OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling

Contribution:

1) A comprehensive large-scale Benchmark (**OptiBench**)
   - Linear Question
   - Nonlinear Question
   - Tabular Data

2) A Data Synthesis Method (**ReSocratic**) Based on this, we further contribute a synthetic dataset, <u>ReSocratic-29k</u>

3) Experimental results show that the <u>ReSocratic-29k</u> significantly improves the performance of open-source models on **OptiBench**.

linear problem without table

Textual Question:

There are two ways to extract a metal from mined ores. The first way is to use process J and the second is process P. Process J can extract 5 units of metal using 8 units of water and produces 3 units of pollution. Process P can extract 9 units of metal using 6 units of water and produces 5 units of pollution. There can be at most 1500 units of water 1350 units of pollution. How many of each type of processes should be performed to maximize the amount of metal extracted?

Optimization Model:

Define j *as* the process number in the type of J, p as the process number in the type of P.

Optimization Target:

$$\max_{j,p\in N} \quad 5j+9p$$

Constraints:

$$s.t. \quad 8j+6p \le 1500, \quad 3j+5p \le 1300$$

nonlinear problem without table

Textual Question:

A piece of cardboard is 1 meter by 1/2 meter. A square is to be cut from each corner and the sides folded up to make an open-top box. What are the dimensions of the box with maximum possible volume?

Optimization Model:

Define W,L *as* the width and length of the cardboard respectively, x as the length of each corner to be cut to make the box.

Optimization Target:

$$\max_{x} \quad (W-2x)(L-2x)x$$

Constraints:

$$s.t. \quad 2x \le W, \quad 2x \le L$$

linear problem with table

Textual Question:

There are six cities (cities 1-6) in Kilroy County. The county must determine where to build fire stations. The county wants to build the minimum number of fire stations needed to ensure that at least one fire station is within 15 minutes (driving time) of each city. The times (in minutes) required to drive between the cities in Kilroy County are shown in the following Table.
Tell Kilroy how many fire stations should be built and where they should be located.

Table (Time Required to Travel between Cities in Kilroy):

From / To	City 1	City 2	City 3	City 4	City 5	City 6
City 1	0	10	20	30	30	20
City 2	10	0	25	35	20	10
City 3	20	25	0	15	30	20
City 4	30	35	15	0	15	25
City 5	30	20	30	15	0	14
City 6	20	10	20	25	14	0

Optimization Model:

Define x_i $(i=1,2,\ldots,6)$ as whether the station should be build in City i.

Optimization Target:

$$\min_{x_1,x_2,\cdots,x_6\in B} \sum_{i=1}^{6} x_i$$

Constraints:

$$\begin{aligned} s.t. \quad & x_1+x_2 \ge 1 \\ & x_1+x_2+x_6 \ge 1 \\ & x_3+x_4 \ge 1 \\ & x_3+x_4+x_5 \ge 1 \\ & x_4+x_5+x_6 \ge 1 \\ & x_2+x_5+x_6 \ge 1 \end{aligned}$$

nonlinear problem with table

Textual Question:

A company is planning to optimize its production of five different products (Product A, Product B, Product C, Product D, and Product E) to maximize profit while considering the environmental impact of production. The profit per unit and the environmental impact per unit for each product are given in the following Table.

Product	Profit per Unit	Environmental Impact per Unit
A	$50	10 units
B	$70	15 units
C	$60	12 units
D	$80	20 units
E	$90	18 units

The company has a total production capacity of 1500 units across all products. The company must produce at least 200 units of Product A and 300 units of Product B to fulfill contractual obligations. The total environmental impact should not exceed 20,000 units. The company wants to maximize the Profit-Impact ratio, where the Profit-Impact ratio is defined as the total profit divided by the total environmental impact.

Optimization Model:

Define $a-e$ as the number of products of A-E that the company should produce.

Optimization Target:

$$\max_{a,b,c,d,e\in N} \quad P/I$$

Constraints:

$$\begin{aligned} & P = 50a+70b+60c+80d+90e \\ & I = 10a+15b+12c+20d+18e \\ s.t. \quad & a+b+c+d+e \le 5000 \\ & a \ge 200 \\ & b \ge 300 \\ & 10a+15b+12c+20d+18e \le 2000 \end{aligned}$$

Compared to former benchmarks, the superiority of our **OptiBench** is as follows:

- **Data Size**: Our benchmark contains 605 instances, which is significantly larger than former benchmarks
- **Data Type**: Our benchmark includes linear, nonlinear problems, and practical tabular format, which are not covered by existing benchmarks.
- **Through Evaluation**: OptiBench will not only examine the mathematical modeling ability of the model but also comprehensively examine whether the numerical results obtained by the model are correct.

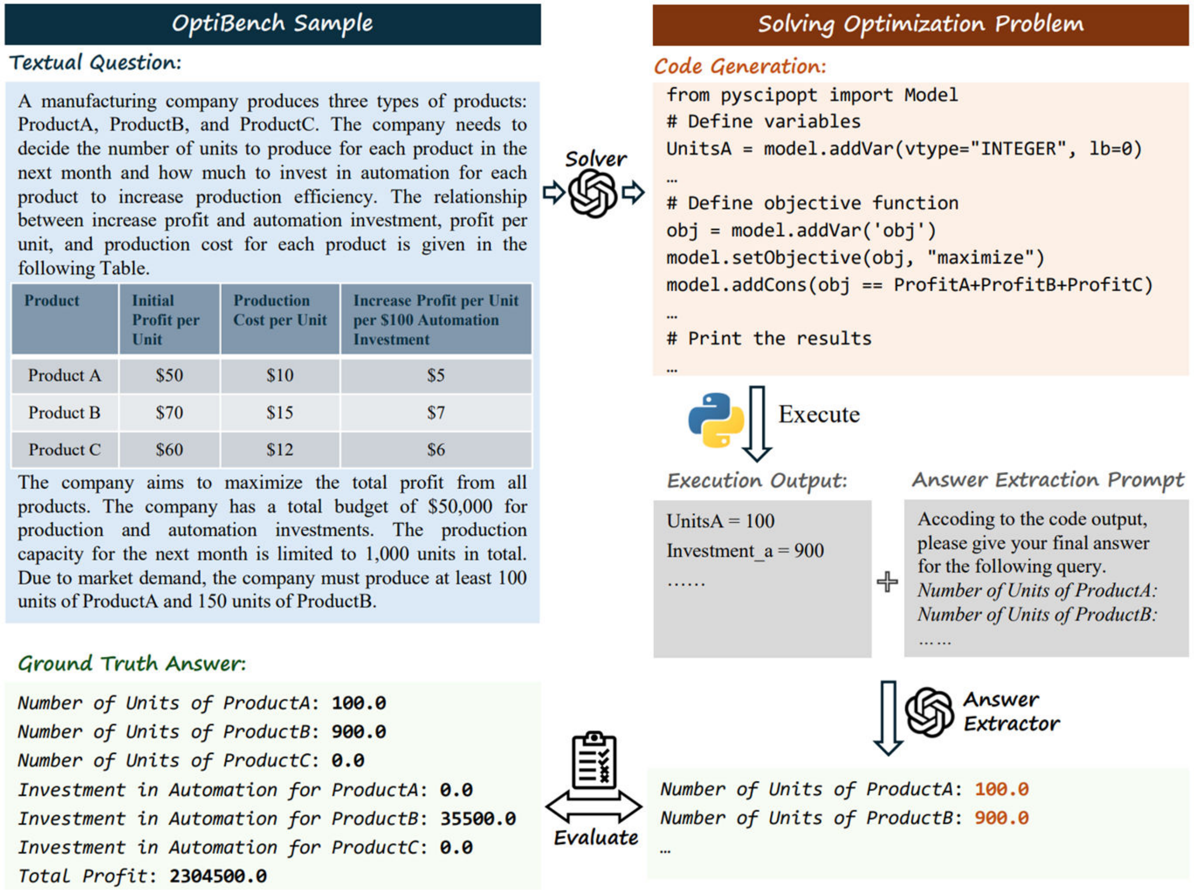
Benchmark	Question Form	Size	End2End	Linear		Nonlinear	
				w/ table	w/o table	w/ table	w/o table
ComplexOR (Xiao et al., 2023)	Implicit	37	√	×	√	×	×
NLP4LP (AhmadiTeshnizi et al., 2024)	Implicit	57	√	×	√	×	×
NL4OPT (Ramamonjison et al., 2022b)	Explicit	289	×	×	√	×	×
OPTIBENCH (**Ours**)	Explicit	**605**	√	√	√	√	√

ICLR

OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling

OptiBench

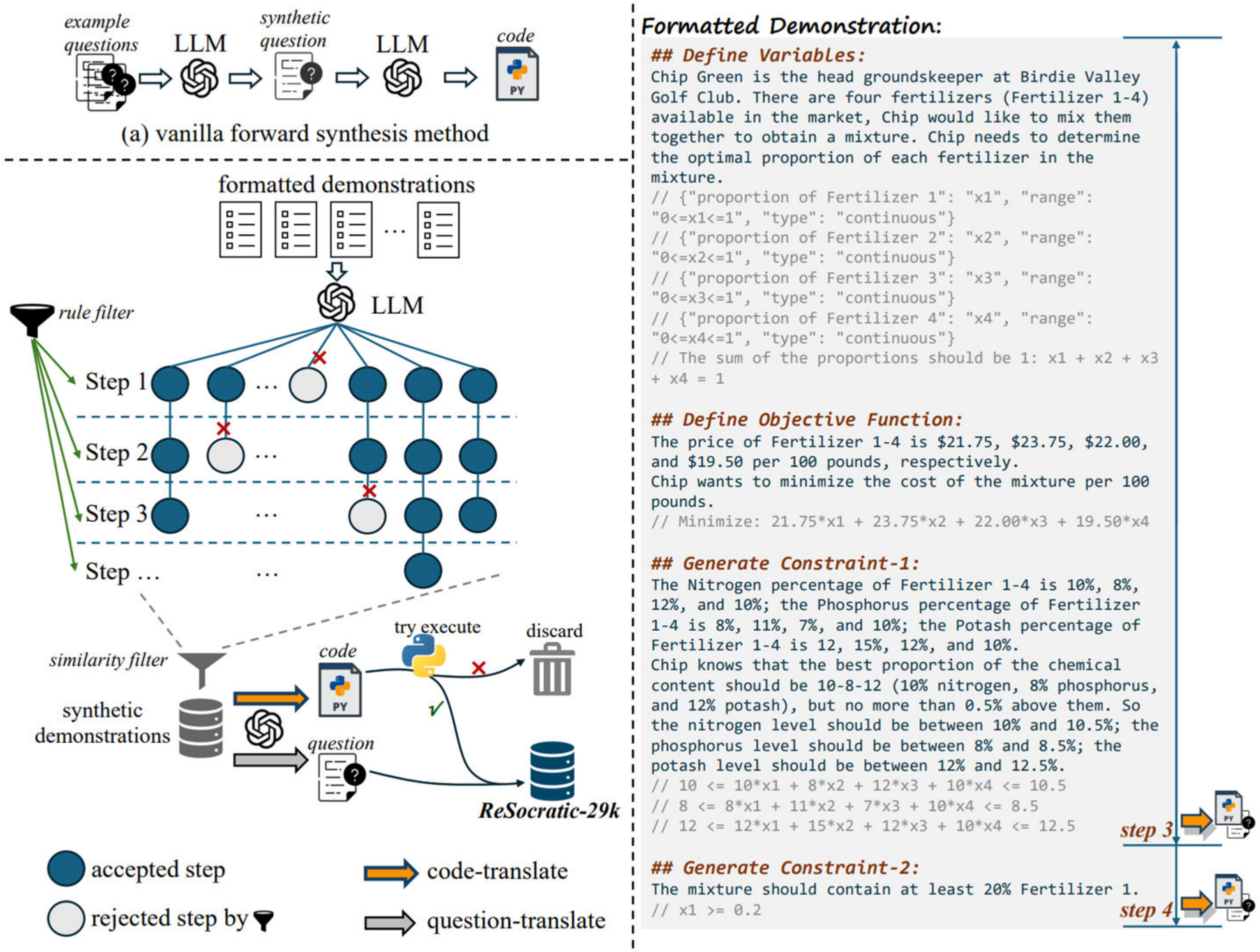
- **Data**: Each sample consists with 1) question text, 2) code solution, 3) numerical results of variables, and objective
- **Evaluation**: During testing, the tested LLM is requested to output code. Then the tested LLM will extract the numerical answers from the execution results.
- **Metric**: A problem is considered solved if and only if all the variables and objectives are correctly matched.



OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling

ReSocratic

- (a) The forward data synthesis method is to synthesize the question first, and then let the LLM generate the answer to the synthetic question.
- (b) In contrast, ReSocratic first synthesizes carefully designed formatted demonstration and then transforms it into code and questions.
- (c) An example of a formatted demonstration.



(a) vanilla forward synthesis method

(b) pipeline overview of our ReSocratic

(c) code & question translation for a 4-step demonstration

OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling

Model	Linear		Nonlinear		All	Code Pass
	w/o Table	w/ Table	w/o Table	w/ Table		
Zero-shot Prompt						
Llama-3-8B-Instruct	0.0%	0.29%	0.0%	0.0%	0.17%	8.8%
Llama-3-70B-Instruct	**76.9%**	50.0%	30.8%	32.0%	59.5%	86.8%
Mistral-7B-Instruct-v0.3	0.6%	0.0%	0.0%	0.0%	0.3%	6.9%
Qwen2-7b-Instruct	3.5%	0.0%	3.0%	0.0%	2.6%	19.2%
DeepSeek-V2	40.4%	27.5%	29.3%	18.0%	34.4%	74.0%
DeepSeek-V2.5	78.4%	**67.5%**	33.1%	24.0%	62.5%	92.7%
GPT-3.5-Turbo	68.1%	37.5%	19.5%	16.0%	49.1%	85.0%
GPT-4	75.4%	62.5%	42.1%	32.0%	62.8%	88.8%
GPT-40-mini	76.0%	48.8%	35.3%	34.0%	60.0%	84.8%
GPT-4o	**78.1%**	65.0%	**45.9%**	**40.0%**	**66.1%**	**90.1%**
Few-shot Prompt						
Llama-3-8B-Instruct	17.8%	2.5%	11.3%	8.0%	13.6%	26.9%
Llama-3-70B-Instruct	79.2%	57.5%	33.8%	32.0%	62.5%	91.2%
Mistral-7B-Instruct-v0.3	40.0%	23.8%	13.5%	18.0%	27.9%	83.8%
Qwen2-7b-Instruct	65.5%	27.5%	18.8%	14.0%	46.0%	87.6%
DeepSeek-V2	79.5%	56.3%	27.1%	32.0%	61.0%	85.5%
DeepSeek-V2.5	79.5%	**71.3%**	40.6%	48.0%	67.3%	91.2%
GPT-3.5-Turbo	75.4%	40.0%	28.6%	26.0%	56.4%	**93.2%**
GPT-4	80.7%	**71.3%**	34.6%	34.0%	65.5%	88.3%
GPT-4o-mini	74.6%	52.5%	14.3%	34.0%	55.0%	74.4%
GPT-4o	**81.0%**	63.8%	**50.4%**	**50.0%**	**69.4%**	91.7%
SFT with Synthetic Data						
Llama-2-7B-Chat	40.6%	11.3%	15.8%	32.0%	30.6%	93.7%
Llama-3-8B-Instruct	**63.5%**	**32.5%**	**33.0%**	**44.0%**	**51.1%**	**96.3%**

OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling

Models	Prompt	Linear w/o Table	Linear w/ Table	Nonlinear w/o Table	Nonlinear w/ Table	Overall Pass
Mistral-7B-Instruct-v0.3	zero-shot	56.4	25.0	23.3	24.0	42.3
Mistral-7B-Instruct-v0.3	few-shot	70.7	38.8	32.3	46.0	56.0
Qwen2-7b-Instruct	zero-shot	71.1	40.0	39.1	26.0	56.2
Qwen2-7b-Instruct	few-shot	78.7	43.8	41.4	46.0	63.1
deepseek-v2.5	zero-shot	91.8	77.5	63.2	56.0	80.7
deepseek-v2.5	few-shot	93.6	81.3	70.7	68.0	84.8
gpt-4o-mini	zero-shot	89.5	72.5	69.2	60.0	80.3
gpt-4o-mini	few-shot	90.1	67.5	72.9	70.0	81.7
gpt-4o	zero-shot	90.6	82.5	80.5	74.0	85.9
gpt-4o	few-shot	92.9	73.8	82.0	70.0	86.1

Model	**SFT Data**		**Linear**		**Nonlinear**		**All**
	Method	Data Acc	w/o Table	w/ Table	w/o Table	w/ Table	
Llama-2-7B-Chat	Self-Instruct (1k responses)	80.0%	16.1%	5.0%	3.0%	4.0%	10.7%
	Evol-Instruct (1k responses)	76.7%	15.5%	**7.5%**	3.8%	6.0%	11.1%
	ReSocratic (1k responses)	**86.7%**	**21.6%**	6.3%	r**5.3%**	**6.0%**	**14.4%**

Reverse data synthesis (answer -> question) has higher accuracy than **forward** data synthesis (question -> answer)