Prompt Smart, Pay Less: Cost-Aware APO for Real-World Applications

Jayesh Choudhari*
Viator, TripAdvisor
London, UK

Piyush Kumar Singh*
Viator, TripAdvisor
London, UK

Douglas McIlwraith
Viator, TripAdvisor
London, UK

Snehal Nair
Viator, TripAdvisor
London, UK

Abstract

Prompt design is a critical factor in the effectiveness of Large Language Models (LLMs), yet remains largely heuristic, manual, and difficult to scale. This paper presents the first comprehensive evaluation of Automatic Prompt Optimization (APO) methods for real-world, high-stakes multiclass classification in a commercial setting, addressing a critical gap in the existing literature where most of the APO frameworks have been validated only on benchmark classification tasks of limited complexity.

We introduce APE-OPRO, a novel hybrid framework that combines the complementary strengths of APE and OPRO, achieving notably better cost-efficiency, around 18% improvement over OPRO, without sacrificing performance. We benchmark APE-OPRO alongside both gradient-free (APE, OPRO) and gradient-based (ProTeGi) methods on a dataset of 2,500 labeled products.

Our results highlight key trade-offs: ProTeGi offers the strongest absolute performance at lower API cost but higher computational time as noted in [6], while APE-OPRO strikes a compelling balance between performance, API efficiency, and scalability. We further conduct ablation studies on depth and breadth hyperparameters, and reveal notable sensitivity to label formatting, indicating implicit sensitivity in LLM behavior. These findings provide actionable insights for implementing APO in commercial applications and establish a foundation for future research in multi-label, vision, and multimodal prompt optimization scenarios.

CCS Concepts

• **Computing methodologies** → **Natural language generation**.

Keywords

LLM, Prompt Optimization, APE, OPRO, ProTeGi, CoT, APE-OPRO

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD 2025 Workshop on Prompt Optimization, Toronto, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3690624

ACM Reference Format:
Jayesh Choudhari, Piyush Kumar Singh, Douglas McIlwraith, and Snehal Nair. 2025. Prompt Smart, Pay Less: Cost-Aware APO for Real-World Applications. In *Proceedings of First International KDD Workshop on Prompt Optimization, 2025 (KDD 2025 Workshop on Prompt Optimization).* ACM, New York, NY, USA, 37 pages. https://doi.org/10.1145/3690624

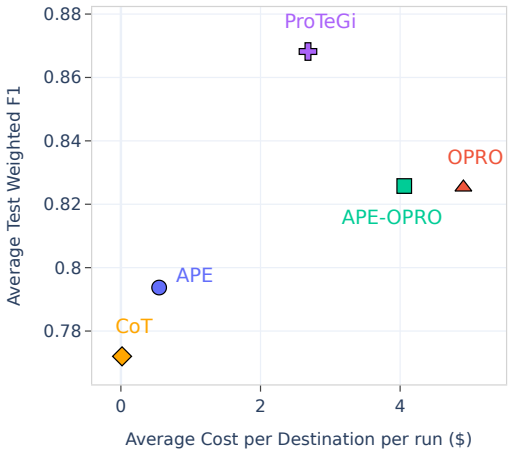

Figure 1: Comparison of evaluated methods based on Mean Test Weighted F1 and Mean Cost per destination. Results are averaged over five runs per destination, spanning 2,500 products across 10 destinations and each run involves an optimization process with 10 iterations and 10 prompts per iteration using GPT-4.1 (optimizer) and GPT-4o-mini (scorer). See Section 3 for details on cost computation.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing across tasks such as summarization, question answering, and classification. However, when used in a prompt-based setting, their real-world performance is highly sensitive to the phrasing and structure of input prompts. Designing effective prompts is often a manual, heuristic-driven process that lacks reproducibility and scalability. Zero-shot prompting, while attractive

for its simplicity, frequently falls short of production-grade performance without prompt optimization. To address the challenge of effective prompt design, recent studies have introduced automatic prompt optimization (APO) techniques that iteratively refine prompts using model feedback or search strategies. Although these approaches have shown promising results, they have primarily been evaluated on tasks such as binary classification or short-form output, often using benchmark datasets with limited complexity. To the best of our knowledge, this work represents the first comprehensive evaluation of APO methods in a complex, real-world single-label multiclass classification setting.

Building on this foundation, our study presents a systematic investigation of prompt optimization strategies for a commercially impactful classification task at Viator (a global platform that aggregates and sells curated travel experiences). We evaluate a range of APO methods and propose a novel hybrid approach APE-OPRO, that integrates the strengths of APE's exploration with OPRO's metaprompt-based approach, and have performed a detailed cost-performance analysis on a curated dataset of 2,500 manually labeled products in the top 10 globally diverse destinations. The techniques explored in this work fall into two main approaches:

- **Gradient-Free Methods** have gained significant attention in recent literature, with several promising approaches such as Dual-Phase Accelerated Prompt Optimization [12], GPO [7], and PE2 [13]. In this work, we focus on two widely adopted gradient-free methods—APE [17] and OPRO [11]. These serve as strong baselines and exemplify contrasting strategies for prompt optimization in the absence of gradient information.
- **Gradient-Based Methods** have also attracted substantial interest, with approaches such as ProTeGi [6], TextGrad [14], and GREATER [2] advancing differentiable prompt optimization. In this study, we specifically focus on ProTeGi [6], which offers a robust and scalable gradient-based framework that aligns well with our experimental design and evaluation objectives.

In addition to gradient-based and gradient-free strategies, there exists a third category of optimization methods inspired by evolutionary algorithms. Evolutionary methods explore the prompt search space using biologically inspired mechanisms such as mutation, selection, and recombination. Notable examples include EvoPrompt [8], PromptBreeder [3] and SPRIG [16]. As noted in the OPRO paper [11], EvoPrompt's optimization trajectory tends to be less stable compared to OPRO, primarily due to its reliance on limited prompt history and the absence of task exemplars during refinement. Consequently, we do not include these methods in our experimental evaluation.

Beyond these methods, there also exist alternative optimization strategies that do not fit neatly into the gradient-based, gradient-free, or evolutionary frameworks. For example, PromptAgent [9] employs Monte Carlo Tree Search to explore and select optimal prompt sequences. CRISPO [4] and SCULPT [5] rely on iterative refinement guided by model generated critiques and suggestions. PREFER [15] adopts an ensemble based feedback-reflect-refine loop to enhance prompt quality. While promising, these strategies are considered complementary and are not included in our current evaluation.

1.1 Contributions

Our work makes the following key contributions:

- **Pushing Beyond Binary: Real-World Evaluation of APO in Multiclass Classification**: We introduce a challenging real-world task of single-label, multiclass classification using our proprietary dataset of 2,500 high-revenue products across ten globally diverse destinations. This departs from the binary classification focus prevalent in most prior work, directly addressing the need for such evaluation highlighted as one of the limitations in [6].
- **APE-OPRO: A hybrid APO framework**: We propose APE-OPRO, a novel hybrid approach that combines APE [17] and OPRO [11]. APE-OPRO matches OPRO's performance while significantly reducing API costs by $\sim 18\%$ as shown in figure 1.
- **Cost-Performance Trade-offs in Prompt Optimization**: We perform a comprehensive cost-sensitive analysis of diverse APO approaches including gradient-based (ProTeGi) and gradient-free (APE, OPRO) and highlight their trade-offs between weighted F1 performance and API cost. Notably, ProTeGi delivers strong cost efficiency despite requiring longer execution time compared to other algorithms.
- **Revealing Implicit LLM sensitivity**: We uncover a previously unreported sensitivity in APE where prompt performance is sensitive to label formatting, emphasizing how the stochastic behavior of large language models can impact automated prompt optimization.

Although this study focuses on single label classification, we outline a generalizable framework for cost-aware APO evaluation and identify promising research frontiers in multi-label, multimodal, and multilingual prompt optimization.

2 Problem Formulation

To improve product discoverability on Viator, we initially adopted a multi-label taxonomy in which individual products could be assigned multiple, overlapping categories (e.g., a tour labeled both 'Historical' and 'Cultural'). While this approach improved overall visibility, it also caused popular products to appear in many categories, which reduced the diversity of products shown to users and limited exposure for less prominent yet relevant offerings. To address this issue, we transitioned to a single label, multiclass taxonomy, where each product is assigned one mutually exclusive label from a predefined set. This structure reduces cognitive load and facilitates easier comparison across distinct categories. In contrast to traditional hierarchical systems, our flat, destination-specific taxonomy is better aligned with how users naturally explore travel experiences and contributes to improved product discoverability, as illustrated in Figure 2.

We frame this task as a single-label classification problem specific to each destination d , where, for a given product description $x \in \mathcal{D}$, the objective is to assign the most appropriate label from a defined, mutually exclusive set $C_d = \{c_1^d, c_2^d, \ldots, c_{N_d}^d\}$.

Rather than relying on conventional supervised training approaches due to limited ground truth, we frame the problem as a prompt optimization task using Large Language Models. Our goal is to find an optimal natural language prompt $p^* \in \mathcal{P}$ such that the

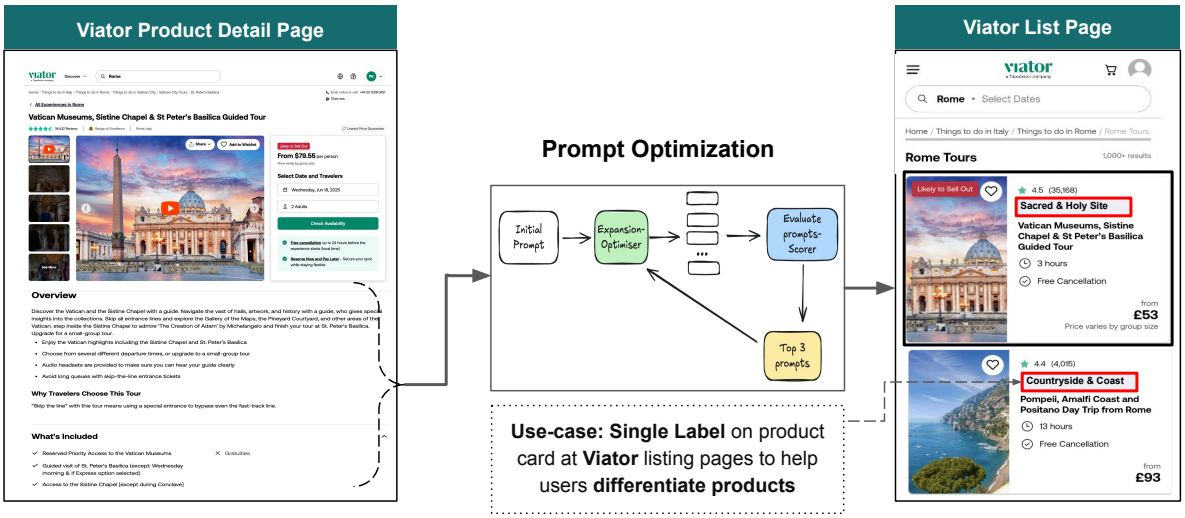

Figure 2: Workflow for single-label classification of Viator products. Product text from Viator's product description pages is passed to a prompt optimization method, which generates a single, destination-specific label to improve product differentiation on listing pages.

LLM, when conditioned on p^* and input x, produces the correct label. Formally, we seek:

$$p^* = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{\mathcal{D}' \sim \mathcal{D}} \left[f(p, \mathcal{D}') \right], \tag{1}$$

where $f(p, \mathcal{D}')$ is an evaluation metric (e.g., weighted f1) calculated over a sample $\mathcal{D}'$ of the training data, assessing the LLM performance using prompt p.

To solve this optimization problem, we investigate several APO techniques, categorizing them as gradient-free (APE [17], OPRO [11]) and gradient-based (ProTeGi [6]). We also introduce a novel hybrid method, APE-OPRO, combining APE's initialization with OPRO's metaprompt guided prompt generation. We unify these methods within a general prompt optimization framework as described in Figure 3.

This framework operates iteratively, starting with an initial prompt. Each iteration involves an **expansion phase**, where an *optimizer model* generates multiple candidate prompts (defining the **width**). These candidates are then evaluated using a computationally less expensive *scorer model*. The top ($k = 3$) performing prompts, are selected and fed back into the optimizer for the next iteration. This cycle continues for a fixed number of iterations, defined as the **depth**. The framework is flexible enough to encapsulate the explored APO strategies. This modular design, which separates prompt generation from evaluation, facilitates experimentation across different techniques.

Each destination uses a custom template for the prompt where the candidate label set C_d is inserted into a fixed template A.5.1 to preserve semantic relevance and improve performance.

3 Cost Computation

We compute the cost of APO for each destination as follows:

- N be the total number of iterations,
- M be the number of prompts generated per iteration,

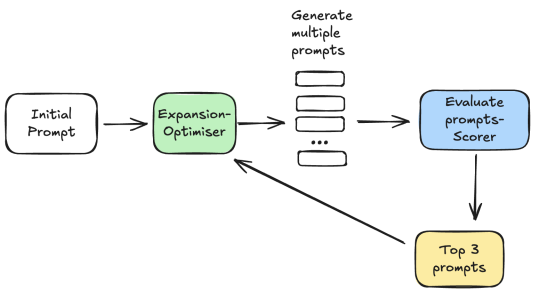

Figure 3: Prompt Optimization Framework

- $C^{\text{op}}_{\text{in}}(i)$ and $C^{\text{op}}_{\text{out}}(i)$ be the input and output token costs for the optimizer model (expansion phase) in iteration i (prompt generation),
- $C^{\text{sc}}_{\text{in}}(i, j)$ and $C^{\text{sc}}_{\text{out}}(i, j)$ be the input and output token costs for scoring (scoring phase) the j-th prompt in iteration i.

Then, the total cost[1] for one run is:

$$\text{Total Cost} = \sum_{i=1}^{N} \left[C^{\text{op}}_{\text{in}}(i) + C^{\text{op}}_{\text{out}}(i) + \sum_{j=1}^{M} \left(C^{\text{sc}}_{\text{in}}(i, j) + C^{\text{sc}}_{\text{out}}(i, j) \right) \right] \tag{2}$$

For each method (APE, OPRO, APE-OPRO, and ProTeGi), the total cost is computed as the cumulative sum of input and output tokens across all optimization iterations, as defined in Equation 2. As shown in Figure 3, each iteration includes two stages: prompt expansion (handled by optimizer model) and prompt evaluation (handled by scorer model). Unless otherwise stated, we use GPT-4.1 for optimization and GPT-4o-mini for scoring[2].

[1]Reported costs do not account for potential savings from prompt caching or batch API calls in offline scenarios

[2]https://openai.com/api/pricing/

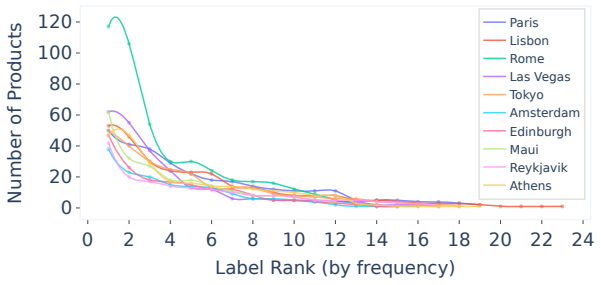

Figure 4: Distribution of labels in Top 10 Destinations

4 Experiment Setup

4.1 Dataset

Our dataset comprises proprietary Viator travel experiences from ten diverse global destinations, selected to capture a wide range of user interests and destination types. These include major European cities (Rome, Paris, Lisbon, Amsterdam, Edinburgh, Athens), Asia (Tokyo), North American entertainment centers (Las Vegas), and island or remote destinations (Maui, Reykjavik), offering geographic and cultural diversity.

Table 1: Overview of the dataset split by destination, including train/test size, number of unique labels, and average token count per product description

Destination	Train	Test	#Labels	Avg. #Tokens
Rome	52	390	14	190
Lisbon	75	213	23	180
Amsterdam	50	109	16	185
Paris	73	229	19	189
Las Vegas	59	186	17	179
Athens	58	185	19	198
Tokyo	65	190	18	183
Edinburgh	51	125	15	179
Reykjavik	63	114	17	183
Maui	48	166	12	173

4.1.1 ***Data Collection and Preprocessing.*** For each destination, we extracted product information from Viator's product description as illustrated in Figure 2. Each entry in the experience contains an average of 185 tokens of descriptive text.

4.1.2 ***Labeling Process.*** To initiate the taxonomy refinement, we manually labeled a subset of top ranked products that cover 80% of the revenue within each destination, focusing on popular and representative experiences. This revenue-weighted sampling approach ensures that our taxonomy captures economically significant experience categories while maintaining destination-specific relevance. Two domain experts independently labeled each experience, achieving an interannotator agreement of 87. 3% (Cohen's κ = 0.81), with disagreements resolved by discussion.

We developed a flat labeling taxonomy (a non-hierarchical set of 66 mutually exclusive categories) to classify travel experience products across 10 major destinations. The number of labels varies by destination (from 12 for Maui to 23 for Lisbon), reflecting the diversity of available experiences. Some labels are destination-specific (e.g., "Volcano Tours" for Reykjavik), while others are universal (e.g., "City Highlights").[3]

4.1.3 ***Dataset Characteristics.*** As illustrated in Figure 4, the distribution of labels follows a long tail pattern, where a few labels represent a large proportion of products, typically representing highly popular and destination-specific activities, while many other labels correspond to niche or very common offerings. For example, in Rome, "Sacred & Holy Site" accounts for 26.47% and "Countryside & Coast" represents 8.82% of all experiences, while in Las Vegas, "Adventure" represents 19%, "City Highlight" represents 3.24%.

The "Other Experiences" label is included as an additional candidate label across all destinations during the training stage to handle products associated with rare or previously unseen categories. When the label is assigned, the product is flagged for manual review by Destination Managers, who periodically assess whether a new dedicated label should be introduced based on product volume.

4.1.4 ***Destination Selection Criteria for Evaluation.*** To provide a representative evaluation of prompt optimization methods, we focus on three key destinations: Rome, Amsterdam, and Lisbon. These were chosen to capture a range of product diversity and complexity in the labeling. Rome represents a destination with well defined tourist attractions and relatively clear categorization. In contrast, Amsterdam offers moderate complexity due to its diverse range of experience types, while Lisbon presents the most challenging case with the highest number of distinct labels (23).

4.2 Train-Test Split

To ensure consistency across methods, we adopt a standardized train-test split strategy. We perform stratified sampling by randomly sampling up to four examples per label within each destination to construct the training set. The remaining examples are reserved for evaluation in the test set. Using four examples per label ensures sufficient representation during training while also minimizing the risk of labels appearing only in the training set and not in the test set, leading to a fairer evaluation. All methods are trained exclusively on the training set and the test set is strictly held out and used only for final evaluation. The number of training and test instances after the split is summarized in Table 1.

4.3 Prompt Optimization Configuration

All methods are executed under a uniform configuration of hyperparameters unless otherwise stated. For APE, OPRO, and APE-OPRO, we fix the optimization depth (number of iterations) to 10, and the breadth (number of prompts generated per iteration) to 10.

For ProTeGi, we similarly set the *search-depth* to 10, aligning it with the other methods for comparability. The *max-expansion-factor*, which controls the number of prompt candidates passed to the selection stage in each iteration, is also set to 10. This parameter

[3]Due to proprietary constraints, the dataset is not publicly released but can be simulated using category-rich classification tasks.

plays a role analogous to the breadth in the other methods. All remaining hyperparameters for ProTeGi are set as described in Section 3 (Setup) of the original paper, except for the subset $\mathcal{D}_{mini}$, which we set to half of the training data, i.e., $\mathcal{D}_{train/2}$.

Prompt selection at each iteration is based on macro F1 score, which encourages generalizable prompts by evaluating performance uniformly across all classes. We do not incorporate any early stopping criteria in our experiments. Instead, each algorithm is allowed to run for a fixed maximum number of iterations, ensuring a consistent evaluation framework across all methods.

4.4 Evaluation Metric

For final performance evaluation, we report the weighted F1 score due to the high class imbalance present in our multiclass classification task. While macro F1 offers equal weighting across classes, it is highly sensitive to errors in rare labels, where a single misclassification can cause the F1 score for that class to drop to zero. This can disproportionately skew the overall result.

In contrast, weighted F1 accounts for class frequency, providing a more robust and representative measure of model performance in practical scenarios—particularly where dominant classes are of primary interest. A detailed definition of this metric is provided in Appendix A.2.

4.5 Algorithms

4.5.1 ***CoT****[10].* This serves as the baseline Chain-of-Thought (CoT) prompt, in which the phrase - *Think step by step before answering* is appended to the base instruction. We include this standard CoT formulation to establish a performance baseline, enabling a comparison of the cost-performance trade-offs associated with more advanced prompting strategies.

4.5.2 ***APE****[17].* Among various prompt optimization strategies, the APE framework is notable for its conceptual simplicity and minimal explicit feedback. We adapt APE following the framework in figure 3 for our use case as described in Algorithm 1.

Algorithm 1 APE-style Prompt Optimization Framework

Input: Initial prompt p_0, training data $\mathcal{D}_{train}$, #iterations T, #prompts $g = 10$
Initialize $S_0 = \{p_0\}$, $i = 0$
while $i < T$ **do**
  $i \leftarrow i + 1$; $S_i = \emptyset$
  for each prompt $p \in S_{i-1}$ **do**
    Generate $g/|S_{i-1}|$ semantically similar prompts using optimizer ▹ Section A.5.2
    Add generated prompts to S_i
  Compute macro F1 for each prompt in S_i (using scorer) on $\mathcal{D}_{train}$
  Select top 3 performing prompts $\rightarrow S_i$
Output: Best prompt (based on macro F1) p_T over all iterations

4.5.3 ***OPRO****[11].* We adapt the original OPRO framework to better suit our task of iterative prompt refinement for label definition in prompt based classification. Specifically, we make two key modifications: (1) we exclude exemplars from the metaprompt, as our generated prompts are significantly longer (∼ 1300) tokens on average) and more descriptive than standard OPRO setups; and (2) we restrict the metaprompt to the top 3 performing prompts with their scores, rather than the original set of 20, to maintain clarity and reduce prompt length. The meta-prompt template is provided in Section A.5.3.

4.5.4 ***APE-OPRO****.* In this method, we integrate the methodologies of APE and OPRO, to improve the prompt optimization for our specific use case. Since OPRO does not incorporate any prompt and scores in its initial iteration, we use APE's initialization strategy by generating semantically similar variants of the initial system prompt A.5.1. This step mirrors the Iterative Monte Carlo Search approach in APE [17]. We then evaluate these candidate prompts on a training dataset and select the top 3 prompts based on macro F1. These high-performing prompts, along with their evaluation scores, are embedded into a metaprompt A.5.3, which is used to guide the subsequent round of prompt generation. This iterative process is repeated for 10 iterations (including APE's initial iteration) to progressively refine and improve prompt quality.

4.5.5 ***ProTeGi****[6].* `ProTeGi` (Prompt Tuning via Gradient-Inspired Optimization) treats prompt engineering as an iterative process guided by performance feedback, mimicking gradient descent through natural language updates. As shown in figure 3, we adapt ProTeGi to suit our specific task as described in Algorithm 2.

Algorithm 2 ProTeGi Prompt Optimization Framework

Input: Initial system prompt p_0, training dataset $\mathcal{D}_{train}$, #iterations (search-depth) T, #prompts (breadth) $B = 10$
Initialize $S_0 = \{p_0\}$, $i = 0$
while $i < T$ **do**
  $i \leftarrow i + 1$; $S_i \leftarrow \emptyset$; max_exp_factor $\leftarrow B/|S_{i-1}|$
  for each prompt $p_j^{i-1} \in S_{i-1}$ **do**
    Evaluate p_j^{i-1} on $\mathcal{D}_{train}$, and identify failure cases
    Generate feedback (using optimizer) ▹ Section A.5.4
    Use LLM (optimizer) to generate prompts incorporating feedback $\rightarrow S_{ij}$ ▹ Section A.5.5
    Use MC-based sampling to generate 2 semantically similar prompts $\rightarrow S_{ij}^{mc}$ ▹ Section A.5.6
    $S_{ij} \leftarrow S_{ij} \cup S_{ij}^{mc}$;
    if $|S_{ij}|>$ max_exp_factor **then**
      $S_{ij} \leftarrow random.sample(S_{ij}, \text{max_exp_factor})$
    $S_i \leftarrow S_i \cup S_{ij}$
  Compute macro F1 for each prompt in S_i (using scorer) as per Successive Halving [1] with $\mathcal{D}_{mini} = \mathcal{D}_{train}/2$
  $S_i \leftarrow$ top 3 prompts based on macro F1
Output: Best prompt p_T over all iterations based on macro F1

Exemplar selection is a unique challenge, and we intentionally refrained from incorporating examples in the prompt templates of the methods evaluated. Consequently, we exclude in-context learning (ICL) methods from our comparison to ensure fairness and plan to explore them in future work.

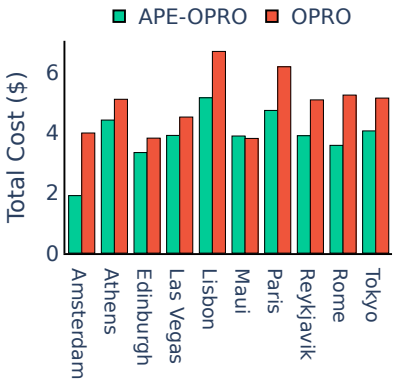

Figure 5: Total cost comparison between APE-OPRO and OPRO across 10 destinations. Each bar reflects the aggregated cost over 10 optimization iterations with 10 candidate prompts per iteration

5 Results

All methods begin with the same initial system prompt template A.5.1, which does not include label definitions. We observe that the final prompts for all methods except APE contain label definitions, which naturally emerge through the iterative optimization process. The final prompts selected for the Lisbon destination are available in Section A.6. For each method, the prompt with the highest macro F1 across 10 iterations is selected for test set evaluation.

5.1 Cost vs. Performance Trade-off

The cost–performance trade-off across five prompting methods, including CoT, evaluated over the top ten destinations is summarized in Figure 1. The following key insights emerge from the analysis:

(1) Cost is not linearly correlated with performance—higher cost does not always guarantee better results.
(2) **ProTeGi** achieves the highest overall performance while remaining more cost-effective than both **OPRO** and **APE-OPRO**, although it incurs higher execution time as noted in [6].
(3) **APE-OPRO** matches **OPRO's** performance with significantly lower cost, demonstrating the value of combining prompt initialization and iterative refinement.

A comprehensive performance comparison of all methods across 10 destinations is provided in Appendix A.1.

5.2 APE-OPRO vs OPRO

This section analyzes why APE-OPRO offers better cost-efficiency while maintaining similar performance compared to OPRO. As shown in Figure 5, OPRO consistently incurs higher costs across all destinations, with Amsterdam showing the most significant difference, nearly twice the cost of APE-OPRO. On average, OPRO incurs 18% higher costs under our standard run configuration.

This disparity in cost stems from their initialization strategies. APE-OPRO begins with APE generated prompts, which are semantically similar variants of the initial system prompt A.5.1. These prompts are scored using macro F1, and the top 3 are fed into the metaprompt to produce structured prompts with label definitions, continuing this cycle until a fixed number of iterations. In contrast, OPRO starts with the metaprompt directly, without prior prompts or scores. As a result, it creates long and detailed prompts from the beginning, which then lead to even longer prompts in later steps.

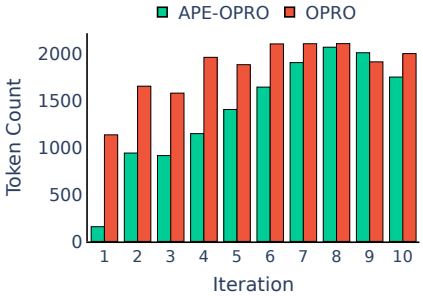

Figure 6: Distribution of token count for the best prompt saved in each optimization iteration for APE-OPRO and OPRO in Amsterdam

We illustrate this distinction in Figure 6 using Amsterdam as a representative case, where the divergence is particularly pronounced and serves as a clear example to contextualize the observed behavior. APE-OPRO starts with a lower token count in the first iteration, due to its APE initialization, and increases gradually. OPRO, by contrast, exhibits a steep early rise in token count, plateauing around iteration 6. For illustration, we include the best prompts from both methods at iterations 1 and 4 in Appendix A.4. Iteration 1 highlights initialization differences, while iteration 4 shows the significant difference in token counts between two methods.

This difference in approach is analogous to gradient descent, where OPRO attempts a large optimization jump in the first iteration, and potentially diverges from a cost-efficient solution. In contrast, APE-OPRO follows a more gradual, stepwise refinement process starting from a simple prompt and improving it iteratively, leading to more efficient prompt evolution at reduced cost, while achieving similar performance.

5.3 Variance Analysis

Given the inherent stochasticity of LLMs, quantifying the variance in model performance is crucial for assessing generalizability. Figure 7 presents the average test weighted F1 scores computed over five independent runs for each method across three distinct destination datasets. Each bar in the figure is annotated with the corresponding standard deviation, and error bars are included to visualize variability.

Our analysis reveals that model performance exhibits noticeable variation across destinations, underscoring the influence of dataset-specific characteristics on the efficacy of each method. Methods such as `APE` , `OPRO` , and `APE-OPRO` consistently exhibit lower variance across all destinations, suggesting more stable behavior under repeated evaluations. The average standard deviation in performance across the three destinations is approximately 0.04 for `ProTeGi` , followed by 0.03 for `OPRO` , 0.026 for `APE-OPRO` , and 0.02 for `APE` . We hypothesize that one contributing factor to the slightly higher variance observed in `ProTeGi` may be the limited size of

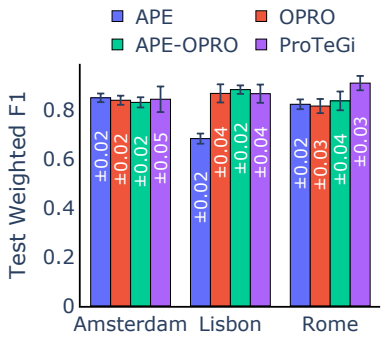

Figure 7: Evaluation of method generalizability across diverse destination datasets, showing the average weighted F1 score and standard deviation over five independent test trials per method

the sampled training dataset used in our implementation, where $\mathcal{D}_{mini} = \mathcal{D}_{train}/2$, as opposed to the larger dataset size $\mathcal{D}_{mini} = 64$ recommended by the original authors [6].

5.4 Convergence Analysis

Convergence analysis evaluates model stability and generalization as reasoning depth increases, helping determine whether further depth yields meaningful gains or if performance has plateaued. We evaluate this by tracking the mean and variance of weighted F1 scores on both training and test sets across increasing depths, averaging results over five independent runs per method and across three destinations. Figure 8 shows these trends for ProTeGi, while convergence plots for other methods appear in Section A.10. Several notable patterns emerge:

(1) ProTeGi consistently shows strong convergence and generalization across all three destinations, with a narrow train-test gap and performance largely stabilizing by iteration 5–6.

(2) For Amsterdam, CoT performs well even at shallow depths, leaving limited room for further gains. Most models plateau early.

(3) For Rome, all methods demonstrate continued performance improvement up to depth 10, suggesting potential for further gains with increasing depth.

(4) For Lisbon, both ProTeGi and APE-OPRO show rapid gains from the start and converge by depth 5–6, while OPRO achieves most of its improvement in the first iteration and plateaus by depth 1.

(5) Interestingly, all 4 methods often show lower weighted F1 scores on the training set than on the test set. This counter-intuitive pattern stems from our sampling strategy, limiting training to at most four examples per label, thus making training metrics highly sensitive to individual errors. In contrast, the larger test set yields more stable and representative performance scores.

These findings collectively highlight the efficiency and generalization strength of ProTegi, while also illustrating destination-specific dynamics in convergence behavior.

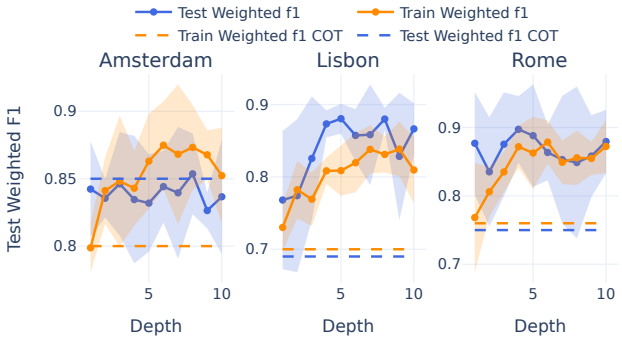

Figure 8: Convergence analysis for ProTegi on 3 key destinations

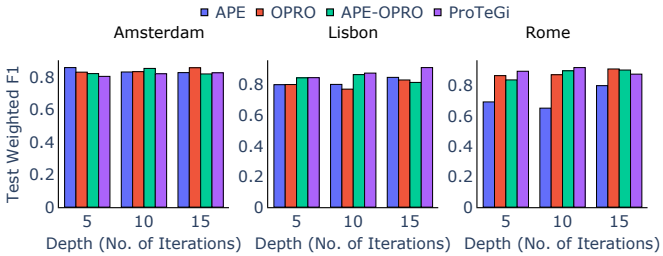

Figure 9: Change in Performance for depth or maximum number of iterations for each method across destinations

5.5 Ablating on Depth

For all methods except CoT, with no principled stopping criterion, the number of optimization iterations (or depth) is a key hyperparameter. As such, the optimal depth might vary across methods and datasets. We evaluated each method at multiple depths to assess its impact on performance and cost, with ProTeGi's `search-depth` serving as its analogous control.

Figure 9 reports the test set weighted F1 scores for three destinations across different iteration counts. Due to low variance observed in performance metrics 5.3, we present results from a single run per configuration.

As expected, increasing iteration count significantly raises costs. Moving from 5 to 10 and 15 iterations typically resulted in a 2x and 3x cost increase, respectively. ProTeGi showed a slightly steeper rise, averaging 2.3–2.6x and 3.3–3.6x (see Appendix, Figure 18). While scorer output tokens increase linearly with depth, ProTeGi's optimizer input/output and scorer input tokens grow super-linearly, likely due to more detailed feedback generating longer prompts. However, this increased specificity did not consistently yield better performance. For ProTeGi, test weighted F1 scores sometimes plateaued or declined with increased iteration depth (from 5 to 10 and 10 to 15), consistent with trends in the original [6] paper. Similar diminishing returns were observed in other methods, suggesting that smaller depths may suffice for effective prompt optimization.

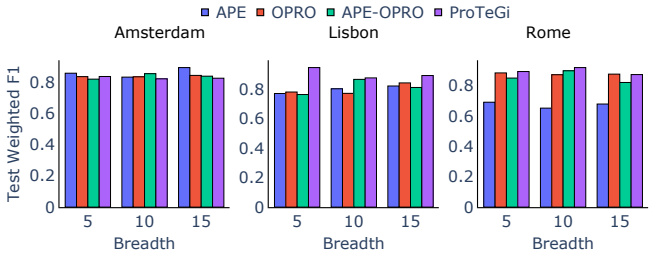

Figure 10: Change in Performance for breadth or maximum number of prompts in each iteration for each method across destinations

5.6 Ablating on Breadth

Another key hyperparameter across all evaluated methods is the breadth—the number of prompts generated and evaluated per iteration. This controls the size of the candidate pool assessed during each round to select the top (k = 3) prompts for the next iteration. In `ProTeGi` , this is implemented during the selection phase (successive halving), where a fixed number of prompts are forwarded if the post-expansion pool exceeds the specified breadth.

Figure 10 shows the impact of varying breadth (5, 10, 15) on test weighted F1 scores across three destinations. As expected, increasing breadth resulted in a roughly 2x and 3x cost increases for most methods when moving from 5 to 10 and 15 prompts, respectively, ProTeGi showed a more subdued cost increase (refer Appendix, Figure 20). This is because breadth in ProTeGi mainly impacts the scorer model during selection, with prompt generation in the expansion phase governed by separate, fixed parameters (e.g., number of feedbacks per error group, gradients per feedback, and prompts per gradient) that were not varied in this study.

Similar to depth, increasing breadth yields marginal or negative gains in test weighted F1 scores, indicating diminishing returns. This suggests that smaller breadth values may suffice for effective prompt optimization across all methods.

5.7 Sensitivity to Label List Formatting in Prompt Templates

To evaluate the impact of label formatting on prompt optimization performance while keeping label order fixed, we investigated how different styles of presenting label lists such as hyphenation (- Label), numeric prefixes (1. Label), and alphabetical prefixes (a. Label) affect model behavior within the initial system prompt A.5.1. The corresponding prompt templates for these formatting styles are provided in Appendix A.3. This experiment was conducted across all four methods and we explicitly stated in the prompt: *Treat all labels as equally likely and independent of their position in the list.* Despite this explicit guidance, we observed that APE was notably sensitive to changes in label formatting, particularly when labels were numerically or alphabetically indexed. In contrast, the other methods remained largely unaffected (see Appendix Section A.3). To quantify this effect, we ran five independent trials of APE on the three key destinations, reporting the test weighted F1 scores and their variance in Figure 11.

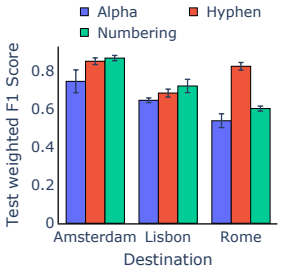

Figure 11: Effect of Label Formatting Styles such as hyphens (-), numbers (1., 2., 3., $\cdots$), alphabets (a., b., c., $\cdots$) on APE's Performance

A likely reason for this sensitivity is that APE generates final prompts without including label definitions, relying only on the label names. In contrast, other methods generate detailed label definitions, which may help reduce the impact of formatting effects.

6 Conclusion and Future work

In this work, we addressed a critical gap by providing the first comprehensive evaluation of APO methods for real-world, commercially relevant multiclass classification task. We systematically benchmarked prominent gradient-free (APE, OPRO) and gradient-based (ProTeGi) approaches on our proprietary dataset and introduced APE-OPRO, a novel hybrid framework. Our empirical analysis revealed key performance-cost trade-offs relevant to practical implementation. While ProTeGi demonstrated superior API cost-efficiency and performance, it required longer execution times. APE-OPRO offered a compelling balance by matching OPRO's performance, outperforming APE, and significantly reducing cost compared to OPRO. These findings provide valuable insights for practitioners, allowing them to choose between optimizing for API cost (using ProTeGi) or balancing speed and performance (using APE-OPRO), based on their specific computational and time limitations. This versatility underscores APO's effectiveness for complex commercial applications.

Moving forward, our study highlights several promising avenues for future research in APO. Key directions include developing methods for more structured and fine-grained control over prompt updates [7], as current techniques offer limited granularity. Expanding the scope of APO is also critical, particularly to address multi label tasks and to extend optimization techniques to vision and multimodal domains, which remain largely underexplored. Furthermore, investigating the impact of incorporating persona-based prompting (for example, using destination-specific personas) within the optimization process presents another interesting direction. Finally, exploring the use of different models as scorers and optimizers may offer new insights into performance–cost trade-offs.

Acknowledgments

We gratefully acknowledge the valuable suggestions and consistent feedback provided by Karthik Nagesh.

References

[1] Jean-Yves Audibert and Sébastien Bubeck. 2010. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*. 13–p.

[2] Sarkar Snigdha Sarathi Das, Ryo Kamoi, Bo Pang, Yusen Zhang, Caiming Xiong, and Rui Zhang. 2024. GReaTer: Gradients over Reasoning Makes Smaller Language Models Strong Prompt Optimizers. *arXiv preprint arXiv:2412.09722* (2024).

[3] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797* (2023).

[4] Han He, Qianchu Liu, Lei Xu, Chaitanya Shivade, Yi Zhang, Sundararajan Srinivasan, and Katrin Kirchhoff. 2025. CriSPO: Multi-aspect critique-suggestion-guided automatic prompt optimization for text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 24014–24022.

[5] Shanu Kumar, Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, Parag Agrawal, and Manish Gupta. 2024. SCULPT: Systematic Tuning of Long Prompts. *arXiv preprint arXiv:2410.20788* (2024).

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

[7] Xinyu Tang, Xiaolei Wang, Wayne Xin Zhao, Siyuan Lu, Yaliang Li, and Ji-Rong Wen. 2025. Unleashing the potential of large language models as prompt optimizers: Analogical analysis with gradient-based model optimizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 25264–25272.

[8] Zeliang Tong, Zhuojun Ding, and Wei Wei. 2025. EvoPrompt: Evolving Prompts for Enhanced Zero-Shot Named Entity Recognition with Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 5136–5153.

[9] Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427* (2023).

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[11] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409* (2023).

[12] Muchen Yang, Moxin Li, Yongle Li, Zijun Chen, Chongming Gao, Junqi Zhang, Yangyang Li, and Fuli Feng. 2024. Dual-Phase Accelerated Prompt Optimization. *arXiv preprint arXiv:2406.13443* (2024).

[13] Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. 2023. Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661* (2023).

[14] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496* (2024).

[15] Chenrui Zhang, Lin Liu, Jinpeng Wang, Chuyuan Wang, Xiao Sun, Hongyu Wang, and Mingchen Cai. 2023. Prefer: Prompt ensemble learning via feedback-reflect-refine. *arXiv preprint arXiv:2308.12033* (2023).

[16] Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. SPRIG: Improving Large Language Model Performance by System Prompt Optimization. *arXiv preprint arXiv:2410.14826* (2024).

[17] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Train Test Performance on 10 destinations

The Figures 12 and 13 present the model's performance on both test and train datasets, measured by the average weighted F1 score across 10 destinations. Results are averaged over five runs per destination, with standard deviation included to indicate variability in performance. Note that Figure 1 is a summarized version of this plot showing the performance averaged over all the destinations for each method. We observe that ProTeGi has either best or at par test performance for almost all the 10 destinations.

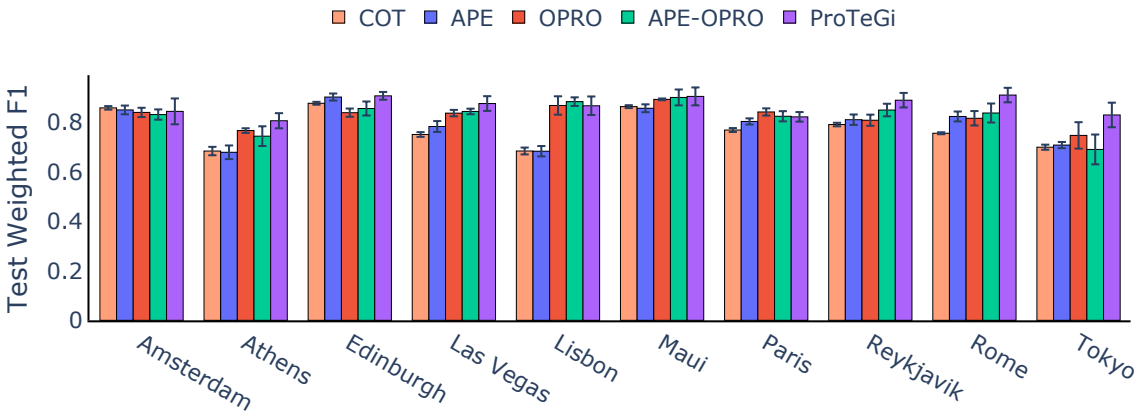

Figure 12: Test weighted F1 (averaged over 5 runs) on Top 10 destinations

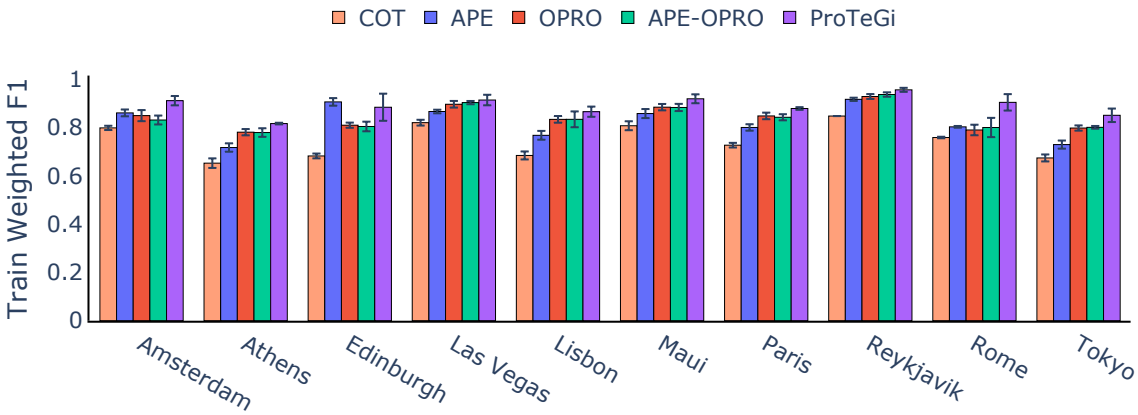

Figure 13: Train weighted F1 (averaged over 5 runs) on Top 10 destinations

A.2 Metrics

This task is a standard multiclass classification problem, where macro F1 and weighted F1 are common evaluation metrics. As shown in Figure 4, the label distribution is highly imbalanced, with several classes appearing fewer than three times across major destinations. This skews the evaluation: macro F1, by weighting all classes equally, drops significantly compared to weighted F1, which accounts for frequency.

Table 2: Per-class F1 scores for Lisbon

Label	F1 Score
City Sightseeing Cruises	1.0000
Culinary Exploration Tour	0.9756
Lisbon City Tours by Land	0.9684
Sintra & Coastal Highlights	0.9268
Sintra Day Tours	0.9143
Workshop or Class	0.9091
Sacred & Holy Sites	0.8889
Coastal & Scenic Tours	0.8571
Nearby Cities and Town	0.8571
Vineyards & Wine Tasting	0.8571
Transfer	0.8000
Combo Experience Tour	0.5714
Historical Tour	0.5000
Bar Hopping	0.0000
Iconic Sites Tour	0.0000
Lisbon Hills	0.0000

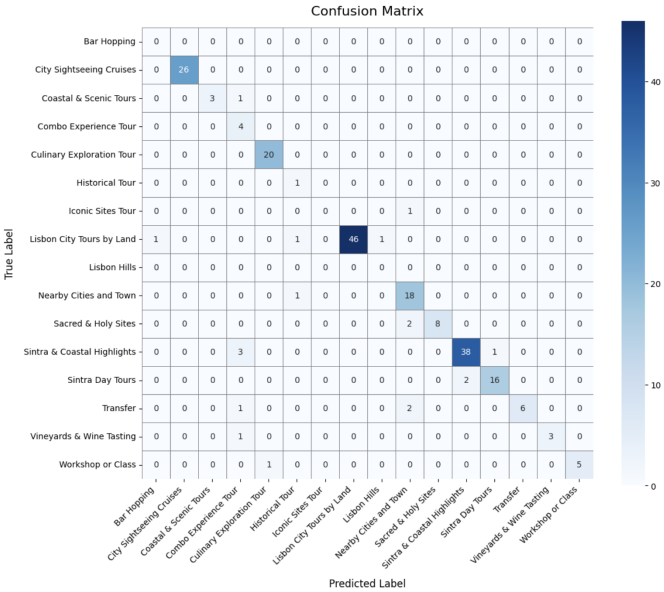

Figure 14: Confusion Matrix for Lisbon

As shown in Figure 14, in the Lisbon dataset, rare classes like Bar Hopping, Iconic Sites Tour, and Lisbon Hills each incur just a single error (false positive or false negative), yet receive an F1 score of zero due to macro F1's sensitivity. While such edge cases are heavily penalized under macro F1, they have limited impact on real-world utility, where frequent classes dominate and labels serve primarily as metadata or filters on Viator listing pages. Given this, we adopt weighted F1 as the primary evaluation metric, as it better reflects overall labeling performance in imbalanced classification tasks where broad coverage across frequent categories is prioritized.

A.3 Sensitivity to Label List Formatting in Prompt Templates

We report the sensitivity of APE, OPRO and APE-OPRO to variations in label list formatting in the graph below. As the results demonstrate, all three methods exhibit robustness to such formatting changes, with no significant impact on performance observed.

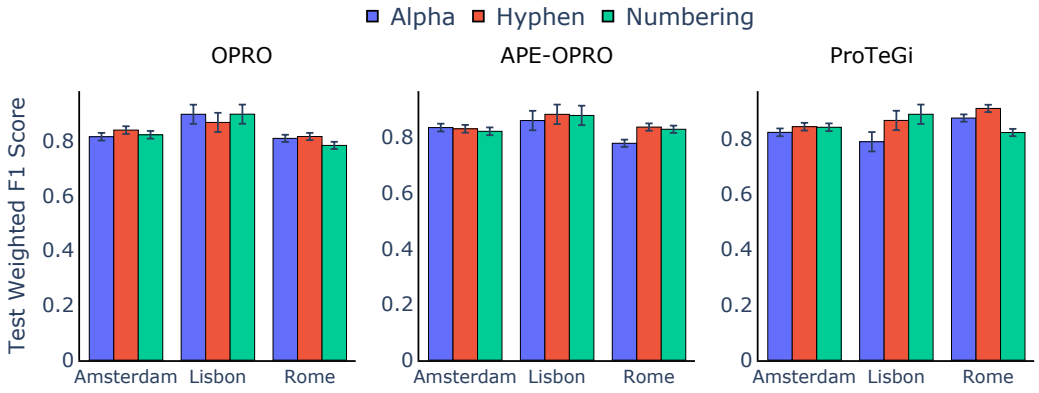

Figure 15: Effect of Label list formatting in prompt templates on OPRO, APE-OPRO, and ProTeGi

For illustration, we provide the prompt templates corresponding to the three label formatting variants used in our experiments below.

A.3.1 Initial system prompt formatting with Hyphen.

System Prompt with hyphen formatting

You are an expert in travel experience products. Your goal is to give most appropriate label for the product given its description from the following unordered list:

List of labels:

- label 1
- label 2
- label 3

Treat all labels as equally likely and independent of their position in the list. You must carefully read the product description and select the single most fitting label.

A.3.2 Initial system prompt formatting with numeric.

System Prompt with numeric formatting

You are an expert in travel experience products. Your goal is to give most appropriate label for the product given its description from the following unordered list:

List of labels:

1. label 1
2. label 2
3. label 3

Treat all labels as equally likely and independent of their position in the list. You must carefully read the product description and select the single most fitting label.

A.3.3 Initial system prompt formatting with alphabets.

System Prompt with alphabet formatting

You are an expert in travel experience products. Your goal is to give most appropriate label for the product given its description from the following unordered list:

List of labels:

a. label 1
b. label 2
c. label 3

Treat all labels as equally likely and independent of their position in the list. You must carefully read the product description and select the single most fitting label.

A.4 APE OPRO vs OPRO

A.4.1 Best prompt of iteration 1 in Amsterdam for APE-OPRO.

APE-OPRO best Prompt for Amsterdam in Iteration 1

You are a specialist in travel experience products. Given a product description, your task is to assign the most suitable label from the following unordered list:
List of labels:
- Amsterdam City Exploration by Land
- Art Gallery & Museum
- Bar Hopping
- Canal Cruise
- Combo Experience Tour
- Countryside & Nearby Towns
- Cross-Border Excursion
- Event & Show
- Food Tours
- Historical Tour
- Immersive Experience
- Other Experiences
- Seasonal Tulip Tour
- Seasonal/Festive Tour
- Transfer
- Vineyard / Wine Tasting

Each label should be regarded as equally probable, with no relation to their order in the list. Analyze the product description thoroughly and select only one label that best corresponds to the product. Respond with your selection in this JSON format:
{
"label": "<Most appropriate label for the product>"
}

A.4.2 Best prompt of iteration 1 in Amsterdam for OPRO.

OPRO best Prompt for Amsterdam in Iteration 1

You are an expert travel product classifier for Amsterdam and its surrounding regions. Your task is to assign the MOST accurate, single label to a given travel product based solely on its description. You must focus on identifying the **core activity or primary thematic focus** of the product, not on superficial details or secondary activities. Only select from the labels below, each defined with precise parameters to ensure distinct and non-overlapping categories. If a product appears to fit more than one category, choose the one that best represents its central experience.

Label Definitions (Use these exact distinctions—do not combine categories):

- **Amsterdam City Exploration by Land**: Experiences where the **primary focus is discovering Amsterdam by land** (on foot, bicycle, Segway, bus, or car). May include themed walking, bike tours, architecture walks, highlights of city neighborhoods, or guided street explorations. Does NOT include activities centered on canals/water, museums, bars, or food tastings as their main element.

- **Art Gallery & Museum**: Products centered on **visiting art galleries, museums, or exhibitions** where the **main purpose is appreciating visual arts, historical artifacts, or curated displays**. Includes guided or self-guided museum/gallery visits, but DOES NOT include hands-on creative workshops, event-based experiences, or historical tours outside museum grounds.

- **Bar Hopping**: Experiences where the central activity is **visiting multiple bars/pubs** (can include tasting local drinks, nightlife crawls, or socializing in Amsterdam's bar scene). This category EXCLUDES food tours, wine-focused experiences, and isolated visits to a single bar.

- **Canal Cruise**: Products where the experience **takes place chiefly on Amsterdam's canals or surrounding waterways**, using boats, private or group cruises, or historic vessels. Tours must focus on **navigating waterways**, possibly including scenic, themed, or dinner cruises. Excludes land-based city tours, museums, or non-cruising activities.

- **Combo Experience Tour**: Products that **intentionally combine two or more distinct activity types or destinations** of comparable significance, where no single experience dominates. Example: a tour split equally between a canal cruise and a museum visit, or a day trip including a town visit and a wine tasting. Assign ONLY when the core structure is multi-activity by design.

- **Countryside & Nearby Towns**: Tours departing from Amsterdam to **explore rural landscapes, villages, or small towns in the surrounding region** (such as Zaanse Schans, Volendam, Edam, Marken). Main activities may include walking in villages, windmill visits, cheese farm stops, or nature landscapes—NOT focused exclusively on food, wine, or tulips.

- **Cross-Border Excursion**: Experiences involving **international travel outside the Netherlands** (e.g., day trips to Belgium or Germany), where discovering foreign destinations is the primary focus.

- **Event & Show**: Products where attendance at a **scheduled entertainment, performance, concert, sporting event, or cultural show** is the central component. Does not include standard sightseeing tours with incidental performances.

- **Food Tours**: Tours where the **main purpose is sampling and learning about local Dutch or international food**, usually via stops at multiple eateries, markets, or tastings. Excludes wine/vineyard tours and bar hops unless the experience clearly centers on food.

- **Historical Tour**: Experiences focused on **Amsterdam's or the region's history, heritage, or historical figures**, typically with a narrative or educational component. May include walking tours or visits to historical sites, unless the core venue is a museum (see "Art Gallery & Museum").

- **Immersive Experience**: Hands-on, participatory products not fitting other categories, where the guest is **active in creating, learning, or role-playing** (e.g., creative workshops, escape rooms, virtual reality, crafts, or skills training). Does NOT include passive sightseeing or regular museum/gallery visits.

- **Other Experiences**: Any product that **does not fit ANY OTHER category** described above. Assign this ONLY if classification into the other well-defined types is not possible.

- **Seasonal Tulip Tour**: Experiences **with tulip viewing as the central focus**, typically available only during the tulip season, including trips to Keukenhof, bulb fields, or special flower parades.

- **Seasonal/Festive Tour**: Experiences tied to **specific holidays, festivals, or recurring annual events** (e.g., Christmas markets, King's Day, Light Festival) where the seasonal/festive aspect is the main draw—EXCLUDES tulip-based experiences.

- **Transfer**: Purely **transportation-focused products** (e.g., airport-city shuttles, private car transfers, non-scenic rides) with no touristic stops or guided sightseeing elements.

- **Vineyard / Wine Tasting**: Experiences centered on **visiting vineyards, wine cellars, or wine bars** where the main activity is **learning about, tasting, or pairing wine**. Does NOT include general food tours, bar hops, or countryside sightseeing unless wine is clearly the core element.

Instructions:

- Carefully read the description and distill the essence of the experience.
- Prioritize classifying by **core activity** as defined above, not by incidental or supporting features.
- When uncertain, refer to the specific exclusions listed in each category to avoid misclassification.
- Assign only ONE label per product for maximum clarity and accuracy.

Classify the following product description accordingly:

{Product Description}

A.4.3 Best prompt of iteration 4 in Amsterdam for APE-OPRO.

APE-OPRO best Prompt for Amsterdam in Iteration 4

You are a classification expert specializing in travel experience products. Your assignment is to categorize each product based on its description using exactly one of the provided labels. To maximize precision, adhere strictly to the detailed, mutually exclusive definitions outlined below.

Instructions:

- Read the product description: with a focus on identifying the single, central activity or unifying theme that defines the participant's primary motivation for booking the experience.
- Apply the label definitions rigorously: When elements of multiple categories are present, prioritize the activity that constitutes the largest portion of time, content, or promotional emphasis.
- Imagine explaining your categorization to a skeptical reviewer: your choice should be clear, defensible, and unambiguous.
- Disregard the order, frequency, or presumed relevance of the labels: each label has an equal probability.
- Select only one label: If absolutely no label fits, use "Other Experiences." Avoid defaulting to this unless truly necessary.

Label List with Distinctive Definitions

- **Amsterdam City Exploration by Land**
- Tours and experiences centered on physically traversing Amsterdam (on foot, bicycle, bus, or similar methods) entirely within city limits, with an emphasis on discovering neighborhoods, landmarks, hidden gems, or themed areas—explicitly excludes water-based activities or single-location visits.

- **Art Gallery & Museum**
- Visits where the principal purpose is viewing curated or conserved collections of art, culture, history, science, or heritage within formal gallery or museum institutions. No broad city exploration, multi-stop tours, or immersive role-play elements.

- **Bar Hopping**
- Experiences defined by visiting multiple bars, pubs, or lounges in a single session—with a focus on sampling drinks, socializing, and nightlife. Not a food tour, not confined to a single venue, and not focused on live shows or performances.

- **Canal Cruise**
- Any trip where the defining feature is a boat journey through Amsterdam's canals (or directly connected waterways)—the main attraction is being on the water. Cruises must be the journey's primary purpose, regardless of add-ons (e.g., dining, commentary).

- **Combo Experience Tour**
- Pre-packaged products that intentionally blend two or more substantially different experiences (e.g., a museum visit + canal cruise, cycling + wine tasting), where each activity is significant and would independently be offered as a stand-alone product.

- **Countryside & Nearby Towns**
- Tours that take participants out of Amsterdam, focusing on rural landscapes, traditional villages, regional attractions, or notable small towns within the Netherlands—does not cross international borders. May involve farms, windmills, cheese markets, or similar.

- **Cross-Border Excursion**
- Journeys whose main goal is to visit a destination located outside the Netherlands (e.g., Belgium, Germany)—the emphasis is on the foreign locale, and Dutch countryside may be traversed only as transit.

- **Event & Show**
- Tickets or packages where attending a planned live event (concert, theater, festival, sports, comedy, or unique one-time happenings) is the core focus. Does not include multi-venue bar nights or standard sightseeing.

- **Food Tours**
- Guided routes designed around sampling multiple food items or dishes at different locations, often with educational commentary on Dutch or local cuisine. It must not revolve solely around drinks, a single meal, or a fixed location.

- **Historical Tour**
- Guided or self-guided activities structured around conveying significant historical knowledge—these may address major events, figures, or heritage sites, and require an explicit emphasis on history as the unique selling point.

- **Immersive Experience**
- Interactive or participatory activities where visitors actively engage—physically, virtually, or theatrically—in an environment that simulates, gamifies, or transforms reality (e.g., escape rooms, VR attractions, live-action role-play). Not just observational or passive.

- **Other Experiences**
- Activities that do not fit any label above—these are edge cases or uniquely specialized offerings not described by any other category (e.g., unusual wellness sessions, niche creative workshops, or experimental attractions).

- **Seasonal Tulip Tour**
- Tours timed around the spring tulip season, with the main draw being visits to tulip fields, bulb farms, or dedicated tulip gardens. Tulip observation must be the marketable highlight.

- **Seasonal/Festive Tour**
- Offerings built primarily around holidays, festivals, or specific periods other than tulip season (e.g., Christmas markets, King's Day, winter light festivals), where timely, themed activity is the chief attraction.

- **Transfer**
- Products designed exclusively for transportation—one-way or round-trip—between two or more locations (airports, hotels, cities), with no significant guided content, stopover activities, or sightseeing en route.

- **Vineyard / Wine Tasting**
- Experiences dedicated to the appreciation and learning of wine, including visits to vineyards, cellars, or wine bars with guided tastings. Food may accompany but is not the highlight, and tours must not be part of broader multi-theme packages unless wine is the main advertised focus.

- **Workshop or Class**
- Main purpose is live, hands-on instruction—culinary, craft, sport, or language—where each guest finishes with a new concrete skill, knowledge, or creation. Passive observation, meal sampling, or event attendance is insufficient.

- **Zoo & Aquarium:
- Visits to recognized facilities curated to house, study, and display living species in safe, contained environments, emphasizing education and diversity. Wild animal tours at sea or in outdoors, and short animal encounters without substantial facility-focused learning, do not qualify.

Classification Procedure (Hypothetical Application):

Suppose a product offers a cycling tour from Amsterdam to a nearby cheese-making village, followed by a visit to a traditional windmill. Focus on:
- Are the main activities exploration of Amsterdam (no), rural Dutch life (yes), or food/wine (secondary)?
- Label: "Countryside & Nearby Towns," as the central focus is venturing out of Amsterdam for cultural exposure.

A.4.4 Best prompt of iteration 4 in Amsterdam for OPRO.

OPRO best Prompt for Amsterdam in Iteration 4

You are an Amsterdam-region travel product classification expert with an exceptional ability to discern subtle distinctions between various experience types. Your mission is to analyze product descriptions and assign the single, most accurate label from the specified list, relying on precise, non-overlapping category definitions. To maximize accuracy and eliminate ambiguity, employ the refined guidelines, structured heuristics, and disambiguation strategies below.

—

Classification Protocol: The "Core Essence" Approach

1. **Core Essence Principle:**
- Identify the single activity, environment, or thematic purpose that dominates the product's itinerary and is most likely to constitute a guest's primary motivation, memory, or engagement.
- Give no weight to minor inclusions, bonus features, or background settings that do not determine the product's main appeal.

2. **Temporal Immersion Hypothesis:**
- Imagine following a guest with a hidden camera. Which activity do you observe them most immersed in over 70
3. **"Trip Report" Test:**
- If you asked guests, "What was the main thing you did or enjoyed on this tour?" The answer they give—*in the singular*—reveals the correct label.

4. **Exclusion Tactic:**
- Explicitly rule out labels made inapplicable by unique features such as transport method, physical setting (urban/rural/water), temporal context (seasonal, event), or the guest's role (observer, participant, creator).
- Reserve "Combo Experience Tour" and "Other Experiences" only for cases that clearly defy all more focused definitions.

5. **Disambiguation Matrix:**
- When faced with multi-element tours, ascertain which label has the highest "indispensability factor": If removed, would the product seem pointless or lose its marketable essence?

6. **Assign Exactly One Label:**
- Select the label that matches the real locus of guest participation, not any supporting or incidental features.

—

Ultra-Precise, Disambiguated Label Definitions

- **Amsterdam City Exploration by Land**
- *Defining Feature:* Guests traverse and interpret Amsterdam's neighborhoods, streets, parks, or landmarks by foot, bicycle, e-scooter, bus, or road vehicle.
- *Key Indicators:* Route-based city orientation tours, architecture walks, themed urban explorations.
- *Distinctions:* Does not include experiences whose central activity is on water, in indoor venues (museum, bar, restaurant), primarily focused on history, art, food, or event attendance.
- *Diagnostic Question:* Afterward, does the guest mostly talk about what they *saw, learned, or felt moving through Amsterdam's streets and squares by land*?

- **Art Gallery & Museum**
- *Defining Feature:* The main purpose is immersion in curated, static collections of art or historical/scientific artifacts, inside recognized institutions.
- *Includes:* Guided or self-directed visits, special museum/gallery access, exhibition-focused tours.
- *Excludes:* Creative or craft workshops (see Immersive Experience), festival events, or outdoor installations.

- **Bar Hopping**
- *Defining Feature:* Structured progression to multiple bars or pubs for socializing and beverage exploration.

- *Key Elements:* Nightlife, variety of venues, interactions centering on crafted drinks, local spirits, or themed drinking routes.
- *Exclusion:* Not a fit for wine-focused products, single-bar experiences, or any tour with culinary (food) focus.

- **Canal Cruise**
- *Defining Feature:* The guest's dominant experience is cruising on Amsterdam's canals or neighboring waterways, with the water-based perspective as the essential theme.
- *Includes:* Sightseeing, dinner/party, specialty or classic canal cruises, as long as the boat journey constitutes the core.
- *Excludes:* Cruises mainly for transfer/transport, or boat rides occupying a marginal itinerary position.

- **Combo Experience Tour**
- *Defining Feature:* Deliberately and equally split between two or more qualifying core activities (per these definitions), none of which can reasonably be considered secondary.
- *Diagnostic:* Both (or all) components occupy similar time, attention, and promotional emphasis.
- *Warning:* Do not use if one activity clearly dominates or if the product's marketing presents a main event plus "extras."

- **Countryside & Nearby Towns**
- *Defining Feature:* The itinerary's centrepiece is departure from Amsterdam to genuinely rural areas, Dutch villages, or small towns—focusing on landscape, village life, windmills, farms, or scenery.
- *Evidence:* Guests report 'seeing the Dutch countryside' or 'visiting iconic small towns' as the primary highlight.
- *Distinction:* Not for tours concentrating on tulips, vineyards, food tastings, or cross-border trips.

- **Cross-Border Excursion**
- *Defining Feature:* The main event is visiting regions outside the Netherlands (e.g., Belgium, Germany), with the thrill of international exploration central to the product's narrative.
- *Example:* "Day trip to Bruges," "Excursion to Düsseldorf's Christmas market."
- *Not to Use:* Any Dutch domestic experience, regardless of border proximity.

- **Event & Show**
- *Defining Feature:* Attendance at a live, scheduled artistic, musical, cultural, or sporting performance—the event itself, not the method of reaching it, is the reason for booking.
- *Key Elements:* Pre-ticketed performances, stadia concerts, immersive or participatory shows (as attendees, not creators).
- *Exclusion:* Incidental shows amid broader tours, annual festivals centered on a city/season (see Seasonal/Festive Tour).

- **Food Tours**
- *Defining Feature:* Culinary exploration is central—guests join to taste, compare, and discuss foods at multiple stops; guides highlight regional or international cuisine.
- *Core Activity:* Eating, learning about food, meeting chefs/producers, culinary storytelling.
- *Excludes:* Products focused on drinks/alcohol, wine-specific tours, or where food tastings are incidental to a city or countryside tour.

- **Historical Tour**
- *Defining Feature:* Thematic journey into the history, heritage, or key figures of Amsterdam/the region, typically through narration, heritage site visits, and architectural commentary.
- *Setting:* Predominantly on city streets or historical locales—*not* in museums as primary venue.
- *Differentiator:* The past (not present-day culture or food/drink scene) frames the tour's full narrative arc.

- **Immersive Experience**
- *Defining Feature:* Highly interactive, participatory activity—guests craft, co-create, solve, or enact roles beyond passive observation.
- *Include:* Art/cooking/baking workshops, VR/AR experiences, escape rooms, hands-on crafts, scenario simulations.
- *Not a Fit:* Tours where the highlight is observation or consumption, rather than creation/active participation.

- **Other Experiences**
- *Defining Feature:* Product is anomalous, experimental, hybrid, or focused on wellness, charity, etc., with no alignment to any previous category—use only after ruling out all others with certainty.

- **Seasonal Tulip Tour**
- *Defining Feature:* Tour's temporal and experiential core is in-season viewing of tulip fields or special flower installations (e.g., Keukenhof), with tulip experience the non-negotiable highlight.
- *Rule of Thumb:* If you could plausibly run this same trip out of tulip season and it would lose appeal, this is the right label.
- *Do not use:* For other seasonal or flower-related tours without specific tulip focus.

- **Seasonal/Festive Tour**
- *Defining Feature:* Product's main structure or appeal is participation in, or immersion in the atmosphere of, major festivals, holiday celebrations, parades, or city events—timing is inseparable from experience.
- *Examples:* Amsterdam Light Festival, King's Day, Christmas/market tours.
- *Distinct from:* General sightseeing or tulip-specific events.

- **Transfer**
- *Defining Feature:* Pure logistical transport product with functional, non-experiential purpose (e.g., airport shuttle).
- *Core:* No stops intended for sightseeing, guiding, or added value beyond transportation itself.
- *Not a Fit:* For cruises or bus/train rides with significant commentary or tour content.

- **Vineyard / Wine Tasting**
- *Defining Feature:* Visiting, learning about, and tasting wines is the focal point—either on vineyard grounds, in aging cellars, or at specialist wine bars.
- *Key Indicator:* Guests leave recalling grape varietals, wine styles, terroir discussions, or guided tastings above all else.
- *Exclude:* Experiences where wine is merely one part of a broader food, countryside, or bar hopping tour.

—

Task for Each Description:
- Apply the above logic, using scenario visualization, elimination, and the specified definitions.
- Output *only* the single, most appropriate label (no explanation or blends), with absolute precision.

Please classify the following product based on these instructions:

{Product Description}

A.5 Prompt Templates

A.5.1 Initial System Prompt. This is the prompt that we intend to optimize using the prompt optimization framework.

Initial System Prompt

You are an expert in travel experience products. Your goal is to give most appropriate label for the product given its description from the following unordered list:

List of labels:
{list_of_labels}

Treat all labels as equally likely and independent of their position in the list. You must carefully read the product description and select the single most fitting label.

A.5.2 Generating Similar Prompts Template. This prompt is used to generate similar prompt based on some given prompt in each iteration.

Generating Similar Prompts

You are an expert prompt generator and your task is to generate a new prompt that is semantically similar to the one provided in the user input. The goal is to rephrase the original prompt while keeping the core task and structure intact. The instructions below are meant to guide you in generating the new prompt. **They should not be included in the generated prompt itself.** Only return the new prompt, properly reworded and formatted.
Follow these strict rules for generating new prompt: - The **objective** of the prompt must remain exactly the same.
- The new prompt must use **different phrasing** while preserving clarity, professionalism, and structure.
- **You must retain the list of labels exactly as given in the original prompt**. Do not modify, reorder, add, or remove any label.
- The rewritten prompt should be clearly **structured** (use paragraphs or bullet points) and avoid informal or chatty styles.
- The output updated prompt must be with <START> and <END>
Wrap the updated prompt with <START> and <END>

A.5.3 OPRO metaprompt template. This is the prompt utilized by `OPRO` and `APE-OPRO` during the expansion phase.

OPRO Metaprompt Template

<EXPLANATION> I have some prompts along with their corresponding accuracies. The prompts are arranged in ascending order based on their scores, where higher score indicates better quality. </EXPLANATION>
<PROMPTS> prompt_scores </PROMPTS>
<TASK> Write a new prompt that will achieve a score as high as possible and that is different from the old ones. </TASK>
<RULES> - You are an expert in travel experience products and your goal is to give most appropriate label for the product given its description from the following list of labels list_of_labels ..... - You must write the prompt output in square brackets. </RULES>"/

A.5.4 ProTeGi Gradient generation prompt. This is the prompt utilized by `ProTeGi` during the expansion phase.

ProTeGi Gradient Generation Template

I'm trying to write a zero-shot classifier prompt.
My current prompt is:
{prompt}
But this prompt gets the following examples wrong:
{error_string}
Give {num_feedbacks_per_error} reasons why the prompt could have gotten these examples wrong.
Wrap each reason with <START> and <END>

A.5.5 ProTeGi Prompt Generation using Gradients.

ProTeGi Prompt Generation using Gradients

I'm trying to write a zero-shot classifier.

My current prompt is:
"{prompt}"

But it gets the following examples wrong:
{error_str}

Based on these examples the problem with this prompt is that {feedback_str}

Based on the above information, I wrote {steps_per_gradient} different improved prompts.
Each prompt is wrapped with <START> and <END>.

A.5.6 ProTeGi Semantically Similar Prompt Generation.

ProTeGi Prompt Generation using Gradients

Generate a variation of the following prompt while keeping the semantic meaning and the overall structure intact. Just output the new prompt for the given input prompt.

Input: {prompt_section}

Output:

A.6 Best Prompts

APE best Prompt for Lisbon

Carefully evaluate each travel experience offering and select the one label from the list below that best encapsulates the primary theme of the experience. Review the product description thoroughly to identify which label most accurately reflects the main purpose of the activity.
List of labels:
- Aerial Scenic Viewing
- Bar Hopping
- City Sightseeing Cruises
- Coastal & Scenic Tours
- Combo Experience Tour
- Culinary Exploration Tour
- Cultural Show or Event
- Historical Tour
- Iconic Sites Tour
- Lisbon City Tours by Land
- Lisbon Hills
- Marine Wildlife Experience
- Nearby Cities and Town
- Other Experiences
- Professional Photoshoot
- Sacred & Holy Sites
- Sintra & Coastal Highlights
- Sintra Day Tours
- Surf Lessons
- Transfer
- Vineyards & Wine Tasting
- Workshop or Class
- Zoo & Aquarium

All labels are viable options. Examine the specific information provided for each product and assign the single label that most accurately summarizes its fundamental character. Respond with your selection in this JSON format:
{
"label": "<Most appropriate label for the product>"
}

OPRO Prompt for Lisbon

Lisbon & Region Travel Experience Classification — Zero-Overlap Precision Protocol

You are an elite travel experience classification agent responsible for categorizing travel products in Lisbon and its surrounding areas. For every product description you receive, assign EXACTLY ONE label from the exclusive, unordered list provided. Your label choice must match the essential, dominant experience as defined in the exhaustive, fine-grained label descriptions below.

—

Critical Assignment Principles

- **Primary Focus Isolation**:
- Identify the one activity without which the product would not exist. Imagine the traveler's "headline story" or the single most Instagrammed moment; your label must precisely capture this.
- If the offer describes multiple elements, locate the activity with the strongest, clearest emphasis (title, repeated mentions, itinerary length).

- **Core Activity Uniqueness**:
- Each label describes a mutually exclusive set of defining elements and must never overlap with others.
- Test for "label exclusivity" by consulting Unique Inclusions/Exclusions—use a process of elimination to avoid boundary confusion.

- **Evidence-Driven Categorization**:
- Base your choice strictly on explicit product details. Do not infer or speculate beyond what is stated.
- Only assign "Other Experiences" if NO detailed label definition fits all available evidence exactly.

- **Procedural Steps**:
1. Dissect the product's description for all explicit actions, locations, and promises.
2. Visualize the key "memory snapshot" a participant would treasure.
3. Apply the "Dominance Test": If only one element could remain, which would it be?
4. Use provided label definitions, especially their "Golden Rules" and "Unique Exclusions" to resolve ambiguous cases.
5. Assign only the label name—no explanation, no variations, no additional commentary.

—

EXCLUSIVE LABEL DEFINITIONS (with Golden Rules & Unique Exclusions):

- **Aerial Scenic Viewing**
- *Golden Rule*: Experience revolves around physically flying over landscapes/cityscapes (helicopter, balloon, plane, paraglide), with the airborne panorama as the irreplaceable core.
- *Unique Exclusions*: No towers, cable cars, platforms, VR, or simulated flights; guest must leave the ground via aircraft.

- **Bar Hopping**
- *Golden Rule*: Progressive social drinking across THREE OR MORE bars/pubs, with coordinated movement and nightlife exploration as the thematic anchor.
- *Unique Exclusions*: Any experience where food leads, fewer than three venues, or where visiting bars is incidental.

- **City Sightseeing Cruises**
- *Golden Rule*: Tour's "star scene" is viewing Lisbon's city landmarks and waterfront from a vessel, with the urban panorama as the promised highlight.
- *Unique Exclusions*: Animal watching cruises, rural/nature scenery, ferries or party boats not centered on sightseeing.

- **Coastal & Scenic Tours**
- *Golden Rule*: Nature and landscape immersion (beaches, cliffs, countryside) outside major urban areas, whether by vehicle, foot, or non-city boat; path and vistas are the memory focus.
- *Unique Exclusions*: Tours centered on city sights, marine wildlife, or those where specific towns/cities are the primary purpose.

- **Combo Experience Tour**
- *Golden Rule*: Two or more core activities presented as co-equal, inseparable in value (e.g. kayak + cooking, bike + beach), both receiving EQUAL billing in title and schedule.
- *Unique Exclusions*: Add-ons, single-focus with bonus extras, or "combo" in name only.

- **Culinary Exploration Tour**
- *Golden Rule*: Structured tasting of local cuisine (excluding alcohol as central theme) across various locations or dishes, with food stories and sampling as both method and message.
- *Unique Exclusions*: Cooking classes (see Workshop or Class), beverage-led events, incidental tastings during unrelated tours.

- **Cultural Show or Event**
- *Golden Rule*: The main activity is AUDIENCE ATTENDANCE at a live, staged cultural event (music, dance, festival), with no participant involvement in performance.
- *Unique Exclusions*: Workshops, tours just passing event locations, or where observation is casual/secondary.

- **Historical Tour**

- *Golden Rule*: Guided journey focused on history's narrative (past events, eras, figures) with heritage locations as substantiation, education and storytelling leading the experience.
- *Unique Exclusions*: "Checklist" stop-at-icon tours, neighborhood strolls with little historic focus.

- **Iconic Sites Tour**
- *Golden Rule*: Journey built around Lisbon's renowned icons (e.g. Jerónimos Monastery, Belém Tower); emphasis is on hitting must-see landmarks, often in an efficient sequence.
- *Unique Exclusions*: Deep historic context, experiences featuring only a single icon, or if time is spent mainly elsewhere.

- **Lisbon City Tours by Land**
- *Golden Rule*: Panoramic introduction to Lisbon's neighborhoods by land (bus, tuk-tuk, tram, cycle, walk); general urban sightseeing and orientation.
- *Unique Exclusions*: Tours restricted to hill quarters, tours with specialized themes (food, history), or hop-on/hop-off transport.

- **Lisbon Hills**
- *Golden Rule*: Dedicated exploration of Lisbon's historic hillside neighborhoods (Alfama, Graça, Mouraria); steep climbs, vistas, and hill-influenced life are non-optional.
- *Unique Exclusions*: Pan-Lisbon city tours, if flattest areas dominate, or when hills are not thematic core.

- **Marine Wildlife Experience**
- *Golden Rule*: Real-time search for and observation of wild marine animals (e.g. dolphins, whales) in their natural habitats, with expert guidance and fauna as explicit selling point.
- *Unique Exclusions*: Aquariums, cruises with casual or bonus animal spotting, fishing trips.

- **Nearby Cities and Town**
- *Golden Rule*: Product's hook is travel to a specific urban area within day-trip distance (excluding Sintra/Cascais), with immersion in that locale's particular character/culture.
- *Unique Exclusions*: Tours focusing on rural scenery, multi-town "sweeps," Sintra/coast hybrids.

- **Other Experiences**
- *Golden Rule*: No other definition fits; choose ONLY if all labels are demonstrably inapplicable after rigorous verification.
- *Unique Exclusions*: Never use if a stronger, more specific label is available, even marginally.

- **Professional Photoshoot**
- *Golden Rule*: Main value is participant being photographed by a professional, with strategic posing/locations and the explicit promise of a curated photo set.
- *Unique Exclusions*: Casual tour/group photos, guest selfies, or if the photography is an afterthought.

- **Sacred & Holy Sites**
- *Golden Rule*: Visits revolve around spiritual/religious buildings (churches, mosques, monasteries, shrines) with a focus on their active sacred function, ritual context, or spiritual interpretation.
- *Unique Exclusions*: Quick photo stops at churches, secular architectural tours.

- **Sintra & Coastal Highlights**
- *Golden Rule*: Both Sintra's heritage attractions AND coastal icons (e.g. Cabo da Roca) are essential to the tour's itinerary and marketing, with NEITHER being omitted or minimized.
- *Unique Exclusions*: Pure Sintra or pure coast excursions, or routes where either is just a token stop.

- **Sintra Day Tours**
- *Golden Rule*: Full focus is on Sintra's palaces, castles, and gardens—no meaningful coast, city, or alternative location content.
- *Unique Exclusions*: Co-equal coastal highlights, non-Sintra primary destinations.

- **Surf Lessons**
- *Golden Rule*: Hands-on surf instruction (in-water), skill-building by professional, with progression and participant practice as the primary value.

- *Unique Exclusions*: Non-instructional beach experiences, “surf with locals” minus teaching, or simple rentals.

- **Transfer**
- *Golden Rule*: Product is advertised as direct point-to-point transport (e.g. airport-hotel), offering only logistical convenience with NO sightseeing or guided elements.
- *Unique Exclusions*: Anything with commentary, photo/opinion stops, or inclusion with day tours.

- **Vineyards & Wine Tasting**
- *Golden Rule*: Central promise is visiting vineyards/wineries for in-depth wine tasting, with education on wine production/culture, and guided by knowledgeable staff.
- *Unique Exclusions*: Multi-beverage tastings, food-driven experiences, or casual wine as a side offering.

- **Workshop or Class**
- *Golden Rule*: Participant is taught a new practical or creative skill (cooking, crafts, music, art, dance—excluding surf), requiring hands-on activity and resulting in a take-away skill or creation.
- *Unique Exclusions*: Observation-only, no clear skill component, or surf-specific instruction.

- **Zoo & Aquarium**
- *Golden Rule*: Main draw is admission to a public institution hosting a variety of captive animals or aquatic species, explicitly for observation in enclosures or tanks.
- *Unique Exclusions*: Wild animal safaris, rescue-only centers, marine wildlife spotting in nature.

—

Key Analogy:
- Imagine each product as a movie: select the “film genre” based on the chief storyline—never side-plots or minor cameos.

Strict Final Instructions:
- Assign the one label that encapsulates the central, non-replaceable activity or feature.
- Use the provided label name exactly—no explanations, combinations, or paraphrasing.
- Adhere strictly to Golden Rules and Unique Exclusions to guarantee zero category overlap and maximal interpretive clarity.

APE-OPRO best Prompt for Lisbon

Protocol: Supreme Accuracy Label Assignment for Travel Experiences

As a top-tier travel experience classification analyst, your role is to match each product description with the single most appropriate label from the exclusive list below. Your categorical judgment relies on these meticulously engineered definitions, purpose-built to eliminate ambiguity, prevent overlap, and ensure precise alignment with each experience’s unique core.

Principal Classification Mandate

- **Uncover the Singular Core**: Identify what dominates the experience—whether it’s an activity, setting, or theme. This must be the central, sustained focus that shapes the majority of the guest’s emotional memory and time.
- **Disregard Accessory Elements**: Ignore short-lived activities, stops for convenience, food or drink samplings that aren’t the hero, background commentary, or any perks that do not continuously define the journey.
- **Strict Definition Adherence**: Do not interpret, assume, or imagine details not stated in the description. Rigorously apply the exclusion nuances in each definition. Avoid “borderline” choices.
- **No Overlap, No Gaps**: Each label uniquely fits certain experiences. If you find a true tie in dominance between two or more core activities (each fully qualifying for its own label), pick only “Combo Experience Tour”. If nothing is a clean match, select “Other Experiences.”
- **Photo/Headline Principle**: Ask yourself, “If guests could only share one photo or social headline to summarize their experience, what would it show?” The answer must dictate your choice.

Label Definitions (With Distinct Differentiators and Boundary Rules)

- **Aerial Scenic Viewing**
- Experiences where guests spend almost all the time in an airborne vehicle (helicopter, plane, balloon, blimp) for the purpose of viewing landscapes, cities, or coastlines from above. The panoramic view is the only meaningful selling point; excludes adrenaline activities (skydiving, flying lessons), air transit, or any flight with substantial ground/water features or activities.

- **Bar Hopping**
- Deliberate, multi-stop journeys through at least two separate bars, pubs, or lounges, explicitly centered on sampling alcoholic drinks and experiencing local drinking culture. The social act of moving from venue to venue for nightlife is paramount; excludes tours focused on food, live entertainment, or single-location socializing.

- **City Sightseeing Cruises**
- Boat trips strictly operating within a city, designed for ongoing observation of city skylines, historic facades, and waterfront districts from the water. The city view is ever-present and promoted above all else. Excludes wildlife, nature, or suburban excursions, and boat rides with food/music as the chief draw.

- **Coastal & Scenic Tours**
- Ground or water journeys prioritizing sustained immersion in natural vistas: coastal cliffs, dramatic landscapes, beaches, or panoramic countryside. At no point does cityscape, cultural history, or wildlife become the prime focus. Brief town, food, or stopovers are acceptable only as minor interruptions.

- **Combo Experience Tour**
- Two or more distinct, label-eligible activities are not only both included, but each receives equal attention, marketing, and time. Only use this if *removing any one component destroys the tour's essence*, and if each segment would otherwise deserve its own label. Unbalanced or predominantly single-focus products do not qualify.

- **Culinary Exploration Tour**
- Tours devoted almost exclusively to discovering and tasting varied, authentic local foods in multiple locations—such as food markets, stalls, or regional specialty spots. Excludes tours with a beverage, nightlife, or instructional cooking foundation; hands-on cooking or wine/beer-centric products are classified elsewhere.

- **Cultural Show or Event**
- Attendance at a live, scheduled performance—dance, music, theater, ceremony, or festival, where the guest is seated or standing as an observer throughout. No hands-on participation or lesson format is present. All other elements (e.g., dinner, sightseeing) are non-central.

- **Historical Tour**
- Guided experiences emphasizing the interpretation, storytelling, and understanding of past people, events, or eras. The narrative thread and site visits revolve around learning history, not general sightseeing, iconic photography, or religious engagement. History must never be secondary to these other themes.

- **Iconic Sites Tour**
- Itineraries structured so the highlight is visiting one or more globally renowned, instantly recognized landmarks (e.g., Eiffel Tower, Christ the Redeemer). "Must-see" status and broad recognition are elevated above deeper cultural context, religion, or scenic immersion. Not valid for broad historical/cultural or city exploration.

- **Lisbon City Tours by Land**
- Exploration by walking, vehicle, bike, or tram within Lisbon's city boundaries—explicitly focusing on its neighborhoods, modern and historic streets, and urban culture. Tours focusing on hills, religion, or a list of icons are excluded unless those do not dominate.

- **Lisbon Hills**
- Experiences with a deliberate, ongoing emphasis on climbing, traversing, or admiring Lisbon's distinctive hills, staircases, and panoramic viewpoints ("miradouros"). If the city's verticality is not repeatedly stressed, do not use this label.

- **Marine Wildlife Experience**
- Journey totally dedicated to observing marine animals (dolphins, whales, seabirds, etc.) in their natural environment. Animal encounters must dominate the itinerary—brief or coincidental sightings on unrelated tours do not count. Controlled environments (aquariums, zoos) or non-wildlife boat trips are excluded.

- **Nearby Cities and Town**
- Excursions whose undiluted focus is intimate, immersive exposure to a named city or town (not Sintra, not Lisbon) within the greater region; includes architecture, local ambiance, markets, and daily urban life. Natural landscapes, vineyards, coastal routes, or multi-stop "best of" itineraries are not covered.

- **Other Experiences**
- A catch-all for any experience clearly outside all other definitions—such as interactive games, VR/AR, sports competitions, escape rooms, or unique hybrid journeys whose primary point cannot be cleanly assigned.

- **Professional Photoshoot**
- The product's crux is to have guests photographed by a professional at curated locations. Posing and receiving a portfolio of photos is the main reason for booking. Any storytelling, touring, or local engagement takes a back seat.

- **Sacred & Holy Sites**
- Structured visits to active or historically significant religious buildings/complexes (churches, mosques, shrines) where spiritual meaning, architecture, or ritual are the main attraction. General history, sightseeing, or non-religious landmarking are excluded; spirituality must be at the foreground.

- **Sintra & Coastal Highlights**
- Balanced, co-primary focus on Sintra's signature palaces/gardens AND on the adjacent coastline (e.g., Cabo da Roca). The two elements are woven inseparably into the itinerary's marketing, time, and narrative. If coverage is lopsided or superficial for either, assign a different label.

- **Sintra Day Tours**
- Tours entirely set within Sintra's territory, deeply exploring its palatial, natural, or historic elements. The itinerary never detours significantly to coastal environments, Lisbon, or other external towns.

- **Surf Lessons**
- Hands-on, in-water surfing instruction led by professionals, aimed at skill acquisition and confidence building. Unrelated activities (beach wellness, sightseeing, surfing as mere recreation) are excluded.

- **Transfer**
- Simple, direct transportation from point A to B without tour guidance, major commentary, or sightseeing stops. Any informative or recreational detour invalidates this label.

- **Vineyards & Wine Tasting**
- Guided visits for the purpose of tasting and learning about wine, set in working vineyards, wineries, or authentic cellars. The essence is discovering wine and its making; associated food is incidental, not central. Bar-based, general gastro, or sightseeing tours with token wine stops do not qualify.

- **Workshop or Class**
- Main purpose is live, hands-on instruction—culinary, craft, sport, or language—where each guest finishes with a new concrete skill, knowledge, or creation. Passive observation, meal sampling, or event attendance is insufficient.

- **Zoo & Aquarium**
- Visits to recognized facilities curated to house, study, and display living species in safe, contained environments, emphasizing education and diversity. Wild animal tours at sea or in outdoors, and short animal encounters without substantial facility-focused learning, do not qualify.

Assignment Directive

- Select the ONE label whose distinct definition best and most specifically matches the main activity, draw, and purpose described.
- Output ONLY the precise label—no explanation, no notes, no formatting.

APE-OPRO best Prompt for Lisbon

You are a specialist in classifying travel activities in the Lisbon area. Your task is to evaluate each product's description and CHOOSE EXACTLY ONE category label from the list below, applying advanced, context-aware JUDGMENT. If the description is subtle or indirect, infer the real emphasis of the experience as a seasoned traveler would. Adhere closely to this STRUCTURED SELECTION PROCESS:
CATEGORY LIST (choose ONE only):
- Aerial Scenic Viewing
- Bar Hopping
- City Sightseeing Cruises
- Coastal & Scenic Tours
- Combo Experience Tour
- Culinary Exploration Tour
- Cultural Show or Event
- Historical Tour
- Iconic Sites Tour
- Lisbon City Tours by Land
- Lisbon Hills
- Marine Wildlife Experience
- Nearby Cities and Town
- Other Experiences
- Professional Photoshoot
- Sacred & Holy Sites
- Sintra & Coastal Highlights
- Sintra Day Tours
- Surf Lessons
- Transfer
- Vineyards & Wine Tasting
- Workshop or Class
- Zoo & Aquarium
DECISION PROCESS:
STEP 1. ASSESS FOR MULTIPLE MAJOR ACTIVITIES:
- Does the offering present TWO OR MORE significant, separate main activities, modes (land+river), or destinations (like a dedicated Sintra DAY plus another day elsewhere)?
- If both are CENTRAL, select "Combo Experience Tour".
- If it's a substantial split between Sintra & the coast/Cascais, choose "Sintra & Coastal Highlights".
- If time is equally divided among multiple off-Sintra towns, select "Nearby Cities and Town".
- If not, continue to step 2.
STEP 2. DETERMINE THE PRIMARY THEME:
- Mostly about Sintra (palaces, gardens, e.g. Pena, Quinta da Regaleira), possibly with a brief coastal stop but not a 50/50 split? —> "Sintra Day Tours"
- Focused on Lisbon's hills, vistas, or elevation-based tours regardless of transport? —> "Lisbon Hills"
- Main experience is a ground tour of Lisbon (bus, van, tuk tuk, cycle, etc.) that's NOT centered on hills? —> "Lisbon City Tours by Land"
- On the river for landmark viewing, sunsets, music, or general city panoramas (even under a "sailing" label), but not primarily for animal spotting? —> "City Sightseeing Cruises"
- Mainly touring outside the city (by road or boat), scenic landscapes, rural/coastal highlights, not focused on Sintra? —> "Coastal & Scenic Tours"
- Intended to feature marine animals? —> "Marine Wildlife Experience"

- Is EATING/DRINKING (tasting/consuming, not being instructed) the main point? —> "Culinary Exploration Tour"
- Principal focus on wine tastings or vineyard settings? —> "Vineyards & Wine Tasting"
- Major purpose is a historical story, covering periods, wars, or historic evolutions, not key icons? —> "Historical Tour"
- Is visiting well-known landmarks (castles, world heritage, key sites) central? For overlap with history, if specific sites are distinguished, prefer 'Iconic Sites Tour'. —> "Iconic Sites Tour"
- Deep dive into sacred, pilgrimage, or holy sites (e.g., Fatima)? —> "Sacred & Holy Sites"
- Main reason is a class, session, or creative skill development (except surfing, which is separate)? —> "Workshop or Class"
- Is the experience primarily about learning to surf? —> "Surf Lessons"
- Main highlight is bar visits, pub crawls, or nightlife exploration? —> "Bar Hopping"
- Special event or live cultural performance at its core? —> "Cultural Show or Event"
- Chief service is providing professional photos? —> "Professional Photoshoot"
- Observing wild animals, zoos, or aquariums is the feature? —> "Zoo & Aquarium"
- Merely a transfer/logistics offering (bus, airport, etc.), not touring? —> "Transfer"
- Clearly doesn't fit anywhere or is totally miscellaneous, select "Other Experiences".

STEP 3. HANDLE SPECIAL CASES:
- If Sintra and one or more coastal areas (ex: Cascais/Estoril) both get major time and emphasis, choose "Sintra & Coastal Highlights".
- If Sintra is a minor side in a broad itinerary, or less important than visits to other nearby towns, pick "Nearby Cities and Town".
- If split by full days (Sintra one, a different town the next), pick "Combo Experience Tour".
- When torn between "Historical Tour" and "Iconic Sites Tour": If visiting specific landmarks or renowned sites is primary, go for "Iconic Sites Tour"; if broader history or era learning stands out, select "Historical Tour".

STEP 4. OUTPUT:
Return ONLY the single chosen label in JSON:
{
"label": "<Most appropriate label>"
}
Do NOT explain your reasoning. Do NOT list choices or reference the description. Give only the plain JSON.
SAMPLES:
1. A sunset cruise for city views and snacks: "label": "City Sightseeing Cruises"
2. Multi-day featuring Sintra and a separate day in Fatima: "label": "Combo Experience Tour"
3. Guided tour centering on historic UNESCO castles: "label": "Iconic Sites Tour"
4. Sintra-focused palaces/gardens tour with a short coastal visit: "label": "Sintra Day Tours"
Before outputting, carefully go through each step and strictly select the ONE most fitting label.

A.7 Variance Analysis

Figure 16 presents the variance in the weighted F1 scores on the training dataset. As anticipated, the standard deviation on the training data is relatively low across most methods when compared to the test data. This aligns with expectations, given that models typically fit better on training data. Notably, APE-OPRO exhibits a slightly higher standard deviation on the training set, averaging ±0.03, in contrast to ±0.0266 observed on the test set.

Figure 17 shows the standard deviation of the macro F1 score on both the test and training datasets. As briefly discussed in Section 4.4, macro F1 is inherently sensitive to class imbalance, particularly due to the label distribution in the test set. This sensitivity contributes to both the lower overall macro F1 scores and the higher observed standard deviation on the test data.

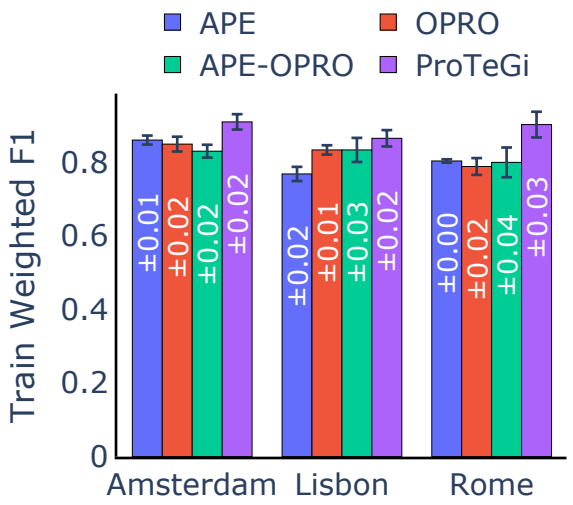

Figure 16: Variance in the train weighted F1 score for all methods, across three different destinations for five independent trials.

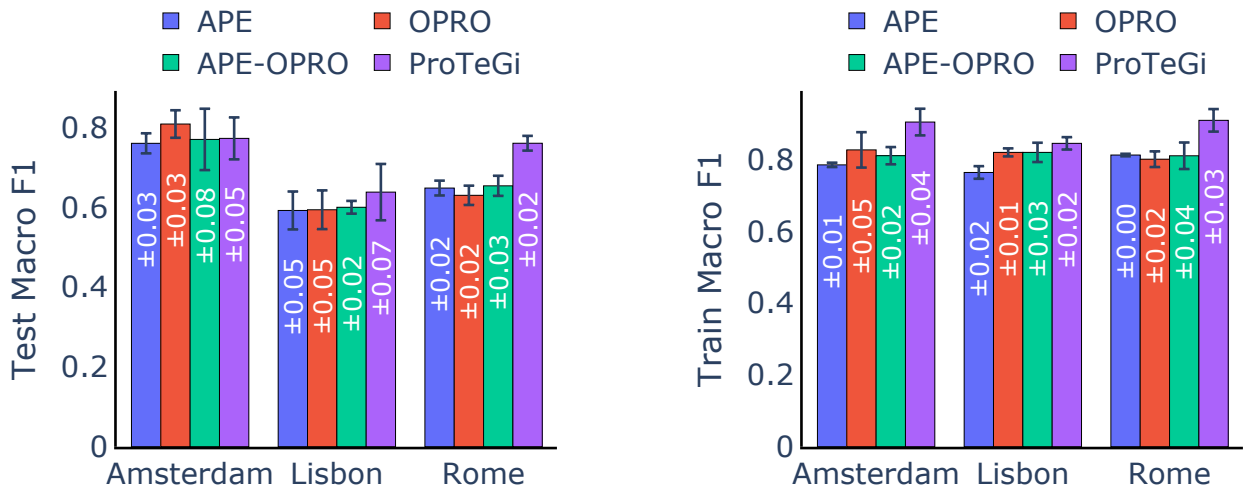

Figure 17: Evaluation of method generalizability across diverse destination datasets, showing the avg. weighted F1 score and standard deviation over five independent test and train trials per method.

A.8 Ablating on Depth

Figure 18 presents the test weighted F1 scores across three key destinations for all four methods, along with the corresponding costs incurred as depth increases. This plot is similar to the plot in section 5.5 with cost information included for each method for each destination.

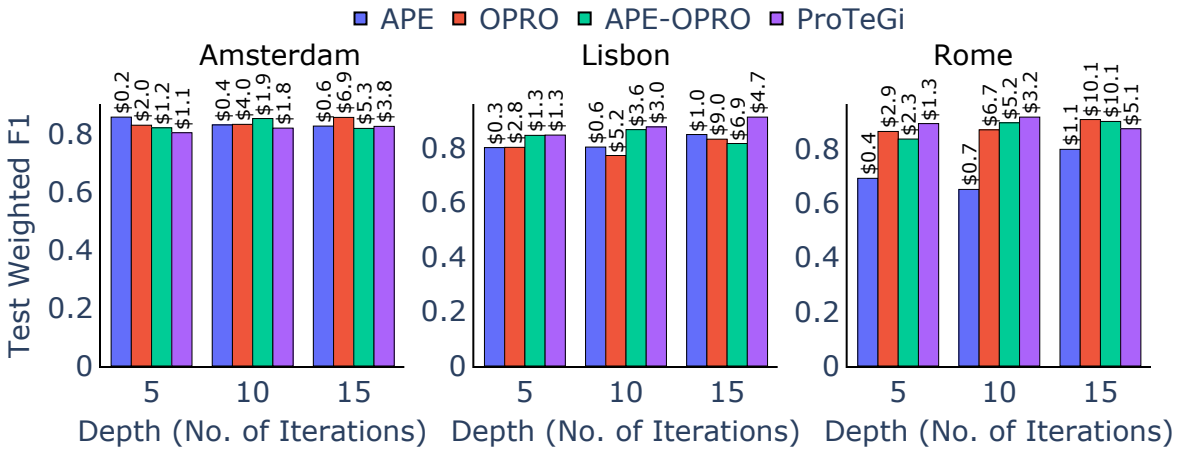

Figure 18: Change in Performance for depth or maximum number of iterations for each method across destinations along with the (API) cost incurred for each run of the respective algorithm.

Figure 19 presents the training weighted F1 scores across three key destinations for all four methods, along with the associated costs as depth increases. Note that the cost values are identical to those reported in Figure 9. Across all methods, diminishing returns in training weighted F1 were observed with increasing depth, suggesting that smaller depths may be sufficient for effective prompt optimization which is consistent with the pattern seen in the test weighted F1 results.

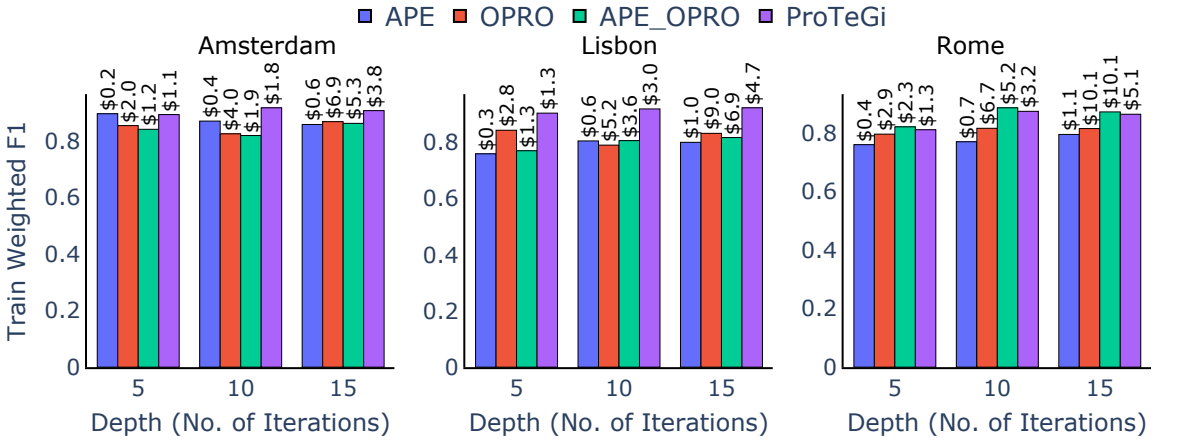

Figure 19: Caption

A.9 Ablating on Breadth

Figure 20 presents the test weighted F1 score and associated cost as breadth increases. This plot is similar to the plot in section 5.6 with cost information included for each method for each destination.

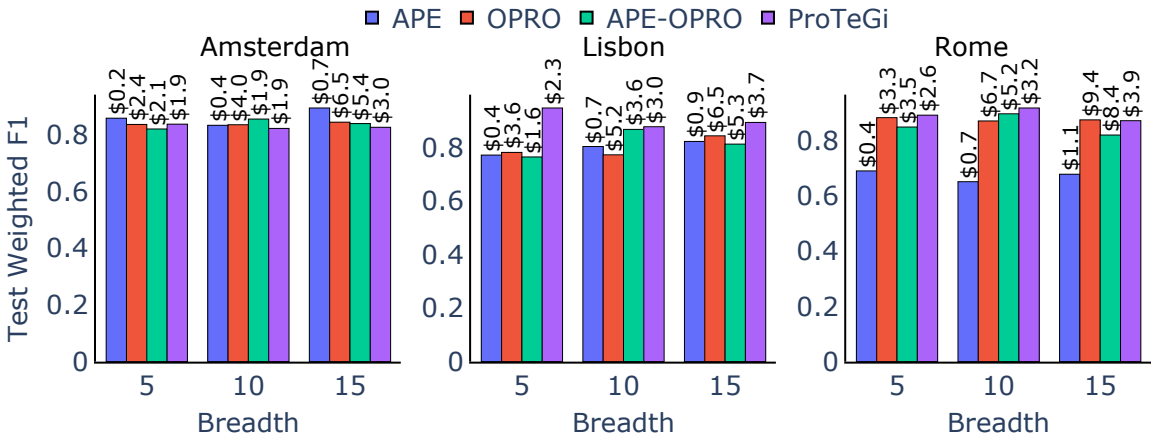

Figure 20: Change in Performance for breadth or maximum number of prompts in each iteration for each method across destinations

Figure 21 displays the training weighted F1 score and corresponding cost as breadth increases. The observed diminishing or negative returns in F1 performance suggest that lower breadth values are likely adequate for effective prompt optimization.

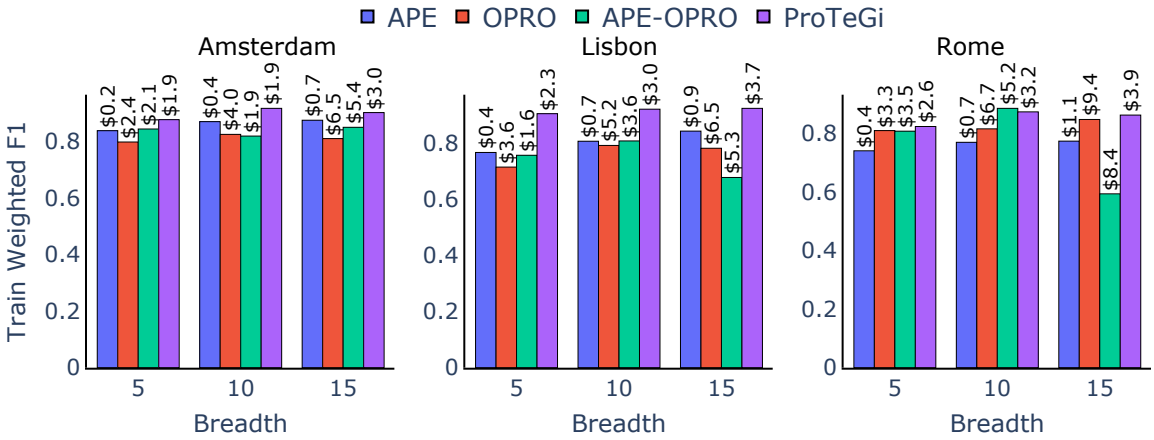

Figure 21: Caption

A.10 Convergence Analysis

The following graphs present the convergence analysis of APE, OPRO, and APE-OPRO for 3 key destinations and these are similar to the plot in section 5.4

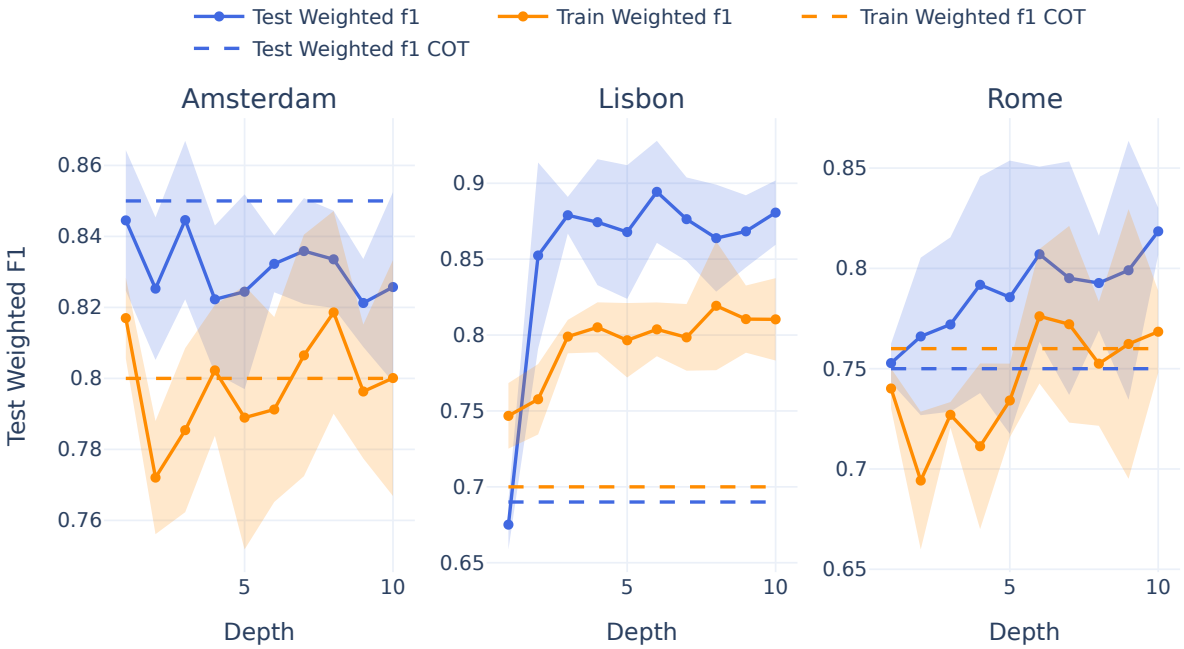

Figure 22: Convergence analysis for APE-OPRO on 3 key destinations

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

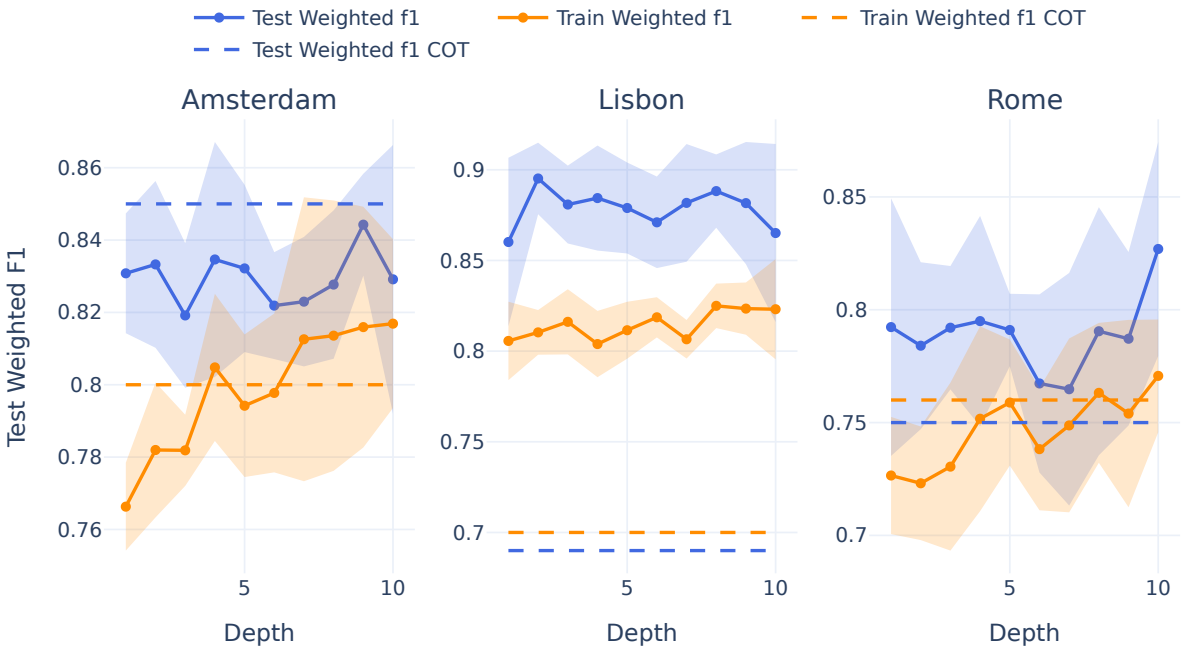

Figure 23: Convergence analysis for OPRO on 3 key destinations

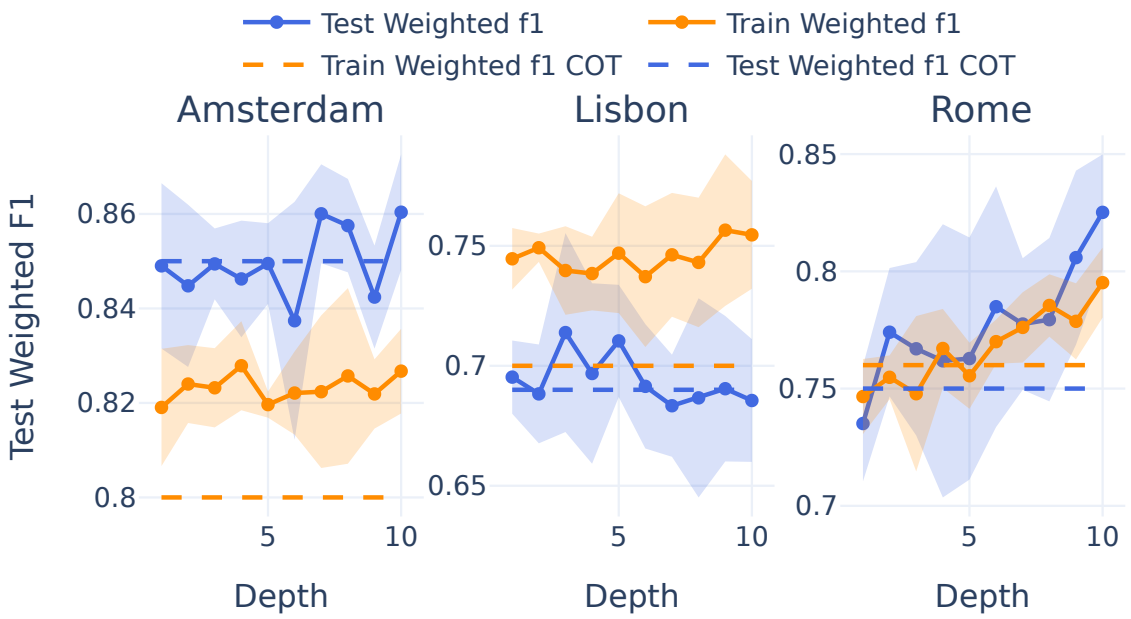

Figure 24: Convergence analysis for APE for 3 key destinations