
Task Facet Learning: A Structured Approach to Prompt Optimization

Gurushaa Juneja

Microsoft Research

gurushajuneja@gmail.com

Nagarajan Natarajan

Microsoft Research

nagarajn@microsoft.com

Hua Li

Microsoft Bing Ads

huli@microsoft.com

Jian Jiao

Microsoft Bing Ads

jian.jiao@microsoft.com

Amit Sharma

Microsoft Research

amshar@microsoft.com

Abstract

Given a task in the form of a basic description and its training examples, prompt optimization is the problem of synthesizing the given information into a text prompt for a large language model (LLM). Humans solve this problem by also considering the different facets that define a task (e.g., counter-examples, explanations, analogies) and including them in the prompt. However, it is unclear whether existing algorithmic approaches, based on iteratively editing a given prompt or automatically selecting a few in-context examples, can cover the multiple facets required to solve a complex task. In this work, we view prompt optimization as that of learning multiple facets of a task from a set of training examples. We identify and exploit structure in the prompt optimization problem — first, we find that prompts can be broken down into loosely coupled semantic sections that have a relatively independent effect on the prompt’s performance; second, we cluster the input space and use clustered batches so that the optimization procedure can learn the different facets of a task across batches. The resulting algorithm, UNIPROMPT, consists of a generative model to generate initial candidates for each prompt section; and a feedback mechanism that aggregates suggested edits from multiple mini-batches into a conceptual description for the section. Empirical evaluation on multiple datasets and a real-world task shows that prompts generated using UNIPROMPT obtain higher accuracy than human-tuned prompts and those from state-of-the-art methods. In particular, our algorithm can generate long, complex prompts that existing methods are unable to generate. Code for UNIPROMPT will be available at <https://aka.ms/uniprompt>.

1 Introduction

Given a task, choosing an input prompt is a key part of optimizing Large Language Model’s (LLM) performance [14, 31]. Minor changes in prompt can lead to performance gains or losses, necessitating prompt engineering [16] for effective model utilization. Typically, manually-developed prompts combine task description with a few in-context examples, along with modifiers like chain-of-thought [14]. For greater accuracy, human prompt engineers spend considerable time to identify errors with a current prompt, consider the different facets of a task (e.g., counter-examples, explanations, analogies) that may fix those errors, and include them in the prompt. For instance, for a hate speech classification task, in addition to the definition, it may be helpful to specify the facets that lead to hate speech: the context of conversation, identifying intent, and differentiating hate speech from opinions or closely-related concepts such as vulgarity and profanity.

To avoid the cumbersome process of manually creating prompts, recent work aims to automate the process of generating natural language prompts that are also interpretable. Since language tokens are discrete, this leads to a challenging discrete optimization problem with a combinatorial space of possible outputs. Techniques for prompt optimization can be divided in two categories: *non-directional*, e.g., random search [36, 35] and genetic algorithms [31, 7], where the sampling of new input is “random” and does not explicitly aim to reduce error on a train set; and *directional*, where the sampling of new input depends on some error measure on a representative train sample. Recently, more complex methods are proposed in the second category including reinforcement learning [33, 6] and updating prompts using auxiliary LLMs to reduce incorrect predictions over a train sample [12, 22]. While all these techniques focus on editing, adding, or deleting parts of a given prompt, they are developed with the goal of obtaining a concise description of the task. None of them focus on ensuring that multiple facets of a task are added to the prompt.

In this paper, we study how to optimize prompts to cover diverse, multiple facets of a task and improve overall accuracy. To develop a suitable algorithm, we analyze properties of the prompt optimization problem. First, we check how sensitive LLMs are to small changes in the input prompt, since directional optimization algorithms may not be as useful if LLMs are highly sensitive. Using a notion of probabilistic Lipschitz continuity, we find that LLMs become less sensitive with size. Across prompt edits with the same semantic meaning, Llama-13B accuracy fluctuates up to 63% of the minimum accuracy whereas the relative accuracy change in GPT-3.5 and GPT-4 models is 6% and 4% respectively. As a result, probabilistic Lipschitz constant for GPT-3.5 and GPT-4 is significantly smaller than that for Llama-13B, indicating that directional optimization methods may be best suited for larger models. Second, we study how LLM accuracy changes as more information is added to the prompt. Ideally, we would want a *diminishing marginal returns* (sub-modularity) property so that greedy addition of different facets can lead to a reasonable approximation of the true optimum. We find that diminishing returns property largely holds when adding different facets as new sections to the prompt, but it does not hold when directly adding in-context examples.

Our analysis provides two insights: 1) directional optimization may be feasible for models like GPT-3.5 having a low probabilistic Lipschitz constant; 2) including facets as additional text sections may lead to a well-behaved optimization problem. Based on these, we propose an algorithm, UNIPROMPT, that can generate the full prompt from scratch, given a one line description for a task. Specifically, we identify a prompt structure that provides different task facets as *sections* in the prompt and initializes them with text generated from a language model. To develop a high-accuracy prompt, we use an auxiliary LLM to obtain feedback about example predictions with the current prompt and decide to make edits at a section level (add, edit or delete) in each iteration. Our key insight is that feedback on a single example or a randomly selected batch of examples does not yield generalizable facet descriptions. Instead, we propose generating edits at a minibatch level and aggregating the edits at the batch level to yield a single edit that conveys a broad concept relevant to the task. To further increase the chances of facet learning, we cluster the input examples beforehand and use clustered mini-batches (see Figure 1).

We evaluate UNIPROMPT on zero-shot (no in-context examples) and few-shot settings on four different tasks. In the zero-shot setting, UNIPROMPT consistently achieves higher accuracy than existing prompt optimization methods. On E ethos, a hate speech dataset, UNIPROMPT obtains 94% accuracy whereas the next best method obtains 82%. In the few-shot setting, we compare UNIPROMPT to MedPrompt [21], a state-of-the-art prompt composition method. We find that UNIPROMPT, requiring only one LLM call at inference time, obtains the same accuracy as MedPrompt that requires five calls. If we allow multiple calls to UNIPROMPT, we obtain over 4% accuracy gains. We also evaluate UNIPROMPT on a real-world semantic matching task in a web search engine. Compared to the best manual prompt, the prompt generated from UNIPROMPT leads to over 5% increase in accuracy on the negative class and nearly 2% accuracy increase overall.

2 Characterizing the Prompt Optimization Problem

Assume access to black-box *solver LLM* $f : \mathcal{X} \rightarrow \mathbb{R}$ that takes as input a prompt $\mathbf{x} \in \mathcal{X}$ and outputs the average accuracy on a validation set D_v . Since the set of prompts is combinatorially large, we assume that all prompts can be embedded in a fixed p -dimensional space such that distance between two prompts in the vector space correspond to their semantic similarity. Thus, each

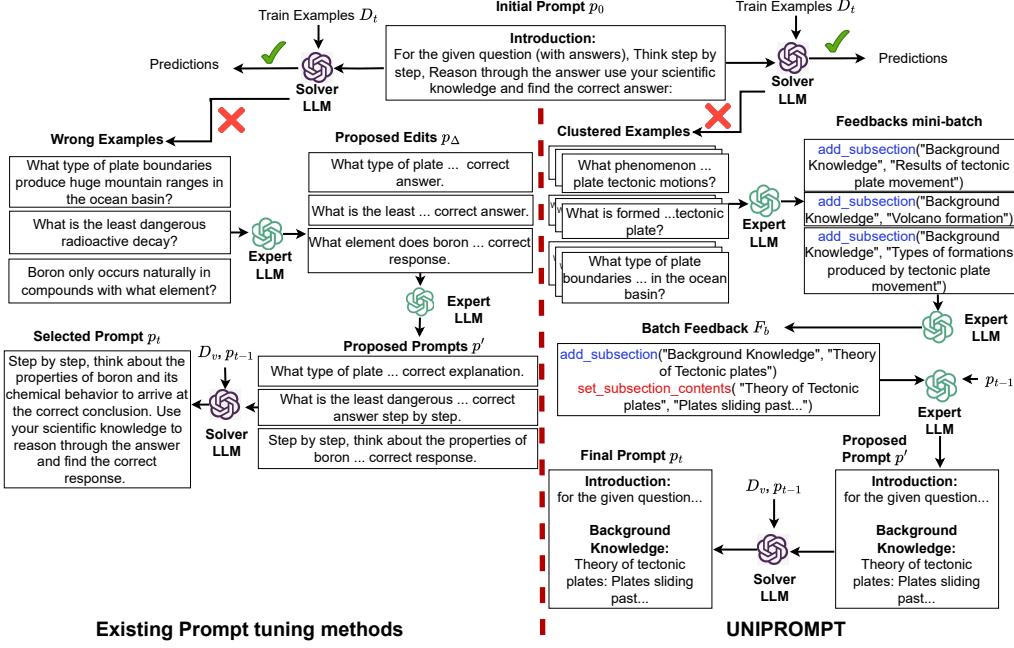


Figure 1: **Existing prompt optimization methods (left) versus UNIPROMPT (right) on the SciQ [29] dataset:** [Left] State-of-the-art prompt optimization methods such as **ProTeGi** [22] sample from the questions wrongly answered by the current prompt, and use an expert LLM (such as GPT-4) to obtain feedback on the mistakes made; This approach tends to overfit to specific examples, as can be seen from the proposed prompts p'_t ; leading to limited generalization. [Right] In contrast, UNIPROMPT identifies key task *facets* that generalize across examples by: (1) clustering examples into groups with similar task facets, and (2) employing a two-tier feedback-based update strategy. The resulting prompt updates extract generalizable concepts from the specific examples seen.

prompt \mathbf{x} can be expressed as a vector in \mathbb{R}^p . The prompt optimization problem can be written as $\arg \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}; D_v)$.

We do not have access to the gradients, hence it is a zeroth-order optimization problem. For optimization, we consider the class of sequential algorithms that decide the next prompt based on all previous \langle prompt, output accuracy \rangle pairs and a fixed task-specific training set that provides (labelled) input-output pairs for the task.

2.1 Is directional optimization feasible?

Previous work has shown that LLMs can be brittle to their input: changing the prompt slightly can create a significant difference in performance [37]. To understand whether optimization algorithms may work, it is important to understand how the surface of f looks. If f is brittle everywhere to small changes, then random search over prompts (such as in [36]) may be a competitive method; it is unclear whether optimization can help. On the other hand, if the change in f given small changes to the input prompt is bounded, then optimization algorithms can hope to perform significantly better than random search.

More formally, we ask whether the prompt optimization problem is well-conditioned. Typically, conditioning can be determined by the Hessian. However, since f is assumed to be black-box, we approximate it by measuring sensitivity, or more specifically, Lipschitz continuity near the optimal solution. Based on prior work on defining continuity of neural networks [18], we use a probabilistic notion.

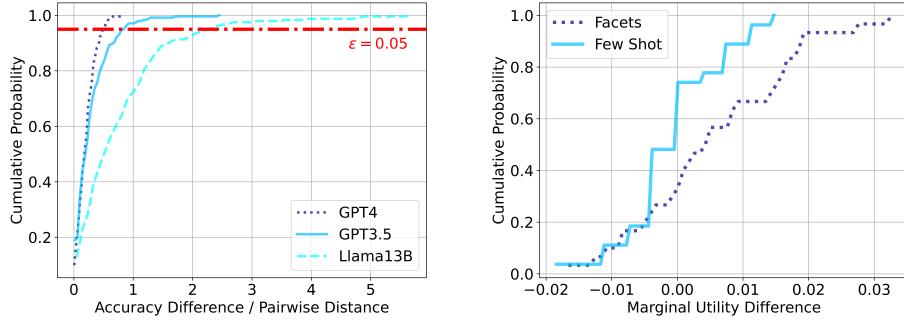


Figure 2: On the Ethos dataset, (a) estimating (probabilistic) Lipschitz constant of models (Definition 1); (b) diminishing marginal utility (submodularity) of GPT-3.5 on two types of prompting strategies.

Definition 1 Probabilistic Lipschitz Continuity [18]. Given a probability distribution over inputs \mathcal{X} , $r \geq 0$, a function $f : \mathcal{X} \rightarrow \mathbb{R}$ is (L, ϵ) -probabilistically Lipschitz with constant $L \geq 0$, if

$$\Pr_{\mathbf{x}, \mathbf{x}' \sim \mathcal{X}} [d(f(\mathbf{x}), f(\mathbf{x}')) \leq L \cdot d(\mathbf{x}, \mathbf{x}') \mid d(\mathbf{x}, \mathbf{x}') \leq r] \geq 1 - \epsilon \quad (1)$$

where d is a distance measure such as ℓ_1 or ℓ_2 norm.

Note the focus on small changes in input through the parameter r . Intuitively, the Lipschitz property bounds the maximum change in f given a small change in input prompt. Typically, the lower bound of error for any sequential optimization algorithm over f is directly proportional to the Lipschitz constant L [17]. Therefore, for faster convergence, it is desirable to have a low L , especially near the optimal solution. Empirically, we estimate L by sampling task-relevant prompts so that they are close to the optimal solution. Then we make small changes to the prompt such that the semantic meaning stays the same and measure the change in f . We show the change in f per change in input for GPT-4, GPT-3.5, and Llama2-13B models in Figure 2(a) for the Ethos dataset commonly used in prompt optimization [22, 19, 32, 23, 10]. Assuming $\epsilon = 0.05$, probabilistic Lipschitz constant L for GPT-3.5 and GPT-4 is < 1 , whereas it is over three times more for Llama2-13B. Thus, as the size of the model increases, the probabilistic Lipschitz constant decreases, indicating that larger models like GPT-3.5 and GPT-4 are more robust to small changes and thus amenable to optimization algorithms. For Llama, the absolute change in accuracy ranges between 0.33 to 0.54, indicating a 63% change in accuracy even across prompts with the same semantic meaning. In contrast, the relative change in accuracy for GPT-3.5 and GPT-4 is 6% and 4% respectively. (See Appendix A.1 for experimental details)

2.2 Can higher-order abstractions of task examples help prompt optimization?

Consider in-context learning [3], where the prompt \mathbf{x} is composed of a simple task description and a set of labeled examples from the task. The optimal prompt should consider all the aspects of the task which can only be discovered by exhaustively looking at the examples. But, in practice, we are constrained by (a) availability of a limited number of training examples, (b) prompt size. It is beneficial for in-context learning to consider the marginal utility of examples [34, 8]. For instance, it would help to add examples where the model fails rather than ones where the model already succeeds (see Table 4 in Appendix). This suggests that one could use a greedy algorithm for iteratively optimizing the prompt by finding failing examples and adding to the prompt. *But would such a procedure converge to a good solution?*

We can view $f(\mathbf{x})$ as a set function over the examples (that compose \mathbf{x}). A natural property to examine is *submodularity* (where greedy optimization provably converges to a good solution under some conditions [15]), i.e., the property of diminishing marginal utility of examples. The marginal utility of adding an example e to the few-shot prompt corresponding to a set of examples A is $\Delta_A = f(A \cup \{e\}) - f(A)$. Let $B \subset A$ denote a subset of A . We define f_{fewshot} to be the validation accuracy of GPT-3.5 using few-shot prompts on the Ethos dataset (see Section 4). In Figure 2 (b), we show the distribution of diminishing marginal utilities for f_{fewshot} (light blue), i.e., we sample multiple

A, B sets of examples from the dataset and compute $\Delta_B - \Delta_A$, which is expected to be non-negative for submodular functions. As we see, the marginal utility differences are nearly uniformly random in $[-0.01, 0.01]$, indicating that there is no submodularity structure to f_{fewshot} .

A recurring theme [1, 2] in the prompts that humans engineer is to organize task information in “sections” (introduction, context, etc.), treating prompt like a structured document. When humans engineer a prompt, they do it systematically by editing specific sections of the prompt. We hypothesize that there are a small number of *facets* that collectively define a task, which can be discovered by inspecting a representative set of examples from the data distribution. Consider the task of tagging hate speech. Identifying intent to harm, recognizing general aggression versus targeted hate, and differentiating hate from other objectionable content are some facets of the task. We ask whether there is submodularity structure in adding different facets (text sections) to a prompt, compared to adding in-context examples. The answer to this question depends on multiple factors: (1) what are the contents of the sections and how are they generated?; (2) how does $f(\mathbf{x})$ behave w.r.t. the sections that make up the prompt \mathbf{x} ?

Figure 2 (b) shows the distribution of marginal utility differences for f_{facets} (dark blue), i.e., we sample multiple A, B sets of *facets* (i.e., short sections) that are generated using a Llama2-13B model, and compute $\Delta_B - \Delta_A$ (defined as above, but the added element e is now a section) using GPT-3.5 as the solver LLM (See Appendix A.2 for experimental details). We see that the diminishing marginal returns property largely holds for f_{facets} , unlike in the case of f_{fewshot} . This suggests that greedy strategies for prompt optimization problem over sections could work — but this crucially relies on how we generate the sections themselves.

The prompt optimization problem can now be posed as one of inferring the facets of a task from a set of training examples. In the following section, we propose a method that iteratively learns the facets of the task as sections of the prompt.

3 UniPrompt: Generating a Prompt that Captures Multiple Facets of a Task

Manual prompt engineering consists broadly of two steps: 1) **generating an initial prompt**; and 2) **identifying errors and editing the prompt**. The initial prompt for a task is constructed based on domain knowledge. Typically, the initial prompt contains implicit or explicit *sections* (e.g., introduction, task description, intended application, edge-cases, challenges, etc.) wherein each section corresponds to a relevant task facet. Given the initial prompt and a solver LLM, a human prompt engineer would look at the incorrect predictions, form hypotheses on which section of the prompt needs to be edited or whether a new section needs to be added, and then propose edits that may fix the errors. Then, the prompt engineer may pick the edit with the highest validation accuracy. These steps are repeated in a loop, until the validation accuracy saturates. A good engineer may also backtrack from an optimization trace and start from some previous prompt, in case the current optimization trajectory yields unsatisfactory results.

We present UNIPROMPT, a prompt optimization algorithm that aims to mimic the above process. Specifically, the algorithm extracts key concepts or facets relevant to a task and updates prompt sections using them, with the overall goal of increasing validation accuracy. Our analysis from Section 2 indicates that such a directional optimization procedure based on iteratively adding task facets can converge to a good solution, especially for solver LLMs such as GPT-3.5 and GPT-4.

We assume that the algorithm receives as input a one-line task description and a set of labelled $\langle \text{input}, \text{output} \rangle$ demonstrations. The algorithm proceeds in two stages:

1) Task Facets Initialization using background knowledge. In the first stage, the algorithm leverages a language model to generate a prompt with multiple facets, i.e., sections focusing on specific task aspects based on the task description. It greedily selects content for each facet from a pool of candidates generated using the language model.

2) Task Facets Refinement using examples. In the second stage, the algorithm refines the prompt using a set of training examples and an auxiliary LLM. Given a candidate prompt, it collects examples with prediction errors by the solver LLM, extracts key concepts or facets required to fix those errors, and makes appropriate edits. An *edit* may create a new section, modify an existing section, or delete an existing section. Additionally, we maintain a beam of best performing prompts to backtrack and restart from a previous prompt if the current optimization path is unsatisfactory.

Algorithm 1: UNIPROMPT

Input: Train set D_t , validation set D_v , initial prompt for the task p_0 , one-line task description T
Output: Optimized prompt P^* for the given task

```
1 Cluster train set, initialize history and validation accuracy arrays:  $C \leftarrow \text{cluster}(D_t)$ ,  $H \leftarrow \{\}$ ,  
     $V \leftarrow []$ ;  
2 Initialize a beam of size 2 with the initial prompt:  $p_1 \leftarrow p_0$  and  $p_2 \leftarrow p_0$  for each cluster c in C do  
3    $B \leftarrow \text{batches}(C)$ ;  
4   for each batch b in B do  
5      $M \leftarrow \text{mini-batches}(B)$   
6      $F \leftarrow []$   
7     for each mini-batch m in M do  
8       Evaluate the best prompt on mini-batch:  $a_m \leftarrow \text{LLM}(m, p_1)$   
9       Get feedback from the expert given history of mini-batch, accuracy and task  
         description:  $f \leftarrow \text{Feedback}(T, a_m, H[m])$   
10       $F.\text{insert}(f)$   
11     Combine feedbacks over a batch:  $F_b \leftarrow \text{Combine}(F)$   
12     Apply feedback to get updated prompts;  $q_1 \leftarrow \text{apply}(F_b, p_1)$ ;  $q_2 \leftarrow \text{apply}(F_b, p_2)$   
13     Update the beam: if not( $p_1 = p_0$ ) then  
14        $p_2 \leftarrow \text{second-high-acc}([p_1, p_2, q_1, q_2], b)$   
15      $p_1 \leftarrow \text{highest-acc}([p_1, q_1, q_2], b)$   
16   Evaluate the best prompt on validation set:  $acc_v \leftarrow \text{evaluate}(p_1, D_v)$   
17    $V \leftarrow V.\text{append}(acc_v)$   
18   Stop based on the validation accuracy:  $c \leftarrow \text{early-stop-criteria}(V)$   
19   if c then  
20     break  
21 return  $p_1$ ;
```

Similar to UNIPROMPT, previous methods [12, 22] propose using an auxiliary LLM to provide feedback on incorrect predictions and propose edits to a given prompt. In our initial experiments, however, we find that collecting feedback over individual examples or randomly sampled batches leads to memorization of individual examples (see Figure 1) rather than recognition of facets that are important to the task. To avoid overfitting to the train examples and to extract generalizable concepts, we make two contributions. First, we follow a two-tier setup of synthesizing feedback for a batch of training examples. We break up a batch into mini-batches, collect feedback on each of the mini-batches and then use a separate prompt to aggregate the different feedback texts into a generalizable concept. Second, to increase chances that a mini-batch corresponds to a coherent facet, we cluster the training data beforehand and ensure that each mini-batch consists of examples from the same cluster. We provide details of the algorithm below (see Algorithm 1).

Notation: We denote the training set with D_t where each example is a question-answer pair of the form $\{q_i, a_i\}$. The validation set is denoted as D_v . We have N training examples and K validation examples. Input to the algorithm is the solver LLM f_{LLM} , train set D_t , validation set D_v , initial prompt for the task p_0 , one-line task description T . In addition, we assume access to an “expert” auxiliary LLM such as GPT-4.

3.1 Task facet initialization using background knowledge

While existing methods for prompt optimization [12, 22] iteratively edit a human-written prompt, our goal is to generate the entire prompt from scratch (specifically, based only on a one line description of the task). Therefore, one option is to initialize the prompt using only the task description. We call this *Task Description Initialization* and initialize the prompt with a single section titled *Introduction* containing the input task description.

In the second kind of initialization, we finetune a language model to generate a prompt with multiple diverse sections such as Introduction, Tricks, and Corner Cases; similar to the initial prompt that

a human prompt engineer may produce. Here we use the world knowledge embedded in language models that are trained on extensive text data. The language model is finetuned to output the entire section contents given the task description and the section title. Specifically, we train a Llama2-13B model using a synthetic dataset generated using GPT-4. Across a wide range of tasks, we first use GPT-4 to identify the important sections given the task, section descriptions and a few examples (See Appendix A.6 for the prompt). We then fine-tune a language model ϕ , that is Llama2-13B, in a supervised setting on a synthetically created dataset of the form $\langle T, P_i, S_i, \mathbf{S}'_{\text{gold}} \rangle$ where T denotes the task description, P_i is the prompt generated using the fine-tuned model sequentially before section S_i (i.e., it would have optimized sections $\{S_1, S_2, \dots, S_{i-1}\}$), and $\mathbf{S}'_{\text{gold}}$ is the gold section contents. ϕ is finetuned on the objective: $\min_{\phi} [-\log(p_{\phi}(\mathbf{S}'_{\text{gold}}|T, P_i, D_t, S_i))]$. We then sample 10 candidates for each section using the fine-tuned model and evaluate each section added incrementally against D_v and select the best performing candidate for each section. More details on data collection and training of SLM are given in the Appendix A.4. Examples of one line task description and the Llama-initialized prompt for different tasks are in Appendix A.3.

3.2 Task facet refinement using examples

Extracting task-relevant concepts from a set of examples to refine a prompt is a complex problem comprising multiple steps. Given a set of incorrect predictions, one needs to analyze what went wrong in each prediction, form hypotheses, aggregate the hypotheses to identify specific concepts that are relevant for the task, and then for each concept, one needs to attribute which facet/section of the current prompt needs to be edited to incorporate the concept. These operations are highly model-specific and are difficult to execute reliably. Therefore, for this stage, we exclusively rely on an “expert” auxiliary LLM such as GPT-4.

The whole process is executed through a series of prompts to the expert LLM. First, we prompt the expert LLM to diagnose mistakes (*feedback*) in each example given the answer and chain-of-thought reasoning produced by the solver LLM. Subsequently, we use this feedback to generate precise edits for the prompt that may fix the error. These individual edits are then aggregated over a mini-batch and fed back into the same LLM, which then identifies a few major edits to be applied to the current prompt. To aid in identifying major edits that correspond to generalizable facets, we propose to cluster the examples as a preprocessing step and create clustered batches, such that each cluster shares some common facet of the task.

3.2.1 Preprocessing step: Clustering examples to facilitate facet identification

We explore two approaches for clustering: *topic based clustering*, and *feedback based clustering*.

Topic-Based Clustering. Given a set of examples, we identify l topics that would span the entire train set. This type of clustering is motivated by the observation that solver LLM may make similar mistakes on examples from the same topic. Hence, for such examples, a common edit to the prompt could improve predictions for all the examples. To obtain the clusters, the expert LLM is prompted (for prompt see Appendix A.7) to provide a broad sub-topic t_i for each question. Then the resultant list of sub-topics $\{t_1, t_2, \dots, t_N\}$ is again clustered into k topics $\{t'_1, t'_2, \dots, t'_l\}$ by prompting the expert LLM. Based on this clustering, each example q_i, a_i is assigned a cluster t'_j corresponding to t_i .

Feedback-Based Clustering. Another way to find examples that share similar task facets is the feedback they receive based on the initial prompt’s predictions. For instance, for a physics-based task, if two examples from different topics obtain the same feedback from the expert LLM to edit the “Rules” section of the prompt and include the statement, “Draw all forces on each body before writing the equations”, then we argue that such examples can be clustered. This type of clustering makes the broad edit identification step easier. To obtain the clusters, we first evaluate all training examples against the initial prompt p_0 and store the feedback f_i from the expert LLM, corresponding to each incorrectly answered example q_i, a_i (all the correctly answered questions form one cluster). We then prompt the expert LLM to cluster these feedbacks $\{f_1, f_2, \dots, f_N\}$ into l clusters $\{f'_1, f'_2, \dots, f'_l\}$ (see Appendix A.8). For each cluster f'_j , we create a batch of examples q_i, a_i corresponding to each f_i in f'_j .

3.2.2 Generating prompt edits that generalize to multiple examples

Two-tier Feedback. To encourage generalizable feedback from the expert LLM that applies to more than one example, we obtain feedback at two levels: mini-batch and batch. Given a batch (created using clustering discussed above), we break it up into mini-batches. For each mini-batch m , we consider the questions for which the solver LLM answered incorrectly with the current prompt. For these questions, we construct a prompt consisting of the questions, the chain-of-thought produced by the solver LLM, their incorrect predicted answers and the ground-truth answers. We ask the expert to provide one feedback for the mini batch (prompt is provided in Appendix A.9). In the prompt, we specify the following types of edits that the expert LLM can suggest: add a section or subsection, delete a section or subsection, and edit a section or subsection.

Given the different feedback texts for mini-batches within a batch b , the next step is use another call to the expert LLM to summarize the given set of edits and provide a single summarized section update. This combination ensures some degree of smoothness at every update which helps stabilize training. To make sure that the expert is able to generate generalizable edits, we additionally provide a random set of incorrect examples that are not in the current batch and ask it to suggest an edit based on the existing edits that can correct the errors. As before, the class of edits allowed is the same.

History for effective exploration: To ensure comprehensive and non-repetitive exploration of prompts, we also provide the batch-level history of edits in the mini-batch-level prompt. Specifically, we provide the expert with the history of edits $H[b]$ [12, 31] for the batch b and the corresponding accuracy difference that each edit caused over the previous edit for the given batch. History is presented in the format $\{e_i, acc_i - acc_{i-1}\}$ where e_i is the edit proposed at the i^{th} update and acc_i is the accuracy of the i^{th} updated prompt (See Appendix A.9 for the full prompt).

3.2.3 Editing the prompt

Once the final set of edits is received for a batch, we use the expert LLM to apply edits to the current prompt (see Appendix A.10 for the prompt). An edit is accepted only if it increases the validation accuracy compared to the current prompt. We call this approach “*Greedy*”. Alternatively, we maintain a beam of l best performing prompts based on validation accuracy, apply the edit to all the l prompts, and update the beam to retain the top l performing prompts. We call this method “*Beam*”. To avoid overfitting on the train examples (or keep adding unnecessary information to the prompt), we employ early stopping in the optimization process (More details given in Section 4).

3.3 Efficiency considerations: Computational Complexity

The computational complexity of UNIPROMPT is dominated by the training phase, and in particular, by the queries made to the solver and the expert LLMs; at inference time, we need a single query to the solver LLM using the learnt prompt and the given test example. This is particularly well-suited for settings such as recommendation systems, where inference time efficiency is critical. In contrast, methods such as MedPrompt [21] require multiple calls at test time for a single test example. Let us now consider the compute complexity in terms of the number of expert or solver LLM calls made per epoch, stage-wise.

Clustering: First, we evaluate all the training examples using the current prompt. Second, for every wrongly predicted example, we obtain feedback from the expert LLM. Third, for the given set of feedbacks, we use a single call to cluster it into l clusters. Each of the above steps incurs $O(N)$ queries, so the total query complexity of the clustering stage is $O(N)$. Finally, for each example, i.e., (question, answer) pair, we simply map it to the l clusters (no LLM calls). This is a one-time cost.

Mini-batch feedback and Batch-level aggregation: At a given epoch, we evaluate every question in the mini-batch using the current prompt and the solver LLM (N queries overall). Next, we obtain one feedback over all the wrong questions in the mini-batch m ($N/|m|$ queries). We use one call to aggregate these feedbacks. For prompt selection, we evaluate $2k$ prompts on the batch b , where k is the beam size ($O(2k|b|)$ queries per batch). Hence overall query complexity is $N + N/|m| + 2kN + 1$ or $O(kN)$. In practice, k is a small number (2 in our experiments).

Assuming LLM throughput of 0.5 queries per second, a training + validation set of 300 examples, $l = 10$ clusters, beam size $k = 2$ and 20 epochs, it takes under 7 hours to train.

4 Experiments

Datasets We perform evaluation on four standard datasets : (1) Ethos [20], (2) ARC [4], (3) MedQA [13], and (4) GSM8K [5]. Ethos, ARC, and MedQA contain multiple choice questions, and GSM8K contains questions with integer answers.

Implementation details We set the initial prompt p_0 for each task as the one line task description and use 200 data points (q_i, a_i) as the train set. We use GPT-3.5-Turbo as the solver model, that is used in a chain-of-thought setting. For Feedback and Combine, we use GPT-4 as the expert. We maintain a beam size of 2. Mini-batch sizes (and batch sizes) are constrained by the context length of GPT-4. We find that mini-batch sizes 3 to 5 and batch sizes 5 to 7 work the best for our datasets. The temperature of the LLMs for our method is set to 0 for reproducibility of results. We employ early stopping at batch-level in UNIPROMPT. We find that a *patience* value of 5 works best for the given tasks.

Baselines We compare UNIPROMPT with the following techniques: (1) **Task Description**: prompt is the one line task description that we use to initialize UNIPROMPT; (2) **Chain-Of-Thought** (or CoT) prompting [14]; (3) **Expert Prompt**: the prompt optimized by humans taken from prior works [21]; (4) **OPRO** [31], that uses LLMs for discrete optimization over text prompts; (5) **ProTeGi** [22] that proposes textual gradients and selects edits to prompts using bandit techniques; (6) **Evoke** [12] that uses two instances of LLM, one that scores the current prompt, and the other that edits the prompt; (7) **EvoPrompt** [7] that uses genetic algorithms to search through the space of prompts; and (8) **MedPrompt** [21] that effectively combines multiple prompting strategies such as kNN-based few-shot selection, CoT reasoning, and ensembling.

Metrics For each dataset, we perform 5-fold validation (we split the train set into 5 folds) and report the final test accuracy as the average accuracy of the 5 runs on the test sets.

4.1 How does UNIPROMPT compare to existing methods in zero-shot setting

In the zero-shot setting, we do not include labeled examples in the prompt for any of the compared methods. We report results for four versions of our method in Table 1; they differ in the initial prompt—task description, expert prompt, or the prompt generated by the fine-tuned Llama model (discussed in Section 3.1)—and the combining strategy (from Section 3.2.3)—beam search vs greedy.

UNIPROMPT variants significantly outperform the baselines including CoT and the state-of-the-art prompt optimization techniques like ProTeGi that crucially leverage LLMs for performing iterative prompt edits. We achieve maximum gains on the Ethos dataset with a 18.2% increase in accuracy over the expert prompt. Further, we see accuracy increases of 4.0% on MedQA, 3.5% on GSM8k, and 7.6% on ARC-Challange datasets.

An example of evolution of prompts using our algorithm is given in Appendix4. It starts with a simple description of task and adds important details like differentiating between hate speech and rudeness. In contrast, **ProTeGi** [22] yields a rather terse prompt on the same dataset: “Does the following text contain language that targets a group of people based on their religion, gender, or other personal characteristics?”. This prompt does not provide various facets of the task, the way our method infers or an expert would typically write.

The training curves in Figure 3 show that our method initially performs edits on the prompt that simultaneously increase the train as well as the validation accuracy. After about 10 or 15 iterations (each batch update is an iteration), validation accuracy decreases while train accuracy continues increasing, indicating overfitting; which we overcome by using early stopping.

4.2 How does UNIPROMPT perform compared to SOTA in the few-shot setting?

We now compare the performance of UNIPROMPT, in the few-shot setting with GPT-4 as the solver model, to the SOTA prompting technique MedPrompt [21] that employs three key components: (1) few-shot prompting, where they dynamically select five relevant examples using k-nearest neighbors (kNN) clustering in the embedding space; (2) Chain-of-Thought (CoT) reasoning on the selected examples; and (3) self-consistency and ensembling with option shuffling. From Table 2, we observe that UNIPROMPT (first col), which requires only one call at inference time, performs almost as well as MedPrompt (last col), which requires five calls, on the MedQA, PubMedQA, and MedMCQA

Table 1: Test accuracies for the compared methods with GPT-3.5-Turbo as the solver model in the zero-shot setting (**best** in bold; second best underlined). The last four rows correspond to our proposed method.

Method	Ethos	ARC	MedQA	GSM8K
Task Description	76.8	79.7	52.7	59.4
Expert Prompt	74.1	<u>78.4</u>	53.1	78.9
Llama Prompt (Section 3.1)	74.0	<u>89.7</u>	52.6	79.5
CoT	72.0	79.4	50.3	76.3
OPRO	65.4	79.1	53.3	77.1
ProTeGi	76.0	78.8	52.9	77.3
Evoke	63.5	89.0	52.8	81.0
EvoPrompt	81.6	89.9	50.3	81.4
UNIPROMPT (Init = Expert Prompt) + Beam	84.0	86.0	52.3	82.4
UNIPROMPT (Init = Task Description) + Beam	<u>92.3</u>	<u>86.0</u>	57.1	82.4
UNIPROMPT (Init = Llama Prompt) + Beam	92.0	90.5	55.5	81.5
UNIPROMPT (Init = Task Description) + Greedy	93.7	90.5	<u>55.5</u>	82.3

datasets. As we incrementally add components such as kNN few-shot, CoT, and Ensemble to our prompt, we see a significant increase in accuracy of 4.35% on average across all datasets.

4.3 Results on a Real-world Task: Search Query Intent

We consider the task of inferring if two search queries share identical intent or not. This is an important task arising in search and recommendation pipelines, and is tackled today using LLMs. It is challenging to tell if two queries share the same intent, as it requires domain knowledge (e.g., brands and product categories), depends on geographical biases (e.g., “cricket” and “cricket game” are likely to mean the same in UK, but unlikely in the US), and even on cultural connotation. So, examples are crucial for understanding the task and engineering a prompt that generalizes well.

We sample real user queries from a proprietary application, rewrite them using ML models, and ask expert judges to label the query-pairs as identical or otherwise based on prescribed guidelines. We use a set of 200 examples as training data, and an additional 50 examples as validation set, to learn a prompt using UNIPROMPT, starting from the one-line description: *Tell if query A and query B have same intent or not*. The dataset is heavily biased towards positive samples, so the metric of success is improvement in accuracy, over the best manually-engineered prompt, on the positive and negative classes individually. For testing, we use a separate labelled set of 2527 examples from two geographies — one where the training data was sampled from, and the other unseen.

The prompt obtained using UNIPROMPT improves over the best manual prompt by 5.77% on the negative (rare) class, by 0.23% on the positive class, and by 1.86% overall on the test set. The learnt prompt captures the following facets of the task: (1) recognizing variations in names and abbreviations, and how they do not necessarily change the context; (2) recognizing the specificity of brands, and how even minor variations do change the context; and (3) recognizing the specificity of terms in queries, and how lack of specific terms can indicate departure of intent.

4.4 Do diverse task facets organized as sections really help?

We now revisit our hypothesis and motivation in Section 2.2. We want to empirically validate if all the diverse task facets that UNIPROMPT learns indeed contribute to the performance gains that we observe in Table 1. We consider two ablations:

- 1) We successively remove each facet (i.e., sections) in the learnt prompt for the task and report the performances of the prompts with fewer facets. In Figure 3, for the Ethos dataset, we see that almost every additional facet contributes to non-trivial gains in accuracy.
- 2) Could we have captured the information differently and retained the performance? We do a simple experiment – we summarize all the facets (i.e., learnt prompt) and evaluate the resulting prompt. In Figure 3 (right) (green line), we see that the summarized prompt has a significant accuracy drop.

4.5 Ablations: Impact of Clustering, Inclusion of History, and Greedy Update

Here we study effect of our design choices, namely clustering (Section 3.2), inclusion of history of past prompts (Section 3.2), and the greedy update rule (Section 3.2.3). The results are shown in Table 3. We see that clustering as well as edit history components are critical for performance of UNIPROMPT in all the datasets. We see a major drop of 14.8% in accuracy in the Ethos dataset when clustering is removed, and a 4.3% drop when history component is removed. In all the datasets except GSM8K, we find clustering is more important than history. This can attributed to variability of question types being less in GSM8K than in others: all questions are grade-8 arithmetic questions in GSM8K.

We also find that the greedy update rule proves to be superior or competitive compared to beam search in relatively easier datasets — where even less exploration produces good results, greedy proves to be a more effective update rule. On the other hand, in more complex datasets like MedQA, greedy appears to be a bad strategy. We also see that clustering examples based on feedback (“Fb Clustering”) is a better strategy than clustering based on topics, except for the Ethos dataset.

5 Related Work

Here, we highlight relevant work that are not addressed in the manuscript so far. Deng et al. [6] present a discrete prompt optimization method, RLPrompt, using reinforcement learning, where a policy network learns to generate effective prompts through reward-based training, with an emphasis on enhancing training efficiency through effective reward stabilization techniques. A drawback of such automatic prompt optimization approaches [22, 36, 6, 31] is that the prompts generated tend to be short, often comprising only one or two sentences, which may not fully encapsulate the complexity of the task at hand.

Prior research [28, 27] has highlighted the significance of specific sections within prompts. For instance, a well-designed prompt typically incorporates elements such as a system prompt and a chain-of-thought section. However, existing methods do not specifically target the optimization of individual sections and their respective contents within the prompts. Hsieh et al. [11] investigate the use of greedy and genetic algorithms to edit lengthy prompts. Their method focuses on paraphrasing one line at a time starting from an existing prompt, compared to our goal of learning facets of a task from scratch, thereby compromising generalization accuracy. Another orthogonal line of work that ha been explored recently for prompt optimization is the algorithmic selection of in-context examples [19, 9, 30, 25, 26].

6 Limitations and Future Work

We presented a method inspired by the human prompt engineering process to generate complex prompts from scratch that include different facets of a task. Our algorithm provides significant improvements over baseline prompt generation methods on multiple standard datasets. However, in our evaluation, we considered only classification tasks and only GPT-4 as the expert LLM. Generalizing to generative tasks, as well as investigating the impact of the expert LLM choice are some future directions.

References

- [1] A developer’s guide to prompt engineering and llms. <https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/>, .
- [2] Full sydney pre-prompt (including rules and limitations and sample chat). https://www.reddit.com/r/bing/comments/11398o3/full_sydney_preprompt_including_rules_and/?rdt=62233, .
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish,

- Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
- [4] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL <https://api.semanticscholar.org/CorpusID:3922816>.
 - [5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
 - [6] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. RL Prompt: Optimizing discrete text prompts with reinforcement learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.222. URL <https://aclanthology.org/2022.emnlp-main.222>.
 - [7] Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*, 2023.
 - [8] Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, 2023.
 - [9] Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.930. URL <https://aclanthology.org/2023.findings-emnlp.930>.
 - [10] Lawrence Han and Hao Tang. Designing of prompts for hate speech recognition with in-context learning. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–320. IEEE, 2022.
 - [11] Cho-Jui Hsieh, Si Si, Felix X. Yu, and Inderjit S. Dhillon. Automatic engineering of long prompts. *ArXiv*, abs/2311.10117, 2023. URL <https://api.semanticscholar.org/CorpusID:265281606>.
 - [12] Xinyu Hu, Pengfei Tang, Simiao Zuo, Zihan Wang, Qiang Lou, Jian Jiao, and Denis Charles. Evoke: Evoking critical thinking abilities in llms via reviewer-author prompt editing. In *ICLR 2024*, May 2024.
 - [13] Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
 - [14] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.
 - [15] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3(71-104):3, 2014.

- [16] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL <https://doi.org/10.1145/3560815>.
- [17] Cédric Malherbe and Nicolas Vayatis. Global optimization of lipschitz functions. In *International Conference on Machine Learning*, pages 2314–2323. PMLR, 2017.
- [18] Ravi Mangal, Kartik Sarangmath, Aditya V Nori, and Alessandro Orso. Probabilistic lipschitz analysis of neural networks. In *International Static Analysis Symposium*, pages 274–309. Springer, 2020.
- [19] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL <https://aclanthology.org/2022.emnlp-main.759>.
- [20] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumacas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16, 2020. URL <https://api.semanticscholar.org/CorpusID:219687112>.
- [21] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3, 2023.
- [22] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [23] Chengwei Qin, Shafiq Joty, Qian Li, and Ruochen Zhao. Learning to initialize: Can meta learning improve cross-task generalization in prompt tuning? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11802–11832, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.659. URL <https://aclanthology.org/2023.acl-long.659>.
- [24] Damien Sileo. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*, 2023. URL <https://arxiv.org/abs/2301.05948>.
- [25] Pragya Srivastava, Satvik Golechha, Amit Deshpande, and Amit Sharma. Nice: To optimize in-context examples or not?, 2024.
- [26] ZhongXiang Sun, Kepu Zhang, Haoyu Wang, Xiao Zhang, and Jun Xu. Effective in-context example selection through data compression. 2024. URL <https://api.semanticscholar.org/CorpusID:269921588>.
- [27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.
- [29] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

- [30] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.79. URL <https://aclanthology.org/2023.acl-long.79>.
- [31] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. 2023.
- [32] Chenrui Zhang, Lin Liu, Chuyuan Wang, Xiao Sun, Hongyu Wang, Jinpeng Wang, and Mingchen Cai. Prefer: Prompt ensemble learning via feedback-reflect-refine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19525–19532, 2024.
- [33] Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. Tempera: Test-time prompting via reinforcement learning. *arXiv preprint arXiv:2211.11890*, 2022.
- [34] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, 2022.
- [35] Zhihan Zhang, Shuohang Wang, Wenhao Yu, Yichong Xu, Dan Iter, Qingkai Zeng, Yang Liu, Chenguang Zhu, and Meng Jiang. Auto-instruct: Automatic instruction generation and ranking for black-box language models, 2023.
- [36] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. 2022.
- [37] Terry Yue Zhuo, Zhuang Li, Yujin Huang, Fatemeh Shiri, Weiqing Wang, Gholamreza Haffari, and Yuan-Fang Li. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex. *arXiv preprint arXiv:2301.12868*, 2023.

A Appendix

Table 2: Comparison of UNIPROMPT (“Ours”) with MedPrompt, with GPT-4 as the solver model.

	Ours	Ours + KNN	Ours + KNN + COT	Ours + KNN+COT+Ensemble	MedPrompt
MedQA	80.9	81.0	83.9	87.0	80.6
PubMedQA	70.3	72.2	74.7	75.6	71.2
MedMCQA	79.2	81.4	82.6	84.5	79.1
MMLU MG	78.0	94.0	96.0	99.0	98.0

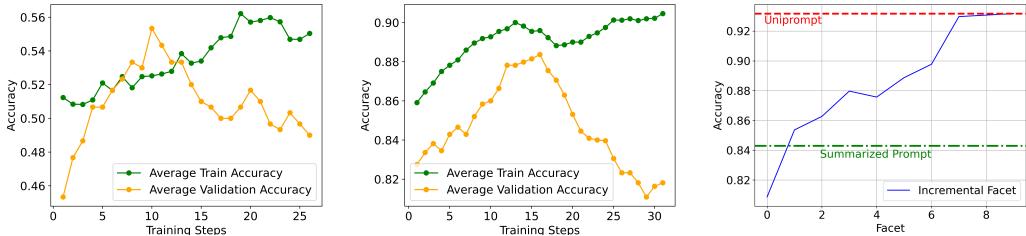


Figure 3: Training curves for MedQA (left) and ARC (middle) datasets when UNIPROMPT is initialized with (published) state-of-the-art prompts; (right) ablation of facets on Ethos (Section 4.4).

Table 3: Ablation of design choices in UNIPROMPT with GPT-3.5-Turbo as the solver model.

	Ethos	ARC	MedQA	GSM8K
UNIPROMPT – History	88.0	84.6	55.3	80.8
UNIPROMPT – Clustering	77.5	82.0	54.1	81.5
UNIPROMPT	92.3	86.0	57.1	82.4
UNIPROMPT + Greedy	93.7	90.5	55.5	82.3
UNIPROMPT + Fb Clustering	87.2	91.2	58.3	82.5

A.1 Details on estimation of Lipschitz constant L

TO calculate the Lipschitz constant for a given LLM and task, we take a human written prompt and generate its paraphrases using GPT-4. We prompt GPT-4 with the following text: “You are given a sentence, you have to generate 30 paraphrases of the sentence, make sure that the cor content of each paraphrase is same, you can use add, subtract or change words”. These paraphrases are then evaluated on the validation set D_v . For a measure of distance between two prompts, we take the cosine similarity between the embeddings of two prompts. We use text-ada-002 for generating the text embeddings for prompts.

A.2 Details on marginal utility difference calculation for prompts

For a given LLM and task, to study the marginal utility of different facets for a prompt, we first sample a few facets (generated using the fine-tuned model) that individually perform better than the chain-of-thought prompt as a baseline. Given these set of facets, we sample three facets f_1, f_2, f_3 and generate four prompts: $A : f_1, B : f_2, A + \Delta : f_1 + f_3, B + \Delta : f_2 + f_3$. We define Δ_A as the accuracy difference between prompt A and prompt $A + \Delta : f_A + f_{A+\Delta}$. Similarly, $\Delta_B = f_B + f_{B+\Delta}$

A.3 Prompt Initialization

One line task descriptions:

1. Ethos: In this task, you have to determine whether a given text is hate speech or not.
2. ARC: You have to solve the following science question.
3. GSM8K: In this task, you are given a math question. You have to solve the question.
4. MedQA: In this task, you are given a medical question. You have to solve the question.

An example of sectioned initialization prompt generated using finetuned Llama Model

Introduction:

Assume the role of a science expert and answer the given question by selecting one of the options A, B, C or D.

1. Understand and solve science questions by selecting the best answer from a given list of options.
2. Identify the logic behind the choices provided and make an informed decision.
3. Use contextual clues to choose the most accurate answer.
4. Be aware of the differences between science and everyday language.

Task Description:

Scientific inquiry: Science is the systematic study of the structure and behavior of the physical and natural world through observation and experiment. The scientific method is a process for acquiring knowledge that has been improved upon since its inception in the 17th century. It involves making observations, formulating hypotheses as to their causes, and experimenting with them to support or refute the

hypotheses.

Real-life Application:

1. Assisting Students in Science Classes:

In the context of science education, the ability to solve science questions can help students to better understand and internalize the concepts. By familiarizing themselves with the basic principles of science, students can develop a stronger foundation of knowledge.

2. Improving Scientific Literacy:

Scientific literacy is a critical skill in today's world, where scientific knowledge is increasingly important. By solving science questions, individuals can improve their understanding of scientific concepts and be more informed about scientific developments.

3. Scientific Questions:

In daily life, there are many questions that require scientific knowledge to answer. For example, understanding the science behind certain phenomena, such as why a magnet sticks to a refrigerator door, can help us in our day-to-day life.

4. Increased Awareness:

By answering scientific questions, we can develop a deeper understanding of the world around us and increase our awareness of scientific phenomena. This can help us in our daily lives and make us more knowledgeable individuals.

Background Knowledge:

1. Understanding of the basic concepts of science and physics, such as the difference between heat, temperature and friction.
2. Basic knowledge of the different types of skin surfaces, such as dry, wet, rough, smooth, etc.
3. Familiarity with the different types of magnets and their properties.
4. Understanding of the different factors that affect the adhesion of magnets to different surfaces.
5. Knowledge of the different types of sedimentary rocks and their properties.

Challenges:

1. Ambiguity in the question:

The question might be ambiguous in nature, and it can be difficult to understand the exact meaning of the question. In such cases, it is important to read the question carefully and identify the key concepts or keywords. This can help in arriving at the correct answer.

2. Scientific terms or concepts:

The question might contain scientific terms or concepts that are unfamiliar to the user. In such cases, it is important to understand the meaning of these terms or concepts and their relationship with the question.

3. Difficulty in understanding the question:

Sometimes, the question might be complex or abstract, making it difficult to understand or interpret.

4. Misleading statements or information:

The question might contain misleading or false information, making it difficult to determine the correct answer.

5. Contradiction:

The answer can be in conflict with well-known scientific facts or principles. In such cases, it is important to make a careful analysis of the evidence and choose the answer that is most consistent with the available

Simplification:

1. Identify the key elements in the question:

Ask yourself, "What is the main question in the question?" Identify the key elements and focus on them to solve the problem.

2. Understand the context:

Understand the context of the question and the background knowledge you need to answer it.

3. Identify the answer choice:

Identify the answer choice that best fits the context and background knowledge.

4. Eliminate the distractors:

Eliminate the distractors that don't fit

Tricks:

1. Read the question carefully: Understand the question and its context. This will help in understanding the information and concepts needed to solve the question.

2. Identify the key concepts: Identify the key concepts and keywords in the question. This will help in understanding the main idea and focus on the relevant information.

3. Understand the question structure: Understand the structure of the question. This will help in identifying the appropriate answer option and avoiding distractions.

4. Look for clues: Look for clues in the question and the answer options

A.4 SLM Training Details

To induce the ability of structured prompt generation in a smaller language model, we curate a section-wise dataset of around 12,000 task-prompt pairs. The tasks for training dataset creation were taken from tasksource library [24] that contains around five hundred classification tasks. We extract the task description from tasksource-instruct, which contains tasksource dataset recasted with instructions. For instance, the task description for BIG-bench Entailed Polarity task is, "*Given a fact, answer the following question with a yes or a no*". The dataset provides diverse tasks and their short description, but not the human-generated prompts for each task. To approximate human-generated prompts, we use GPT-4 as a teacher model.

By prompting GPT-4 with the task description and section description, we ask it to generate the contents of the section. To ensure that the generated section-wise prompts are concise and relevant, we prompt GPT-4 to not generate more than five lines of content for each section. We use LLAMA2-13B model, which we finetune using LoRA adapters as the auxiliary LM that generates sections.

A.5 Prompt to Llama2-13B for fine-tuning

Instruction:

You are a prompt engineer, you have to write a structured prompt. For the given task description, examples and section description, write the contents of the section that align with section description.

Task Description:

{data_point['task_description']}

Section Description:

{data_point['section']}: {section_descriptions[data_point['section']]}

```
### Response:  
{data_point['prompt']}
```

A.5.1 Examples of training dataset

Task: super glue boolq

Task Description: With no explanation, label the following with either False or True.

Section: description

Prompt: As a prompt engineer, your task is to create a prompt that accurately labels statements as either True or False without providing any explanations. The statements may involve general knowledge, trivia, or facts about events, places, people, or objects. Ensure you have a strong understanding of the subject matter and write a concise and specific task description related to the given statements.

=====

Task: glue qnli

Task Description: With no explanation, label A to B with either entailment or not entailment

Section: background

Prompt: 1. Entailment means that the information in statement B can be inferred directly from statement A.

2. Not entailment means that the information in statement B cannot be inferred directly from statement A or is unrelated.

3. Understanding the context and relationship between the two statements is crucial for accurate classification.

=====

Task: bigbench undo permutation

Task Description: In the following sentences with shuffled words, find the correct order of word swaps to unscramble the sentence.

Section: tricks

Prompt: 1. Identify the key words or phrases in the task to understand the context of the sentence. Look for nouns, verbs, and adjectives that seem related or could logically fit together.

2. Start by solving the problem step by step and focus on one swap at a time. Breaking the problem into smaller sub-problems will make it easier to manage.

3. To make the task more manageable, first focus on swapping the words that are clearly out of place, such as words that should be at the beginning or end of the sentence.

A.6 Prompt for identifying important facets

you are given a task, along with it's description, some examples of how to solve the task and section descriptions.

What do you think would be the most important sections to include for the given task.

```
## Task
{task}
## Task Description
{tas_description}
## Examples
{Examples_string}
## Section Descriptions
{sections}
```

A.7 Clustering Type 1

You are given a science question, you need to tell which broad topic is this question from.

Question: {train_questions_new[ij]}

Answer: {answer}

Give your answer as a single word, between <Answer></Answer> tags like: <Answer>Thermodynamics</Answer> or

<Answer>Botany</Answer>.

Subtopic:

A.8 Clustering Type 2

You are given a set of feedbacks, you need to cluster them into five groups based on similarity, and then provide a summary of each group. You can use the following feedbacks to cluster: \n {feedback}

provide each cluster explnation within the following tags:
<Cluster></Cluster>

You are given a feedback and a set of clusters, you need to tell which cluster this feedback belongs to.

The clusters are: \n {string_of_clusters}

The feedback is: {feedback}

give your final answer as the number of the correct cluster between <Answer></Answer> tags like: <Answer>1</Answer>.''

A.9 Feedback Prompts

Feedback over mini-batch

You are a teacher and you have to give feedback to your students on their answers.

You are teaching how to solve math problems to your students.

You are given a question, it's true answer and answer given by student.

You are also given the explanations written by your students while solving the questions.

The questions are answered wrong by the students.
You have to tell why is the solution wrong and what information
is can be added to the in the Background Knowledge part that
would have helped the student to write better explanations.

IMPORTANT: You are also given a history of changes you made
to the background knowledge part and the change in student's
accuracy after making the change. You have to use this history
to make your feedback.

Be explicit and tell the exact information that can be added
without further modification / addition.

IMPORTANT: Give feedback in form of instructions like add a
section, add a subsection, set the content of a section, set the
content of a subsection, delete a section or delete a subsection
in the background knowledge part.

Give very granular feedbacks, like if the student has made a
mistake in the calculation, then tell what is the mistake in the
calculation and how to correct it, if the student has made a
mistake in the concept, then tell what is the mistake in the
concept and how to correct it.

```
## Background Knowledge  
{current_prompt}
```

```
## History  
{history_string}
```

Now, it is your turn to give feedbacks to the students.
You can only provide a one line feedback.

Feedback over batch

You are given a set of feedbacks for some problems. The set
feedbacks for each problem separated by ===== symbol.
You have to summarize the feedbacks into a final feedback.
You are also given a set of wrong questions. You need to tell
which edit can be applied to aid the student in solving the wrong question.

To achieve your task, try to follow the following steps;
1. Identify the general problem that is being solved by all the
feedbacks.
2. Once you have identified the problem, try to make a new
feedback that covers most of the
feedbacks given.

Let's say the problem in the first feedback is the absence of
methods to solve linear equation and in the second feedback it
is the method to inverse a matrix.

You know that both of these problems can be caused by adding how
to solve convert a matrix into row reduced echelon form. So,
add that.

3. Try and validate your feedback. Once, you have a feedback try
to see if it covers every
feedback, if it does not cover any feedback, add that to your
new feedback.
4. See the wrong questions and try to identify what is the

problem in the question.

If the problem is not covered by your feedback, add that to your feedback.

5. You can add specifics like examples, definitions etc make sure that the feedback is enough to be directly added without any modification.

You may use the following function templates-

```
add_section(sectionname)
add_subsection(section_name, subsection_name)
set_section_content(section_name, new_content)
set_subsection_content(section_name, subsection_name, new_content)
delete_section(section_name)
delete_subsection(section_name, subsection_name)
```

Your summary cannot include more than four functions. Make sure that the content is useful, not just a very general statement. Something specific.

Instructions:

```
{edits}
```

Wrong Questions:

```
{wrong_examples_string}
```

Summary:

A.10 Editing Prompt

You are given an input prompt and a feedback, you have to incorporate the feedback into the input prompt and output the final prompt.

An example of the task is given below

Input Prompt

Introduction: In this task you have to answer the given question.

Feedback

The background knowledge is incomplete, it does not include what are the factors that affect the water usage and how many water sources are there.

```
\\"add_subsection("Background Knowledge")
\\"add_subsection_content(water usage depends on the population,
climate, economic development, and availability of water
sources. There are two sources of water, surface water and
groundwater.)
```

Final Prompt

Introduction: In this task you have to answer the given question.

Background Knowledge: water usage depends on the population, climate, economic development, and availability of water sources. There are two sources of water, surface water and groundwater.

Only output the final prompt nothing else.

```
### INPUT PROMPT
{current_prompt}
```

```
### FEEDBACK
{edits}
```

```
### FINAL PROMPT
```

A.11 Example of prompt evolution using our method

See example in Figure 4.

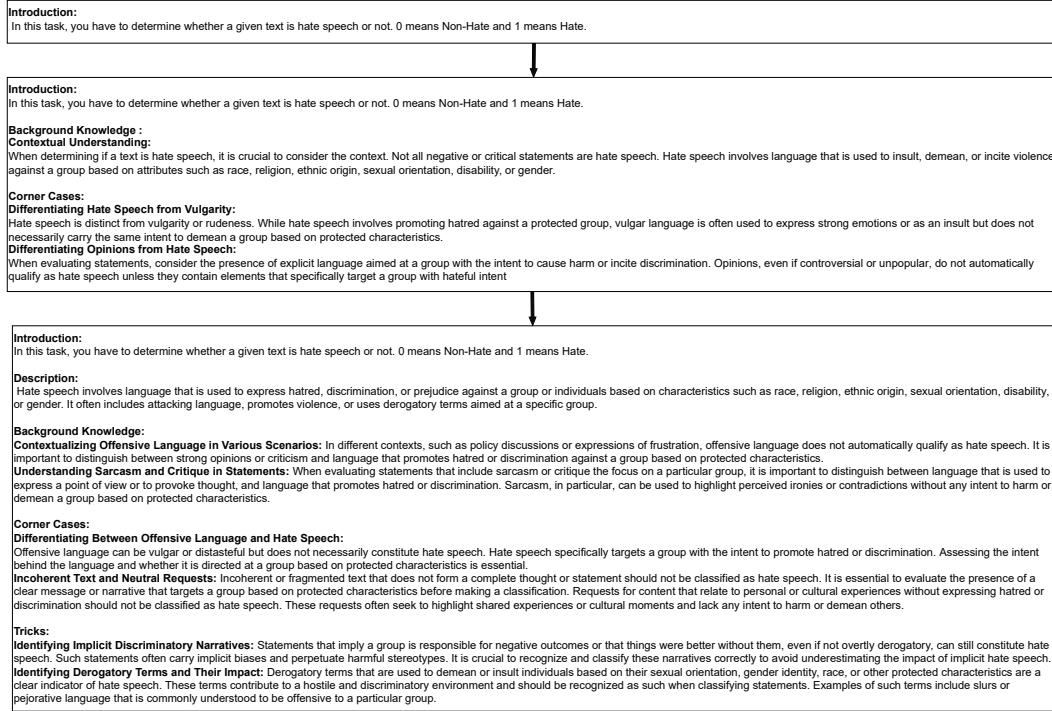


Figure 4: Evolution of prompts through iterations of UNIPROMPT on the Ethos dataset. Starting from a simple one-line prompt having an accuracy of 82%, UNIPROMPT adds background knowledge, corner cases, and additional sub-sections yielding a prompt with accuracy 88%. After further iterations, our algorithm converges to a detailed, human-like longform prompt that achieves accuracy of 92%.

A.12 Comparision of our method with existing methods

See Figure 5.

A.13 Effect of length on performance of prompt

Here we answer the question: *How much does only length contribute to UNIPROMPT's success?*. To answer this, we replace the prompt with in-context examples of the same context length and compare the accuracies in Table 4. We also compare the case where we include only the examples that the solver LLM gives incorrect prediction on, denoted as “Wrong ICL” row in the table. We see that there is a slight increase in accuracy when wrong examples are included in the prompt over randomly including examples. But, overall, UNIPROMPT performs much better than including in-context examples. This shows that length is not the only factor contributing to UNIPROMPT’s success.

Human Prompt

Let's differentiate using step by step reasoning like a medical expert.

Our Prompt

Introduction: In this task, you are given a medical question. You have to solve the question.

Description: To solve medical questions effectively, it is important to understand various medical conditions, their progression, and associated clinical features.

Background Knowledge: Differential Diagnosis of Subcutaneous Nodules:

When evaluating subcutaneous nodules, consider mobility, consistency, and skin adherence. Epidermoid cysts are firm, non-tender, and the skin cannot be pinched over them. Lipomas are soft, mobile, and have pinchable skin.

Corner Cases: Antiretroviral Therapy Complications:

Doctor should be aware of the common side effects of antiretroviral drugs, with specific attention to the association between didanosine and pancreatitis, and the recommended management strategies, such as replacing didanosine with lamivudine.

Figure 5: Comparison of human-written Prompt and prompt produced by UNIPROMPT on MedQA dataset.

OPRO optimized prompt

Start by dissecting the problem to highlight important numbers and their relations. Decide on the necessary mathematical operations like addition, subtraction, multiplication, or division, required for resolution. Implement these operations, keeping in mind any units or conditions. Round off by ensuring your solution fits the context of the problem to ensure accuracy

Our Prompt

Introduction: In this task, you are given a math question. You have to solve the question.

Strategies for Word Problems:

1. Understanding Word Problems: When solving word problems, it is crucial to read each sentence carefully and comprehend the time periods and quantities involved. Avoid incorrect multiplication or addition by paying close attention to whether a quantity remains constant over a period or changes. If a quantity is consistent, it does not need to be multiplied by the number of days or weeks unless the problem specifies otherwise.

2. Calculating Averages: To calculate the average of a set of numbers, add all the numbers together and then divide by the number of items. In word problems, ensure you have the correct total before dividing by the number of periods, such as weeks, to find the average for each period.

3. Understanding Past and Future Events in Word Problems: Distinguish between past and future events by identifying the starting and ending points. To calculate the time interval between two events, determine the direction of time from past to future and compute the interval accordingly. This understanding is essential when dealing with problems that ask for the time since a past event or until a future event.

Figure 6: Comparison of prompt produced by the state-of-the-art ORPO [31] and by UNIPROMPT on the GSM8K dataset.

Table 4: Analysis of the effect of length and contents on the performance of UNIPROMPT

	Ethos	ARC	GSM8K
UNIPROMPT	93.7	90.5	82.4
ICL Prompt	63.0	86.7	76.3
Wrong ICL	70.4	87.1	78.2
Summarized Prompt	84.3	85.5	66.0