, at each iteration t, maps the history H_{t-1} to a new prompt π_t, i.e.,

$$\pi_t = \mathcal{A}_t(H_{t-1}).$$

Operationally, given access to the history H_{t-1}, a self-improving AI algorithm $\mathcal{A}$ goes through following three steps each iteration t:

1. *Analysis:* Analyze the previous history H_{t-1}.
2. *Generation:* Propose the next prompts π_t based on the analysis.
3. *Evaluation:* The system outcome $\Phi(\pi_t)$ is deployed, and its performance is measured, yielding a score s_t. The new observation is added to the history: $H_t = H_{t-1} \cup \{(\pi_t, \Phi\left(\pi_t\right), s_t)\}$.

Optimization objective: In this paper, we formulate the goal of self-improving AI algorithms $\mathcal{A}$ as improving prompts $(\pi_t)_{t=1}^{T}$ over the evaluation budget T to efficiently optimize the objective $J(\pi)$. Specifically, we discuss evaluation efficiency using two standard regret criteria. Let $\pi^\star \in \arg\max_{\pi\in\Pi} J(\pi)$. Then,

- The (expected) cumulative regret is $\mathbb{E}\left[\sum_{t=1}^{T}\left(J\left(\pi^\star\right) - J\left(\pi_t\right)\right)\right]$.
- The (expected) simple regret is $\mathbb{E}\left[J\left(\pi^\star\right) - \max_{1\leq t\leq T} J\left(\pi_t\right)\right]$.

In the next section, we show how a self-improving AI algorithm can be viewed as UCB-based Bayesian Optimization, which is optimal under both regret criteria (Whitehouse et al. 2023; Wang et al. 2023).

3 Prompt Optimization as Bayesian Optimization in Language Space

Canonical UCB Bayesian Optimization. To fix ideas, we start with a quick overview of BO. Classical BO considers a black-box function $f : \mathcal{Z} \to \mathbb{R}$ defined over a continuous input space $\mathcal{Z}$. The algorithm maintains a surrogate model—typically a Gaussian Process or related probabilistic regressor—that, given a history of evaluations $H_{t-1} = \{(z_\tau, y_\tau)\}_{\tau=1}^{t-1}$, produces for each $z \in \mathcal{Z}$ a posterior mean $\mu_{t-1}(z)$ and an uncertainty estimate $\sigma_{t-1}(z)$. Rather than maximizing $\mu_{t-1}(z)$ directly, BO selects the next query by maximizing an *acquisition function* that trades off exploitation (high predicted value) and exploration (high uncertainty). A widely used acquisition is the Upper Confidence Bound (UCB)

$$A_{\beta_t}(z) \;=\; \mu_{t-1}(z) \;+\; \beta_t\, \sigma_{t-1}(z), \tag{4}$$

where $\beta_t > 0$ controls the strength of exploration (Srinivas et al. 2012). Intuitively, $A_{\beta_t}(z)$ implements an "optimism in the face of uncertainty" rule: points with large uncertainty receive a positive bonus even if their current mean is modest. Under mild conditions, UCB-BO achieves sublinear regret and enjoys strong evaluation-efficiency guarantees: no algorithm can perform uniformly better in terms of regret scaling (Srinivas et al. 2012; Whitehouse et al. 2023; Wang et al. 2023).

In many practical settings, the global maximizer of A_{β_t} is not available in closed form, and the acquisition is highly non-convex. A standard remedy is *parallel gradient-based BO* (Wilson et al. 2018; Papenmeier et al. 2025): one starts from multiple initial points, performs gradient ascent on A_{β_t} from each initialization to find local maxima, and then evaluates f at these local optima in parallel. This procedure has shown surprising empirical effectiveness in high-dimensional numerical optimization (Papenmeier et al. 2025).

Mapping self-improving AI to UCB-BO. We now conceptually align our iterative prompt-optimization problem with the BO template above. The correspondence is:

- The BO input space $\mathcal{Z}$ corresponds to the prompt space Π.
- The black-box function $f(z)$ corresponds to the system-level objective $f(\pi) \equiv J(\pi) := \mathbb{E}[r(y(\Phi(\pi), x))]$, where $x \sim \mathcal{D}_{\mathcal{X}}(\pi)$ and r is the scoring rule defined in § 2.
- The evaluation of $f(\pi)$ is the costly step: it requires deploying $\Phi(\pi)$, interacting with the environment (e.g., ad platform users), and computing s_t as an empirical estimate of $J(\pi)$.
- The cheap computation is everything the LLM-based system can do *without* new evaluations: analyzing past history, proposing textual edits, or internally comparing candidate ads.

If we had an explicit surrogate model over prompts, providing $\mu_t(\pi)$ and $\sigma_t(\pi)$ for each $\pi \in \Pi$, we could in principle construct a UCB acquisition

$$A_{\beta_t}(\pi) \;=\; \mu_t(\pi) \;+\; \beta_t\, \sigma_t(\pi), \tag{5}$$

and then run parallel gradient-based BO over Π: start from multiple initial prompts, ascend the gradient $\nabla_\pi A_{\beta_t}(\pi)$ to obtain local optima, and evaluate $J(\pi)$ only at those local optima. This would give us a self-improving AI algorithm that is, by construction, evaluation-efficient.

4 BEST-OF-N GRADIENT AND THEORY

We show that BEST-OF-N gradient, the combination of (i) *textual gradients* and (ii) *BEST-OF-N* selection, induces gradient ascent on a UCB acquisition function in the LLM's implicit embedding space.

4.1 Critic LLM

We treat the *critic LLM* as a natural, practically realistic component that provides the basic operations required to implement Bayesian-optimization-style updates in the language space. The use of a separate LLM as a critic is now a standard pattern in the prompt-optimization and self-improving AI literature: critic models are routinely used to provide structured feedback, propose improvements, or act as reward models and judges that guide search over prompts and outputs (Zhou et al. 2022; Yuksekgonul et al. 2025; Agrawal et al. 2025; Zhang et al. 2025; Wang et al. 2025; Shi et al. 2024).

Recall that we denote a prompt as π and the corresponding candidate solution as $\Phi(\pi)$, the optimization history up to iteration t as $H_t = \{(\pi_\tau, \Phi(\pi_\tau), s_\tau)\}_{\tau \le t}$, where s_τ is the evaluated score of π_τ. Then, let M_{critic} denote the critic LLM, which can perform the following four operations:

1. *Pairwise Judge Operator.* Given two prompts $\pi^{(1)}$ and $\pi^{(2)}$, and their corresponding outcomes $\Phi(\pi^{(1)})$ and $\Phi(\pi^{(2)})$, predict which is better given reflection R:

$$\textsc{PairwiseJudge}(\Phi(\pi^{(1)}), \Phi(\pi^{(2)}); R, M_{\text{critic}}) \in \{\Phi(\pi^{(1)}), \Phi(\pi^{(2)})\}. \tag{6}$$

2. *Meta-reflection Operator.* Compress the entire history into a small set of natural-language rules or guidelines:

$$R_t = \textsc{MetaReflect}(H_t; M_{\text{critic}}). \tag{7}$$

3. *Textual Gradient Operator.* Stochastically generate a local edit δ intended to improve π under reflection R:

$$\delta \sim \nabla_{\text{text}}(\pi; R, M_{\text{critic}}). \tag{8}$$

4. *Apply Operator.* Apply an edit δ to obtain a new prompt:

$$\pi' = \text{Apply}(\pi, \delta; M_{\text{critic}}). \tag{9}$$

4.2 BEST-OF-N GRADIENT

Formally, given a current prompt π and reflection R, a single BEST-OF-N gradient step proceeds as follows:

1) *Textual gradient generation.* Use M_{critic} to generate N independent textual edits:

$$\delta^{(i)} \sim \nabla_{\text{text}}(\pi; R, M_{\text{critic}}), \qquad i = 1, \ldots, N. \tag{10}$$

2) *Candidate prompt generation.* Apply each edit to produce N candidate prompts:

$$\pi^{(i)} = \mathrm{Apply}(\pi, \delta^{(i)}; M_{\text{critic}}). \tag{11}$$

3) *Candidate system outcome generation.* Produce the associated system outcomes:

$$\Phi(\pi^{(i)}), \qquad i = 1, \ldots, N. \tag{12}$$

4) *BEST-OF-N selection.* Identify the most promising candidate using the critic:

$$i^\star = \textsc{Best-of-N}\big(\{(\pi^{(i)}, \Phi(\pi^{(i)})\}_{i=1}^N; R, M_{\text{critic}}\big). \tag{13}$$

The resulting prompt $\pi^{(i^\star)}$ defines the update made by the BEST-OF-N gradient for a given step. Intuitively, each textual edit represents a noisy proposal for a local improvement direction; BEST-OF-N filters these proposals through an optimism-based rule, selecting the direction most likely to improve performance under the critic LLM's current understanding.

4.3 Theory: BEST-OF-N gradient's Asymptotic Equivalence to UCB gradient

Let Π be the set of syntactically valid prompts. Suppose that there exists a text embedding $\mathsf{E} : \Pi \to \mathbb{R}^d$, where $e := \mathsf{E}(\pi) \in \mathbb{R}^d$ is the embedding of a prompt $\pi \in \Pi$. We first state the assumptions motivated by how language model embeddings behave in practice, and then state the main theorem, Theorem 2. At each gradient step of TEXTBO, we sample edits that correspond to $u_1, \ldots, u_N \overset{\text{i.i.d.}}{\sim} \rho$ on the embedding space $\subset \mathbb{R}^d$, where ρ is assumed to have a density bounded below by a positive constant with respect to surface measure.

Lemma 1 (Gaussian maxima and spacing (Vershynin 2018))**.** *Let $\xi_1, \ldots, \xi_N$ be i.i.d. mean-zero, unit-variance Gaussian, $M_N = \max_{i \le N} \xi_i$, S_N the second largest, and $q_N := F^{-1}(1 - 1/N)$ the $(1 - 1/N)$-quantile of ξ_1. Then $q_N = \Theta(\sqrt{\ln N})$, $M_N - q_N \to 0$ in probability, and $M_N - S_N = O_p(1/q_N)$.*

Theorem 1 (BEST-OF-N asymptotically induces a UCB gradient direction)**.** *Under mild assumptions 1-4, define $\beta_N := q_N$ from Lemma 2. Let $i^\star \in \arg\max_i Y_i$ and set $\widehat{u}_N := u_{i^\star}$. Define the step*

$$\Delta e^{(N)} := \mathsf{E}\big(\mathrm{Apply}(\pi, \varepsilon \widehat{u}_N)\big) - e = \varepsilon \widehat{u}_N + r(\varepsilon, \widehat{u}_N). \tag{14}$$

Then, with the limit order $N \to \infty$ first and $\varepsilon \downarrow 0$ second,

$$\frac{\Delta e^{(N)}}{\|\Delta e^{(N)}\|} = \widehat{u}_N \xrightarrow{p} \frac{\nabla A_{\beta_N}(e)}{\|\nabla A_{\beta_N}(e)\|} = \frac{g + \beta_N h}{\|g + \beta_N h\|}. \tag{15}$$

Moreover,

$$A_{\beta_N}(e + \Delta e^{(N)}) \ge A_{\beta_N}(e) + \varepsilon \|\nabla A_{\beta_N}(e)\| - o_p(\varepsilon). \tag{16}$$

Thus, each BEST-OF-N gradient step asymptotically performs a first-order improvement on an implicit UCB acquisition function, with exploration strength β_N controlled entirely by N. The detailed theory and the proof is given in Appendix C and D.

5 TEXTBO Algorithm

For each iteration $t \in [1, T]$ and each trajectory $j \in [1, J]$, we evaluate one prompt; to determine the one prompt to evaluate, we take G steps of BEST-OF-N gradients beforehand, without any evaluation. We present the pseudocode of the TEXTBO algorithm (Algorithm 1) and provide the line-by-line description of the details of the procedure below.

Initialization ($t = 0$). At iteration $t = 0$, we first evaluate all initial prompts. For each trajectory $j = 1, \ldots, J$, we first generate the initial system outcome $\Phi(\pi_0^j)$. Next, we evaluate it on the corresponding input distribution $\mathcal{D}_{\mathcal{X}}(\pi_0^j)$ to obtain a scalar score s_0^j. The shared history is then initialized as $H_0 = \bigcup_{j=1}^J \{(\pi_0^j, \Phi(\pi_0^j), s_0^j)\}$, and and the shared reflection is initialized as $R_0 = \emptyset$.

Algorithm 1: TEXTBO

Input: System Φ; $\{\pi_0^j\}_{j=1}^J$; critic M_{critic}; candidates per step N; gradient steps G; iterations T; J trajectories

Output: Optimized triples $\{(\pi_T^j, c_T^j, s_T^j)\}_{j=1}^J$

for $j = 1$ **to** J **do**
  $c_0^j \leftarrow \Phi(\pi_0^j)$
  $s_0^j \leftarrow \text{Eval}(c_0^j, \mathcal{D}_{\mathcal{X}}(\pi_0))$
  $H_0 \leftarrow \{(\pi_0^j, c_0^j, s_0^j)\}$, $R \leftarrow \emptyset$
for $t = 1$ **to** T **do**
  for $j = 1$ **to** J **do**
    $\pi_{t,0}^j \leftarrow \pi_{t-1}^j$
    for $g = 1$ **to** G **do**
      for $i = 1$ **to** N **do**
        $\delta_{t,g}^{j,(i)} \leftarrow \nabla_{\text{text}}(\pi_{t,g-1}^j; R, M_{\text{critic}})$
        $\pi_{t,g}^{j,(i)} \leftarrow \text{Apply}(\pi_{t,g-1}^j, \delta_{t,g}^{j,(i)})$
        $c_{t,g}^{j,(i)} \leftarrow \Phi(\pi_{t,g}^{j,(i)})$
      $(\pi_{t,g}^j, c_{t,g}^j) \leftarrow$ BEST-OF-N$(\{(\pi_{t,g}^{j,(i)}, c_{t,g}^{j,(i)})\}, R_{t-1}, M_{\text{critic}})$
    $c_t^j \leftarrow c_{t,G}^j$, $\pi_t^j \leftarrow \pi_{t,G}^j$
    $s_t^j \leftarrow \text{Eval}(c_t^j, \mathcal{D}_{\mathcal{X}}(\pi_t))$
    $H_t \leftarrow H_{t-1} \cup \{(\pi_t^j, c_t^j, s_t^j)\}$
    if $s_t^j < s_{t-1}^j$ **then**
      $\pi_t^j \leftarrow \pi_{t-1}^j$, $c_t^j \leftarrow c_{t-1}^j$, $s_t^j \leftarrow s_{t-1}^j$
  $R_t \leftarrow$ METAREFLECT(H_t)
return $\{c_t^{j_t^*}\}_{t=1}^T$, *where* $j_t^* := \text{argmax}_j\, s_t^j$

Per-iteration update ($t \geq 1$). At each optimization iteration $t \in \{1, \ldots, T\}$, the algorithm updates all trajectories in parallel. For each trajectory $j \in \{1, \ldots, J\}$, we set the starting prompt for this iteration to be the final prompt from the previous iteration, i.e., $\pi_{t,0}^j \leftarrow \pi_{t-1}^j$. We then perform the following three steps.

1. **BEST-OF-N gradients phase.** (Detailed in §4.2)
2. **Evaluation phase.** For each trajectory j:
   1) Set the trajectory's candidate for evaluation to the final output of the BEST-OF-N gradient phase:
   $$\pi_t^j := \pi_{t,G}^j$$
   2) Evaluate c_t^j to obtain the scalar score
   $$s_t^j := \text{Eval}(\Phi(\pi_t^j), \mathcal{D}_{\mathcal{X}}(\pi_t^j)) = \frac{1}{L}\sum_{l=1}^{L} r\Big(y\Big(\Phi(\pi_t^j), x_t^{j,(l)}\Big)\Big), \quad \text{where } x_t^{j,(l)} \sim \mathcal{D}_{\mathcal{X}}(\pi_t^j). \tag{17}$$
   Note that this is the step where the evaluation happens; and in each iteration, we conduct J evaluations – one for each trajectory.
   3) Append the new triple $(\pi_t^j, \Phi(\pi_t^j), s_t^j)$ to the shared history:
   $$H_t \leftarrow H_{t-1} \cup \{(\pi_t^j, \Phi(\pi_t^j), s_t^j)\}.$$
   4) *Acceptance rule.* If the new score does not improve, revert this trajectory to its previous state:
   $$\text{if } s_t^j \leq s_{t-1}^j \text{ then } (\pi_t^j, \Phi(\pi_t^j), s_t^j) \leftarrow (\pi_{t-1}^j, \Phi(\pi_{t-1}^j), s_{t-1}^j).$$

This acceptance rule follows Yuksekgonul et al. (2025): it enforces monotone non-decreasing scores along each trajectory by rolling back any non-improving update.

3. **Shared meta-reflection update.**
   At the end of iteration t, after all J trajectories have been evaluated and the history updated, we recompute the global reflection:
   $$R_t \leftarrow \textsc{MetaReflect}(H_t; M_{\text{critic}}).$$
   The operator $\textsc{MetaReflect}(H_t; M_{\text{critic}})$ compresses the full history H_t—prompts, system outcomes, and scores—into a concise natural-language reflection R_t.
   After T optimization iterations, we report as the algorithm's output the sequence of best-performing outcomes
   $$\{\Phi(\pi_t^{j_t^*})\}_{t=1}^{T}, \qquad j_t^* := \operatorname{argmax}_j s_t^j,$$
   i.e., the best trajectory at each iteration, according to the evaluation scores.

6 Experiments for the Digital Ad optimization

In this section, we discuss the performance of TextBO for the ad-optimization (detailed setup discussed in Appendix §E). In §6.1, we discuss the baseline algorithms we compare against, and then in §6.2, we describe the experiment design. We present the main results in §6.3.

6.1 Baselines

We consider two state-of-the-art algorithms in the AI literature as baselines – (1) Best-of-N (Snell et al. 2024) and (2) GEPA (Agrawal et al. 2025). Best-of-N is one of the most popular and well-performing test-time alignment methods that has been reported to outperform reinforcement learning fine-tuning methods under large enough N (Gao et al. 2023; Mudgal et al. 2023; Eisenstein et al. 2023; Beirami et al. 2024). GEPA is a reflective and evolutionary prompt optimization algorithm that has demonstrated superior performance compared to state-of-the-art reinforcement fine-tuning techniques such as GRPO (Liu et al. 2024a). In our implementation of GEPA, we use the same set of initial prompts, critic models, ad generation modules, and evaluation operators we use for TextBO (see §E.2 for details).

6.2 Experiment Design

Our experiment is designed to test whether we can start from ex-post least-aligned ads (Worst5-of-64) and apply TextBO to outperform the ex-post best-aligned ad (Best-of-64). Here, "alignment" is measured by the evaluation score induced by our persona-based evaluation environment (see §E.2). This design choice matters because the Best-of-64 baseline is already a strong, ex-post selection benchmark: it is the best-performing ad within a reasonably rich, human-like initial pool of 64 creatives generated from the same creative brief (Snell et al. 2024; Beirami et al. 2024). By initializing optimization from the opposite end of the same pool (Worst5-of-64), we attribute improvements to the algorithm's performance rather than to a favorable starting prompt.

6.3 Results and Analysis

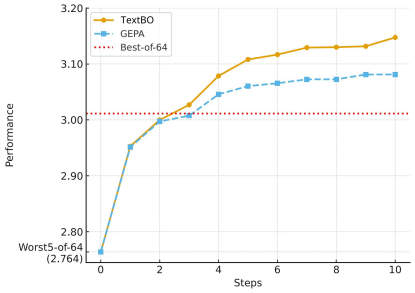
Main results and trend analysis. Figure 2 shows the results of our numerical experiments over 10 iterations: it plots how much TextBO and GEPA improved their performance from the initial Worst5-of-N score across eight scenarios. (For the detailed data used to generate this plot, see Tables 2 and 3 in Appendix G.2.) At each step (the x-axis), the performance score (the y-axis) is the average of eight scenarios' mean evaluation score for the test persona dataset.[3] For example, the step 2 score is the average of the eight step 2 mean evaluation scores.

Statistical analysis Figure 2 only shows the mean of means (i.e., scores are averaged over personas in test dataset and then over the eight scenarios). We also consider the Two-way Fixed Effect (TWFE) model:

$$\text{score} = \alpha + \beta_1 \mathbf{1}\{t = 10\} + \beta_2 \mathbf{1}\{\text{TextBO}\}\, \mathbf{1}\{t = 10\} + \text{persona FE} + \text{scenario FE} + \varepsilon.$$

The results from this TWFE model are shown on the right-hand side of Figure 2. We see that both TextBO's and GEPA show statistically significant gains from the starting point (Worst5-of-64) after ten iterations. In addition, at $T = 10$, TextBO shows a 23.5% more gain in score compared to GEPA.

[3] Best-of-N remains constant because, as discussed earlier, it is simply the best among the initial pool of ads and this baseline does not involve generation of new ads.



Effect	Estimate	Std. err	T-stat
β_1: GEPA gain ($t = 10$)	0.3168***	0.0325	9.752
β_2: TextBO gain - GEPA gain ($t = 10$)	0.0743***	0.0248	2.991
$\beta_1 + \beta_2$: TextBO gain ($t = 10$)	0.3911***	0.0355	11.011

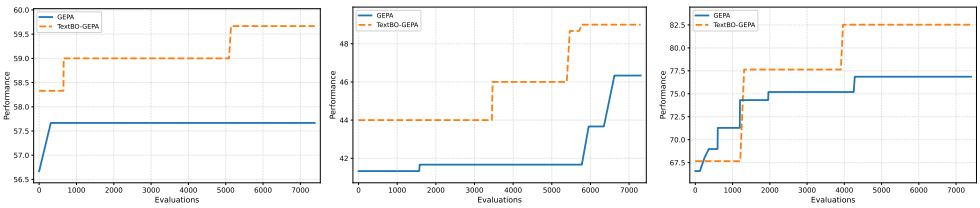
Notes: Two-way cluster-robust standard errors.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 2: Progress comparison of TextBO and GEPA with Best-of-64 baseline. TextBO implements parallel TextBO with $J = 5$. The performance score (the y-axis) is the average of eight scenarios' mean evaluation score for the testing persona set. Table reports TWFE analysis results.

7 Experiments for Agentic AI Benchmarks

The natural follow-up question is whether Best-of-N gradient sampling and multiple gradient steps will also improve the performance of standard GEPA used in the agentic AI benchmarks considered in Agrawal et al. (2025). Therefore, we now present a series of experiments, where we augment GEPA with the key theoretical and algorithmic innovations from our approach – Best-of-N gradient sampling and multiple gradient steps (as described in §4) . In our augmented version of GEPA, denoted as TextBO-GEPA, at each iteration, we take $G = 5$ gradient steps, and at each step, we employ Best-of-N textual gradients with $N = 5$. (In standard GEPA, each time we improve a prompt, we only take a single textual gradient step.) See Appendix §H for detailed experiment setup.

Figure 3 and Table 1 present the results of agentic AI benchmark experiments. We plot the improvement from their initial scores for both TextBO-GPEA and GEPA. Relative to the baseline, TextBO-GEPA delivers larger gains than GEPA across all datasets.



(a) HotpotQA, Qwen3 8B (b) HoVer, Qwen3 8B (c) PUPA, Qwen3 8B

Figure 3: Experiment results for three agentic AI benchmarks considered in Agrawal et al. (2025).

Model	**HotpotQA**	**Hover**	**PUPA**	**Aggregate**	**Improvement**
Qwen3-8B					
Baseline	56.67	41.33	66.58	54.86	—
GEPA	57.67	46.33	76.86	60.29	+5.43
TextBO-GEPA	**59.67**	**49.00**	**82.53**	**63.73**	**+8.87**

Table 1: Results on HotpotQA, Hover, and PUPA for Qwen3-8B after $\sim 7,500$ evaluations. Aggregate is the mean across the three datasets; Improvement is vs. Baseline aggregate.

8 Conclusion

Evaluation-efficient self-improving AI is increasingly important in scientific and societal settings where real-world feedback (e.g., human responses, engineering/field experiments) is slow and costly, even as candidate generation and analysis have become cheap. In this paper, we address this gap by proposing TextBO, a simple, provably evaluation-efficient prompt-optimization framework that combines textual gradients with Best-of-N gradient selection.

References

Emre Can Acikgoz, Cheng Qian, Heng Ji, Dilek Hakkani-Tür, and Gokhan Tur. Self-improving llm agents at test-time. *arXiv preprint arXiv:2510.07841*, 2025.

Virginia Aglietti, Ira Ktena, Jessica Schrouff, Eleni Sgouritsa, Francisco Ruiz, Alan Malek, Alexis Bellot, and Silvia Chiappa. Funbo: Discovering acquisition functions for bayesian optimization with funsearch. In *Forty-second International Conference on Machine Learning*, 2025.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.

Nicolás Aramayo, Mario Schiappacasse, and Marcel Goic. A multiarmed bandit approach for house ads recommendations. *Marketing Science*, 42(2):271–292, 2023.

Wenjia Ba, J Michael Harrison, and Harikesh S Nair. Advertising media and target audience optimization via high-dimensional bandits. *arXiv preprint arXiv:2209.08403*, 2022.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Shirsho Biswas. Investigating the effects of including discount information in advertising. Working Paper, 2025.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, et al. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *arXiv preprint arXiv:2506.13651*, 2025.

Abdoulatif Cissé, Xenophon Evangelopoulos, Vladimir V Gusev, and Andrew I Cooper. Language-based bayesian optimization research assistant (bora). *arXiv preprint arXiv:2501.16224*, 2025.

Li Ding, Jenny Zhang, Jeff Clune, Lee Spector, and Joel Lehman. Quality diversity through human feedback: Towards open-ended diversity-driven optimization. *arXiv preprint arXiv:2310.12103*, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104, 2024.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rockt""aschel. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Tong Geng, Xiliang Lin, and Harikesh S Nair. Online evaluation of audiences for targeted advertising via bandit experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

Anna Goldie, Azalia Mirhoseini, and Jeff Dean. That chip has sailed: A critique of unfounded skepticism around ai for chip design. *arXiv preprint arXiv:2411.10053*, 2024.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*, 2024a. URL `https://openreview.net/forum?id=ZG3RaNIsO8`.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024b.

Jochen Hartmann, Yannick Exner, and Samuel Domdey. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31, 2025.

Neil He, Jiahong Liu, Buze Zhang, Ngoc Bui, Ali Maatouk, Menglin Yang, Irwin King, Melanie Weber, and Rex Ying. Position: Beyond euclidean–foundation models should embrace non-euclidean geometries. *arXiv preprint arXiv:2504.08896*, 2025.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. *arXiv preprint arXiv:2408.08435*, 2024.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1051–1068, Singapore,

December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.67. URL https://aclanthology.org/2023.emnlp-main.67/.

Tijmen Jansen, Mark Heitmann, Martin Reisenbichler, and David A Schweidel. Automated alignment: Engaging customers with visual generative ai. *Available at SSRN 4656622*, 2024.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3441–3460, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.findings-emnlp.309. URL https://aclanthology.org/2020.findings-emnlp.309/.

Garrett A Johnson. Inferno: A guide to field experiments in online display advertising. *Journal of economics & management strategy*, 32(3):469–490, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Enoch Hyunwook Kang. Llm personas as a substitute for field experiments in method benchmarking, 2025. URL https://arxiv.org/abs/2512.21080.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Mingze Kong, Zhiyong Wang, Yao Shu, and Zhongxiang Dai. Meta-prompt optimization for llm-based sequential decision making, 2025. URL https://arxiv.org/abs/2502.00728.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Randall A Lewis and Justin M Rao. The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973, 2015.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025a.

Siyan Li, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. PAPILLON: Privacy preservation from Internet-based and local language model ensembles. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3390, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi:10.18653/v1/2025.naacl-long.173. URL https://aclanthology.org/2025.naacl-long.173/.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. *arXiv preprint arXiv:2401.02051*, 2024b.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024c.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Pairjudge rm: Perform best-of-n sampling with knockout tournament, 2025. URL https://arxiv.org/abs/2501.13007.

Deborah J MacInnis, Ambar G Rao, and Allen M Weiss. Assessing when increased media weight of real-world advertisements helps sales. *Journal of Marketing Research*, 39(4):391–407, 2002.

Linda Mantia, Surojit Chatterjee, and Vivian S. Lee. Designing a successful agentic ai system. *Harvard Business Review*, 2025. URL https://hbr.org/2025/10/designing-a-successful-agentic-ai-system. Accessed: 2025-11-01.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding innovation engines: Automated creativity and improved stochastic optimization via deep learning. *Evolutionary computation*, 24(3):545–572, 2016.

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodola. Language models are injective and hence invertible. *arXiv preprint arXiv:2510.15511*, 2025.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.

Leonard Papenmeier, Matthias Poloczek, and Luigi Nardi. Understanding high-dimensional bayesian optimization. *arXiv preprint arXiv:2502.09198*, 2025.

Tiany Peng, George Gui, Daniel J Merlau, Grace Jiarui Fan, Malek Ben Sliman, Melanie Brucks, Eric J Johnson, Vicki Morwitz, Abdullah Althenayyan, Silvia Bellezza, et al. A mega-study of digital twins reveals strengths, weaknesses and opportunities for further improvement. *arXiv preprint arXiv:2509.19088*, 2025.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025. URL https://arxiv.org/abs/2507.02833.

Jiahao Qiu, Xuan Qi, Hongru Wang, Xinzhe Juan, Yimin Wang, Zelin Zhao, Jiayi Geng, Jiacheng Guo, Peihang Li, Jingzhe Shi, et al. Alita-g: Self-evolving generative agent for agent generation. *arXiv preprint arXiv:2510.23601*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Sara Rosengren, Martin Eisend, Scott Koslow, and Micael Dahlen. A meta-analysis of when and how advertising creativity works. *Journal of Marketing*, 84(6):39–56, 2020.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on learning theory*, pp. 1417–1418. PMLR, 2016.

Oliver J Rutz, Garrett P Sonnier, and Michael Trusov. A new method to aid copy testing of paid search text advertisements. *Journal of Marketing Research*, 54(6):885–900, 2017.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pp. 1723–1742. PMLR, 2017.

Lennart Schneider, Martin Wistuba, Aaron Klein, Jacek Golebiowski, Giovanni Zappella, and Felice Antonio Merra. Hyperband-based bayesian optimization for black-box prompt selection. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Lm9DXFrcHD.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv*, 2024. URL https://arxiv.org/abs/2406.07791. arXiv:2406.07791 [cs.CL].

David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 1, 2025.

Joykirat Singh, Subhabrata Dutta, and Tanmoy Chakraborty. Mechanistic behavior editing of language models, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Dilara Soylu, Christopher Potts, and Omar Khattab. Fine-tuning and prompt optimization: Two great steps that work better together, 2024. URL https://arxiv.org/abs/2407.10930.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE transactions on information theory*, 58 (5):3250–3265, 2012.

Eric Tang, Bangding Yang, and Xingyou Song. Understanding LLM embeddings for regression. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Wt6Iz5XNIO.

Olivier Toubia, George Z Gui, Tianyi Peng, Daniel J Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *arXiv preprint arXiv:2505.17479*, 2025.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wenjia Wang, Xiaowei Zhang, and Lu Zou. Regret optimality of gp-ucb. *arXiv preprint arXiv:2312.01386*, 2023.

Wenxiao Wang, Priyatham Kattakinda, and Soheil Feizi. Maestro: Joint graph & config optimization for reliable ai agents. *arXiv preprint arXiv:2509.04642*, 2025.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory, 2024. URL https://arxiv.org/abs/2409.07429.

Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. On the sublinear regret of gp-ucb. *Advances in Neural Information Processing Systems*, 36:35266–35276, 2023.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

James Wilson, Frank Hutter, and Marc Deisenroth. Maximizing acquisition functions for bayesian optimization. *Advances in neural information processing systems*, 31, 2018.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers, 2024. URL https://arxiv.org/abs/2309.03409.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639:609–616, 2025.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*, 2025.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022.

A Related Literature

The literature on self-improving AI has often been motivated by an important observation: progress in AI based solely on human data is approaching limits, so AI must create and learn from its own data. In other words, powerful AI "should have its own stream of experience that progresses, like humans, over a long time-scale" (Silver & Sutton 2025). That is, we need to design the self-improving AI that iteratively 1) generate data of outcomes, evaluate outcomes with grounded signals, and 2) self-tune the three knobs – *models*, *tools*, and *prompts* – in the system to achieve long-term objectives under real-world feedback (Wang et al. 2025). Here, tuning models is about altering language models' input-output correspondences (Huang et al. 2023; Lee et al. 2023; Acikgoz et al. 2025); tuning tools is about specifying the tools the AI has access to, including memory tools (Hou et al. 2025; Ouyang et al. 2025; Qiu et al. 2025); tuning prompts is about engineering prompts in the model's context window, including system messages, task instructions, constraints, schema, and few-shot exemplars (Zhou et al. 2022; Khattab et al. 2023; Pryzant et al. 2023; Yang et al. 2024; Yuksekgonul et al. 2025; Agrawal et al. 2025; Zhang et al. 2025; Wang et al. 2025).

In particular, iterative prompt-optimization based self-improving AI have become increasingly popular. In a recent paper, Agrawal et al. (2025) demonstrate their method, GEPA (Genetic-Pareto) is able to outperform state-of-the-art model fine-tuning techniques such as GRPO (Liu et al. 2024a) under a fixed budget of LLM generation calls across a range of challenging tasks including multi-document QA, constrained instruction following, retrieval-augmented fact verification, and privacy-related objectives. At a high-level, both GEPA and other prompt-based self-improvement (Agrawal et al. 2025; Zhang et al. 2025; Wang et al. 2025) is specified as procedural heuristic search: maintain a pool of prompts, generate variants via LLM-driven mutation/crossover/reflection, score candidates on held-out data or rollouts, retain top performers, and iterate (Fernando et al. 2024; Guo et al. 2024a; Agrawal et al. 2025; Soylu et al. 2024). This mirrors classical heuristic and evolutionary optimization, where performance is assessed as a function of *generation efficiency* rather than *evaluation efficiency* (Nguyen et al. 2016; Ding et al. 2023; Hu et al. 2024; Liu et al. 2024b). Yet, as discussed earlier, in many real-world deployments, the scarce resource is not candidate generation but evaluation. Thus, in this paper, we focus on developing an iterative prompt-optimization framework that treat evaluation calls as the budget to allocate and optimize.

Our paper also relates to the literature on relating LLMs to Bayesian Optimization (BO). A series of papers focuses on applying LLMs to enhance BO for numerical optimization problems. Liu et al. (2024c) and Cissé et al. (2025) focus on leveraging LLMs' contextual understanding and few-shot learning capabilities to improve BO's efficiency, enabling better warm-starting and surrogate modeling; Aglietti et al. (2025) uses an LLM to discover new acquisition functions for Bayesian optimization. The LLM writes candidate acquisition function formulas as code, which are evaluated on various optimization problems. Singh et al. (2025) applies BO to enhance LLM's internal behavior to boost the model's zero-shot and few-shot performance. In contrast to these papers, we focus on the prompt optimization problem that arises when formulating LLM-based self-improving AI system design as a Bayesian Optimization problem. In line with this focus, Schneider et al. (2025) and Kong et al. (2025) formulate prompt optimization among the fixed set of candidate prompts as Bayesian optimization. While Schneider et al. (2025) utilizes ideas from adversarial bandits, Kong et al. (2025) feeds each prompt text into an embedding model, and the GP surrogate operates on those fixed embedding features to guide selection. However, unlike these papers, we actively *generate* new candidate prompts as the process proceeds, rather than pre-specifying a fixed set of candidate prompts.

Our paper relates to the literature on efficient ad evaluation in digital marketing using bandit methods. Here, the goal is to iteratively improve the ads served, given a fixed set of candidate ads. Schwartz et al. (2017) adopts Thompson Sampling with a hierarchical general linear model for a real display advertising campaign for a financial-services firm. Geng et al. (2020) partitions JD.com's platform users into disjoint sub-populations and applies a contextual Thompson sampling algorithm to maximize the expected advertiser payoff. Aramayo et al. (2023) developed a contextual Thompson sampling algorithm to dynamically determine the list of house ads to display on the homepage of an electronic retailer to maximize the accumulated click-through rate of the ads. Ba et al. (2022) used parametric Thompson Sampling, where a parametric structure links features (media, audience attributes) to conversion likelihood to optimally allocate ad-media and target-audience combinations in high-dimensional settings with low success rates. The main difference between

these bandit-based papers and ours is the existence of a pre-defined set of candidate arms. Our paper focuses on creating new solutions by analyzing previous results. On the other hand, bandit methods aim to balance exploration of a pre-defined set of candidate ads optimally. The scope of our problem is also quite different. Our self-improving AI iterates over three steps–generation, evaluation, and analysis–whereas, bandits focus on efficiently executing the evaluation step. Indeed, one could employ a bandit method, such as best-arm-identification, in our evaluation step (instead of using A/B tests).

Finally, our paper relates to the literature on ad content. Researchers have studied many different features of ad creatives and their effect on consumer response, e.g., creativity (Rosengren et al. 2020), discount information (Biswas 2025), emotionality of content (MacInnis et al. 2002), and specific text phrases (Rutz et al. 2017). However, a limitation of this literature is that each study is designed to test one specific ad feature in one specific advertising format/context (e.g., TV ads, search ads), and there is not much consensus on how various features and estimated effects should be combined to optimize the ad creative and whether these effects can be ported to other formats and audiences. In this paper, we move away from manually manipulating individual ad features and instead treat the ad-creative optimization problem as a high-dimensional optimization problem in the prompt-space; and our self-improving AI framework can be used to automate and optimize ad-creative generation at scale for any given combination product, ad-format, and audience.

B GP-UCB and Its Evaluation Efficiency

Setup. Let f be a unknown, black-box function that maps a compact and convex feasible region $\mathcal{D} \subset \mathbb{R}^d$ to $\mathbb{R}$. At each time t, a sampling algorithm $\mathcal{A}$ selects a point $x_t \in \mathcal{D}$ to evaluate. This selection is adaptive, based on the history of previously chosen points and their observed outcomes. The evaluation at x_t yields a noisy observation $y_t = f(x_t) + \epsilon_t$, where ϵ_t is a random noise term, where $\{\epsilon_t : t = 1, 2, \ldots\}$ are independent, sub-Gaussian random variables. The unknown function f is assumed to be an element of a Reproducing Kernel Hilbert Space (RKHS), $\mathcal{N}_\Psi(\mathcal{D})$, induced by a stationary kernel Ψ. Functions within this RKHS possess a specific smoothness property, the nature of which is determined by the kernel used. The space $\mathcal{N}_\Psi(\mathcal{D})$ is endowed with an inner product and is formally defined as the set of functions

$$\mathcal{N}_\Psi(\mathcal{D}) := \left\{ g : \mathcal{D} \mapsto \mathbb{R} \mid g(\mathbf{x}) = \sum_{j=1}^{\infty} c_j \Psi(\mathbf{x} - \mathbf{x}_j), \text{ for } \{c_j\} \subset \mathbb{R} \text{ and } \{\mathbf{x}_j\} \subset \mathcal{D} \text{ such that } \|g\|_{\mathcal{N}_\Psi(\mathcal{D})} < \infty \right\}$$

where the norm is given by: $\|g\|_{\mathcal{N}_\Psi(\mathcal{D})} := \left(\sum_{j,l=1}^{\infty} c_j c_l \Psi(\mathbf{x}_j - \mathbf{x}_l) \right)^{\frac{1}{2}}$. For any two functions $g_1(\mathbf{x}) = \sum_{j=1}^{\infty} a_j \Psi(\mathbf{x} - \mathbf{x}_j)$ and $g_2(\mathbf{x}) = \sum_{l=1}^{\infty} b_l \Psi(\mathbf{x} - \mathbf{x}_l)$ in the space, their inner product is $\langle g_1, g_2 \rangle_{\mathcal{N}_\Psi(\mathcal{D})} := \sum_{j,l=1}^{\infty} a_j b_l \Psi(\mathbf{x}_j - \mathbf{x}_l)$. A key feature of RKHS space is the reproducing property, $\langle g, \Psi(\mathbf{x} - \cdot) \rangle_{\mathcal{N}_\Psi(\mathcal{D})} = g(\mathbf{x})$, which holds for all $\mathbf{x} \in \mathcal{D}$. The norm of the function f in this space is assumed to be bounded such that $\|f\|_{\mathcal{N}_\Psi(\mathcal{D})} \leq B$ for some constant $B > 0$.

The objectives. The performance of $\mathcal{A}$ is often evaluated through two types of regret over a time horizon T. One is *cumulative regret*, defined as

$$\mathcal{R}_C(T; f, \mathcal{A}) := \sum_{t=1}^{T} \left[\max_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) - f(\mathbf{x}_t) \right] \tag{18}$$

The other is *simple regret*, defined as

$$\mathcal{R}_S(T; f, \mathcal{A}) := \max_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) - f\left(\mathbf{x}^{(T)}\right) \tag{19}$$

where $\mathbf{x}^{(T)}$ is the output of the algorithm $\mathcal{A}$ after taking T function evaluations.

The lower bound. Scarlett et al. (2017) showed that a lower bound exists on the regret of *any* algorithm for $f \in \mathcal{N}_\Psi(\mathcal{D})$. Specifically, for any constant $B > 0$,

$$\inf_{\mathcal{A}} \sup_{\|f\|_{\mathcal{N}_\Psi(\mathcal{D})} \leq B} \mathbb{E}\left[\mathcal{R}_C(T; f, \mathcal{A})\right] = \begin{cases} \Omega(T^{\frac{\nu+d}{2\nu+d}}), & \text{for Matérn kernels,} \\ \Omega(T^{\frac{1}{2}} \ln^{\frac{d}{2}}(T)), & \text{for SE kernels.} \end{cases} \tag{20}$$

$$\inf_{\mathcal{A}} \sup_{\|f\|_{\mathcal{N}_\Psi(\mathcal{D})} \leq B} T(\epsilon, f, \mathcal{A}) = \begin{cases} \Omega\left(\left(\frac{1}{\epsilon}\right)^{2+d/\nu}\right), & \text{for Matérn kernels,} \\ \Omega\left(\frac{1}{\epsilon^2}\left(\log \frac{1}{\epsilon}\right)^{d/2}\right), & \text{for SE kernels.} \end{cases} \tag{21}$$

where Matérn kernels are defined as

$$\Psi_{\mathrm{M}}(\mathbf{x} - \mathbf{x}') := \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{2\sqrt{\nu} \|\mathbf{x} - \mathbf{x}'\|_2}{\ell} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu} \|\mathbf{x} - \mathbf{x}'\|_2}{\ell} \right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{D} \tag{22}$$

and SE kernels are defined as

$$\Psi_{\mathrm{SE}}(\mathbf{x} - \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2} \right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{D} \tag{23}$$

where $\nu > 0$ is the smoothness parameter, $\ell > 0$ is the length-scale parameter, $\Gamma(\cdot)$ is the gamma function, $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν, and $\|\cdot\|_2$ denotes the Euclidean norm.

GP-UCB algorithm. The Gaussian Process Upper Confidence Bound (GP-UCB) algorithm, introduced in the seminal work of Srinivas et al. (2012), operates from a Bayesian perspective by placing a Gaussian Process (GP) prior on the unknown objective function f. This GP is characterized by a zero-mean function and a symmetric, positive-definite covariance function, or kernel, $k : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$, where $k(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x} - \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{D}$. For any finite set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\} \subset \mathcal{D}$, the corresponding function values $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_t))^\top$ are assumed to follow a multivariate normal distribution with a zero mean vector and a covariance matrix with entries $[\mathbf{K}_t]_{jl} = k(\mathbf{x}_j, \mathbf{x}_l)$. Under the assumption of independent, zero-mean Gaussian observation noise with variance σ^2, conditioning the GP prior on a set of observations $\mathbf{D}_t := \{(\mathbf{x}_j, y_j) : j = 1, \ldots, t\}$ yields a closed-form posterior distribution. The posterior mean and variance at any point $\mathbf{x} \in \mathcal{D}$ are given by:

$$\mathbb{E}[f(\mathbf{x}) \mid \mathbf{D}_t] = \mathbf{k}_t^\top(\mathbf{x}) \left(\mathbf{K}_t + \sigma^2 \mathbf{I}_t\right)^{-1} \mathbf{y}_t, \mathrm{Var}[f(\mathbf{x}) \mid \mathbf{D}_t] = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^\top \left(\mathbf{K}_t + \sigma^2 \mathbf{I}_t\right)^{-1} \mathbf{k}_t(\mathbf{x}), \tag{24}$$

where $\mathbf{k}_t(\mathbf{x})$ is the vector of covariances between $\mathbf{x}$ and the observed points, $\mathbf{K}_t$ is the covariance matrix of the observed points, $\mathbf{y}_t$ is the vector of observed values, and $\mathbf{I}_t$ is the identity matrix. Inspired by upper confidence bound methods in the multi-armed bandit literature (Auer et al., 2002), GP-UCB employs an optimistic acquisition strategy to select subsequent evaluation points. At each step t, the next point $\mathbf{x}_{t+1}$ is chosen by maximizing an upper confidence bound on the function's value $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{D}} \left(\mathbb{E}[f(\mathbf{x}) \mid \mathbf{D}_t] + \sqrt{\beta_t \mathrm{Var}[f(\mathbf{x}) \mid \mathbf{D}_t]}\right)$. This acquisition function naturally balances exploitation, driven by the posterior mean $\mathbb{E}[f(\mathbf{x}) \mid \mathbf{D}_t]$, with exploration, driven by the posterior standard deviation $\sqrt{\mathrm{Var}[f(\mathbf{x}) \mid \mathbf{D}_t]}$. The tunable parameter $\beta_t > 0$ explicitly governs this trade-off, making its specification critical to the algorithm's performance.

The evaluation efficiency of GP-UCB. Whitehouse et al. (2023); Wang et al. (2023) showed that GP-UCB Srinivas et al. (2012) regret upper bound for both expected cumulative regret and simple regret almost tightly matches the lower bound provided by Scarlett et al. (2017). Specifically, for any constant $B > 0$, the GP-UCB algorithm $\mathcal{A}_{\text{GPUCB}}$ achieves

$$\sup_{\|f\|_{\mathcal{N}_\Psi(\mathcal{D})} \leq B} \mathbb{E}[\mathcal{R}_C(T; f, \mathcal{A}_{\text{GPUCB}})] = \begin{cases} O(T^{\frac{\nu+d}{2\nu+d}}), & \text{for Matérn kernels,} \\ O(T^{\frac{1}{2}} \ln^{\frac{d}{2}}(T)) & \text{for SE kernels.} \end{cases} \tag{25}$$

$$\sup_{\|f\|_{\mathcal{N}_\Psi(\mathcal{D})} \leq B} T(\epsilon, f, \mathcal{A}_{\text{GPUCB}}) = \begin{cases} O\left(\left(\frac{1}{\epsilon}\right)^{2+d/\nu}\right), & \text{for Matérn kernels,} \\ O\left(\frac{1}{\epsilon^2} \cdot \left(\log \frac{1}{\epsilon}\right)^{d+3}\right), & \text{for SE kernels.} \end{cases} \tag{26}$$

Wang et al. (2023)'s approach to proving the regret optimality of GP-UCB follows the two-step framework. The first critical component is to construct a high-probability uniform error bound that quantifies the difference between the true objective function $f(x)$ and its posterior mean estimate $\mu_t(x)$ at any given step t. This bound takes the form $|f(x) - \mu_t(x)| \leq \sqrt{\beta_t}\sigma_t(x)$ for all $x \in \mathcal{D}$, where $\sigma_t(x)$ is the posterior standard deviation and β_t is a carefully chosen exploration parameter. The second component involves bounding the cumulative sum of the instantaneous regrets, which, under the uniform error bound, can be related to the sum of the posterior variances $\sum_{t=1}^{T} \sigma_{t-1}^2(x_t)$. This sum is, in turn, bounded by the maximal information gain, γ_T. The final regret bound is thus a function of T, β_T, and an upper bound on γ_T. Wang et al. (2023) employs tools from empirical process theory and decomposes the estimation error, $f(x) - \mu_t(x)$, into a bias term and a random error term. By leveraging the connection between Gaussian process regression and kernel ridge regression, along with properties of the Fourier transform for stationary kernels, they show that the bias term is bounded by $||f||_{\mathcal{N}_\Psi(\mathcal{D})}\sigma_t(x)$. The more challenging random error term is handled by viewing it as an empirical process indexed by a class of functions and bounding its supremum. They bound the ϵ-entropy of this function class, which allows for a high-probability bound on the random error. With the new uniform error bound, we can select a much smaller, dimension-independent exploration parameter β_t that grows only logarithmically with T. Combining this sharper β_t with the tightest known bounds on the maximal information gain γ_T for Matérn and SE kernels, we can achieve the tight cumulative regret of GP-UCB.

C Theory: BEST-OF-N gradient's Asymptotic Equivalence to UCB gradient

We now prove that the BEST-OF-N over textual gradients selects, in probability, the ascent direction of the UCB acquisition with exploration weight β scaling as $\Theta(\sqrt{\ln N})$ (Theorem 2). What allows us to do rigorous theoretical analysis is that LLM prompt embeddings are invertible and bi-Lipschitz: LLM's prompt embeddings are known to be injective, and therefore invertible (Nikolaou et al. 2025); LLM embeddings are Lipschitz (Tang et al. 2025), implying they are bi-Lipschitz. Hence, small textual edits can be represented as small changes in the embedding space and vice versa. Consequently, small textual edits correspond to small perturbations in the embedding space, allowing us to analyze BEST-OF-N gradient steps in $\mathbb{R}^d$ and then translate the theoretical guarantees back to the prompt space.

C.1 Intuitive Explanation

Denote e to be an embedding of a prompt. Pick a small radius $\varepsilon > 0$ around e and sample N nearby edits $e_i = e + \varepsilon u_i$ with u_i on the unit sphere. Next, denote the critic LLM's (M_{critic} from §4.1) internal scoring of each edit, Y_i, as follows:

$$Y_i = \mu(e_i) + \xi_i \, \sigma(e_i), \tag{27}$$

where μ is the (implicit) posterior mean, σ is the (implicit) epistemic uncertainty, and ξ_i is an i.i.d. sub-Gaussian noise variable independent of u_i. We note that Y_i is in line with the BO scoring function from §3.

Let $M_N = \max_i \xi_i$ and let $q_N = F^{-1}(1 - 1/N)$ be the $(1 - 1/N)$-quantile of the noise distribution. For Gaussian noise, classical extreme-value theory implies that M_N concentrates around q_N and that $q_N = \Theta(\sqrt{\ln N})$. In other words, as N grows large, the largest noise term among $\{\xi_i\}_{i=1}^N$ behaves like q_N up to lower-order fluctuations. Consequently, the index $i^\star$ selected by BEST-OF-N approximately maximizes

$$\mu(e_i) + q_N \, \sigma(e_i). \tag{28}$$

Consequently, for large N, selecting the BEST-OF-N index $i^\star$ is asymptotically equivalent to selecting the maximizer of UCB-style scores with exploration weight q_N.

Now consider a small sphere around e in embedding space and the UCB acquisition $A_\beta(e') = \mu(e') + \beta \, \sigma(e')$. On this sphere, there is a single dominant ascent direction for A_β, which we denote by v_β (the direction in which A_β increases the fastest to first order). Draw a narrow cone around v_β. For sufficiently large N, at least one of the sampled directions u_i falls inside this cone with high probability, and the BEST-OF-N winner $e_{i^\star}$ comes from this cone. Thus, among the locally sampled perturbations $e_i = e + \varepsilon u_i$, the BEST-OF-N rule selects a candidate whose direction is (with high probability, for large enough N) very close to the ascent direction of the UCB acquisition function with $\beta_N := q_N$.

C.2 Rigorous Derivation

Let Π be the set of syntactically valid prompts. Suppose that there exists a text embedding $\mathsf{E} : \Pi \to \mathbb{R}^d$, where $e := \mathsf{E}(\pi) \in \mathbb{R}^d$ is the embedding of a prompt $\pi \in \Pi$. We first state the assumptions motivated by how language model embeddings behave in practice, and then state the main theorem, Theorem 2.

At each gradient step of TEXTBO, we sample edits that correspond to $u_1, \ldots, u_N \overset{\text{i.i.d.}}{\sim} \rho$ on the embedding space $\subset \mathbb{R}^d$, where ρ is assumed to have a density bounded below by a positive constant with respect to surface measure. Assumption 1 implies that small textual edits induce near-linear moves in $\mathbb{R}^d$, consistent with the recent findings that prompt embeddings preserve phrase similarity under small changes (Reimers & Gurevych 2019; He et al. 2025; Tang et al. 2025; Nikolaou et al. 2025).

Assumption 1 (Embedding map and edit realization)**.** *For the current prompt π with $e = \mathsf{E}(\pi)$, constants $C < \infty$ and $\varepsilon_0 > 0$ such that for every unit vector $u \in \mathbb{R}^d$ and every $\varepsilon \in (0, \varepsilon_0]$ the edit operator* Apply *realizes a first–order move:*

$$\mathsf{E}\big(\mathrm{Apply}(\pi, \varepsilon u)\big) = e + \varepsilon u + r(\varepsilon, u), \qquad \sup_{\|u\|=1} \|r(\varepsilon, u)\| \le C \, \varepsilon^2. \tag{29}$$

Next, Assumption 2 is a standard regularity condition: function $f : \mathbb{R}^d \to \mathbb{R}$ is of $C^{1,1}$ if it is continuously differentiable and its gradient is Lipschitz: $\exists L < \infty$ such that $\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\|$ for all y, z.

Assumption 2 (μ, σ are of $C^{1,1}$)**.** *Let $\mu(e)$ and $\sigma(e)$ denote the mean and standard deviation of the critic LLM's internal score of embedding e. We assume μ, σ are $C^{1,1}$ locally. We use $g = \nabla\mu(e)$ and $h = \nabla\sigma(e)$.*

Assumption 3 specifies a heteroskedastic sub-Gaussian model for candidate scores (with i.i.d. mean-zero, unit-variance, unbounded-support noise independent of the sampled directions) and an ideal BEST-OF-N choice oracle.

Assumption 3 (Noise model and choice oracle)**.** *Let $\{\xi_i\}_{i=1}^N$ be i.i.d. mean-zero, unit-variance, Gaussian random variables with unbounded support, independent of $\{u_i\}$. For candidates $e_i := x + \varepsilon u_i$, the critic LLM observes its internal scoring*

$$Y_i \;=\; \mu(e_i) \;+\; \sigma(e_i)\,\xi_i, \tag{30}$$

and the BEST-OF-N selection oracle (often built using pairwise tournaments in practice; see §4.2) returns corresponding $i^\star \in \arg\max_{i\in[N]} Y_i$.

Assumption 4 (Small-step regime and limit order)**.** *We take limits with $N \to \infty$ first, then $\varepsilon \downarrow 0$.*

Lemma 2 (Gaussian maxima and spacing (Vershynin 2018))**.** *Let $\xi_1, \ldots, \xi_N$ be i.i.d. mean-zero, unit-variance Gaussian, $M_N = \max_{i\leq N} \xi_i$, S_N the second largest, and $q_N := F^{-1}(1 - 1/N)$ the $(1 - 1/N)$-quantile of ξ_1. Then $q_N = \Theta(\sqrt{\ln N})$, $M_N - q_N \to 0$ in probability, and $M_N - S_N = O_p(1/q_N)$.*

Theorem 2 (BEST-OF-N asymptotically induces a UCB gradient direction)**.** *Under Assumptions 1-4, define $\beta_N := q_N$ from Lemma 2. Let $i^\star \in \arg\max_i Y_i$ and set $\widehat{u}_N := u_{i^\star}$. Define the step*

$$\Delta e^{(N)} \;:=\; \mathsf{E}\big(\mathrm{Apply}(\pi, \varepsilon\widehat{u}_N)\big) - e \;=\; \varepsilon\widehat{u}_N + r(\varepsilon, \widehat{u}_N). \tag{31}$$

Then, with the limit order $N \to \infty$ first and $\varepsilon \downarrow 0$ second,

$$\frac{\Delta e^{(N)}}{\|\Delta e^{(N)}\|} \;=\; \widehat{u}_N \;\xrightarrow{p}\; \frac{\nabla A_{\beta_N}(e)}{\|\nabla A_{\beta_N}(e)\|} \;=\; \frac{g + \beta_N h}{\|g + \beta_N h\|}. \tag{32}$$

Moreover,

$$A_{\beta_N}(e + \Delta e^{(N)}) \;\geq\; A_{\beta_N}(e) \;+\; \varepsilon\,\|\nabla A_{\beta_N}(e)\| \;-\; o_p(\varepsilon). \tag{33}$$

Thus, each BEST-OF-N gradient step asymptotically performs a first-order improvement on an implicit UCB acquisition function, with exploration strength β_N controlled entirely by N. The full proof is given in Web Appendix D.

Theorem 2 provides the central conceptual and technical contribution of this paper: it shows that a procedure that looks like a purely heuristic rule in language space *implicitly implements* the same optimism-driven search direction as UCB BO in the model's embedding space. This is novel and important because it removes the main obstacle to bringing UCB BO's evaluation-efficiency guarantees into prompt optimization: we never construct an explicit surrogate model, never compute $\sigma(\cdot)$, and never differentiate a closed-form acquisition function, yet the effective update direction converges to the UCB gradient direction. Moreover, the result establishes a new use case for the ubiquitous BEST-OF-N principle as an *interpretable exploration mechanism*, where N becomes a principled "exploration knob" for the exploration weight $\beta_N = \Theta(\sqrt{\ln N})$. In short, Theorem 2 is what turns BEST-OF-N gradient from an empirical recipe into a principled BO-style optimizer: it justifies viewing our language-space self-improvement loop as parallel gradient-based UCB-BO and is the reason TEXTBO can inherit BO's evaluation-efficiency benefits.

D Proofs

D.1 Auxiliary lemmas

Lemma 3 (Uniform first-order expansion)**.** *Let $g = \nabla\mu(e)$ and $h = \nabla\sigma(e)$. There exist $C > 0$ and $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$ and all $u \in \mathbb{R}^d$,*

$$\mu(e + \varepsilon u) = \mu(e) + \varepsilon\, g^\top u \;+\; r_\mu(\varepsilon, u), \qquad |r_\mu(\varepsilon, u)| \leq C\varepsilon^2, \tag{34}$$

$$\sigma(e + \varepsilon u) = \sigma(e) + \varepsilon\, h^\top u \;+\; r_\sigma(\varepsilon, u), \qquad |r_\sigma(\varepsilon, u)| \leq C\varepsilon^2. \tag{35}$$

Proof. By the mean-value theorem with Lipschitz gradients, for some ξ between e and $e+\varepsilon u$, $\mu(e+\varepsilon u) = \mu(e) + \nabla\mu(\xi)^\top(\varepsilon u)$. Hence $\mu(e+\varepsilon u) = \mu(e) + \varepsilon g^\top u + \varepsilon(\nabla\mu(\xi) - \nabla\mu(e))^\top u$, and $|\nabla\mu(\xi) - \nabla\mu(e)| \le L_\mu\|\xi - e\| \le L_\mu\varepsilon$, giving $|r_\mu| \le L_\mu\varepsilon^2$. The proof for σ is identical with L_σ. Taking $C := \max\{L_\mu, L_\sigma\}$ yields the stated bounds with the same C. □

Lemma 4 (Max-stability under bounded perturbations)**.** *Let $(f_i)_{i\le N}$ and $(g_i)_{i\le N}$ be real arrays. Let $i_f \in \arg\max_i f_i$ and $i_g \in \arg\max_i g_i$. If for some $\Delta \ge 0$,*

$$g_{i_f} \ge f_{i_f} - \Delta \quad \textit{and} \quad f_{i_g} \ge g_{i_g} - \Delta, \tag{36}$$

then $g_{i_f} \ge \max_i g_i - \Delta$ and $f_{i_g} \ge \max_i f_i - \Delta$. In particular, $g_{i_f} \ge \max_i g_i - \Delta$ implies that any maximizer of f is a Δ-approximate maximizer of g.

Proof. Since $g_{i_f} \ge f_{i_f} - \Delta = \max_i f_i - \Delta \ge g_{i_g} - \Delta = \max_i g_i - \Delta$, the first claim holds. The second is symmetric. □

Lemma 5 (Spherical cap coverage)**.** *For $v \in \mathbb{S}^{d-1}$ and $\eta \in (0,1)$, define $C(v,\eta) := \{u \in \mathbb{S}^{d-1} : v^\top u \ge 1-\eta\}$. Let $U_1, \dots, U_N \overset{i.i.d.}{\sim} \rho$ as above. Then $\max_{i\le N} v^\top U_i \to 1$ almost surely as $N \to \infty$. Moreover, for any deterministic sequence $\eta_N \downarrow 0$ with $N \cdot \rho\{u : v^\top u \ge 1-\eta_N\} \to \infty$, we have*

$$\mathbb{P}\left(\max_{i\le N} v^\top U_i \ge 1 - \eta_N\right) \to 1. \tag{37}$$

Proof. Let $C(\eta) = \{u \in \mathbb{S}^{d-1} : v^\top u \ge 1-\eta\}$ be a spherical cap. Since ρ has a density bounded below, $\rho(C(\eta)) \asymp \eta^{\frac{d-1}{2}}$ as $\eta \downarrow 0$. Then $\mathbb{P}(\max_i v^\top U_i < 1-\eta) = (1-\rho(C(\eta)))^N \le \exp(-N\rho(C(\eta)))$. For any fixed $\eta \in (0,1)$, $\sum_{N\ge1} \mathbb{P}(\max_i v^\top U_i < 1-\eta) \le \sum_{N\ge1} e^{-N\rho(C(\eta))} < \infty$, so by Borel–Cantelli, $\mathbb{P}(\max_i v^\top U_i < 1-\eta \text{ i.o.}) = 0$. Intersecting over a countable sequence $\eta_m \downarrow 0$ yields $\max_{i\le N} v^\top U_i \to 1$ almost surely. Choosing $\eta_N \downarrow 0$ so that $N\rho(C(\eta_N)) \to \infty$ yields the result, and taking $\eta = \eta_N \to 0$ proves the first statement. □

Lemma 6 (Near-maximum coupling in a thin band)**.** *Let $M_N = \max_i \xi_i$ and define $v_N := (g + M_N h)/\|g + M_N h\|$. Also, $S_i := a_i + b_i M_N$, $Y_i^{(0)} := a_i + b_i\xi_i$, $b_i := \sigma(e) + \varepsilon h^\top U_i$. Fix any sequence $c_N \uparrow \infty$ with $c_N/q_N \to 0$ and set $\delta_N := c_N/q_N$. Then with probability $\to 1$ there exists an index j such that*

$$\xi_j \ge M_N - \delta_N \quad \textit{and} \quad v_N^\top U_j \ge 1 - \eta_N, \tag{38}$$

whenever $\eta_N \downarrow 0$ satisfies

$$N\,\mathbb{P}(\xi \ge M_N - \delta_N)\,\rho\{u : v_N^\top u \ge 1-\eta_N\} \to \infty. \tag{39}$$

Moreover, for this j,

$$S_j - Y_j^{(0)} = b_j(M_N - \xi_j) \le b_{\max}\,\delta_N. \tag{40}$$

Proof. Define $p_{\mathrm{cap},N} := \rho\{u : v_N^\top u \ge 1-\eta_N\}$ and the event

$$E_N := \{M_N \ge q_N - \delta_N\}. \tag{41}$$

By Lemma 2, $M_N - q_N \to 0$ in probability and $\delta_N \to 0$, hence $\mathbb{P}(E_N) \to 1$.

Set the deterministic threshold $t_N := q_N - 2\delta_N$. On E_N we have $M_N - \delta_N \ge q_N - 2\delta_N = t_N$, hence by monotonicity of the tail,

$$\mathbb{P}(\xi \ge t_N) \;\ge\; \mathbb{P}(\xi \ge M_N - \delta_N) \quad \text{on } E_N. \tag{42}$$

Let

$$K_N := \sum_{i=1}^N \mathbf{1}\{\xi_i \ge t_N\} \sim \mathrm{Binomial}\big(N,\, p_N\big), \qquad p_N := \mathbb{P}(\xi \ge t_N). \tag{43}$$

A Chernoff bound gives

$$\mathbb{P}\big(K_N < \tfrac{1}{2}Np_N\big) \;\le\; \exp\big(-\tfrac{1}{8}Np_N\big)\,. \tag{44}$$

Consider

$$L_N \;:=\; \sum_{i=1}^{N} \mathbf{1}\{\xi_i \ge t_N\}\, \mathbf{1}\{v_N^\top U_i \ge 1-\eta_N\}. \tag{45}$$

Conditional on $\{\xi_i\}_{i=1}^N$ (and hence on K_N and v_N), the variables $\{\mathbf{1}\{v_N^\top U_i \ge 1-\eta_N\}\}_{i=1}^N$ are i.i.d. Bernoulli$(p_{\mathrm{cap},N})$ and independent of $\{\xi_i\}$. Therefore,

$$L_N \,\big|\, \{\xi_i\} \;\sim\; \mathrm{Binomial}\big(K_N,\, p_{\mathrm{cap},N}\big), \qquad \mathbb{P}\big(L_N = 0 \,\big|\, \{\xi_i\}\big) \;=\; (1-p_{\mathrm{cap},N})^{K_N} \;\le\; \exp\big(-K_N p_{\mathrm{cap},N}\big). \tag{46}$$

We now bound $\mathbb{P}(L_N = 0)$ by splitting on $\{K_N \ge \frac{1}{2}Np_N\}$ and E_N:

$$\begin{aligned}
\mathbb{P}(L_N = 0) &\le \mathbb{E}\Big[\exp\big(-K_N p_{\mathrm{cap},N}\big)\, \mathbf{1}_{\{K_N \ge \frac{1}{2}Np_N\}\cap E_N}\Big] \;+\; \mathbb{P}\big(K_N < \tfrac{1}{2}Np_N\big) \;+\; \mathbb{P}(E_N^c) \\
&\le \mathbb{E}\big[\exp\big(-\tfrac{1}{2}Np_N\, p_{\mathrm{cap},N}\big)\, \mathbf{1}_{E_N}\big] \;+\; \exp\big(-\tfrac{1}{8}Np_N\big) \;+\; o(1) \\
&\le \mathbb{E}\big[\exp\big(-\tfrac{1}{2}Np_N\, p_{\mathrm{cap},N}\big)\big] \;+\; \exp\big(-\tfrac{1}{8}Np_N\big) \;+\; o(1).
\end{aligned} \tag{47}$$

On E_N, equation 42 yields $p_N \ge \mathbb{P}(\xi \ge M_N - \delta_N)$. By the lemma's hypothesis,

$$Z_N := Np_N\, p_{\mathrm{cap},N} \;\ge\; N\,\mathbb{P}(\xi \ge M_N - \delta_N)\, p_{\mathrm{cap},N} \;\xrightarrow{p}\; \infty.$$

Since $0 \le e^{-\frac{1}{2}Z_N} \le 1$, for any fixed $T > 0$,

$$\mathbb{E}\Big[e^{-\frac{1}{2}Z_N}\Big] \le e^{-\frac{1}{2}T} + \mathbb{P}(Z_N \le T) \;\longrightarrow\; e^{-\frac{1}{2}T}.$$

Letting $T \to \infty$ gives $\mathbb{E}[e^{-\frac{1}{2}Z_N}] \to 0$. Moreover, $Z_N \le Np_N$ since $p_{\mathrm{cap},N} \le 1$. If Np_N were bounded along a subsequence, then Z_N would be bounded along that subsequence, contradicting $Z_N \xrightarrow{p} \infty$. Hence $Np_N \to \infty$, so $\exp(-\frac{1}{8}Np_N) \to 0$. Therefore the second term in equation 47 vanishes, and $\mathbb{P}(L_N = 0) \to 0$.

Therefore, with probability tending to one there exists j such that $\xi_j \ge t_N$ and $v_N^\top U_j \ge 1-\eta_N$. Since $t_N \le M_N - \delta_N$ on E_N, this j also satisfies $\xi_j \ge M_N - \delta_N$. Finally, $S_j - Y_j^{(0)} = b_j(M_N - \xi_j) \le b_{\max}\delta_N$ holds deterministically by $M_N - \xi_j \le \delta_N$ and the definition of $b_{\max}$. □

D.2 Proof of Theorem 2

Proof.
By Lemma 3,

$$\begin{aligned}
Y_i &= \mu(e) + \varepsilon g^\top u_i + r_\mu(\varepsilon, u_i) + \big(\sigma(e) + \varepsilon h^\top u_i + r_\sigma(\varepsilon, u_i)\big)\xi_i \\
&= a_i + b_i\xi_i + R_i,
\end{aligned} \tag{48}$$

where $a_i := \mu(e) + \varepsilon g^\top u_i$, $b_i := \sigma(e) + \varepsilon h^\top u_i$, and $R_i := r_\mu(\varepsilon, u_i) + \xi_i r_\sigma(\varepsilon, u_i)$. Then $|R_i| \le C\varepsilon^2(1 + |\xi_i|)$ for all i. Define $S_i := a_i + b_i M_N$, $b_{\max} := \max_{1\le i\le N} |b_i|$, and $U_i := u_i$.

Now let $Y_i^{(0)} := a_i + b_i\xi_i$ and $i_0 \in \arg\max_i Y_i^{(0)}$. From equation 48,

$$Y_{i^\star} \ge Y_{i_0} \implies Y_{i^\star}^{(0)} \ge Y_{i_0}^{(0)} - |R_{i^\star}| - |R_{i_0}|.$$

By the bound on R_i and $|\xi_i| \le M_N := \max_j \xi_j$,

$$Y_{i^\star}^{(0)} \ge \max_i Y_i^{(0)} - 2C\varepsilon^2(1 + M_N). \tag{49}$$

Symmetrically, $\max_i Y_i^{(0)} \ge Y_{i^\star}^{(0)} \ge \max_i Y_i^{(0)} - 2C\varepsilon^2(1 + M_N)$. By Lemma 4, $i^\star$ is a $2C\varepsilon^2(1 + M_N)$-approximate maximizer of $Y^{(0)}$.

Now let j be the index from Lemma 6 so that $\xi_j \ge M_N - \delta_N$ and $v_N^\top U_j \ge 1 - \eta_N$. By Lemma 3,

$$\max_{i\le N} S_i - S_j \;\le\; \varepsilon\|g + M_N h\|\,(1 - v_N^\top U_j) + 2C\varepsilon^2(1+M_N) \;\le\; \varepsilon\|g + M_N h\|\,\eta_N + 2C\varepsilon^2(1+M_N).$$

Also,

$$S_j - S_{i^\star} = (S_j - Y_j^{(0)}) + (Y_j^{(0)} - Y_{i^\star}^{(0)}) + (Y_{i^\star}^{(0)} - S_{i^\star}) \;\le\; b_{\max}\delta_N + 2C\varepsilon^2(1+M_N) + \underbrace{(Y_{i^\star}^{(0)} - S_{i^\star})}_{\le 0},$$

where the last term is ≤ 0 provided $b_i \ge 0$ for all i, which holds by Assumption A4 for $\varepsilon \le \varepsilon_0 := \sigma(e)/(2\|h\|)$ when $\sigma(e) > 0$ since $b_i = \sigma(e) + \varepsilon h^\top U_i \ge \sigma(e) - \varepsilon\|h\| \ge \frac{1}{2}\sigma(e) > 0$. Therefore,

$$\max_{i \le N} S_i - S_{i^\star} \;\le\; \varepsilon\|g + M_N h\|\,\eta_N \;+\; b_{\max}\delta_N \;+\; 4C\varepsilon^2(1+M_N). \tag{50}$$

Set:

$$\Delta'_N(\varepsilon) \;:=\; b_{\max}\,\frac{c_N}{q_N} \;+\; 4C\varepsilon^2(1+M_N).$$

By Lemma 3,

$$S_i \;=\; \mu(e) + M_N\sigma(e) + \varepsilon(g + M_N h)^\top u_i + \tilde{r}_i, \qquad |\tilde{r}_i| \le C\varepsilon^2(1+M_N).$$

Hence for any i,

$$S_i - S_{i^\star} = \varepsilon\|g + M_N h\|\,(v_N^\top U_i - v_N^\top \widehat{u}_N) + (\tilde{r}_i - \tilde{r}_{i^\star}),$$

so

$$\big|\,(S_i - S_{i^\star}) - \varepsilon\|g + M_N h\|\,(v_N^\top U_i - v_N^\top \widehat{u}_N)\,\big| \;\le\; 2C\varepsilon^2(1+M_N). \tag{51}$$

Taking i that attains $\max_i S_i$ and using equation 50 yields

$$\max_{i \le N} v_N^\top U_i - v_N^\top \widehat{u}_N \;\le\; \eta_N \;+\; \frac{\Delta'_N(\varepsilon)}{\varepsilon\|g + M_N h\|}. \tag{52}$$

Thus

$$1 - v_N^\top \widehat{u}_N \;\le\; \underbrace{1 - \max_{i \le N} v_N^\top U_i}_{\to 0 \text{ a.s. by Lemma 5}} + \eta_N \;+\; \frac{\Delta'_N(\varepsilon)}{\varepsilon\|g + M_N h\|}. \tag{53}$$

Assume $\|h\| > 0$ (the $\|h\| = 0$ case is addressed in Remark 1). Using

$$\|g + M_N h\| \;\ge\; M_N\|h\| - \|g\|$$

and Lemma 2, $M_N \to \infty$ and $M_N \asymp q_N$. Also $b_{\max} \le \sigma(e) + \varepsilon\|h\| + C\varepsilon^2$. Hence

$$\frac{\Delta'_N(\varepsilon)}{\varepsilon\|g + M_N h\|} = O_p\Big(\tfrac{c_N}{\varepsilon q_N^2\|h\|}\Big) + O\Big(\tfrac{\varepsilon}{\|h\|}\Big).$$

Choosing, e.g., $c_N = \frac{1}{2}\log\log N$ and any $\eta_N \downarrow 0$ with

$$N\,e^{c_N}\,\rho\{u : v_N^\top u \ge 1 - \eta_N\} \to \infty$$

makes the RHS of equation 53 equal to $o_p(1) + O(\varepsilon)$ for fixed $\varepsilon > 0$; then let $\varepsilon \downarrow 0$.

With $N \to \infty$ first and fixed $\varepsilon > 0$, equation 53 implies $1 - v_N^\top \widehat{u}_N \xrightarrow{p} 0$. Then $\varepsilon \downarrow 0$ removes the $O(\varepsilon)$ residual, yielding $v_N^\top \widehat{u}_N \xrightarrow{p} 1$. Since $M_N - q_N \to 0$ in probability by Lemma 2, the unit vectors $v_N = (g + M_N h)/\|g + M_N h\|$ and $\tilde{v}_N := (g + q_N h)/\|g + q_N h\|$ satisfy $\|v_N - \tilde{v}_N\| \to 0$ in probability. Hence $\widehat{u}_N \xrightarrow{p} \tilde{v}_N$, which is the claimed gradient direction of A_{β_N} with $\beta_N = q_N$.

By Lemma 3,

$$A_{\beta_N}(e + \Delta e^{(N)}) = A_{\beta_N}(e) + \varepsilon\nabla A_{\beta_N}(e)^\top \widehat{u}_N + O(\varepsilon^2).$$

Write $\nabla A_{\beta_N}(e) = g + q_N h$. Using $\tilde{v}_N^\top \widehat{u}_N \xrightarrow{p} 1$ gives $\nabla A_{\beta_N}(e)^\top \widehat{u}_N = \|\nabla A_{\beta_N}(e)\|\,\tilde{v}_N^\top \widehat{u}_N = \|\nabla A_{\beta_N}(e)\|(1 - o_p(1))$. Thus $A_{\beta_N}(e + \Delta e^{(N)}) \ge A_{\beta_N}(e) + \varepsilon\|\nabla A_{\beta_N}(e)\| - o_p(\varepsilon)$, as claimed. □

Remark 1 (Edge cases). *If $\|h\| = 0$, then $\nabla A_{\beta_N}(e) = g$ and the direction reduces to $\nabla\mu(e)$. The Bayesian Optimization setup does not allow $\sigma(e) = 0$.*

E Case Study: Digital Ad Optimization

We now instantiate our framework in a digital advertising setting, showing how an advertiser (or ad platform acting on behalf of an advertiser) can use TextBO to automate the ad generation–analysis–evaluation loop. Here, the goal is to refine the ad-generation prompts so that the resulting creatives are better aligned with a target population's preference distribution, under a limited budget of costly ad evaluations. We specialize the general setup from §2 to ad optimization as follows:

- Φ is an ad-generation AI system that can generate ads given a prompt π.[4]
- For a prompt π, $\Phi(\pi)$ is the resulting ad creative deployed on the platform.
- $\mathcal{X}$ is the population of platform users, and $\mathcal{D}_{\mathcal{X}}(\pi)$ is the (unknown) distribution of users who are exposed to $\Phi(\pi)$ when this ad is run.
- $\mathcal{Y}$ is the space of observed ad responses (clicks, conversions, spend). For each user $x \in \mathcal{X}$, $y(\Phi(\pi), x)$ is the random ad response when x is shown ad $\Phi(\pi)$.
- $r : \mathcal{Y} \to [0, 1]$ is a scoring rule that maps responses to a normalized effectiveness score (e.g., a weighted function of click, conversion, and spend).

The campaign objective is the same as the general objective introduced in §2: $J(\pi) = \mathbb{E}\big[\, r\big(y(\Phi(\pi), x)\big)\big]$, where the expectation taken over $x \sim \mathcal{D}_{\mathcal{X}}(\pi)$ and the stochasticity in $y(\Phi(\pi), x)$.

E.1 Ad Campaign Scenarios

We consider eight synthetic ad campaign scenarios that cover a diverse range of products across distinct categories, each defined by a creative brief that outlines the strategic and creative direction (see Web Appendix F.1 for the full creative briefs):

- Scenario 1: "GreenBite," a new plant-based burger patty.
- Scenario 2: "AuraSonics X1," high-end, noise-canceling wireless earbuds.
- Scenario 3: "Odyssey E-SUV," a new all-electric family SUV.
- Scenario 4: "Oasis Eco-Lodge," a secluded, luxury resort with beautiful natural surroundings.
- Scenario 5: "Momentum," a mobile-first banking app for freelancers and the gig economy.
- Scenario 6: "MindGarden," a subscription-based meditation and mindfulness app.
- Scenario 7: "Aeterno," a classic, automatic Swiss-made wristwatch with a heritage design.
- Scenario 8: "SyncFlow," a project management and collaboration software platform for remote teams.

E.2 Implementation Details

There are five main components of the algorithm: (1) initial prompt (π_0), and (2) Ad-generation module (Φ), (3) Critic LLM and its four operators defined in §4.1, (4) the population distribution $\mathcal{D}_{\mathcal{X}}(\pi)$ and the evaluation operator $\mathrm{Eval}()$, as defined in Equation equation 17. We now describe how each of these components is implemented in our experiments.

Initial prompts (π_0^j). For each scenario, we first created 64 high-quality prompts from its creative brief using a meta-prompt. In Web Appendix F.2, see Figure 4 for the details of the meta-prompt and Figure 5 for an example of two prompts in scenario 1. Among those 64 ads, we chose five as initial prompts; we will describe the details of the choice rule in §E.3.

Ad-generation module (Φ). We utilize a state-of-the-art image generation model, Google Imagen 4,[5] Imagen 4 delivers high-fidelity, photorealistic text-to-image generation at up to 2K resolution, and excels in rendering intricate details such as realistic textures, lighting, and camera effects. Its features include typography accuracy for legible in-image text and precise adherence to complex prompts.

[4] In practice, this image-generation model can be fine-tuned for a particular brand and/or customized to the platform. Our approach takes any pre-training as given.

[5] See Vertex AI's description (`https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/imagen-4.0-generate-preview-06-06`) for details.

Critic LLM and its Operators. Unless otherwise noted, we use Gemini 2.5 Flash [6] as the multimodal critic model M_{critic} and implement the four operators discussed in §4.1 using the meta-prompts described in Web Appendix F.3: (1) PAIRWISEJUDGE operation; see meta-prompt in Figure 6, (2) METAREFLECT operation; see meta-prompt in Figure 7, (3) Textual gradient operation ∇_{text}; see meta-prompt in Figure 8, and (4) Rewrite operation APPLY; see meta-prompt in Figure 9.

Evaluation module using persona data and LLM-based simulation. We now describe how we implement the evaluation operator Eval in this case study. To evaluate how well an ad performs on the target population and to update prompts, we need (i) a target population ($\mathcal{D}_{\mathcal{X}}(\pi).$) and (ii) a way to measure that population's preferences for a given ad ($\mathrm{Eval}(\cdot)$). In principle, there are three ways define the evaluation operator – (1) field experiments/deployment in a real ad platform, (2) consumer surveys, and (3) simulated responses from a pre-defined target population. The former two options (field deployment and consumer surveys) are costly in time and resources, and can require large samples to identify effective differences across ads given low CTRs in digital ad settings; see (Lewis & Rao 2015; Johnson 2023). Thus, using these approaches for algorithm evaluation is not ideal.

Therefore, in our experiments, we use the Twin-2k-500 persona distribution of Toubia et al. (2025) as the target population and LLM-generated preferences over this distribution as the evaluation module. The Twin-2K-500 dataset consists of "digital twins" or "digital persona" of $2,058$ representative people from the United States. The dataset was assembled by giving a comprehensive multi-wave survey (consisting of a total of 500 questions) to each person in the sample, spanning demographics, personality, cognitive ability, economic preferences, classic heuristics-and-biases tasks, and a pricing study. Thus, the resulting personas are quite rich and detailed. This persona distribution forms our target population $\mathcal{D}_{\mathcal{X}}(\pi)$. Further, consistent with a randomized experiment design, we assume that the ad does not influence the personas who are shown the ad, i.e., $\mathcal{D}_{\mathcal{X}}(\pi)$ is the same for all πs. For each persona, we pass their survey answers to a multi-modal LLM (Gemini 2.5 Flash) as context, and then ask the LLM to assess the effectiveness of a given ad for that persona. Concretely, for a given persona and ad, we ask the LLM (see Figure 10 for the detailed prompt in Web Appendix F.4) to rate the ad's "effectiveness" on a scale of 1–5, where 1 is least effective and 5 is most effective. Because the LLM returns token-level log probabilities over the discrete scores,[7] we convert these into a mean score and treat this mean (a real number such as 3.41) as the ad's evaluation score for that persona. Averaging over personas then yields the final evaluation score used by $\mathrm{Eval}()$.

We split the Twin-2k-500 personas into 80% (1,647 personas) for training and 20% (411 personas) for testing. During algorithm training, each evaluation of an ad $\Phi(\pi)$ proceeds as follows: we randomly sample 200 personas from the training set and simulate ad effectiveness for each persona via the LLM. When we test a final ad (e.g., after running TEXTBO), we simulate its effectiveness for all personas in the test set.

We note that important advantage of our approach: using this persona distribution with LLM-generated preferences enables fast and controlled comparisons of different algorithms without large-scale field experiments or expensive surveys.[8] Indeed, our approach can serve as a template for algorithm comparisons in many other settings where researchers require evaluations that reflect the variance/heterogeneity of human preferences. Further, we note that while LLM-based preference simulation may not reflect true human preferences (Li et al. 2025a; Peng et al. 2025), it still induces a well-defined preference distribution on which we can benchmark optimization methods and algorithms. Specifically, Kang (2025) proves that swapping humans for personas is equivalent to changing the evaluation population (e.g., from the New York population to the Jakarta population), under two benchmark hygiene conditions: (i) methods observe only the aggregate outcome (called *aggregate-only observation* condition) and (ii) evaluation depends only on the submitted artifact and not on the algorithm's identity or provenance (called *algorithm-blind evaluation* condition). Furthermore, they show that simply increasing the size of the persona dataset is sufficient to guarantee that persona

[6] See Vertex AI's description (`https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash`) for details.

[7] See Vertex AI's logprobs description (`https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/inference#logprobs`) for details.

[8] Naive field deployment cannot be used to test the relative performance of algorithms, since an algorithm's measured advantage would then mix its own effect with distributional effects induced by the ad platform's targeting systems.

simulation becomes as useful a benchmark as field experimentation. Nevertheless, we also provide evaluations of our approach that do not leverage persona datasets; see §G.3 for details.

E.3 Application of TextBO

We now describe how the TextBO algorithm is instantiated for the ad-optimization problem. Here, we choose the hyperparameters as: iterations $T = 10$; trajectories $J = 5$; gradient steps $G = 5$; and candidates per gradient step $N = 5$.

Initialization. As described in §E.2, for each scenario, we first generate 64 high-quality prompts from its creative brief using a meta-prompt. Among the 64 ads for each scenario, we identify the best initial ad, Best-of-64 (which will be used as one of the baselines but is not directly used in our algorithm; see §6.1 for details), and the worst five, Worst5-of-64, by running a best-arm identification bandit algorithm (Russo 2016). Specifically, we alternate between a best-arm identification objective and a worst 5-arm identification objective over 5,000 sequential random samples from the training persona set. These Worst5-of-64 ads serve as starting points $\{\pi_0^j\}_{j=1}^5$ for TextBO procedure. This starting point allows us to attribute observed gains to the algorithms rather than to favorable starting prompts.[9] Let $\{s_0^j\}_{j=1}^J$ denote the evaluation scores for these five ads. Then, the tuples $\{(\pi_0^j, \Phi(\pi_0^j), s_0^j)\}_{j=1}^J$ form the initial history H_0 for the algorithm.

Best-of-N textual-gradient steps. Within each optimization iteration t, TextBO refines each trajectory's prompt π_{t-1}^j through five steps of Best-of-N textual-gradients before evaluation (see Figure 2). We follow the generic procedure from §5, but here specialize the critic's context and edits to the ad domain.

For a given trajectory j and inner step g, the critic model M_{critic} (Gemini 2.5 Flash) receives as context: (1) the current prompt $\pi_{t,g-1}^j$ and (2) the current meta-reflection R_{t-1}. Conditioned on this context, using operation ∇_{text} , it independently samples N textual gradients $\{\delta_{t,g}^{j,(i)}\}_{i=1}^N$, applies them using operation Apply to obtain candidate prompts $\{\pi_{t,g}^{j,(i)}\}_{i=1}^N$. These candidate prompts are then fed to the ad generation module Φ to generate the corresponding 5 ads $\{c_{t,g}^{j,(i)}\}_{i=1}^N = \{\Phi(\pi_{t,g}^{j,(i)})\}_{i=1}^N$. For example, suppose that the current prompt $\pi_{t,g-1}^j$ is "`A simple, photorealistic image of a plant-based burger patty.`", if R_{t-1} emphasizes that "social occasions and visible grill marks tend to perform better than plain pack shots," then we may query the operation ∇_{text} (see Figure 8 in Web Appendix F.3 for the meta-prompt) to propose textual-gradient edits such as:

- *Emphasize social context and enjoyment:* "Show the burger being enjoyed at an outdoor barbecue with friends, not alone on a plate."
- *Highlight sensory appeal and indulgence:* "Make the patty look extra juicy with clear grill marks, a toasted bun, and melty toppings."
- *Make sustainability salient but secondary:* "Subtly include eco-friendly cues (like a small 'plant-based' tag or greenery in the background) without overpowering the food."

We then query these edits for the operation Apply operation to an initial prompt (see Figure 9 in Web Appendix F.3 for the meta-prompt) may yield candidates such as:

```
"A photorealistic image of a juicy plant-based burger with grill
marks, served at a vibrant summer barbecue with friends, everyone
smiling and reaching for food."

"A close-up, photorealistic shot of a plant-based burger with deep
grill marks, a toasted brioche bun, melty vegan cheese, fresh
toppings, and steam rising to suggest warmth and juiciness."

"A photorealistic image of a sizzling plant-based burger with rich
grill marks, served on a rustic wooden table with subtle greenery
```

[9] To maintain a rigorous baseline, we make these 64 prompts detailed enough to yield strong creatives. This ensures that subsequent improvements are not merely due to trivial gains from starting with an under-specified prompts; instead, improvements reflect better alignment with the underlying target audience distribution.

```
in the background and a small plant-based tag, conveying indulgence
without sacrificing sustainability."
```

To select which candidate to keep for the next gradient step, TEXTBO applies the BEST-OF-N rule from §4.2 to the ads $\{\Phi(\pi_{t,g}^{j,(i)})\}_{i=1}^{N}$. Specifically, following Liu et al. (2025), we implement BEST-OF-N via a pairwise tournament: in each match, the critic compares two creatives side by side and predicts which will perform better using operator PAIRWISEJUDGE, given the campaign goals and the current reflection R_{t-1} (see Figure 6 in Web Appendix F.3 for the meta-prompt). The tournament winner prompt and its corresponding ad, $\left(\pi_{t,g}^{j}, \Phi\left(\pi_{t,g}^{j}\right)\right)$, are carried forward to the next gradient step. Repeating this for $g = 1, \ldots, G$ yields the final refined prompt π_t^j and ad $\Phi(\pi_t^j)$ for trajectory j at iteration t.

Evaluation as the costly step. After the G gradient steps, TEXTBO performs a single expensive evaluation per trajectory (each of which consists of assessing the ad for 200 randomly sampled personas from the training persona set) by deploying $\Phi(\pi_t^j)$ to get the evaluation result s_t^j using the evaluation operator Eval from Equation equation 17 (instantiated using the evaluation module discussed in §E.2). The new triple $(\pi_t^j, \Phi(\pi_t^j), s_t^j)$ is appended to H_t, and the usual acceptance rule is applied: if $s_t^j \leq s_{t-1}^j$, the trajectory reverts to $(\pi_{t-1}^j, \Phi(\pi_{t-1}^j), s_{t-1}^j)$, ensuring non-decreasing scores along each trajectory.

Meta-reflection shared across trajectories. After all J trajectories have been evaluated at iteration t, the algorithm updates the global meta-reflection $R_t = \textsc{MetaReflect}(H_t; M_{\text{critic}})$. Concretely, M_{critic} is prompted to review a subset of the best- and worst-performing ads and prompts in H_t and to summarize the key patterns that distinguish successful creatives from unsuccessful ones (see Figure 7 in Web Appendix F.3 for the meta-prompt). For example, it may infer that scenes with social eating and visible grill marks outperform isolated pack shots, or that emphasizing "satisfying" and "juicy" tends to improve effectiveness. This reflection is then injected into the context of subsequent textual-gradient generation and BEST-OF-N comparisons, allowing trajectories to share accumulated knowledge even as they explore different regions of the prompt space.

Reported outcome. After $T = 10$ iterations, the system reports the sequence of best-performing ads $\{\Phi\left(\pi_t^{j_t^*}\right)\}_{t=1}^{T}$, where $j_t^* = \arg\max_j s_t^j$ is the trajectory that achieved the highest score at iteration t.

F Detailed setup for Section E

F.1 Creative Briefs for the Scenarios

Below are the creative briefs for the eight ad campaigns. Each creative brief outlines the strategic and creative direction for each scenario. The selected campaigns cover a diverse range of products across distinct categories. Each campaign addresses a unique consumer need—whether related to sustainability, luxury, technology, or lifestyle—ensuring comprehensive coverage of various market segments. From environmentally-conscious millennials to affluent, experience-driven travelers, the approach accounts for the full spectrum of consumer interests. By focusing on different consumer needs, the structure enables a holistic marketing strategy that captures both niche and broad demands, providing a well-rounded framework for the study.

F.1.1 Scenario 1: GreenBite Plant-Based Burger

- *Product:* "GreenBite," a new plant-based burger patty.
- *Background:* The market for plant-based food is growing, but many consumers remain skeptical about taste, believing they must sacrifice flavor for health.
- *The Challenge:* Convince flexitarians that GreenBite offers the juicy, satisfying experience of a traditional beef burger with no compromise.
- *Core Insight:* Consumers want to eat better for themselves and the planet, but they fear it means giving up the foods they love.

- *Single-Minded Proposition (SMP):* The delicious, no-compromise burger experience that's better for you and the planet.
- *Reasons to Believe (RTB):* Made with savory pea protein; sears and tastes like real beef; free of soy and GMOs.
- *Desired Response: Think*: "I can finally have a plant-based burger that tastes like the real thing.", *Feel:* Satisfied, vibrant, and guilt-free. *Do:* Purchase GreenBite for their next barbecue.
- *Brand Personality & Tone:* Positive, confident, and mouth-watering.

F.1.2 Scenario 2: AuraSonics X1 Earbuds

- *Product:* "AuraSonics X1," high-end, noise-canceling wireless earbuds.
- *Background:* The rise of hybrid work and urban density has increased daily auditory overload, making focus a precious commodity.
- *The Challenge:* Cut through the saturated audio market by positioning the X1 not as a gadget, but as an essential tool for mental clarity.
- *Core Insight:* In a world of constant noise, true luxury is the ability to control your own soundscape.
- *Single-Minded Proposition (SMP):* AuraSonics X1 creates a personal sanctuary for focus and immersion.
- *Reasons to Believe (RTB):* Adaptive Active Noise Cancellation; studio-quality audio; all-day battery life; minimalist aluminum design.
- *Desired Response: Think:* "This is the solution I need to escape the daily chaos." *Feel:* Calm, empowered, and sophisticated. *Do:* Visit the website to learn more.
- *Brand Personality & Tone:* Minimalist, intelligent, and calm.

F.1.3 Scenario 3: Odyssey E-SUV

- *Product:* The "Odyssey E-SUV," a new all-electric family SUV.
- *Background:* Many families want to switch to electric vehicles but are concerned about sacrificing space, safety, and range for sustainability.
- *The Challenge:* Position the Odyssey E-SUV as the first EV that meets all the practical and adventurous needs of a modern family, making the switch to electric feel like an upgrade, not a compromise.
- *Single-Minded Proposition (SMP):* The future of family adventure is here: all-electric, zero compromise.
- *Reasons to Believe (RTB):* 500km range; top-tier safety rating; spacious three-row seating; all-wheel drive capability.
- *Desired Response: Think:* "An electric car can actually handle my family's active lifestyle." *Feel:* Adventurous, optimistic, and secure. *Do:* Schedule a test drive.
- *Brand Personality & Tone:* Inspiring, capable, and forward-thinking.

F.1.4 Scenario 4: Oasis Eco-Lodge

- *Product:* "Oasis Eco-Lodge," a secluded, luxury resort operating in harmony with its natural surroundings.
- *Background:* The luxury travel market often equates opulence with excess. A growing segment of affluent travelers seeks experiences that are both exclusive and responsible.
- *The Challenge:* Redefine luxury as a seamless integration with nature, promising an experience that is restorative for both the guest and the environment.
- *Core Insight:* For those who have everything, true luxury is not more things, but a deeper connection to something pure and serene.
- *Single-Minded Proposition (SMP):* Rediscover tranquility in a luxury that respects nature.
- *Reasons to Believe (RTB):* Secluded private bungalows; farm-to-table dining with locally sourced ingredients; carbon-neutral operations.
- *Desired Response: Think:* "This is a truly special escape, not just another five-star hotel." *Feel:* Serene, exclusive, and rejuvenated. *Do:* Book a stay.
- *Brand Personality & Tone:* Elegant, peaceful, and understated.

F.1.5 Scenario 5: Momentum Digital Banking

- *Product:* "Momentum," a mobile-first banking app for freelancers and the gig economy.
- *Background:* Traditional banks are not built for the fluctuating incomes and unique business needs of self-employed professionals.
- *The Challenge:* Position Momentum as the essential financial co-pilot for the independent worker, simplifying complexity and providing stability.
- *Core Insight:* Freelancers love their freedom but feel anxious about their financial instability. They need a tool that brings order to their financial chaos.
- *Single-Minded Proposition (SMP):* Banking that is as flexible and entrepreneurial as you are.
- *Reasons to Believe (RTB):* Automated tax-saving tools; integrated invoicing and payment tracking; instant business expense categorization.
- *Desired Response: Think:* "Finally, a bank that gets the way I work." *Feel:* Empowered, organized, and financially confident. *Do:* Download the app and open an account.
- *Brand Personality & Tone:* Modern, empowering, and simple.

F.1.6 Scenario 6: MindGarden Meditation App

- *Product:* "MindGarden," a subscription-based meditation and mindfulness app.
- *Background:* While many are interested in mindfulness for stress-relief, they are often intimidated by the practice, seeing it as difficult or time-consuming.
- *The Challenge:* Make meditation feel accessible and achievable for absolute beginners, removing the barriers of time and perceived difficulty.
- *Core Insight:* People want the benefits of mindfulness but are convinced they don't have the time or ability to practice it correctly.
- *Single-Minded Proposition (SMP):* Find your calm in just 5 minutes a day.
- *Reasons to Believe (RTB):* Guided 5-minute sessions for specific needs (anxiety, focus); progress tracking to build a habit; simple, jargon-free instructions.
- *Desired Response: Think:* "I can do this. 5 minutes is easy." *Feel:* Calm, supported, and hopeful. *Do:* Start a free trial.
- *Brand Personality & Tone:* Gentle, approachable, and encouraging.

F.1.7 Scenario 7: Aeterno Watch

- *Product:* The "Aeterno," a classic, automatic Swiss-made wristwatch with a heritage design.
- *Background:* In an age of smartwatches that are obsolete in a few years, there is a renewed appreciation for objects with permanence and enduring value.
- *The Challenge:* Reassert the relevance of the classic mechanical watch as a symbol of taste, craftsmanship, and timelessness.
- *Core Insight:* In a disposable world, true status comes from owning something permanent that tells a story.
- *Single-Minded Proposition (SMP):* A legacy on your wrist. Craftsmanship that transcends time.
- *Reasons to Believe (RTB):* Swiss-made automatic movement; sapphire crystal glass; timeless design inspired by 1950s classics.
- *Desired Response: Think:* "This is a beautiful object that I will own forever." *Feel:* Prestigious, sophisticated, and discerning. *Do:* Locate an authorized dealer.
- *Brand Personality & Tone:* Elegant, timeless, and confident.

F.1.8 Scenario 8: SyncFlow B2B Software

- *Product:* "SyncFlow," a project management and collaboration software platform for remote teams.
- *Background:* Remote work has increased flexibility but also created challenges in team alignment, communication, and project visibility.
- *The Challenge:* Cut through a crowded SaaS market by focusing on the outcome of "effortless collaboration" rather than just listing features.

- *Core Insight:* Managers don't want another tool to manage; they want the feeling of clarity and momentum that comes from a team working in perfect sync.
- *Single-Minded Proposition (SMP):* Bring your remote team together for effortless collaboration and remarkable results.
- *Reasons to Believe (RTB):* Centralized dashboards with real-time project status; integrated communication channels; automated workflows and reporting.
- *Desired Response: Think:* "This could finally solve our remote work chaos and get everyone on the same page." *Feel:* Organized, in control, and successful. *Do:* Sign up for a team demo.
- *Brand Personality & Tone:* Professional, efficient, and innovative.

F.2 Prompts for Initial Ad Generation

Meta-prompt used for generating prompts for initial 64 ads generation

CREATIVE BRIEF:
`{creative_brief}`

Based on this creative brief, generate a creative, structured, and descriptive prompt for a generative AI model (e.g., Imagen) that will produce a brand-aligned and scroll-stopping advertisement image suitable for an Instagram feed.
A successful prompt must be constructed using the following components and principles:
0. (Most important!) Prompt NEVER includes any description of kids. Kids are not allowed to generate in Imagen 4. Even if creative_brief includes description of "family" or "kids", never include kids in the prompt. Only create adults if humans are included.
1. Key Message (The "Why"): This is the foundational element that guides all other components. It defines the core idea or feeling the ad must communicate. Before writing the rest of the prompt, clearly articulate the message the ad will deliver. This message will act as the "North Star" for all subsequent creative choices. It may be desirable to put the message as the text overlay.
2. Core Components (The "What"):
- This should completely depend on the key message.
- Scene & Environment: Based on the key message, establish a clear, relatable setting that aligns with the brand's lifestyle appeal.
- Action & Narrative: Based on the key message and the scene, describe a dynamic but clear action or interaction to create a sense of a captured moment and tell a micro-story.
- Composition & Framing: Specify the camera shot, angle, and framing (e.g., "low-angle shot," "dynamic medium shot," "close-up on the shoe").
3. Stylistic Qualities (The "How"):
- This should completely depend on the key message and the scene.
- Photography Style: Define the overall aesthetic (e.g., "photorealistic," "cinematic," "professional product photography," "lifestyle action shot").
- Lighting: Be specific about the lighting to set the mood (e.g., "warm golden hour light," "bright morning sun," "dramatic side-lighting").
- Color Palette & Tone: Guide the color scheme and emotional feel (e.g., "vibrant and energetic colors," "empowering and motivational tones," "clean and modern palette").
- Atmosphere & Feeling: Aim to evoke a specific feeling aligned with the brand (e.g., "a feeling of effortless performance," "an atmosphere of vibrant energy," "a sense of supreme comfort").
4. What to Avoid:
- Vagueness: Use specific, descriptive terms instead of "nice" or "good."
- Contradictory Elements: Ensure all elements work together harmoniously.
- Over-stuffing: Focus on a single, clear message without too many competing objects.
5. What to include:
- Logo: create a logo based on creative brief, and naturally place it.
- Text: If you are including a text message, prompt for "negative space for text overlay".

Figure 4: Meta-prompt for generating the prompts for the initial 64 ads.

Worst-of-64 prompt for GreenBite (plant-based burger patty) ad scenario.

Key Message: To visually communicate the core idea: "Experience the juicy, no-compromise satisfaction of a real burger, made better for you and the planet."
Image Generation Prompt: A dynamic, commercial lifestyle photograph for an Instagram feed ad.
Core Components:
* **Scene & Environment:** A stylish, modern outdoor patio during a late afternoon barbecue. The background is softly blurred, showing contemporary patio furniture, lush green potted plants, and the warm glow of string lights just beginning to turn on.
* **Subject & Action:** The focal point is a vibrant, stylish woman in her early thirties, captured in a candid moment of pure satisfaction. She is taking a bite of a huge, delicious-looking plant-based burger. Her eyes are closed in bliss, and a small, genuine smile plays on her lips, conveying an authentic "this is amazing" reaction. Juice from the burger is subtly visible on her hand to emphasize its succulence. She is dressed in a casual but chic linen shirt.
* **Product:** The GreenBite burger is the hero. It's a thick, perfectly seared patty with grill marks, nestled in a glossy brioche bun. It's layered with glistening melted cheddar cheese, crisp green lettuce, a thick slice of ripe red tomato, and caramelized onions spilling out slightly. The burger is partially wrapped in a simple, brown kraft paper sheet which has a small, clean logo on it: "GreenBite" in a modern sans-serif font next to a stylized green leaf.
* **Composition & Framing:** A medium close-up shot, angled slightly upwards to make the moment feel triumphant and empowering. The camera uses a shallow depth of field, keeping the woman's expression and the burger in sharp focus while the background is beautifully out of focus (bokeh). The composition is balanced, with a clear area of soft background on the upper right side, creating negative space for a text overlay.
Stylistic Qualities:
* **Photography Style:** High-end, photorealistic commercial food photography with a warm, authentic lifestyle feel. The image should be incredibly sharp and detailed, especially the texture of the burger patty and the fresh ingredients.
* **Lighting:** Bathed in the warm, dreamy glow of the golden hour. The sunlight comes from the side, creating soft highlights and long, gentle shadows, emphasizing the textures of the food and giving her skin a radiant look.
* **Color Palette & Tone:** A rich and mouth-watering color palette. Saturated and vibrant, but natural. Warm tones of orange, gold, and red from the sun and the food are balanced by the deep greens of the background foliage and the crispness of the lettuce.
* **Atmosphere & Feeling:** The image radiates a feeling of guilt-free indulgence, vibrant well-being, and ultimate satisfaction. The atmosphere is confident, positive, and irresistibly appetizing, making the viewer crave that exact experience.

Best-of-64 prompt for GreenBite (plant-based burger patty) ad scenario.

1. Key Message (The "Why"): To visually capture the peak moment of guilt-free indulgence, proving that a plant-based burger can deliver an unbelievably juicy, satisfying, and no-compromise flavor experience identical to a classic beef burger. The text overlay will read: "The No-Compromise Burger."
2. Core Components (The "What"):
* **Scene & Environment:** A vibrant, sun-drenched backyard barbecue scene, shot with a very shallow depth of field. The background is a beautiful, soft-focus bokeh of friends laughing (adults only, 25-40 age range), string lights, and lush greenery, suggesting a warm, sociable atmosphere without distracting from the main subject.
* **Action & Narrative:** A stylish woman in her early 30s, with a look of pure, blissful satisfaction, is taking the first, epic bite of a burger. Her eyes are closed in enjoyment. Her hands grip the burger firmly but gently. A single, perfect drop of sauce is starting to fall from the side of the burger, caught in mid-air, emphasizing its juiciness.
* **Composition & Framing:** A dramatic, macro close-up, food photography style. The shot is framed tightly on the lower half of the woman's face and her hands holding the burger, making the burger the undeniable hero. The angle is slightly low to give the burger an epic, monumental feel. The right side of the frame has clean, out-of-focus background elements, creating intentional negative space for a text overlay.
3. Stylistic Qualities (The "How"):
* **Photography Style:** Hyper-realistic, professional commercial food photography. Every detail is rendered with extreme clarity: the char marks on the patty, the condensation on the lettuce, the glistening melt of the plant-based cheese, and the texture of the toasted brioche bun. Shot with a high-end DSLR camera and a macro lens.
* **Lighting:** Warm, radiant golden hour lighting. The sun acts as a dramatic backlight, creating a halo effect around the burger and the woman's hands, and creating glistening specular highlights on the juicy patty and sauce. The lighting is bright, positive, and makes the food look incredibly appetizing.
* **Color Palette & Tone:** A mouth-watering and vibrant palette. Rich, savory browns of the seared "GreenBite" patty, deep reds of a ripe tomato slice, vibrant green of crisp lettuce, and a warm golden-yellow of the melted cheese and toasted bun. The overall tone is confident, delicious, and energetic.
* **Atmosphere & Feeling:** The atmosphere is one of ultimate satisfaction and pure, unadulterated foodie bliss. It feels warm, authentic, and completely satisfying, eliminating any doubt about the quality and taste of the plant-based option.
4. Inclusions & Details:
* **The Burger Details:** The GreenBite patty is thick, with a perfect, glistening sear and visible char marks, looking indistinguishable from high-quality ground beef. It's topped with melting vegan cheddar cheese, a leaf of crisp butter lettuce, a thick slice of heirloom tomato, and a creamy aioli sauce.
* **Logo Placement:** The burger is partially wrapped at the base with a small square of branded, unbleached butcher paper. The "GreenBite" logo—a simple, modern green leaf icon next to the word "GreenBite" in a clean sans-serif font—is subtly visible on the paper.
* **Negative Space:** Ensure the composition leaves a clean, uncluttered area of soft-focus background on the right third of the image for text overlay.

Figure 5: Prompts for the best and worst ad among the initial 64 ads generated using the meta-prompt in Figure 4.

F.3 Prompts used for Critic LLM Operators

Best-of-N Gradient steps' pairwise tournament prompt for ads.

contents = [image1, image2, comparison_prompt]
comparison_prompt =
"You are evaluating two advertisement images for mobile Instagram ads. Which image would be more effective at engaging users and driving clicks?
RESPOND WITH ONLY THE NUMBER 1 OR 2. NOTHING ELSE.
1 = first image is better
2 = second image is better "

Figure 6: M_{critic}'s pairwise tournament prompt for digital marketing example.

Meta reflection prompt for ads

You are an expert at analyzing visual patterns in advertising performance.

VISUAL ANALYSIS TASK:
I will show you images from the lowest-scoring and highest-scoring ad iterations. Your task is to identify specific visual patterns that distinguish effective from ineffective ads.

VISUAL EXAMPLES – BEST VS WORST PERFORMING:

RANK 1/10 WORST PERFORMING (Score: 2.14/5.0):
Prompt excerpt: <prompt>...
[Image]

RANK 2/10 LOWER HALF (Score: 2.52/5.0):
Prompt excerpt: <prompt>...
[Image]

. . .

RANK 9/10 UPPER HALF (Score: 3.81/5.0):
Prompt excerpt: <prompt>...
[Image]

RANK 10/10 BEST PERFORMING (Score: 4.15/5.0):
Prompt excerpt: <prompt>...
[Image]

Based on your visual analysis, identify patterns that correlate with higher effectiveness scores:

1. Visual composition and framing differences
2. Lighting conditions and mood variations
3. Color palettes and visual tone patterns
4. Subject positioning and action effectiveness
5. Brand integration approaches
6. Environmental and atmospheric elements

RESPONSE FORMAT:
Provide a structured analysis of visual patterns observed, focusing on what distinguishes high-performing from low-performing ads.

Figure 7: Meta reflection prompt for digital marketing example.

Textual gradient generating meta-prompt for ads (at $t > 0$)

You are an expert at optimizing image generation prompts for advertising effectiveness.

CURRENT PROMPT TO IMPROVE:
`{current_prompt}`

PERFORMANCE ANALYSIS FROM PREVIOUS ITERATIONS:
`{performance_analysis}`

TASK: Based on the performance analysis above, generate specific, actionable improvements to make the current prompt more effective.
Focus on implementing the successful visual patterns identified in the analysis while avoiding the ineffective elements.

Provide **3–5** specific, implementable suggestions for improvement. (It can be addition of a new prompt part, deletion of existing prompt part, or rewriting a prompt part). Each suggestion should reference insights from the performance analysis.

RESPONSE FORMAT:

1. [Specific improvement suggestion based on performance analysis]
2. [Specific improvement suggestion based on performance analysis]
3. . . .

Be concrete and actionable. Focus on changes that will meaningfully improve ad effectiveness based on the analysis.

Textual gradient generating meta-prompt for ads (at $t = 0$)

You are an expert at optimizing image generation prompts for advertising effectiveness.

CURRENT PROMPT TO IMPROVE:
`{current_prompt}`

TASK: Generate specific, actionable improvements to make this ad prompt more effective. Focus on elements that will increase engagement and appeal to the target audience.

Consider improvements in:

1. Visual composition and framing
2. Emotional appeal and messaging
3. Color palette and lighting
4. Subject positioning and action
5. Brand integration and logo placement
6. Text overlay space and readability

Provide **3–5** specific, implementable suggestions for improvement. Each suggestion should be concrete and actionable.

RESPONSE FORMAT:

1. [Specific improvement suggestion based on performance analysis]
2. [Specific improvement suggestion based on performance analysis]
3. . . .

Be concise but specific. Focus on changes that will meaningfully improve ad effectiveness.

Figure 8: Meta prompts for improving ad image-generation prompts (with and without performance analysis).

Meta-prompt for applying textual gradients

You are an expert at revising image generation prompts based on improvement suggestions.

ORIGINAL PROMPT:
`{current_prompt}`

IMPROVEMENT SUGGESTIONS:
`{gradient}`

TASK: Rewrite the prompt incorporating the improvement suggestions while maintaining the core message and structure.

REQUIREMENTS:

- Keep the prompt structure and format similar to the original
- Integrate the improvement suggestions naturally
- Maintain coherence and readability
- Ensure the prompt is optimized for image generation
- Keep the prompt length reasonable (not too long)

OUTPUT CONSTRAINT:
Return ONLY the revised prompt, no explanations or additional text.

Figure 9: Meta-prompt for applying improvement suggestions to revise an image-generation ad prompt.

F.4 Prompt for Evaluation Based on Persona

Persona prompt for simulating ad effectiveness.

SYSTEM: {
You are an AI assistant. Your task is to answer the TASK as if you are the individual described in the 'Persona Profile' (which contains their past survey responses). Remain consistent with the persona's past survey responses and stated characteristics. Carefully follow any instructions provided for the new question, including formatting requirements.
}
PERSONA DATA: {
Which part of the United States do you currently live in?
Question Type: Single Choice
Options:
1 - Northeast (PA, NY, NJ, RI, CT, MA, VT, NH, ME)
2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)
3 - South (TX, OK, AR, LA, KY, TN, MS, AL, WV, DC, MD, DE, VA, NC, SC, GA, FL)
4 - West (WA, OR, ID, MT, WY, CA, NV, UT, CO, AZ, NM)
5 - Pacific (HI, AK)
Answer: 2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)
What is the highest level of schooling or degree that you have completed?
Question Type: Single Choice
Options:
1 - Less than high school
2 - High school graduate
3 - Some college, no degree
4 - Associate's degree
5 - College graduate/some postgrad
6 - Postgraduate
Answer: 3 - Some college, no degree
. . . (Many other survey questions and answers) . . .
Suppose you were given $5 and had to offer to another (anonymous) person a way to split the money. The other person can either accept or reject your offer. If the other person accepts your offer, you would each receive the amount you proposed. If the other person rejects your offer, you would both receive $0. How much would you offer to the other person?
Question Type: Single Choice
Options:
1 - $0
2 - $1
3 - $2
4 - $3
5 - $4
6 - $5
Answer: 3 - $2
. . . (Many other survey questions and answers) . . .
}

AD IMAGE: [image]

TASK:
Return only one item from ["1","2","3","4","5"] for ad effectiveness.
Effective Score Scale Definition:
1: Extremely Unlikely. The persona would actively ignore or be annoyed by this ad.
2: Unlikely. The persona would likely scroll past without a second thought.
3: Mediocre. It is hard to decide whether the personal would click or don't click.
4: Likely. The persona is intrigued and has a good chance of clicking to learn more.
5: Extremely Likely. The persona is the ideal target; a click is almost certain.
No explanation. Just the score.

Figure 10: Prompt for simulating the effectiveness of a given ad-persona combination.

G Appendix for Section 6

G.1 Implementation Details of GEPA for digital ad optimization

Implementation overview GEPA (Agrawal et al. 2025) is an iterative, prompt-optimization-based, self-improving AI method. The key idea of GEPA is to "combine reflection with evolution". Here, an evolution is the same concept as taking a textual gradient in Yuksekgonul et al. (2025), and reflection is the same reflection TextBO ($5) uses. GEPA iterates over three steps in each optimization iteration: Step 1) choosing a prompt and evaluating it; Step 2) updating the meta-reflection; and Step 3) improving the prompt.

GEPA optimizes the prompt for a set of *problem instances*. In standard agentic benchmarks considered in Agrawal et al. (2025), a problem instance is a single task in a benchmark (together with whatever information is needed to evaluate it) that can be stored and revisited. In our digital ad setting, a problem instance corresponds to an individual ad viewer (a persona): each evaluation asks how effective a given ad is for a specific ad viewer (i.e., an ad–ad viewer combination), and the overall score for an ad is computed by averaging effectiveness ratings over a random sample of ad viewers (personas).

Unlike the agentic benchmarks, in ad optimization, ad platforms *do not* have the ability to store and retrieve a particular ad viewer at will. Instead, ad platforms (or managers) test an ad to a population of ad viewers and receive an aggregated evaluation signal (e.g., an average effectiveness rating across a sampled set of personas), which prevents us from maintaining instance-level records, such as which prompts win on which ad viewer (in our case, a persona). Therefore, to apply GEPA for digital ad optimization, we need to adapt it.

For Step 1), in agentic AI experiments considered in §7 and Agrawal et al. (2025)), where we can store, retrieve, or evaluate individual problem instances), it uses instance-level information to (i) maintain a Pareto archive of which prompts in the candidate prompt set currently win on which problem instances, (ii) stochastically sample a non-dominated candidate prompt (favoring candidates that win on more instances), and (iii) evaluate that sampled candidate by rolling it out on a minibatch of problem instances. On the other hand, in the digital ad optimization experiment (§E), where algorithms have access only to aggregated evaluation results, this step's procedure simplifies to that of TextBO: evaluate the prompt across the entire test persona set.

In Step 2), we employ the same meta-reflection procedure as used for TextBO.

For Step 3), we take one step of evolution (or equivalently, one step of textual gradient) of the prompt we evaluated in Step 1) and accept it only if it has a better evaluation than the original prompt. For agentic AI experiments considered in §7 and Agrawal et al. (2025)) where we have access to individual problem instances, when the offspring prompt is accepted, we append it to the candidate prompt set and don't throw away the original prompt from the candidate set. On the other hand, in digital ad optimization in §E, where we cannot maintain the instance-level Pareto archive and thus cannot construct the frequency-weighted, non-dominated sampling distribution, we discard the original prompt from the candidate set. This is because, if we were to keep appending accepted prompts, the candidate prompt set would grow without bound, and the evaluation budget per prompt would be uniformly diluted. Therefore, when the offspring prompt is accepted, we discard the original prompt, keeping the candidate prompt set size constant.

Notably, after adapting GEPA to the digital ad optimization problem setup, where we cannot store and retrieve personas freely, it effectively reduces to TextBO with $G = 1$ and $N = 1$; we describe the pseudocode of the adapted version of GEPA in Algorithm 2.

Shared components with TextBO. To ensure a fair comparison, we keep the non-algorithmic components identical to those used for TextBO (§ E.2): (i) the same scenario-specific initial pool of 64 prompts generated from the creative brief, (ii) the same ad-generation module Φ (Imagen 4), (iii) the same critic model M_{critic} (Gemini 2.5 Flash) and the same meta-prompts for reflection and prompt rewriting, and (iv) the same persona-based evaluation operator $\mathrm{Eval}(\cdot)$ induced by Twin-2k-500 personas and LLM-scored effectiveness ratings.

Candidate prompt pool ("population"). In the ad setting, each prompt π produces a single ad image $\Phi(\pi)$ and a single scalar score $s = \mathrm{Eval}(\Phi(\pi), \mathcal{D}_{\mathcal{X}})$. As we described earlier, we implement

GEPA with a fixed-size candidate pool (population) of size $J = 5$. We initialize this pool using the same WORST5-OF-64 prompts described in §E.3. For each candidate $j \in \{1, \ldots, J\}$ we maintain a triple $(\pi_t^j, \Phi(\pi_t^j), s_t^j)$ where s_t^j is the empirical evaluation score.

Evaluation Each time we evaluate an ad image $\Phi(\pi)$ during training, we follow the same protocol as in § E.2: we randomly sample 200 personas from the *training* persona set and compute the mean effectiveness score across those personas (using the same 1–5 rubric and the same logprob-to-expectation conversion). This yields the scalar feedback used by GEPA for selection. For reporting final performance (plots/tables), we evaluate selected ads on the held-out test personas.

Meta-reflection update. After collecting evaluations, GEPA constructs a global meta-reflection R_t using the same METAREFLECT operator as TEXTBO (§ 5 and § E.2). Concretely, the critic is prompted to compare a subset of the best- and worst-performing creatives observed so far (subject to context-length limits) and to distill recurring patterns into a small set of actionable rules (e.g., "prefer warm lighting," "show social context," "avoid clutter"). This reflection is then used to guide subsequent prompt evolution steps.

Evolution (mutation) step and acceptance rule. Given a current candidate prompt π_t^j and reflection R_t, we generate a single offspring prompt by (i) sampling a targeted textual improvement using the critic (a TextGrad-style local edit (Yuksekgonul et al. 2025)) and (ii) rewriting the prompt accordingly:

$$\delta_t^j \sim \nabla_{\text{text}}(\pi_t^j; R_t, M_{\text{critic}}), \qquad \tilde{\pi}_t^j = \text{Apply}(\pi_t^j, \delta_t^j).$$

We then generate the corresponding offspring creative $\Phi(\tilde{\pi}_t^j)$ and evaluate it to obtain $\tilde{s}_t^j = \text{Eval}(\Phi(\tilde{\pi}_t^j), \mathcal{D}_{\mathcal{X}})$. We use an elitist acceptance rule: if $\tilde{s}_t^j > s_t^j$, the offspring replaces the parent in the pool; otherwise the parent is kept. This keeps the pool size constant across iterations (preventing unbounded growth of candidates under a fixed evaluation budget) while ensuring that each candidate lineage is non-decreasing in its accepted scores.

Full procedure and hyperparameters. As in TEXTBO, we run GEPA for $T = 10$ iterations with pool size $J = 5$ per scenario. At each iteration, we evolve each of the J candidates once (one offspring per candidate), evaluate the offspring using 200 sampled training personas, apply the acceptance rule above, and then update the shared meta-reflection. For comparison plots and tables, we track the best-performing candidate in the pool at each iteration, $j_t^\star := \arg\max_j s_t^j$, and report the corresponding ad $\Phi(\pi_t^{j_t^\star})$.

Algorithm 2: GEPA (adapted for ad optimization)

Input: System Φ; initial prompts $\{\pi_0^j\}_{j=1}^J$; critic M_{critic}; iterations T; pool size J

Output: Best ads $\{\Phi(\pi_t^{j_t^\star})\}_{t=0}^T$, where $j_t^\star = \arg\max_j s_t^j$

for $j = 1$ **to** J **do**
  $c_0^j \leftarrow \Phi(\pi_0^j)$
  $s_0^j \leftarrow \text{Eval}(c_0^j, \mathcal{D}_\mathcal{X})$
$H_0 \leftarrow \{(\pi_0^j, c_0^j, s_0^j)\}_{j=1}^J, \quad R_0 \leftarrow \emptyset$
for $t = 0$ **to** $T-1$ **do**
  $R_t \leftarrow \textsc{MetaReflect}(H_t; M_{\text{critic}})$
  for $j = 1$ **to** J **do**
    $\delta_t^j \leftarrow \nabla_{\text{text}}(\pi_t^j; R_t, M_{\text{critic}})$
    $\tilde{\pi}_{t+1}^j \leftarrow \text{Apply}(\pi_t^j, \delta_t^j)$
    $\tilde{c}_{t+1}^j \leftarrow \Phi(\tilde{\pi}_{t+1}^j)$
    $\tilde{s}_{t+1}^j \leftarrow \text{Eval}(\tilde{c}_{t+1}^j, \mathcal{D}_\mathcal{X})$
    if $\tilde{s}_{t+1}^j > s_t^j$ **then**
      $(\pi_{t+1}^j, c_{t+1}^j, s_{t+1}^j) \leftarrow (\tilde{\pi}_{t+1}^j, \tilde{c}_{t+1}^j, \tilde{s}_{t+1}^j)$
    else
      $(\pi_{t+1}^j, c_{t+1}^j, s_{t+1}^j) \leftarrow (\pi_t^j, c_t^j, s_t^j)$
  $H_{t+1} \leftarrow H_t \cup \{(\pi_{t+1}^j, c_{t+1}^j, s_{t+1}^j)\}_{j=1}^J$
return $\{\Phi(\pi_t^{j_t^\star})\}_{t=0}^T$

G.2 Detailed Experiment Results

TextBO

		Aeterno	Aurasonics	Greenbite	Mindgarden	Momentum	Oasis	Odyssey	Syncflow	Average
WORST5-OF-N		2.942 (0.025)	2.793 (0.023)	2.706 (0.024)	2.722 (0.024)	2.644 (0.024)	2.839 (0.025)	2.828 (0.025)	2.641 (0.023)	2.739
Steps	1	3.098 (0.028)	2.813 (0.022)	3.001 (0.027)	2.980 (0.027)	2.799 (0.027)	3.062 (0.028)	2.943 (0.027)	2.927 (0.023)	2.932
	2	3.125 (0.026)	2.969 (0.027)	3.041 (0.026)	2.980 (0.027)	2.907 (0.025)	3.107 (0.026)	2.943 (0.027)	2.927 (0.023)	2.982
	3	3.125 (0.026)	3.041 (0.025)	3.041 (0.026)	2.980 (0.027)	2.907 (0.025)	3.131 (0.019)	3.063 (0.019)	2.927 (0.023)	3.013
	4	3.194 (0.029)	3.041 (0.025)	3.041 (0.026)	3.132 (0.022)	2.907 (0.025)	3.324 (0.028)	3.063 (0.019)	2.927 (0.023)	3.062
	5	3.194 (0.029)	3.041 (0.025)	3.076 (0.026)	3.132 (0.022)	2.907 (0.025)	3.324 (0.028)	3.063 (0.019)	3.127 (0.022)	3.096
	6	3.257 (0.023)	3.041 (0.025)	3.076 (0.026)	3.132 (0.022)	2.907 (0.025)	3.324 (0.028)	3.070 (0.022)	3.127 (0.022)	3.097
	7	3.257 (0.023)	3.104 (0.028)	3.076 (0.026)	3.132 (0.022)	2.907 (0.025)	3.324 (0.028)	3.107 (0.027)	3.127 (0.022)	3.111
	8	3.257 (0.023)	3.104 (0.028)	3.076 (0.026)	3.132 (0.022)	2.913 (0.026)	3.324 (0.028)	3.107 (0.027)	3.127 (0.022)	3.112
	9	3.257 (0.023)	3.104 (0.028)	3.076 (0.026)	3.132 (0.022)	2.926 (0.026)	3.324 (0.028)	3.107 (0.027)	3.127 (0.022)	3.114
	10	3.257 (0.023)	3.104 (0.028)	3.168 (0.025)	3.132 (0.022)	3.024 (0.025)	3.324 (0.028)	3.107 (0.027)	3.127 (0.022)	3.155
BEST-OF-N		2.982 (0.025)	2.934 (0.027)	2.898 (0.025)	2.930 (0.028)	3.030 (0.027)	3.265 (0.022)	2.966 (0.027)	3.086 (0.023)	3.011

Table 2: Progress of TEXTBO across 10 optimization steps from its starting point (WORST5-OF-N) and comparison with the BEST-OF-N baseline. Each data point, in the mean (standard error) format, indicates the mean and standard error across 411 personas in the test persona set. The last column, labeled "Average", reports the averages of the means across the eight scenarios.

GEPA

		Aeterno	Aurasonics	Greenbite	Mindgarden	Momentum	Oasis	Odyssey	Syncflow	Average
WORST5-OF-N		2.942 (0.025)	2.793 (0.023)	2.706 (0.024)	2.722 (0.024)	2.644 (0.024)	2.839 (0.025)	2.828 (0.025)	2.641 (0.023)	2.739
Steps	1	3.036 (0.025)	2.942 (0.025)	3.029 (0.025)	2.841 (0.024)	2.816 (0.024)	3.151 (0.024)	2.915 (0.025)	2.880 (0.025)	2.939
	2	3.036 (0.025)	2.942 (0.025)	3.046 (0.024)	2.945 (0.023)	2.907 (0.023)	3.151 (0.024)	3.071 (0.028)	2.880 (0.025)	2.992
	3	3.036 (0.025)	2.942 (0.025)	3.046 (0.024)	2.945 (0.023)	2.907 (0.023)	3.182 (0.025)	3.088 (0.026)	2.914 (0.018)	3.003
	4	3.138 (0.025)	3.028 (0.022)	3.046 (0.024)	2.954 (0.024)	2.976 (0.026)	3.187 (0.023)	3.088 (0.026)	2.949 (0.028)	3.033
	5	3.138 (0.025)	3.028 (0.022)	3.046 (0.024)	3.012 (0.022)	2.976 (0.026)	3.187 (0.023)	3.088 (0.026)	3.008 (0.026)	3.049
	6	3.157 (0.023)	3.028 (0.022)	3.046 (0.024)	3.012 (0.022)	2.976 (0.026)	3.209 (0.020)	3.088 (0.026)	3.008 (0.026)	3.052
	7	3.157 (0.023)	3.028 (0.022)	3.046 (0.024)	3.068 (0.027)	2.976 (0.026)	3.209 (0.020)	3.088 (0.026)	3.008 (0.026)	3.060
	8	3.157 (0.023)	3.028 (0.022)	3.046 (0.024)	3.068 (0.027)	2.976 (0.026)	3.209 (0.020)	3.088 (0.026)	3.008 (0.026)	3.060
	9	3.157 (0.023)	3.028 (0.022)	3.046 (0.024)	3.068 (0.027)	2.976 (0.026)	3.278 (0.023)	3.088 (0.026)	3.008 (0.026)	3.070
	10	3.157 (0.023)	3.028 (0.022)	3.046 (0.024)	3.068 (0.027)	2.976 (0.026)	3.278 (0.023)	3.088 (0.026)	3.008 (0.026)	3.070
BEST-OF-N		2.982 (0.025)	2.934 (0.027)	2.898 (0.025)	2.930 (0.028)	3.030 (0.027)	3.265 (0.022)	2.966 (0.027)	3.086 (0.023)	3.011

Table 3: Progress of GEPA across 10 optimization steps from its starting point (WORST5-OF-N) and comparison with the BEST-OF-N baseline. Each data point, in the mean (standard error) format, indicates the mean and standard error across 411 personas in the test persona set. The last column, labeled "Average", reports the averages of the means across the eight scenarios.

Qualitative analysis: So far, we have shown that TEXTBO is able to self-improve and generate ads that align well with the persona distribution within a few iterations. However, one could argue that this can be simply due to a weak set of initial prompts, i.e., a poor set of WORST5-OF-N prompts and corresponding ad images. We examine this alternative explanation by examining the prompts and ad creatives/images.

First, one natural way to rule out this explanation is to compare the quality of the prompts at the starting point and end point of the algorithm, and show that the prompts at the starting point are grammatically correct, meaningful, and sensible and that the differences in the prompts at $t = 0$ and $t = 10$ stem from differences in the message and focus of the prompt rather than grammar/structure. We confirm that this in indeed the case by comparing the prompts at the starting and end points of the algorithm for all the scenarios and trajectories. For example, see the prompt for the WORST-OF-64 prompt for Scenario 1 in Figure 5 in Appendix F.2 and the prompt for TEXTBO at the last step in Figure 14 in Appendix G.4.

A second way to rule this out this alternative explanation is to compare the generated ad images and verify that the differences among ads across algorithms and (over the iterations, within a algorithm) stem from the message and representation, and not from the image quality. To that end, in Figure 11, we show the ad images at the start and end points for both GEPA and TEXTBO, as well the BEST-OF-N baseline (for three scenarios). There is no clear visual quality differences among these ad images; the differences are mainly in the ads' message and how it is represented. This further suggests that the improvements from TEXTBO (and GEPA) over the WORST5-OF-N starting point can be attributed

to alignment, and are not due to trivial quality gains from correcting poor or under-specified initial prompts.

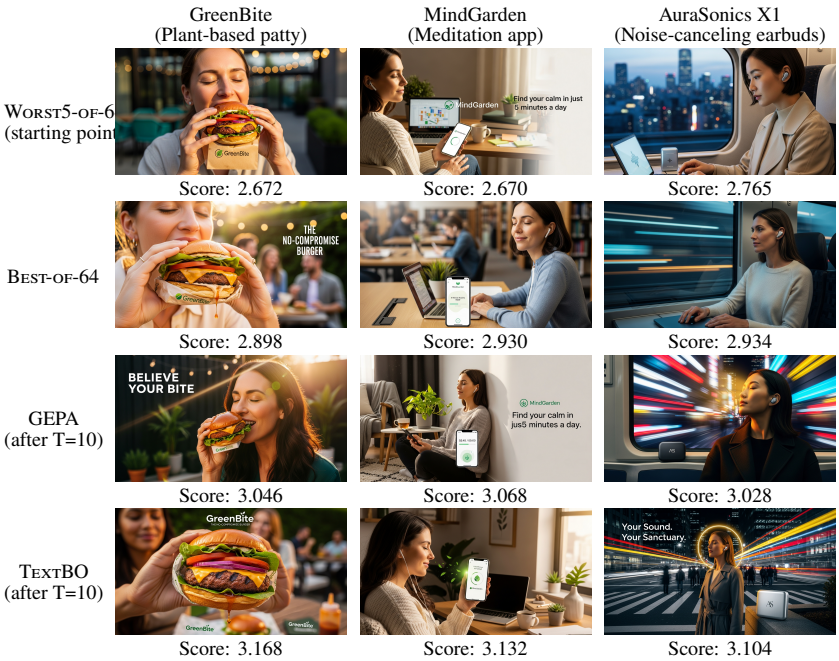


Figure 11: TEXTBO and baselines' generated ad images for three fictional brands: GreenBite burger patty (Scenario 1), MindGarden meditation app (Scenario 6), and AuraSonics noise-canceling earbuds (Scenario 2).

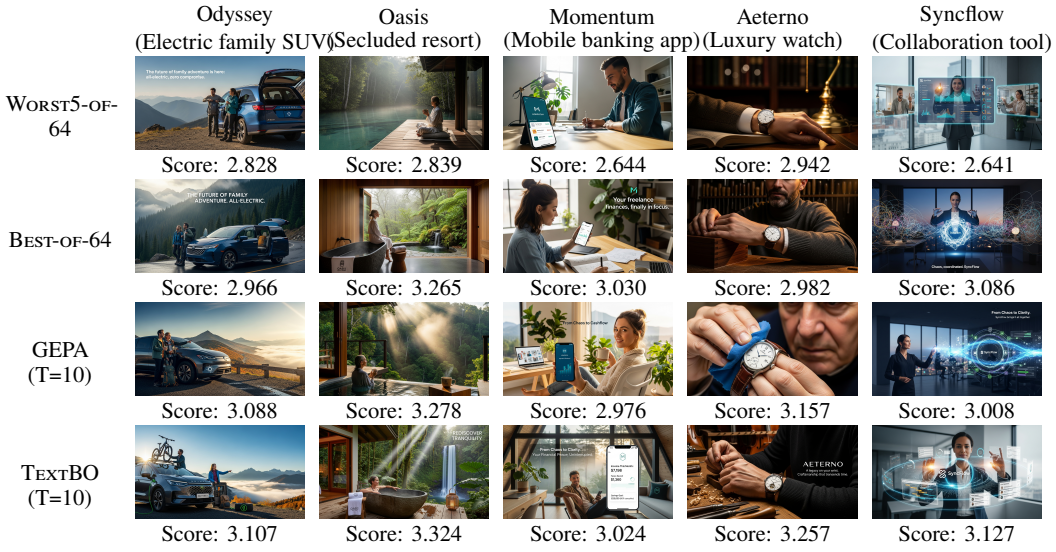


Figure 12: TEXTBO and baseline's generated Ad images for five fictional brands: Odyssey electric family SUV (Scenario 3), Oasis eco-lodge (Scenario 4), Momentum Digital Banking (Scenario 5), Aeterno luxury watch (Scenario 7), and SyncFlow B2B software (Scenario 8)

G.3 Ablation Study

So far, in all our experiments, we used the persona dataset for evaluation and the goal was to optimize the ads serve to the target population captured by the persona dataset (Toubia et al. 2025). However, a

natural question here is whether TextBO can still perform well if we simply use a different target population/evaluator without access to extensive persona information during training. (Recall that each persona consists of a large amount of information, with 500 survey questions accounting for around 50,000 tokens.) Therefore, we run a no-persona ablation experiment in which the evaluator is a single LLM (Gemini 2.5 Flash) with no auxiliary persona context. Concretely, at each iteration, we present only the ad image candidate to a LLM judge (Gemini 2.5 Flash) together with the same 1–5 effectiveness rubric, set the temperature to 0, and compute the same scalar score as the log-probability–weighted expectation over $\{1, \ldots, 5\}$ as before.[10] This collapses the evaluation to a deterministic mapping between ads and scores; it removes heterogeneity from personas and stochasticity from sampling, allowing us to test whether TextBO's performance depends on specific persona information-related setup. To ensure that there is no leakage of preferences, we used GPT-5 instead of Gemini 2.5 Flash for critic model M_{critic} (i.e., the operators used for generating meta-reflection, generating textual gradients, and making the Best-of-N choices.).

We again compare TextBO (J=5 trajectories) against the Best-of-64 initial baseline. The results from this exercise are shown in Figure 13 plots TextBO across 10 optimization steps starting from the Worst5-of-64 prompts under the ablation setup.[11] As before, we see that TextBO outperforms Best-of-64 baseline from optimization step 3 onwards, and the overall trend in Figure 13 is similar to that in Figure 2. The main difference is that the evaluation scores without persona information are more optimistic than scores with persona information for both TextBO and Best-of-64. This is understandable, since Peng et al. (2025) find that persona information does not significantly improve LLMs' overall ability to predict preferences and mainly adds variance to the predictions. Since variance makes the learning problem harder, it is natural that the main experiment, which uses personas, shows slower improvement than the pure-LLM case here. Overall, these findings suggest that when learning the target population's preferences is difficult, the method may require more iterations and larger training samples to optimize effectively.

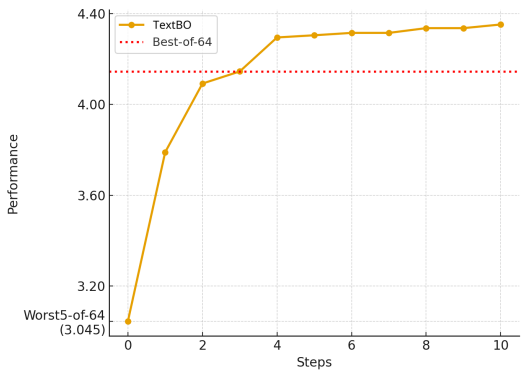


Figure 13: Ablation study with TextBO with $J = 5$ and Best-of-64 for Gemini 2.5 Flash as the evaluator, with temperature 0. The performance score is the average of eight scenarios' mean evaluation score at that step.

[10] With temperature = 0, the judge's output is deterministic for a fixed image and rubric; the logprob vector is still available, so the expected score is stable across repeated calls.

[11] Here, Worst5-of-64 and Best-of-64 are different from those of the main experiments, as the preference distribution is different.

G.4 Final Prompts in TextBO and GEPA

Prompt for generating the TextBO ($T = 10$) ad for GreenBite (plant-based burger patty)

"**Key Message:**
"The No-Compromise Burger"—to visually prove that a plant-based burger can be just as juicy, delicious, and satisfying as a traditional beef burger, making it the hero of the ultimate backyard BBQ.
*Core Components:**
* **Scene & Environment:** A vibrant, meticulously designed modern backyard patio, perfect for a relaxed, aspirational late-afternoon get-together. The background is a soft-focus composition of stylish outdoor furniture, lush green garden plants, and subtly visible, engaged friends enjoying themselves and partaking in the shared experience. Occasional hints of other complementary gourmet side dishes or refreshing beverages on the blurred patio table enhance the sense of a complete, desirable BBQ experience, creating a rich, warm, and inviting social atmosphere.
* **Action & Narrative:** A pair of adult hands are holding up a perfectly constructed, gourmet-style cheeseburger. The action captures either: (A) a moment of serene, authentic anticipation just before the first bite, with the subject's face (in soft focus) conveying genuine contentment or subtle bliss, possibly with eyes gently closed, savoring the aroma; OR (B) the burger held out slightly towards the camera, inviting the viewer into the experience, with the subject's soft-focus expression conveying a shared, authentic delight or playful 'aha!' moment. The emphasis is on genuine human emotion and connection, avoiding any forced, exaggerated, or unnatural expressions (especially if a bite is implied). A single, glistening drop of rich, savory juice is visible, about to fall from the patty, powerfully emphasizing its irresistible succulence and indulgent quality without appearing messy. * **Central Subject:** The star of the image is the GreenBite burger. It's a thick, plant-based patty with a deeply browned, flawlessly seared crust and visible grill marks, looking indistinguishable from premium ground beef. It's topped with a layer of glistening, perfectly melted cheddar cheese, a crisp piece of bright green leaf lettuce, a thick, ruby-red slice of a heirloom tomato, and a few rings of sharp red onion, all between a fluffy, toasted brioche bun. * **Composition & Framing:** An enticing, slightly low-angle medium close-up shot where the GreenBite burger is the undeniable hero, prominently filling a significant portion of the frame and drawing the viewer's eye directly to its appetizing qualities. The focus is tack-sharp on the GreenBite patty's seared texture and the glistening juice. The hands holding the burger, and the subject's face (if present), along with the backyard scene, are rendered in a beautiful, soft-focus bokeh, creating a strong emotional connection and aspirational appeal while ensuring the burger remains supremely appetizing. There is significant negative space with soft, blurred greenery in the upper portion of the frame for a text overlay.
Stylistic Qualities:
* **Photography Style:** High-end, photorealistic commercial food photography with a cinematic lifestyle feel. The image should be incredibly crisp, detailed, and textured. * **Lighting:** Bathed in the warm, dreamy glow of golden hour sunlight. The light source comes from the side, casting soft shadows that accentuate the burger's shape and highlight the glistening textures of the melted cheese and juicy patty. A slight, elegant lens flare bleeds into the frame, adding to the warm, aspirational feel. * **Color Palette & Tone:** Rich, saturated, and mouth-watering. A color palette dominated by the warm browns of the seared patty, the vibrant orange of the cheese, the fresh greens of the lettuce, and the deep reds of the tomato, all set against the warm, earthy tones of the background.
* **Atmosphere & Feeling:** Evokes a powerful feeling of shared satisfaction, vibrant well-being, and guilt-free indulgence. The atmosphere is confidently positive, supremely delicious, and distinctly aspirational, conveying a desirable lifestyle of effortless enjoyment. **Brand Elements:** *
Logo: The simple, modern "GreenBite" logo (clean sans-serif font with a subtle leaf element) is integrated prominently yet naturally within the scene. It is elegantly printed on the branded wax paper partially wrapping the bottom of the burger, and also subtly visible on a branded napkin or a stylish sauce bottle on the blurred patio table in the background, reinforcing brand presence contextually within the aspirational setting. * In addition to the physical logo placements, a clean, prominent text overlay of the 'GreenBite' brand name, and optionally a relevant tagline (e.g., 'Believe Your Bite'), should be integrated into the significant negative space in the upper portion of the frame. This overlay should be highly legible and reinforce clear brand recognition without detracting from the visual appeal."

Figure 14: Prompts for TextBO ($T = 10$)

Prompt for generating the GEPA ($T = 10$) ad for GreenBite (plant-based burger patty)

"**Key Message:**
Visually communicate the "no-compromise" promise of GreenBite by capturing the exact moment of pure, blissful satisfaction of a consumer realizing a plant-based burger can be just as juicy and delicious as its beef counterpart.
Core Components:
* **Scene & Environment:** A vibrant, sun-drenched modern backyard patio during a relaxed get-together of friends in their late 20s and early 30s. The setting is stylish and aspirational, featuring a natural wood deck, comfortable outdoor furniture, festoon string lights, and lush, green potted plants in the background.
Action & Narrative: The central focus is a woman with her eyes open, full of genuine, surprised satisfaction, capturing the peak moment of consumption. Her mouth is wide open, actively biting into the fully-loaded GreenBite burger. The patty is clearly visible within her mouth, emphasizing the powerful 'A-ha!' moment of discovery and the sheer indulgence of the bite, directly communicating the 'no-compromise' promise.
* **Composition & Framing:** Dynamic, slightly low-angle medium shot focusing on the woman and the burger. The burger is the hero; the depth of field is shallow, rendering the juicy, perfectly seared patty, melted plant-based cheddar, crisp lettuce, and glossy brioche bun in hyper-detailed, sharp focus. The background is beautifully out of focus with a pleasing bokeh effect. There is negative space in the upper left corner suitable for a text overlay.
Stylistic Qualities: * **Photography Style:** High-end, commercial lifestyle food photography. Polished, authentic, and incredibly photorealistic. Looks like an image from a top-tier food publication. * **Lighting:** Bathed in warm, cinematic golden hour sunlight. The light creates long, soft shadows and a beautiful, inviting glow, highlighting the textures of the food. Crucially, integrate a strong, warm sun flare or haze into the scene, creating a dramatic, aspirational glow that enhances the sense of satisfaction and makes the overall atmosphere feel more vibrant and special. * **Color Palette & Tone:** Rich, warm, and saturated colors. Deep browns of the seared patty, vibrant greens of the lettuce and surrounding plants, warm yellows and oranges from the cheese and sunlight. The overall tone is confident, energetic, and mouth-watering.
* **Atmosphere & Feeling:** A powerful feeling of guilt-free indulgence and vibrant satisfaction. The atmosphere is warm, happy, and effortlessly cool, capturing the joy of sharing great food with friends. **Inclusions:**
* **Logo:** Ensure the 'GreenBite' logo is subtly yet clearly visible on the burger wrapper, or on a small, non-distracting element within the scene, to provide product context without pulling focus from the main action.
* **Text Overlay:** A prominent, clean white text overlay, featuring a strong, benefit-driven headline such as 'THE NO-COMPROMISE BURGER' or 'BEYOND EXPECTATION.', is strategically placed in the negative space (e.g., upper left corner). This headline should be clearly legible, visually impactful, and act as a primary visual element, prioritizing the core message and value proposition over just generic brand reinforcement."

Figure 15: Prompts for GEPA ($T = 10$)

H Experiments for Agentic AI Benchmarks

In §6, we saw that TextBO significantly outperforms GEPA for ad optimization experiments. Recall that the adapted version of GEPA in the ads settings effectively reduced to TextBO with $N = 1$ and $G = 1$. As such, one of the main takeaways from those experiments is that Best-of-N gradient sampling and multiple gradient steps can effectively improve the performance of self-improving AI under limited evaluation budget.

The natural follow-up question is whether Best-of-N gradient sampling and multiple gradient steps will also improve the performance of standard GEPA used in the agentic AI benchmarks considered in Agrawal et al. (2025). Therefore, we now present a series of experiments, where we augment GEPA with the key theoretical and algorithmic innovations from our approach – Best-of-N gradient sampling and multiple gradient steps (as described in §4) . In our augmented version of GEPA, denoted as TextBO-GEPA, at each iteration, we take $G = 5$ gradient steps, and at each step, we employ Best-of-N textual gradients with $N = 5$. (In standard GEPA, each time we improve a prompt, we only take a single textual gradient step.) See Web Appendix §H.2 for details.

H.1 Experiment Design

To rigorously evaluate GEPA and TextBO-GEPA, we closely replicate three agentic AI benchmarks and their corresponding experimental setups from Agrawal et al. (2025).[12] In all the experiments, we used Qwen3-8B (Yang et al. 2025), the same model and setup used in Agrawal et al. (2025) (decoding temperature of 0.6, top-p of 0.95, and top-k of 20 for training as well as inference). As reported in Agrawal et al. (2025), GEPA saturates after $7,500$ evaluations (rollouts), so we conduct experiments up to $7,500$ evaluations.

We now briefly describe the three benchmark tasks:
HotpotQA Yang et al. (2018) provides HotpotQA, a set of diverse, explainable multi-hop question answering tasks comprising approximately 113k Wikipedia-based examples. This task requires reasoning over multiple supporting documents and providing sentence-level supporting facts to enable intense supervision. We follow the implementation details described in Agrawal et al. (2025), except for the retriever module, which was not specified in the paper. So we conservatively set it to BM25 (Robertson et al. 2009), one of the most widely used retrievers. We use 150 examples for training, 300 for validation, and 300 for testing.

HoVer Jiang et al. (2020) provides HoVer, a set of many-hop fact-extraction and claim-verification tasks, constructed from the same 113k Wikipedia-based examples used in HotpotQA. It tests complex reasoning that requires aggregating evidence from multiple sentences across disparate documents, often involving diverse reasoning graphs and long-range dependencies. Again, we follow the implementation details described in Agrawal et al. (2025), except for the retriever module, which was again not specified in the paper, and was set to BM25 (Robertson et al. 2009). Again, we use 150 examples for training, 300 for validation, and 300 for testing.

PUPA Li et al. (2025b) provides PUPA, a set of privacy-conscious delegation tasks which address real-world user queries by orchestrating an ensemble of trusted and untrusted models. The primary objective is to maintain a high response quality, comparable to that of proprietary models, while minimizing exposure of personally identifiable information (PII) to the untrusted backend. Again, we follow the implementation details described in Agrawal et al. (2025), except that we use case-insensitive substring matching instead of LLM-as-a-judge to check whether each PII exists in the output. We use 111 training examples, 111 for validation, and 221 for testing.

H.2 Implementation Details of GEPA and TextBO-GEPA for agentic AI experiments

GEPA GEPA (Agrawal et al. 2025) is an iterative, prompt-optimization-based, self-improving AI method. The key idea of GEPA is to "combine reflection with evolution". Here, an evolution is the same concept as taking a textual gradient in Yuksekgonul et al. (2025). GEPA iterates over three steps in each optimization iteration: Step 1) choosing a prompt and evaluating it; Step 2) updating the meta-reflection; and Step 3) improving the prompt.

[12] Agrawal et al. (2025) also included IFBench Pyatkin et al. (2025), for which they report that none of the algorithm show improvement. Therefore we exclude it in our analysis.

For Step 1), it uses instance-level information to (i) maintain a Pareto archive of which prompts in the candidate prompt set currently win on which problem instances, (ii) stochastically sample a non-dominated candidate prompt (favoring candidates that win on more instances), and (iii) evaluate that sampled candidate by rolling it out on a minibatch of problem instances.

In Step 2), GEPA updates a global meta-reflection R_t that summarizes what has tended to work well or poorly so far. Concretely, after collecting rollouts and their scores, it applies a reflection operator to the accumulated history, $R_t := \texttt{MetaReflect}(H_t; M_{\text{critic}})$, where the critic is shown a context-limited subset of representative high-performing and low-performing rollouts (including the prompts and outcomes) and asked to (i) identify recurring success patterns and failure modes and (ii) distill them into a short list of actionable, prompt-level guidelines. This shared reflection is carried forward and provided as context to subsequent prompt-evolution (textual-gradient) steps. This step is fundamentally the same as TextBO's meta-reflection steps.

For Step 3), we take one step of evolution (or equivalently, one step of textual gradient) of the prompt we evaluated in Step 1) and accept it only if it has a better evaluation than the original prompt. When the offspring prompt is accepted, we append it to the candidate prompt set without removing the original prompt from the candidate set.

TextBO-GEPA TextBO-GEPA is identical to GEPA except for the prompt-improvement step (Step 3): whenever GEPA would take a single evolution step (one textual-gradient update) to propose an offspring prompt, TextBO-GEPA instead performs multiple successive Best-of-N textual-gradient steps (as in §4) before evaluating and applying the same acceptance rule as GEPA. That is, at each inner step, it samples N local textual edits, applies them to form N candidate prompts, generates the corresponding candidate outcomes, and uses the critic (conditioned on the current reflection) to select the most promising candidate; repeating this for several steps yields a refined offspring prompt, which is then accepted and appended to the candidate prompt set only when it improves upon its parent under the evaluation procedure.