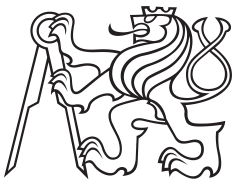
Bachelor Project



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Artificial Intelligence Center

Meta-prompts for LLM Prompt Optimization

abcd

Vojtěch Klouda

Supervisor: Ing. Jan Drchal PhD.
Field of study: Artificial Intelligence
Subfield: Natural Language Processing
May 2025

Acknowledgements

=)))

Declaration

Prohlašuji, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškerou použitou literaturu.

V Praze, 10. May 2025

Abstract

TODO

Keywords: language model, optimization

Supervisor: Ing. Jan Drchal PhD. Resslova 307/9 Praha, E-322

Abstrakt

TODO

Klíčová slova: jazykový model, optimalizace

Překlad názvu: abcd — abcd

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Background

The central focus of this work is an instance of an LLM, denoted $\mathcal{M}$. When appropriate, we can differentiate between instances with a lower index, specifying its purpose. For example, when using separate LLM instances for optimizing and task-solving, we will denote them $\mathcal{M}_{optim}$ and $\mathcal{M}_{solve}$ respectively. This way, we put emphasis on the fact we can choose a different LLM provider and hyperparameters for each instance.

In general, $\mathcal{M} : \mathbb{T} \times \mathbb{T}$ is a stochastic (for a positive sampling temperature) mapping on the space of text sequences $\mathbb{T}$. A prompt $p \in \mathbb{T}$ is a text sequence that, when inputted into an LLM, produces an output

$$y \sim \mathcal{M}(p). \tag{1.1}$$

We can use LLMs to solve a general task

$$t \in \mathcal{D} \,|\, \mathcal{D} = \{(q_1, g_1), (q_2, g_2), ..., (q_n, g_n), \}, \tag{1.2}$$

where $\mathbb{D}$ is a dataset consisting of n pairs of queries q and gold-labels g. For open-ended tasks, the gold-label does not exist.

We can further define a prompt as

$$p = \mathbf{i}(q), \tag{1.3}$$

where $\mathbf{i} \in \mathbf{T}$ is a set of text instructions into which a task query q is inserted.

Algorithm 1: General optimization loop

Input: Initialization Operator $\mathcal{O}_I$, Selection Operator $\mathcal{O}_S$, Expansion Operator $\mathcal{O}_E$, Termination Condition Φ_{stop}
Output: Optimized Population $\mathcal{P}$
Data: $\mathcal{P} \leftarrow \mathcal{O}_I$
`// Initialize the population`
1 **while** $\neg\Phi_{stop}(\mathcal{P})$ **do**
  `// Selection and Expansion Steps`
2 $\mathcal{P}_{\text{selected}} \leftarrow \mathcal{O}_S(\mathcal{P})$
3 $\mathcal{P}_{\text{expanded}} \leftarrow \mathcal{O}_E(\mathcal{P}_{\text{selected}})$
4 $\mathcal{P} \leftarrow \mathcal{P}_{\text{expanded}}$ `// Update the population`
5 **return** $\mathcal{P}$ `// Return the optimized population`

Next, we move onto the optimization notation. In Algorithm 1 we can see the general outline of a population-based optimization method. The

initialization operator $\mathcal{O}_I$ creates an initial population of individuals $\mathcal{P}$. Then, in each step, a selection operator $\mathcal{O}_S$ selects a portion of the population according to some criteria. These selected individuals are then used by the expansion operator $\mathcal{O}_E$ to create new individuals. This process continues until a termination condition Φ_{stop} is reached.

Chapter 2

Literature

2.1 Inference-time scaling

Inference-time scaling or test-time scaling is a paradigm that has gained traction in the recent years with the advent of dedicated reasoning models **cite some model cards / deepseek**. As opposed to training-time scaling, where the performance of models scales with training times, model parameter counts and dataset sizes **cite smth about training scaling**, inference-time scaling aims to improve performance by dedicating more resources to each inference call.

At their heart, LLMs are probabilistic models over sequences and to generate a sequence, they employ generation algorithms. Welleck et al.[1] provide an overview of these generation algorithms and then frame more advanced inference-time techniques as meta-generations, or strategies that employ sub-generators. Most generation algorithms attempt to find either highly probable sequences (MAP algorithms) or sample from the model's distribution. The simplest MAP algorithm is greedy decoding, which recursively finds the next token with the highest probability in the distribution.

A generalization of greedy decoding is the beam search algorithm which maintains a structure of possible prefixes and each step expands them and scores them. An example of a beam search algorithm[2] can identify decoding branches where the model employs a reasoning chain to solve a given task. Authors of this algorithm found that answer tokens found in the decoding paths with a reasoning chains have greater token probabilities, meaning the model shows greater confidence in its answer having reasoned about it beforehand. In general beam search improves on simple greedy decoding but at a high computational cost.

An interpolation between greedy decoding and uniform sampling is temperature sampling, which outperforms other adapters in input-output tasks like code generation and translation. An example of algorithms that sample from the model's distribution is the ancestral sampling algorithm. Interpolating between ancestral sampling and simple greedy sampling gave rise to decding algorithms such as nucleus, top-k and η- and ϵ-sampling. When we require a structured output, for example a JSON data structure following a JSON schema, we can utilize parser-based decoding, which enforce a structural requirement. This can however come at worsened performance when using inflexible templates.

These strategies can be divided into the categories of chained, parallel, step-level, and refinement-based meta-generators[1].

2.1.1 Chained meta-generation

Chained meta-generation is the composition of several subgenerators in sequence. These can be LLM calls or other functions that use previous inputs, such as code execution function cite program of thought. The subgenerators can be implemented as several LLM calls or with a single call given sufficient instructions in the prompt. [3] Some examples include Program-of-thought, Plan-and-Solve and Chain-of-Thought techniques.

Chain-of-thought (CoT)

Chain-of-Thought (CoT) is a LLM prompting technique that works by inducing a coherent series of intermediate reasoning steps that lead to the final answer for a problem[4]. Existing work suggest LLMs falter in a direct-QA scenarios (without inducing CoT), where the greedy decoding path mostly does not contain a reasoning chain[2]. In its essence, the model is a left-to-right text completion engine. We can make the analogy with human thinking modes, where it is said that humans have a fast automatic "System 1" mode and a slow and deliberate "System 2" mode[5]. In direct-QA mode, the LLM can underestimate the difficulty of the task[2] and stay in the "System 1" thinking mode. By crafting a good prompt that instructs the model to reason we can shift the model from "System 1" to "System 2" thinking. Furthermore, CoT allows models to allocate additional computation to problems with more reasoning steps[4] Prystawski et al.[6] also speculate that direct prediction fails for tasks where the relevant variables are rarely seen together in training, whereas CoT reasoning can incrementally chain known dependencies.

CoT can been elicited by prompting techniques - few-shot with steps demonstrations or zero-shot with specific instructions[2] First CoT methods[4] involved one/few-shot prompting, Although effective, this requires human engineering of multi-step reasoning prompts. This method is also highly sensitive to prompt design with performance deteriorating for mismatched prompt example and task question types[7]. For this method, authors found that CoT is an emergent capability of model scale and did not observe benefits for small models[4]. where the prompt included examples of CoT reasoning in the prompt in facilitate a reasoning chain response.

On the other hand, zero-shot prompting can induce a reasoning chain with a simple prompt like "Let's think step-by-step", making it versatile and task-agnostic[7]. Similar prompts also improve reasoning performance and

some research[tady OPRO? nebo kde hledali cot prefixy] has been been done on finding the optimal CoT prefix prompt.

Apart from prompting, CoT can been elicited by model training or tuning. This method, requiring a significant amount of reasoning data[2], has gained traction with the development of dedicated reasoning models like OpenAI's o1 or Deepseek-R1[cite o1 deepseek]. Using methods such as supervised fine-tuning (SFT) or reinforcement learning (RL), the model is trained to automatically produce longer reasoning chains, often bound in dedicated "thought" tags or tokens. These models have shown significant performance boosts on reasoning benchmarks [cite]. Models similar to o1 all primarily extend solution length by self-revision[8] After finishing a thought process, the model tries to self-revise, which is marked by words such as "Wait" or "Alternatively". The model then tries to spot mistakes or inconsistencies in its reasoning or propose an alternative solution. Self-revision ability is thus a key factor in the effectiveness of sequential scaling for reasoning models. [8]

Prompting techniques like chain-of-thought can increase answer quality at the cost of longer and more computationally expensive outputs. [9] Performance gains are observed mainly on arithmetic and coding tasks with more performance gains being observed for more complicated problems[4]. Further research by Liu et al.[10] suggests that for some tasks CoT can be detrimental. Their experiments proved their hypothesis that CoT hurts performance on tasks where humans do better without deliberation and where the nature of LLM, like the much greater context memory, does not provide an advantage over human thinking. This phenomenon was observed on tasks like facial recognition, implicit statistical learning or pattern recognition. Limited performance gains were noticed on commonsense reasoning tasks[7].

Longer reasoning chains mean more computing power spent at inference. How far can we take this sequential scaling? In their study, Zeng at al.[8] argue that longer CoTs do not consistently improve accuracy of reasoning models. Furthermore, they find that the average length of correct solutions is shorter than that of incorrect ones. Because self-revision accounts for most of the CoT length, the effectiveness of the method relies on the model's ability to self-revise. Authors of this paper argue that the self-revision ability of models is insufficient as they demonstrate limited capacity to correct their answers during self-revision. Some models on some tasks are even more likely to change a correct answer to an incorrect one than vice-versa.

2.1.2 Parallel meta-generation

Parallel meta-generation involves multiple generations concurrently. The final answer can then be chosen - with a reward model or with voting - or constructed from the ensemble of generations [1].

One of the simplest such techniques is self-consistency[11] (SC), a method which builds upon CoT to aggregate answers from diverse reasoning chains and selects the best one based on majority voting. It significantly improves accuracy in a range of arithmetic and commonsense reasoning tasks at the cost of increased computation expenditure[11]. The effectiveness of SC comes from the fact that, for tasks with objective answers, there are more ways to be right than to be wrong. For our next discussion of SC and related methods we will compare the terms *coverage* $\mathrm{C}_{\mathbb{D}}$ and *accuracy* $\mathrm{A}_{\mathbb{D}}$ for a dataset $\mathbb{D}$. Given a language model $\mathcal{M}$, a task query $q_k \in \mathbb{D}$ and a task instruction $\mathbf{i}$, we can define the generation collection of length n as

$$Y_k = \{y_{jk} \mid j \in 1, ..., n\}, \tag{2.1}$$

$$y_{jk} \sim \mathcal{M}(\mathbf{i}(q_k)). \tag{2.2}$$

For objective tasks we can check the correctness with a metric $\mathcal{G}$

$$\mathcal{G}_k(y_{jk}, q_k) = \begin{cases} 1.0 & y_{jk} \text{ is the correct answer for } q_k \\ 0.0 & y_{jk} \text{ is an incorrect answer for } q_k. \end{cases} \tag{2.3}$$

To choose the final answer, we will define a answer selection function $\mathcal{S}(Y)$. This can be a majority vote selection function or some reward-based method. We can now define *coverage* $\mathrm{C}_{\mathbb{D}}$ and *accuracy* $\mathrm{A}_{\mathbb{D}}$ as

$$\mathrm{C}_{\mathbb{D}} = \frac{1}{|\mathbb{D}|} \sum_{q_k \in \mathbb{D}} \max_{j=1,...,n} \mathcal{G}_k(y_{jk}, q_k) \tag{2.4}$$

$$\mathrm{A}_{\mathbb{D}} = \frac{1}{|\mathbb{D}|} \sum_{q_k \in \mathbb{D}} \mathcal{G}_k\left(\mathcal{S}(Y_k), q_k\right). \tag{2.5}$$

It is easy to see why coverage, or the fraction of tasks where at least one sample resulted in a correct answer, might rise as we increase the amount of samples in SC generation. Indeed research[9] has found that the relationship of coverage and the number of samples can be modeled by an exponentiated power law, suggesting a scaling law for inference similar to the training scaling laws[cite smth about training laws]. However coverage alone is no enough to paint to complete picture as generating large sample collections is only useful if the correct samples in a collection can be identified[9] and Practical parallel scaling methods must be able to select the best final answer from the set of candidates[8]. The accuracy gain of SC tends to saturate quickly as we increase the number of paths[11] with simple majority voting as coverage and

accuracy diverge[9], highlighting the need for automatic answer verification. Zeng et al.[8] make use of the fact that the correct solutions have shorter CoT on average and develop a length-weighted majority vote that outperformes simple majority voting on the AIME[cite] benchmark.

Parallel inference-time scaling provides ample room for further research as authors[8] that it offers a significantly higher improvement in coverage compared to sequential scaling for the same amount of processed tokens. Brown et al.[9] find that parallel scaling allows weaker models to outperform single-sampling with bigger and more expensive models, sometimes reducing cost. This is helped by the fact that parallel sampling can make use of batching and other system throughput optimization available for parallel inference[9].

2.1.3 Step-level meta-generation

Step-level meta-generation implements search algorithms on the generation state-space, which can be made up of tokens or longer sequences. Many search algorithms and state evaluation functions are possible. [1]

Multi-turn inference-time methods with a fixed width exhibit diminishing gains when computational budget is increased, failing to leverage the vast output space of LLMs.[12]

Unlike the standard tasks typically tackled by tree search algorithms where the number of possible actions at each node is finite, each call to LLM can yield a new output even for the same input, making each node's branching factor theoretically infinite.[12]

Tree-of-thought

Similar to CoT with self-consistency but the reasoning chain is split up into steps creating a tree. This reasoning tree can then be explored using a graph search algorithm such as DFS.

2.1.4 Refinement meta-generation

Refinement meta-generation generates a revised version of the output based on past versions and additional information such as intrinsic or extrinsic feedback or environment observations. [1] For extrinsic refinement, it is plausible that there are information sources which add new information, and hence lead to a potential gain with refinement but the efficacy of intrinsic refinement has been mixed. [1]

Inference-Time Scaling by training models to think before responding is insufficient because these methods include Reinforcement learning with verifiers, making them unsuitable for open-ended tasks. [13]

Authors train dedicated Feedback and Edit models that can be used at inference time to improve model responses to open-ended general domain tasks. [13]

Efficacy of LLMs in providing feedback and making edits to their own responses is unclear as using LLMs that were not specifically trained to provide feedback is ineffective compared to using high quality feedback. [13]

Reflexion

Reflexion converts binary or scalar feedback from the environment into verbal feedback in the form of a textual summary, which is then added as additional context for the LLM agent, e.g. CoT or ReAct module, in the next episode. [14]

The model can go through several steps of using the tools, which generates "observation". The model uses these observations to generate a final answer and leaves the ReAct chain when ready using a "finish" function. Reflections go into a long-term memory context limited to a sliding window with maximum capacity. [14]

Improves performance over strong baselines on sequential decision making, reasoning and programming tasks. [14]

ReAct

Multi-turn prompting technique that forms the basis of agentic LLMs. The model is given a set of tools, such as a Wikipedia search function or a math expression evaluator.

2.2 Prompting techniques

Prompting techniques encode human priors, making it difficult to assess a language model's intrinsic reasoning abilities [2]

2.2.1 Prompt engineering

In many modern LLM applications, prompts have become programs themselves. [15] Motivation for prompt engineering is to improve the model's capabilities not by changing the underlying weights with training on data but by crafting an optimal instruction string, or a prompt. This can be done by providing examples of the task as a part of the prompt or by instructing the model how to solve the task.

In-context learning

Prompts are distinguished based on the number of included examples. Research[16]

Prompting Type	Description
Zero-shot Prompting	Prompt has no examples. Model relies on its pre-trained knowledge.
One-shot Prompting	Prompt has one example to guide the model.
Few-shot Prompting	Prompt includes a few examples (typically 2 to 5).

Table 2.1: Comparison of Zero-shot, One-shot, and Few-shot Prompting

has shown that with growing model size the knowledge-generalizing ability of the model increases. Instead of expensive fine-tuning models can reuse knowledge from pre-training and solve many tasks when provided just by a few examples.

Few-shot prompting highlights that LLMs can be seen as powerful pattern-completion engines. [17]

Providing a prompt of examples from a distribution can condition the LLM to generate further high-probability examples from that distribution [17]

2.2.2 Prompting techniques

talk about how we can achieve meta-generation just by updating the prompt

2.3 Prompt optimization

Prompt engineering is a language generation task requiring complex reasoning to identify model error's and remedy them by modifying the prompt [18] Creating effective prompts often requires substantial trial-and-error experimentation and deep task-specific knowledge [19] Compile-time optimization is carried out only once before deployment thus amortizing the optimization cost over multiple uses of the prompt program. [15]

2.3.1 Soft prompt tuning

Prompts for models which allow access to gradients, which is not the case for proprietary models accessed via APIs, can be optimized in the high-dimensional embedding space.

This makes the optimization problem continous. Soft prompts however pose the problem of interpretability and are non-transferable across different LLMs [20].

Continuous prompt-optimization techniques, although effective, require parameters of LLMs inaccessible to black-box APIs and often fall short of interpretability. [21]

2.3.2 Discrete prompt tuning

The area of optimizing prompts discretely while utilizing language models as optimization operators has attracted significant research interest in recent years.

Natural language prompt engineering is particularly interesting because it is a natural interface for humans to communicate with machines, but plain language prompts do not always produce the desired result. [22]

Natural language program synthesis search space is infinitely large. [22]

Discrete tokens are not amenable to gradient-based optimization and brute-force search has an exponential complexity. [20]

Unlike perturbing e.g. network weights continuously which predictably generates small changes in functionality, perturbing code requires discrete changes which often dramatically change functionality. [23] **Reinforcement learning**

Heuristics based on “enumeration (e.g., paraphrasing)-then-selection” do not explore the prompt space systematically [20]

RLPrompt trains a policy network which is inserted as a MLP layer into a frozen compact model. [20]

Agent chooses the next token at each step using the previous tokens according to a learned policy. When the prompt is completed, the agent receives the task reward. [20]

The policy network can also take another inputs, leading to input-specific prompts. [20]

Meta-prompts are flexible but studies lack principled guidelines about their design. [24]

Reproduces key model parameter learning factors - update direction and update method - in LLMs to seek theoretical foundations. [24]

OPRO[25] and APO[26] introduced analogical "gradient" forms. [24]

Analogical momentum forms inspired by the momentum method involve including the optimization trajectory in the meta-prompt. To fit into the context limit and reduce noise, trajectory can be summarized or k most recent/relevant/important gradients can be retrieved. [24]

To mimic effects or learning rate, prompt variation can be limited by edit distance (maximum words to be changed). Warm-up and decay strategies can be applied to this constraint. [24]

New prompt can be created by editing a previous prompt or generate a

new one by following a demonstration. [24]

In an experiment on BBH, authors found that optimization without reflection performs better and the best momentum method being relevance. For prompt variation control, the best combination was cosine decay and no warm-up. [24]

Summarization-based trajectory is less helpful because it tends to only capture common elements. [24]

Task input-output examples are beneficial in the meta-prompt to provide additional context to the LLM to understand the task. [24]

GPT-4 can consistently find better task prompts than GPT-3.5-turbo, which suggests the need for a capable model as the prompt optimizer [24]

Trajectory-based methods perform very well possible because trajectory helps the prompt optimizer pay more attention to the important information instead of the noise in the current step. [24]

APE LLMs are used to construct a good set of candidate solutions by inferring the most likely instructions from input/output demonstrations. [22]

Local search around the best candidates by resampling - asking the LLM to paraphrase the candidate prompt - this however only provides marginal improvements over just choosing the best-performing prompt from instruction induction. [22]

APE was used to improve on Zero-Shot-CoT [7] universal "Let's think by step" prompt"on GSM8k.[22]

Prompt to the LLM optimizer is called the meta-prompt and includes previous prompts with their training accuracies sorted in ascending order along with the task description and training set samples. [25]

The main advantage of LLMs for optimization is their ability of understanding natural language, which allows people to describe their optimization tasks without formal specifications. [25]

Motivated by linear regression and TSP and on small-scale traveling salesman problems, OPRO performs on par with some hand-crafted heuristic algorithms. [25]

Optimization stability can be improved by generating multiple solutions when relying on random ICL samples. [25]

To balance between exploration and exploitation, LLM sampling temperature can be tuned. Lower temperature encourages exploitation in the local solution space and higher temperature allows more aggressive exploration of different solutions. [25]

Only the top instructions are kept in the meta-prompt to fit in the LLM context limit. [25]

New outstanding solution is usually found only all the prompts are of similar quality: first all the worse prompts are purged and substituted by a prompt similar to the current best. [25]

Semantically similar instructions have vastly different performance on GSM8k: "Let's think step by step." achieves accuracy 71.8, "Let's solve the problem together." has accuracy 60.5, while the accuracy of "Let's work together to solve this problem step by step." is only 49.4. [25]

Symbolic search With increasing complexity of prompt structure, many prompt optimization techniques are no longer applicable. [15] Symbolic prompt programs (SPPs) can be represented as directed acyclic graphs where nodes are functions (subprograms) and edges indicate call dependencies. [15]

Search space can be defined in two ways, as an enumerative search, where a small number of options is known beforehand, and iterative search, where a large search space is explored with iterative search strategies. [15]

Textual gradients

Naturally there are no gradients in the text space but some researchers try to emulate them using reflection-based operators.

APO mirrors the steps of gradient descent within a text-based Socratic dialogue substituting differentiation with LLM feedback and backpropagation with LLM editing [26]

Beam search is an iterative optimization process where in current prompt is expanded into many more candidates in each iteration and a selection process decides which will be used in the next iteration. [26]

Expansion first uses gradients to edit the current prompt and then explores the local monte-carlo search space by paraphrasing the editions [26]

To limit the computation used on evaluating prompts, an approach inspired by best arm identification in bandit optimization is utilized. [26]

Applying previous iterative prompt optimization methods, based on prompt+score pairs, to text generation tasks is challenging due to the lack of effective optimization signals. [27]

Critiques and suggestions, written in natural language, are more helpful for prompt improvement than a single score.[27]

CriSPO uses prompt+score+critique triples for next candidate generation. [27] Unlike APE [26] prompt generation is decoupled from suggestions and a history of critiques and suggestions as packed into the optimizer for a more stable optimization. [27]

CoT is applied in optimization by first asking to compare high-score prompts to low-score ones and draft general ideas. [27]

Critique-based optimization explores a larger space, which is indicated by lower similarity of the prompts in lexicons and semantics.[27]

CriSPO outperforms OPRO [25] both on summarization and QA tasks and metaprompt allows for creating ICL and RAG template prompts. [27]

DSPy optimizers

Most prompt optimizer approaches do not apply to multi-stage LLM programs where we lack gold labels or evaluation metrics for individual LLM

calls. [28]

Proposing a few high-quality instructions is essential due to the intractably large search space. [28]

Uses a surrogate Bayesian optimization model, which is updated periodically by evaluating the program on batches, to sample instructions and demonstrations for each stage of the LLM program [28]

Optimizing demonstrations alone usually yields better performance than just optimizing instructions, but optimizing both yield the best performance. [28]

Optimizing instructions is most valuable for tasks with subtle conditional rules not expressible by a few examples. [28]

For LLM programs, it is beneficial to alternate between optimizing weights (fine-tuning) and optimizing prompts. [29]

Evolutionary optimization

There is considerable synergy potential between the fields of evolutionary computation and deep learning. [23]

LLM-based variation operator

LLMs trained on code can suggest intelligent mutations and thus sidestepping many of the challenges in evolving programs. [23]

Genetic programming still offers an advantage when the programming task is far from the training distribution of the LLM. [23]

Genetic programming can in principle evolve in any space. [23] this can be connected with the idea

Topic of mutation is guided by previously chosen "commit message" this is like a pseudogradient. [23]

Evolution with language models can be used as a way of generating domain data for downstream deep learning where it did not previously exist. [23]

The pattern-completion ability of few-shot prompting can be leveraged to create a form of intelligent evolutionary crossover. [17]

Performance at in-context learning improves with model scale, implying that methods relying upon this capability will benefit from continuing progress in LLM training. [17]

Advent of large, pretrained foundation models marked a significant step in the paradigm of evolutionary recombination with deep generative models, where these new models can be directly leveraged without additional training and allow for searching more abstract spaces. [17]

Next-token prediction naturally lends itself to creating an evolutionary variation operator. [17]

LLM variation operator should be capable of generating meaningful variation for any text representation that has moderate support in the training set, meaning it is basically domain-independent. [17]

LLM crossover acts like an EDA in that it builds a probabilistic model of parents from which the children are sampled. [17]

LLM crossover becomes more similar to an EDA as the number of parents increases. [17]

LLM crossover can express any genetic operator even with small parent sets by fine-tuning or through effective prompting schemes. [17]

Simultaneously evolving the solution x along with the LLM and the prompting mechanism could be a powerful paradigm for more open-ended systems. [17]

Input arranged in ascending fitness order prompts the model to generate output that follows an ascending fitness trend. [17]

Evolutionary algorithms have been shown to be effective in search spaces

with millions and billions of variables, which are inaccessible to LLM crossover due to the size of the LLM context window. [17]

The level of stochasticity in LMX can be controlled by the softmax temperature parameter, which can be seen as analogous to a mutation rate parameter in a traditional EA [17]

Frameworks Building upon the inherent ability of LLMs to paraphrase (mutation) and combine (crossover) text, an interesting intersection of traditional evolutionary algorithms and modern LLMs has formed.

Sequences of phrases can be regarded as gene sequences in typical Evolutionary algorithms. [21]

Considers two widely used EAs: Genetic Algorithm and Differential Evolution with DE outperforming GA on most tasks [21]

ELM uses a MAP-elites Quality Diversity algorithm. [23]

Initial population consists of manually-written prompts to leverage human knowledge as well as some prompts generated by LLMs to reflect the fact that EAs start from random solutions to avoid local optima. [21]

DE-inspired approached builds on the idea that the common elements of the current best prompts need to be preserved [21]

Evoprompt performs best with roulette selection when compared with tournament and random selection. [21]

Similar results are achieved when population is initialized with the best and with random prompts, hinting that the crafted design of initial prompts is not essential. [21]

Previous research optimized zero-shot instructions and examples separately, overlooking their interplay and resulting in sub-optimal performance. [30]

There is a prevailing notion that prompt engineering sacrifices efficiency for performance due to the lengthening of prompts, but PhaseEvo actively shortens the prompts [30]

Current EA applications to prompt optimization suffer from extremely high computational cost and slow convergence speed due to the complexity of the high-dimensional search space. [30]

PhaseEvo alternates between two phases: exploration with evolution operators and exploitation using a feedback "gradient". [30]

TABLE 1 **recreate** compares all 5 operators. [30]

4 phases: initialization - lamarck or manual, local feedback mutation, global evolution with EDA and CR operators, local semantic mutation (paraphrasing) [30]

Candidates for evolution operators are selected based on a "performance vector", combining prompts that do not make the same mistakes. [30]

When the performance improvement with an operator stagnates up to some operator-specific tolerance, the current phase is terminated. [30]

Evolution in phases outperforms random operator selection. [30]

PhaseEvo is the most cost-effective but still needs around 12 iterations and 4000 API calls. [30]

APE [22] ran into problems with diminishing returns and abandoning the iterative approach entirely, Promptbreeder aims to solve this with a diversity-maintaining evolutionary algorithm for self-referential self-improvement of prompts [31]

Prompt optimization techniques utilize the fact that LLMs are effective at generating mutations from examples and can encode human notions of interestingness and can be used to quantify novelty. [31]

Self-referential system should improve the way it is improving, thus Promptbreeder used a "hyper-prompt" to optimize its meta-prompt [31]

Uses a binary tournament genetic algorithm. [31]

Uses a random uniformly sampled mutation operators out of 9 total from 5 broad categories for each replication event. [31]

Zero-order mutation (creating a prompt from task description) generates new task prompts more aligned with the task description in the event the evolution diverges. [31]

LLMs tend to be biased to examples found later in EDA mutation lists. Lying to the LLM and telling it that the prompts are sorted by performance in a descending order improves diversity. [31]

Removing any self-referential operator in ablation is harmful under nearly all circumstances [31]

Prior works do not provide sufficient guidance in the meta-prompt. [18]

PE2 improves on APE and APO with a back-tracking search procedure and a more complete metaprompt with a two-step task description, prompt layout specification and a step-by-step reflection template. [18]

Experiments with step size and momentum and also a prompt engineering tutorial in the metaprompt to mixed results. [18]

Optimized prompts do not seem to be generalizable across different models. [18]

Gold-agnostic methods

Current prompt optimization methods often depend heavily on external references for evaluation which are often unavailable or unpractical to define in many applications, especially for open-ended tasks [19] Gold-agnostic evaluation is important because we ultimately expect LLMs to solve problems for which answers are not already known [32]

Two primary sources can be used for evaluation: LLM-generated outputs and task-specific truths. These can then be evaluated using either a predefined metric, LLM-as-a-judge or by a human judge to produce an optimization signal based on a numeric score or a textual feedback. [19]

In each iteration, SPO generates new prompts, executes them, and performs pair-wise evaluations of outputs to assess their adherence [19]

SPO achieves high efficiency, requiring only 8 LLM calls per iteration with three samples, significantly lower than existing methods [19]

Outputs of LLMs inherently contain rich quality information that directly reflects prompt effectiveness [19]

LLMs exhibit human-like task comprehension [19]

With score-based feedback a large sample size is needed to ensure scoring stability, which can be avoided by pairwise comparison of outputs [19]

LLM-as-a-judge biases do not affect the overall optimization trend because eval's feedback merely serves as a reference for the next round of optimization [19]

While maintaining comparable performance with other ground truth-dependent prompt optimization methods, SPO requires only 1.1% to 5.6% of their optimization costs [19]

Self-consistency can by used as a metric instead of accuracy as correct answers generally exhibit greater self-consistency than incorrect ones. However it can overestimate prompts that produce consistent incorrect answers. [32]

Mutual-consistency refinement refines self-consistency scores based on the self-consistency scores of other prompts [32]

Gold-agnostic evaluation methods like GLaPE are robust metrics akin to accuracy and are able to produce effective prompts similarly to gold-label-based methods like OPRO[25]. [32]

If all prompts produce consistent but incorrect answers it is challenging to discern the error without external resources. This happens in some datasets, leading to diminished correlation between GLaPE and accuracy. [32]

Metaprompting

Metaprompting or "prompting to create prompts"

Meta-prompts are task-agnostic, meaning they will return the relevant outputs for an arbitrary task, provided a task description is provided as an input. [33].

It is possible to construct a general-purpose meta-prompt. [33].

Meta-prompts will perform better than standard prompts at executing a wide range of tasks. [33].

In tests meta-generated prompts were ranked as more useful than the baselines as well as producing content that was rated more suitable. [33].

Chapter 3

Methodology

3.1 Datasets

Datasets were chosen according to the following requirements:

1. The task is challenging for modern LLMs using a standard CoT prompt but has non-zero accuracy
2. Complex output (no multiple-choice answers)
3. Easy to check programmatically to avoid human/AI judges

3.1.1 External dataset

We found the Livebench datasets to meet our requirements.

3.1.2 Custom dataset design

Sequences dataset challenges the pattern recognition and algebraic capabilities of the model

3.2 Optimization methods

3.2.1 Optimization operators

Metaprompts that define the transition between optimizer generations.

3.3 Experimental setup

Language model used for solving is gpt-4o-mini. Prompts for the solver model are optimized by the optimized model, for which we use the gpt-4o. To encourage diversity and exploration in the optimization process, a temperature of 0.7 is used for the optimizer model. The solver model uses temperature 0.0 to keep the outputs deterministic.

Chapter 4

Experiments

4.1 Comparative analysis of optimization operators

Appendix A

Bibliography

[1] S. Welleck, A. Bertsch, M. Finlayson, H. Schoelkopf, A. Xie, G. Neubig, I. Kulikov, and Z. Harchaoui, "From decoding to meta-generation: Inference-time algorithms for large language models," 2024.

[2] X. Wang and D. Zhou, "Chain-of-thought reasoning without prompting," 2024.

[3] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, and C. Potts, "Dspy: Compiling declarative language model calls into self-improving pipelines," 2023.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[5] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023.

[6] B. Prystawski, M. Y. Li, and N. D. Goodman, "Why think step by step? reasoning emerges from the locality of experience," 2023.

[7] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 22199–22213, Curran Associates, Inc., 2022.

[8] Z. Zeng, Q. Cheng, Z. Yin, Y. Zhou, and X. Qiu, "Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities?," 2025.

[9] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini, "Large language monkeys: Scaling inference compute with repeated sampling," 2024.

[10] R. Liu, J. Geng, A. J. Wu, I. Sucholutsky, T. Lombrozo, and T. L. Griffiths, "Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse," 2024.

[11] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," 2023.

[12] K. Misaki, Y. Inoue, Y. Imajuku, S. Kuroki, T. Nakamura, and T. Akiba, "Wider or deeper? scaling llm inference-time compute with adaptive branching tree search," 2025.

[13] Z. Wang, J. Zeng, O. Delalleau, D. Egert, E. Evans, H.-C. Shin, F. Soares, Y. Dong, and O. Kuchaiev, "Dedicated feedback and edit models empower inference-time scaling for open-ended general-domain tasks," 2025.

[14] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," 2023.

[15] T. Schnabel and J. Neville, "Symbolic prompt program search: A structure-aware approach to efficient compile-time prompt optimization," 2024.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[17] E. Meyerson, M. J. Nelson, H. Bradley, A. Gaier, A. Moradi, A. K. Hoover, and J. Lehman, "Language model crossover: Variation through few-shot prompting," 2024.

[18] Q. Ye, M. Axmed, R. Pryzant, and F. Khani, "Prompt engineering a prompt engineer," 2024.

[19] J. Xiang, J. Zhang, Z. Yu, F. Teng, J. Tu, X. Liang, S. Hong, C. Wu, and Y. Luo, "Self-supervised prompt optimization," 2025.

[20] M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. P. Xing, and Z. Hu, "Rlprompt: Optimizing discrete text prompts with reinforcement learning," 2022.

[21] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang, "Connecting large language models with evolutionary algorithms yields powerful prompt optimizers," 2024.

[22] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," 2023.

[23] J. Lehman, J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley, "Evolution through large models," 2022.

[24] X. Tang, X. Wang, W. X. Zhao, S. Lu, Y. Li, and J.-R. Wen, "Unleashing the potential of large language models as prompt optimizers: An analogical analysis with gradient-based model optimizers," 2024.

[25] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen, "Large language models as optimizers," 2024.

[26] R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng, "Automatic prompt optimization with "gradient descent" and beam search," 2023.

[27] H. He, Q. Liu, L. Xu, C. Shivade, Y. Zhang, S. Srinivasan, and K. Kirchhoff, "Crispo: Multi-aspect critique-suggestion-guided automatic prompt optimization for text generation," 2024.

[28] K. Opsahl-Ong, M. J. Ryan, J. Purtell, D. Broman, C. Potts, M. Zaharia, and O. Khattab, "Optimizing instructions and demonstrations for multi-stage language model programs," 2024.

[29] D. Soylu, C. Potts, and O. Khattab, "Fine-tuning and prompt optimization: Two great steps that work better together," 2024.

[30] W. Cui, J. Zhang, Z. Li, H. Sun, D. Lopez, K. Das, B. Malin, and S. Kumar, "Phaseevo: Towards unified in-context prompt optimization for large language models," 2024.

[31] C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel, "Promptbreeder: Self-referential self-improvement via prompt evolution," 2023.

[32] X. Zhang, Z. Zhang, and H. Zhao, "Glape: Gold label-agnostic prompt evaluation and optimization for large language model," 2024.

[33] A. de Wynter, X. Wang, Q. Gu, and S.-Q. Chen, "On meta-prompting," 2024.