

A Practical Survey on Zero-shot Prompt Design for In-context Learning

Yinheng Li

Columbia University / New York City

li.yinheng@columbia.edu

Abstract

The remarkable advancements in large language models (LLMs) have brought about significant improvements in Natural Language Processing (NLP) tasks. This paper presents a comprehensive review of in-context learning techniques, focusing on different types of prompts, including discrete, continuous, few-shot, and zero-shot, and their impact on LLM performance. We explore various approaches to prompt design, such as manual design, optimization algorithms, and evaluation methods, to optimize LLM performance across diverse tasks. Our review covers key research studies in prompt engineering, discussing their methodologies and contributions to the field. We also delve into the challenges faced in evaluating prompt performance, given the absence of a single "best" prompt and the importance of considering multiple metrics. In conclusion, the paper highlights the critical role of prompt design in harnessing the full potential of LLMs and provides insights into the combination of manual design, optimization techniques, and rigorous evaluation for more effective and efficient use of LLMs in various NLP tasks.

1 Introduction

In recent years, transformer-based language models (such as (Raffel et al., 2019), (Lewis et al., 2019), (Brown et al., 2020), (Devlin et al., 2018)) have emerged as a transformative force in the field of artificial intelligence, revolutionizing Natural Language Understanding (NLU) and Generation (NLG). As model size and training data have evolved, the GPT series has exhibited extraordinary capabilities in a wide range of natural language tasks by relying on a paradigm known as in-context learning. According to (Brown et al., 2020), in-context learning harnesses the context provided by input data to generate appropriate responses or predictions, contrasting with traditional methods that necessitate

explicit task-specific training and fine-tuning on labeled datasets. In-context learning enables large language models to capitalize on vast amounts of data and adapt to various tasks in a flexible and dynamic manner. There are several categories of in-context learning, including zero-shot, one-shot, and few-shot learning. In all types of in-context learning, the key to success lies in effective prompt design, which is occasionally referred to as an "art." This survey paper aims to categorize each type of in-context learning, discuss the core principles, examine state-of-the-art design techniques, and explore recent advancements in in-context learning, with a particular focus on zero-shot discrete in-context learning.

2 Definition

Although there is no formal definition for prompt design optimization, we follow the principle from (Brown et al., 2020) and provide the definition in (1) for prompt design in in-context learning:

$$P^* = \arg \max_P \mathbb{E}_{x_i, y_i \in \mathcal{D}} [S(f_\theta(P, x_i), y_i)] \quad (1)$$

Here, x_i represents input sentences and features, while y_i denotes the target labels. θ signifies the parameters for any Large Language Models (LLMs) or Pretrained Language Models (PLMs), which remain frozen in the case of in-context learning. f_θ represents the output from LLMs given input x_i and prompt P . S is a scoring function that measures the performance of the model output in relation to the ground truth label y_i . The objective of in-context learning (or prompt engineering) is to identify the optimal prompt P^* that maximizes the score S in the test distribution.

Based on the structure of P , in-context learning can be further classified into discrete (hard) prompt when P consists of a list of tokens or continuous

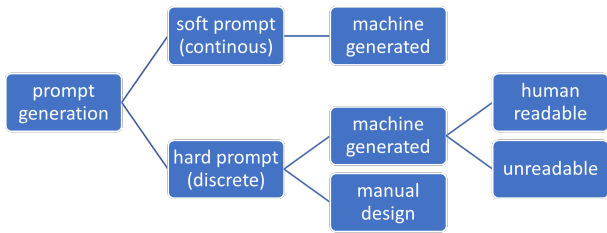


Figure 1: Prompt categorization by prompt form

prompt (soft) where P represents an embedding vector (see Figure 1). Additionally, for zero-shot in-context learning, P is independent of x_i , whereas for one-shot or few-shot in-context learning, P can be a function of x_i (from training data). This survey focuses on zero-shot in-context learning with discrete prompts and examines its application exclusively in decoder-only LLMs, such as the GPTx series.

3 Relevant Work

3.1 Prompts for Encoder-only Transformer Models (BERT)

Before the advent of in-context learning, some research efforts have been devoted to studying how to design effective prompts to enhance the performance of BERT models. As depicted in Figure 2, prompts in BERT are usually combined with input to form a cloze-style structure, while for transformer decoder-based models, prompts are more flexible.

Numerous studies have investigated prompt design in BERT. In the work by (Jiang et al., 2020), the authors proposed heuristic-based approaches for designing discrete prompts. Dependency parsing is employed to identify useful prompts from Wikipedia. In (Gao et al., 2021), the authors utilized T5 as a prompt generator with a beam search to create a set of diversified prompts. They then used D_{dev} to select a single prompt with the best performance. In (Shin et al., 2020), a gradient-based prompt search approach was proposed, wherein each prompt token is learned by directly optimizing LMs on the downstream task.

In addition to prompt designing strategies, other research work focuses on enriching the prompt can-

didates and ensembling the output from multiple prompts for the same input. To enrich prompts, (Jiang et al., 2020) employed back-translation to paraphrase prompts. Building on this work, (Haviv et al., 2021) trained a separate BERT model to rewrite prompts using the nearest BERT vector embedding.

The concept of in-context learning originates from the work by (Brown et al., 2020). However, BERT models can also perform similar tasks by using a single token as output. For example,

France’s capital is [MASK].

Only the output for the [MASK] position is used for inference. This characteristic enables the ensemble of answers from different prompts, although it is not apparent for similar practices in GPT-style models. In (Jiang et al., 2020), the authors proposed rank-based ensemble and optimized ensemble methods to aggregate answers generated from different prompts.

Among the studies designing prompts for BERT models, the majority focus on discrete prompts (i.e., hard prompts). To the best of our knowledge, we did not find any work attempting to generate continuous prompts. In general, optimizing prompts in BERT brings only marginal improvements to the original model. Given the size and structure of BERT, it is more favorable to fine-tune on downstream tasks.

3.2 Prompts for Decoder-only Transformer (GPT)

3.2.1 Continuous Prompt

Another line of research has focused on optimizing soft prompts, which eliminate the constraint that prompts have to be natural language. Soft prompts can be learned and optimized directly within the same language model. The key difference between soft prompt tuning and fine-tuning is that prompt tuning typically fixes the weights of the language model and only performs gradient updates on the network that generates the prompt. Prefix-Tuning (Li and Liang, 2021) is one of the early works that tunes prompts on GPT-2 with a small amount of data per task, achieving comparable performance to the full data fine-tuning setting. Prefix-Tuning does not use a separate network; instead, it utilizes the same transformer network but only optimizes the input embedding of the prompt. In P-Tuning V1 (Liu et al., 2021b) and V2 (Liu et al., 2022),

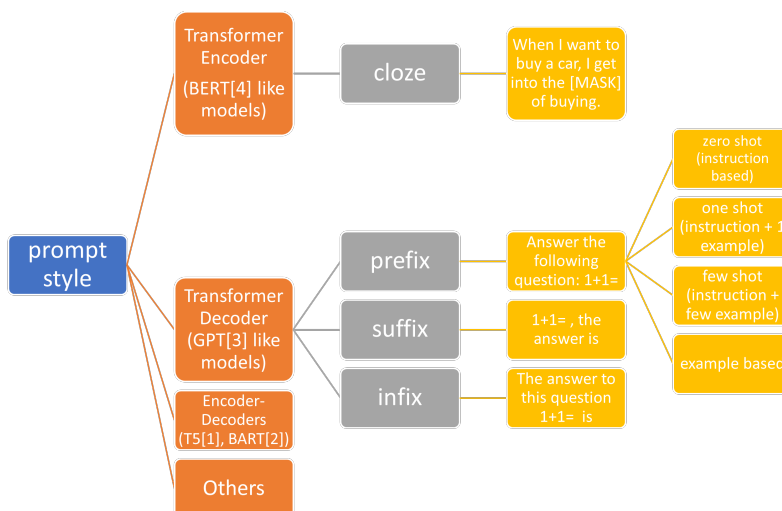


Figure 2: Prompt categorization by model types

the authors employ a separate LSTM network to generate the input prompt for the language model. While using soft prompts provides more flexibility in prompt design, it requires access to either the weights of language models or the ability to input vectors into language models. As recent language models are hosted as cloud services and large language models are difficult to access via vector inputs, this practice becomes less feasible when using GPT-3 or PaLM (Chowdhery et al., 2022).

3.2.2 Few-Shot Learning

In the GPT paper (Brown et al., 2020), few-shot learning demonstrates strong NLP capabilities across various benchmarks. As the title suggests, Language Models are Few-Shot Learners. In the few-shot setting, a task description along with a few examples are presented to the model, which is then asked to complete the task for an unseen example. Numerous studies have been conducted to optimize few-shot examples and prompts to enhance performance. In (Liu et al., 2021a), the authors discovered that GPT-3 generally performs better when in-context examples are similar to the test examples. As a result, they proposed an in-context example algorithm based on example similarities. Similarity is measured using RoBERTa embedding distance in Euclidean space or cosine distance. Other works, such as (Rubin et al., 2021) and (Gutierrez et al., 2022), have adopted similar example selection logic and demonstrated better performance over randomly selected examples. In addition to example selection methods, research efforts like (Wu et al., 2022) and (Kumar and Talukdar, 2021) have been made to optimize the rank

and order of retrieved examples.

While few-shot learning exhibits remarkable performance, according to the no free lunch(NFL) theorem (Wolpert and Macready, 1995, 1997), providing examples inevitably introduces bias to the prediction algorithm. In cases where out-of-distribution samples occur, applying few-shot learning can hinder the inference process.

4 Zero-Shot Discrete Prompts

With the recent success of Large Language Models such as GPTs, designing zero-shot discrete prompts has become increasingly popular in practice. In the experiments conducted by (Reynolds and McDonnell, 2021), the authors demonstrate that carefully engineered zero-shot prompts can actually outperform few-shot prompts. They argue that providing examples does not always help because examples tend to be interpreted as part of a narrative rather than serving as categorical guidance.

On the other hand, the advantages of using zero-shot discrete prompts can be listed as follows: (1) zero-shot prompts are highly interpretable, (2) few training data or examples are required, (3) the designing process is more straightforward as we only need to deal with task instructions, and (4) the prompt structure is flexible, allowing us to insert our input wherever needed. Zero-shot discrete prompts are also known as task instructions. There are two primary approaches to obtaining a good discrete prompt. The first is heuristic-based manual design, while the second relies on an optimization algorithm to find the optimal prompt. In this section, we focus on reviewing research on prompt

design for transformer decoder style models (e.g., GPT), which has been the focus of a majority of research efforts.

4.1 Manual Design

In their work (Reynolds and McDonell, 2021), the authors argue that GPT (or other LLMs) resemble a superposition of human authors. Therefore, it can be helpful to ask GPT to pretend to be a character in the prompt or use the prompt to signify a dialogue between people (i.e., task specification by memetic proxy). The authors also discuss the idea of MetaPrompts, which encapsulate a general intention that will develop towards specific meanings when additional information, such as a task question, is provided. The example prompts they provide, such as "Let's solve this problem by splitting it into steps," have been proven to be significantly helpful by subsequent works.

In the work (Mishra et al., 2021), the authors propose five principles for designing prompts for GPT-3 based on their observations of GPT-3's failures. These principles include: (1) using simple patterns to specify expected output, (2) using bulleted lists and assertions, (3) breaking down complex tasks into multiple simpler ones, (4) adding explicit textual statements of output constraints, and (5) customizing the instructions so that the model can directly output the results. These principles can be a good starting point for manual design.

Another line of work focuses on improving the reasoning capabilities of large language models via prompt design. The work Chain-of-Thought (CoT) (Wei et al., 2022) was initially proposed in few-shot learning, where the reasoning steps were presented as part of the solution for several few-shot examples. The zero-shot version of CoT was later proposed in (Kojima et al., 2022), which demonstrates that inserting the single prompt "let's think step by step" into the task instruction significantly improves performance on mathematical reasoning. The authors also experimented with different templates for prompts and found that instructive prompts help improve the model's performance in mathematical reasoning, while misleading or irrelevant prompts do not contribute to performance enhancement.

4.2 Prompt Optimization

Finding the optimal prompt can also be treated as an optimization process, where the goal is to optimize the performance of the target task. Similar

to finding the best soft prompt or finding the optimal examples for few-shot learning, algorithms can be implemented to find the best zero-shot prompt. However, such work typically requires a small set of evaluation data to assess the prompt performance. In the work by (Zhou et al., 2022), the authors proposed Automatic Prompt Engineer (APE) for zero-shot prompt design. A LLM is used to generate a group of prompts given the task example or human description, and an iterative Monte Carlo search method is used to search for the optimal prompt given the objective function. In addition to using Monte Carlo search for prompt optimization, a gradient-free, edit-based search approach called Gradientfree Instructional Prompt Search (GRIPS) is introduced in (Prasad et al., 2022). GRIPS starts from a manually designed instruction and iteratively searches among generated prompts from four operations (delete, add, swap, paraphrase) to find the optimal prompt for a target task.

Another line of research uses gradient-based methods but to generate discrete zero-shot prompts. The work FluentPrompt (Shi et al., 2022) follows the idea from AutoPrompt (Shin et al., 2020), using a gradient-based method to generate discrete prompts. They also use a fluency constraint to encourage human-readable prompt outcomes, which helps improve performance. Another gradient-based prompt generation method RLPROMPT is introduced in (Deng et al., 2022). This work uses a reinforcement learning structure to generate prompts that optimize the task-based reward function. The prompts generated from this framework are often incoherent gibberish but are claimed to achieve significant performance improvement.

4.3 Evaluation

Evaluating prompt design is very challenging. As there is no ground truth dataset for prompt generation, there is no "best" prompt but only better prompts. Therefore, the evaluation of the prompt performance for in-context learning usually falls into the following categories.

Conditional Probability (Likelihood): To evaluate the performance of a text generation model, we can measure the probability of the generated text. In our case, we can calculate the conditional probability of ground truth(y) given prompt (p), input(x) or calculate the joint probability of x, y, p averaging over the training data, as shown in (2)

$$Prob(y|x, p)_{x, y \in X, Y} \quad (2)$$

This is a simple strategy because the models for in-context learning are generative language models which will generate the joint probability (likelihood) automatically. However, this metric sometimes fails to represent the actual performance of the downstream task.

Execution Accuracy: A more direct method to measure the performance of a prompt is to use metrics from the target task (Zhou et al., 2022), as ultimately the performance on the task is what we care about. In addition to measuring the execution accuracy directly on the entire training set, there are ways to efficiently estimate the performance on a subset of training data to save computational cost (Zhou et al., 2022), (Li et al., 2022).

Prompt Transferability is another evaluation metric reported in (Zhou et al., 2022), (Deng et al., 2022) which is used to prove the quality of the prompt generation methods. However, this metric is more useful in selecting the prompt designing method than evaluating the performance of a single prompt.

General Metrics for Language Models should be used when using large language models via zero-shot in-context learning. It is also important to measure the performance from additional aspects. For example, if we are to build a Question-Answering system, we need to measure the risk of hallucination (Ji et al., 2022). If we are to build an email generation system, we may need to measure the toxicity and prevent generating any aggressive content. The work of Holistic Evaluation of Language Models (HELM) (Liang et al., 2022) provides a great example in evaluating the performance for language models via in-context learning. Although various metrics have been reported in HELM for existing models, it is worth noting that the design of our prompt will directly impact the models' performance.

5 Conclusion

The rapid development of large language models (LLMs) has significantly influenced various NLP tasks. Among the techniques to harness their capabilities, in-context learning with different types of prompts—discrete, continuous, few-shot, and zero-shot—has shown remarkable promise.

Discrete prompt engineering emphasizes human-readable prompts that can enhance model performance, while continuous prompt optimization involves soft prompts that can be learned and opti-

mized directly in the same language model. Few-shot learning leverages a small number of examples to guide the model in the right direction, whereas zero-shot discrete prompts only require task instructions, offering a more straightforward design process.

Manual design of prompts can be guided by principles based on model behavior, and optimization algorithms can be used to find optimal prompts. Evaluating the performance of prompts is challenging, as there is no single "best" prompt, and various metrics need to be considered.

In conclusion, as LLMs continue to evolve, prompt design remains a crucial factor in harnessing their full potential across a wide range of applications. A combination of manual design, optimization techniques, and rigorous evaluation can lead to more effective and efficient use of LLMs in diverse NLP tasks.

References

- 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. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov,

- and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723.
- Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Conference on Empirical Methods in Natural Language Processing*.
- Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. Bertese: Learning to speak to bert. *ArXiv*, abs/2103.05327.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. [How can we know what language models know?](#) *Transactions of the Association for Computational Linguistics*, 8:423–438.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.
- Sawan Kumar and Partha P. Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. *ArXiv*, abs/2106.01751.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. [BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *CoRR*, abs/1910.13461.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. [Competition-level code generation with alpha-code](#). *Science*, 378(6624):1092–1097.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekogul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. [Holistic evaluation of language models](#).
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. [P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. [GPT understands, too](#). *CoRR*, abs/2103.10385.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021. Reframing instructional prompts to gptk’s language. In *Findings*.
- Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *ArXiv*, abs/2203.07281.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *CoRR*, abs/1910.10683.
- Laria Reynolds and Kyle McDonell. 2021. [Prompt programming for large language models: Beyond the few-shot paradigm](#).

- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *ArXiv*, abs/2112.08633.
- Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. 2022. Toward human readable prompt tuning: Kubrick’s the shining is a good movie, and a good prompt too? *ArXiv*, abs/2212.10539.
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. *ArXiv*, abs/2010.15980.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#).
- David H. Wolpert and William G. Macready. 1995. No free lunch theorems for search.
- D.H. Wolpert and W.G. Macready. 1997. [No free lunch theorems for optimization](#). *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.
- Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2022. Self-adaptive in-context learning. *ArXiv*, abs/2212.10375.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *ArXiv*, abs/2211.01910.