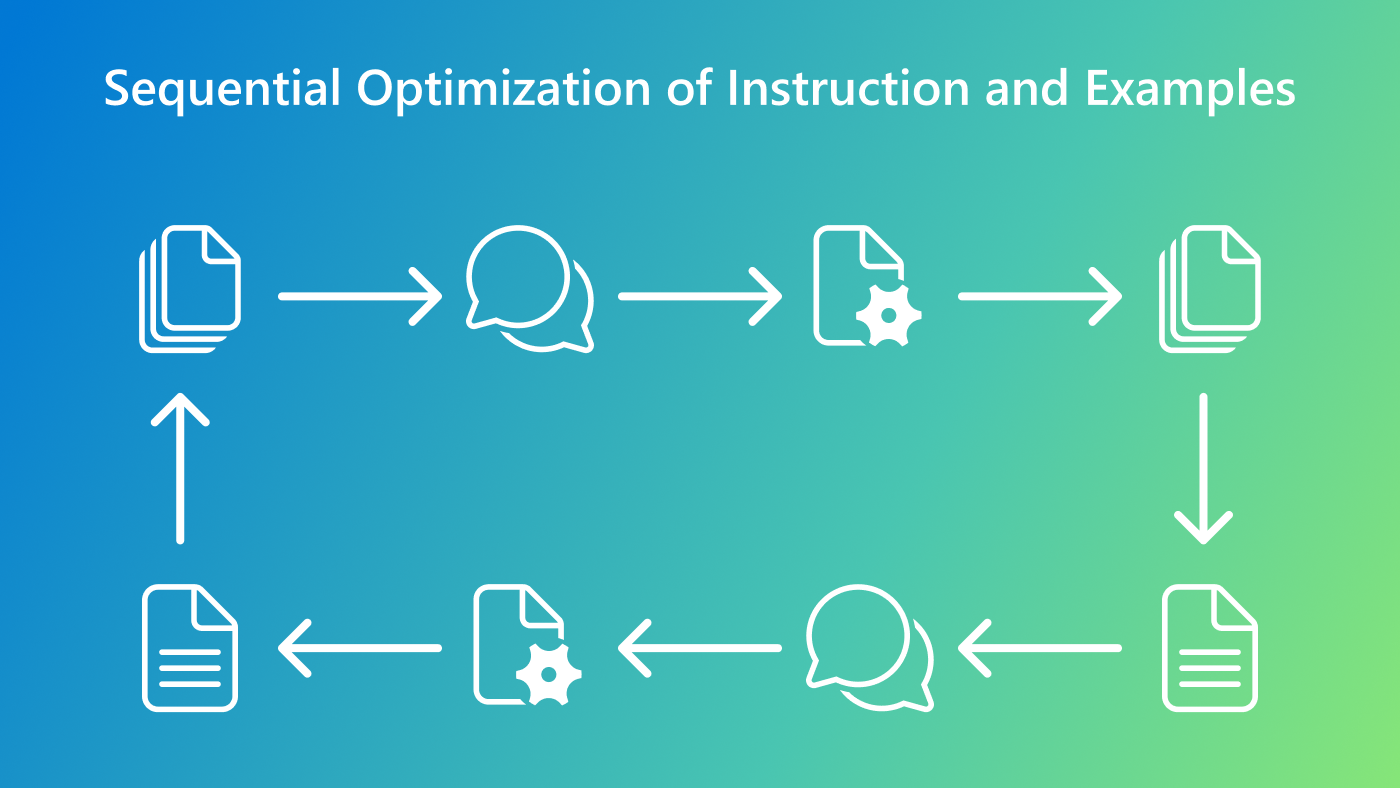
PromptWizard: The future of prompt optimization through feedback-driven self-evolving prompts



# The challenge of effective prompting

AI is reshaping industries—from education to healthcare—thanks to advancements in large language models (LLMs). These models rely on prompts, carefully crafted inputs that guide them to produce relevant and meaningful outputs. While the impact of prompts is profound, creating prompts that can help with complex tasks is a time-intensive and expertise-heavy process, often involving months of trial and error.

This challenge grows as new tasks arise and models evolve rapidly, making manual methods for prompt engineering increasingly unsustainable. The question then becomes: **How can we make prompt optimization faster, more accessible, and more adaptable across diverse tasks?**

To address this challenge, we developed PromptWizard (PW), a research framework that automates and streamlines the process of prompt optimization. We are open sourcing the [PromptWizard codebase(opens in new tab)](https://github.com/microsoft/PromptWizard) to foster collaboration and innovation within the research and development community.

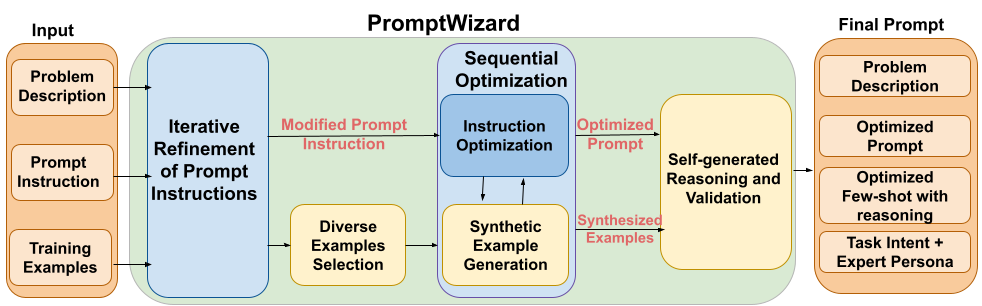
# Introducing PromptWizard

PromptWizard (PW) is designed to **automate and simplify prompt optimization**. It combines iterative feedback from LLMs with efficient exploration and refinement techniques to **create highly effective prompts within minutes**.

**PromptWizard optimizes both the instruction and the in-context learning examples.** Central to PW is its self-evolving and self-adaptive mechanism, where the LLM iteratively generates, critiques, and refines prompts and examples in tandem. This process ensures continuous improvement through feedback and synthesis, achieving a holistic optimization tailored to the specific task at hand. By evolving both instructions and examples simultaneously, PW ensures significant gains in task performance.

## Three key insights behind PromptWizard:

* *Feedback-driven refinement:*[TODO 1]
* *Joint optimization and synthesis of diverse examples:*[TODO 2]
* *Self-generated chain-of-thought (CoT) steps*: [TODO 3]

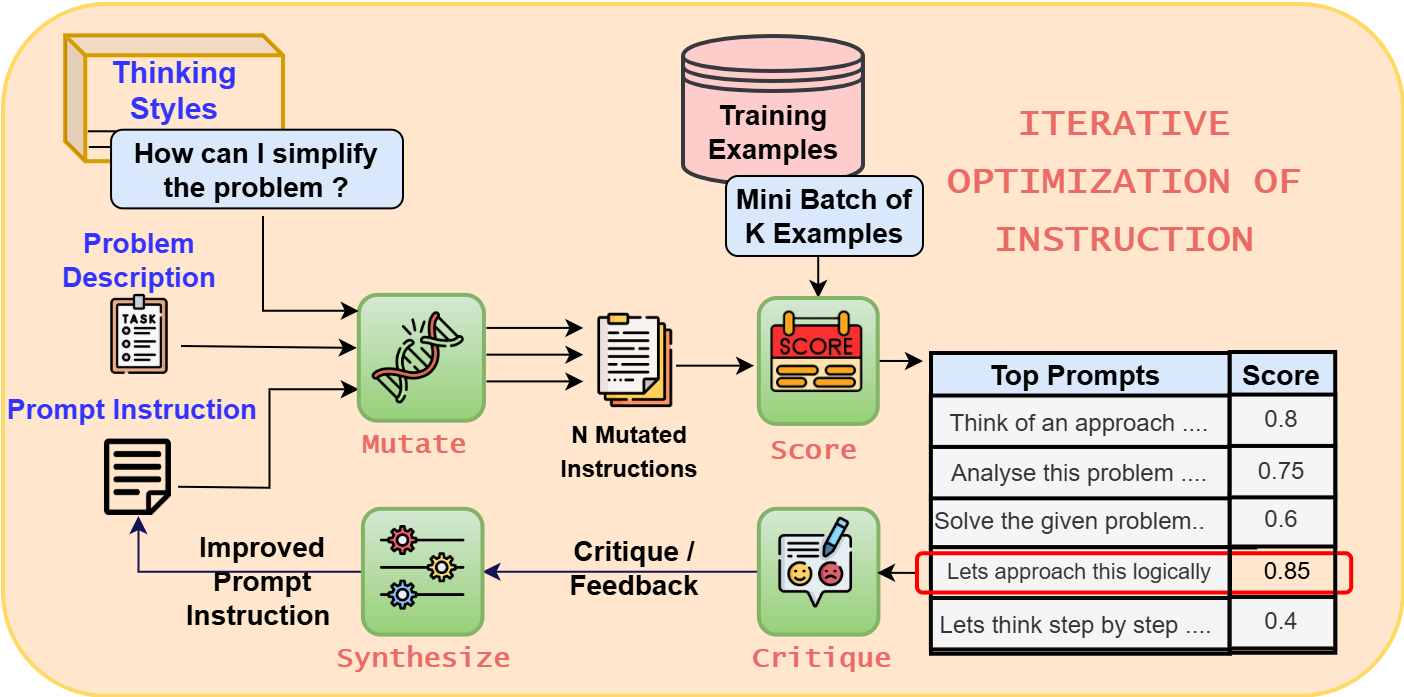
Figure 1. Overview of PromptWizard

# How PromptWizard works

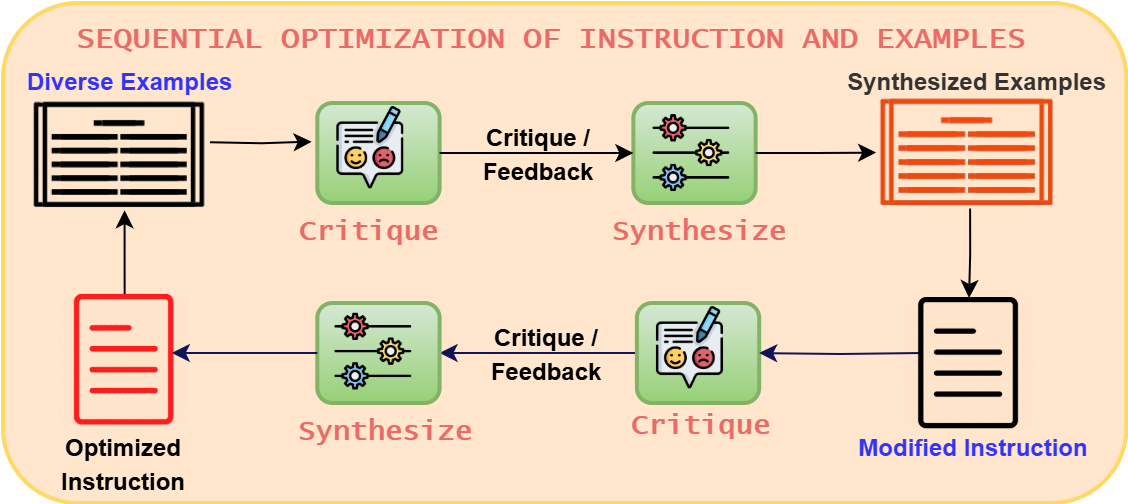
PromptWizard begins with a user input: a problem description, an initial prompt instruction, and a few training examples that serve as a foundation for the task at hand.

Its output is a refined, optimized set of prompt instructions paired with carefully curated in-context few-shot examples. These outputs are enriched with detailed reasoning chains, task intent, and an expert profile that bridges human-like reasoning with the AI’s responses.

## Stage 1: Refinement of prompt instruction

[TODO 4]Figure 2. Refinement of prompt instruction

## Stage 2: Joint optimization of instructions and examples

[TODO 5]Figure 3. Joint optimization of instructions and examples

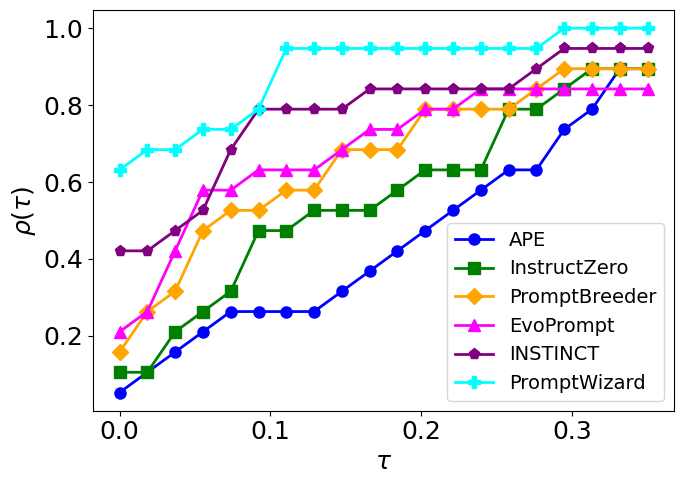
# Results

PromptWizard stands out for its feedback-driven refinement and systematic exploration, delivering exceptional results across a wide variety of tasks while maintaining computational efficiency.

## Comprehensive evaluation across tasks

PromptWizard was rigorously evaluated on over 45 tasks, spanning both general and domain-specific challenges. Benchmarked against state-of-the-art techniques—including Instinct, InstructZero, APE, PromptBreeder, EvoPrompt, DSPy, APO, and PromptAgent—PW consistently outperformed competitors in accuracy, efficiency, and adaptability. Please see detailed results in [our paper](https://www.microsoft.com/en-us/research/publication/promptwizard-task-aware-agent-driven-prompt-optimization-framework/).

[TODO 6]

Figure 4. Performance Profile curve on BBII dataset

|  |  |  |
| --- | --- | --- |
| **Methods** | **API calls** | **Total tokens** |
| Instinct | 1730 | 115k |
| PromptBreeder | 18600 | 1488k |
| EvoPrompt | 5000 | 400k |
| **PW** | **69** | **24k** |

Table 1. Cost analysis on BBII dataset

## Resilience with limited data

[TODO 7]

|  |  |  |
| --- | --- | --- |
| **Datasets** | **5 Examples** | **25 Examples** |
| MMLU | 80.4 | 89.5 |
| GSM8k | 94 | 95.4 |
| Ethos | 86.4 | 89.4 |
| PubMedQA | 68 | 78.2 |
| MedQA | 80.4 | 82.9 |
| **Average** | **81.9** | **87** |

Table 2. PW’s performance with varying number of examples

## Leveraging smaller models for optimization

[TODO 8]

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Prompt Gen: Llama-70B** | **Prompt Gen: GPT4** |
| GSM8k | 94.6 | 95.4 |
| Ethos | 89.2 | 89.4 |
| **Average** | **91.9** | **92.4** |

Table 3. Performance with smaller LLMs for prompt generation

# Conclusion

Whether you are a researcher addressing cutting-edge challenges or an organization looking to streamline workflows, PromptWizard provides a practical, scalable, and impactful solution for enhancing model performance.