DSPy meets HELM

Structured Prompting Enables More Robust, Holistic Evaluation of Language Models

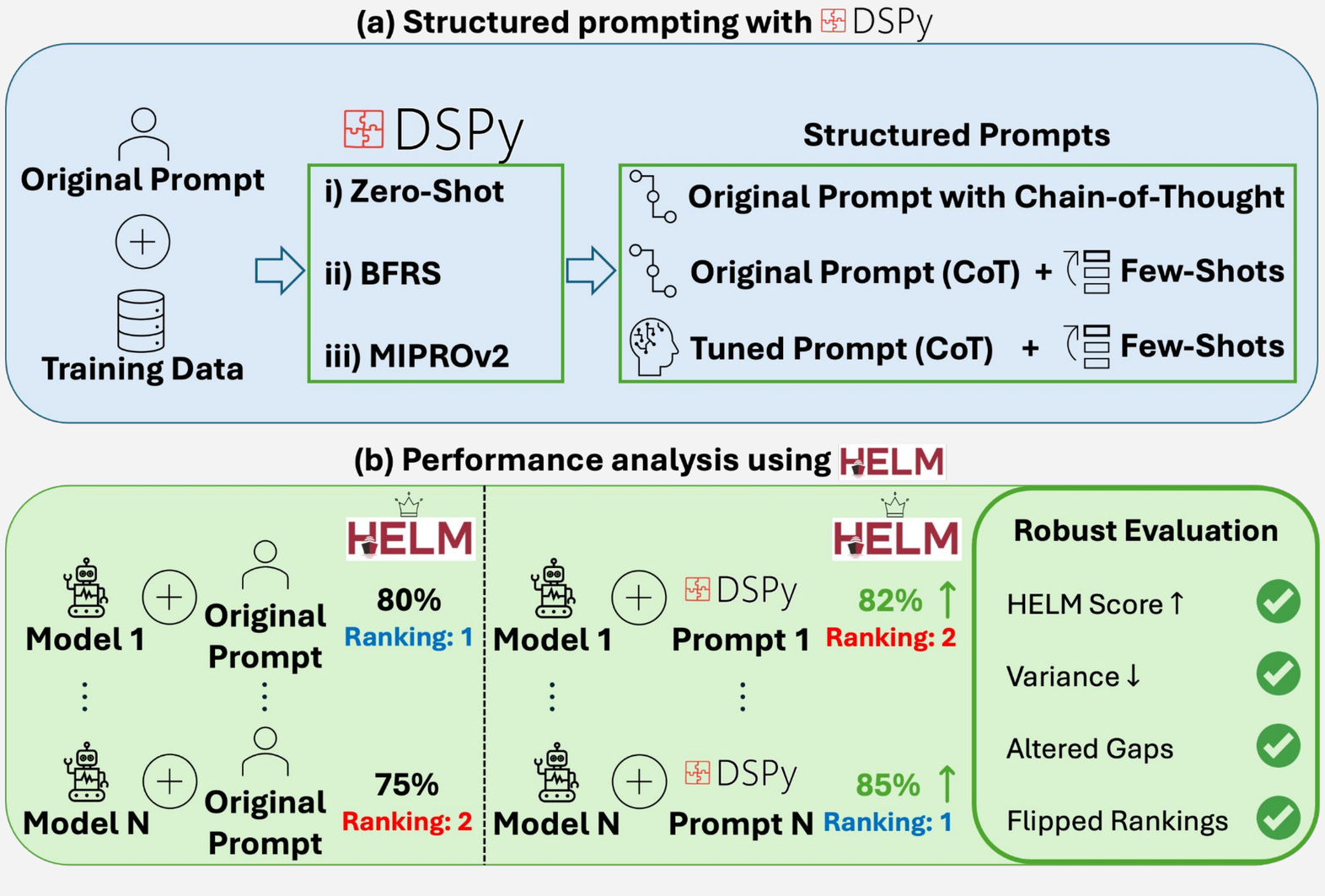
Asad Aali, Muhammad Ahmed Mohsin, Vasiliki Bikia, Arnav Singhvi, Richard Gaus, Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Yifan Mai, Jordan Cahoon, Michael Adam Pfeffer, Roxana Daneshjou, Sanmi Koyejo, Emily Alsentzer, Percy Liang, Christopher Potts, Nigam H. Shah, Akshay S. Chaudhari

Stanford

Corresponding Author: ***Asad Aali*** (*asadaali.com*)

Paper

DSPy+HELM Pipeline



Structured Prompting Methods

Prompt 1: HELM Baseline

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————

Prompt 2: Zero-Shot CoT

Your input fields are: "INPUTS"
Your output fields are: "REASONING" and "OUTPUT"
Your objective is: Given the fields "INPUTS", produce the fields "OUTPUT"

INPUTS:

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————

Respond with the corresponding output fields, starting with "REASONING", then "OUTPUT".

Prompt 3: BFRS (Few-Shot Optimized)

Your input fields are: "INPUTS"
Your output fields are: "REASONING" and "OUTPUT"
Your objective is: Given the fields "INPUTS", produce the fields "OUTPUT"

IN-CONTEXT EXAMPLES (K Demos):

INPUTS: <input text> → REASONING: <steps>, OUTPUT: <output text>

INPUTS:

Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————

Respond with the corresponding output fields, starting with "REASONING", then "OUTPUT".

Prompt 4: MIPROv2 (Instruction + Few-Shot Optimized)

Your input fields are: "INPUTS"
Your output fields are: "REASONING" and "OUTPUT"
Your objective is: You are a highly skilled medical expert working in a busy emergency room. A patient presents with a complex medical history and concerning symptoms. The attending physician needs your immediate assistance in calculating a critical risk score to guide treatment decisions. The patient's life may depend on your accuracy.

IN-CONTEXT EXAMPLES (K Demos):

INPUTS: <input text> → REASONING: <steps>, OUTPUT: <output text>

INPUTS:

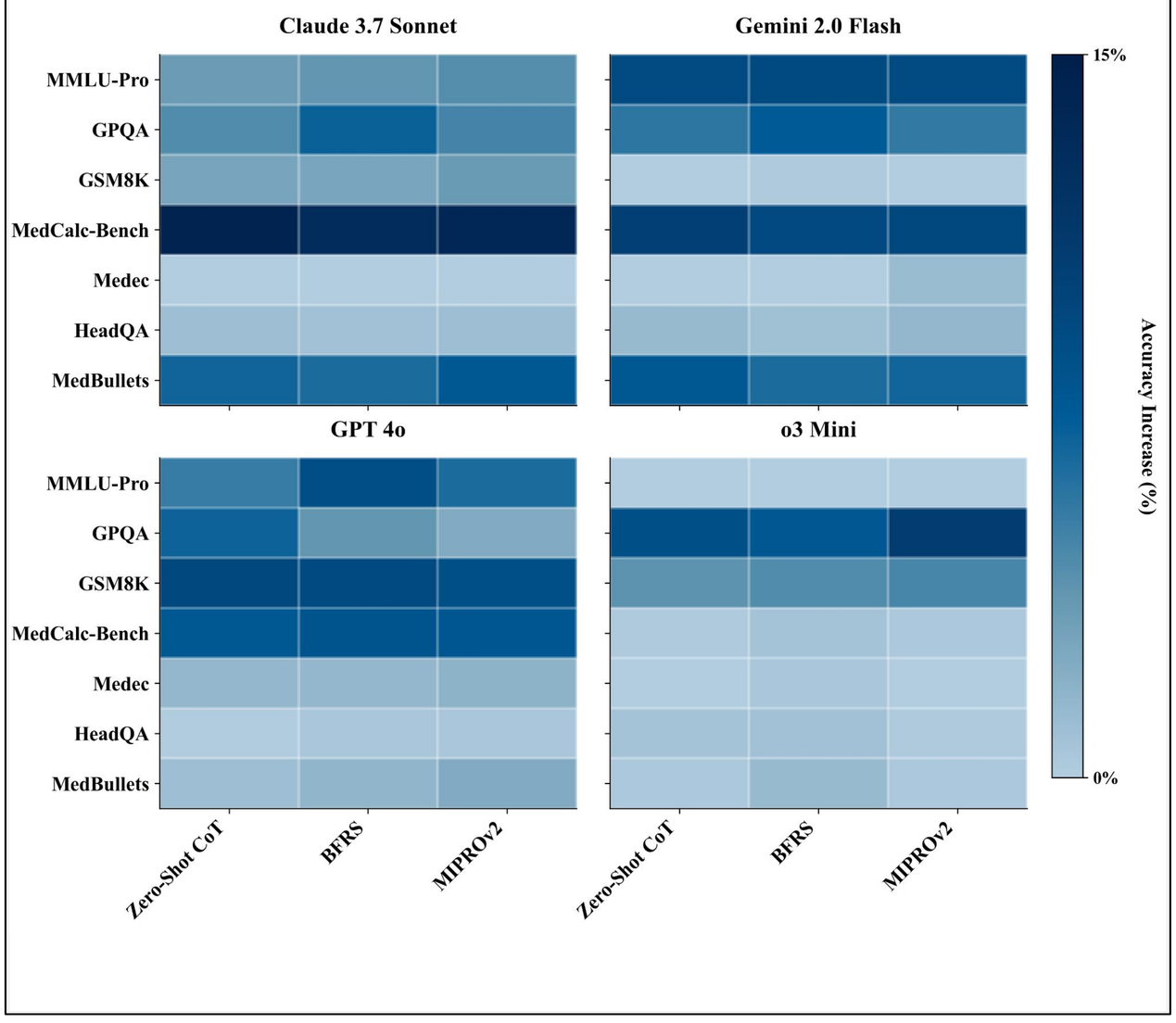
Given a patient note and a clinical question, compute the requested medical value.
Patient Note and Question: ————————————————————

Respond with the corresponding output fields, starting with "REASONING", then "OUTPUT".

HELM Leaderboard (Macro-Averaged)

Prompting Method	Claude 3.7 Sonnet	Gemini 2.0 Flash	GPT 4o	o3 Mini
HELM Baseline	64.81% ± 22.6	61.41% ± 23.8	61.04% ± 23.9	70.93% ± 19.7
Zero-Shot Predict	65.10% ± 22.6	61.69% ± 22.7	59.69% ± 25.0	**73.24% ± 20.3**
Zero-Shot CoT	69.36% ± 18.8	**66.21% ± 20.9**	65.67% ± 22.5	72.73% ± 19.7
BFRS	69.34% ± 19.0	66.19% ± 21.2	**65.87% ± 22.9**	73.07% ± 19.7
MIPROv2	**69.80% ± 19.0**	66.19% ± 21.1	65.34% ± 23.0	73.07% ± 19.6
Ceiling − Baseline (Δ)	+4.99%	+4.80%	+4.83%	+2.31%

Increase in Accuracy over HELM Baseline



Accuracy vs Cost Tradeoff