Data-Driven Automatic Prompt-Optimization for Robust Agentic AI

Zhiyuan Peng[1*], Liyi Zhang[2], Tin Nguyen[3], Chen Zhang[4], Chris Cholete[4], Ilan Twig[4], Itamar Kahn[5]

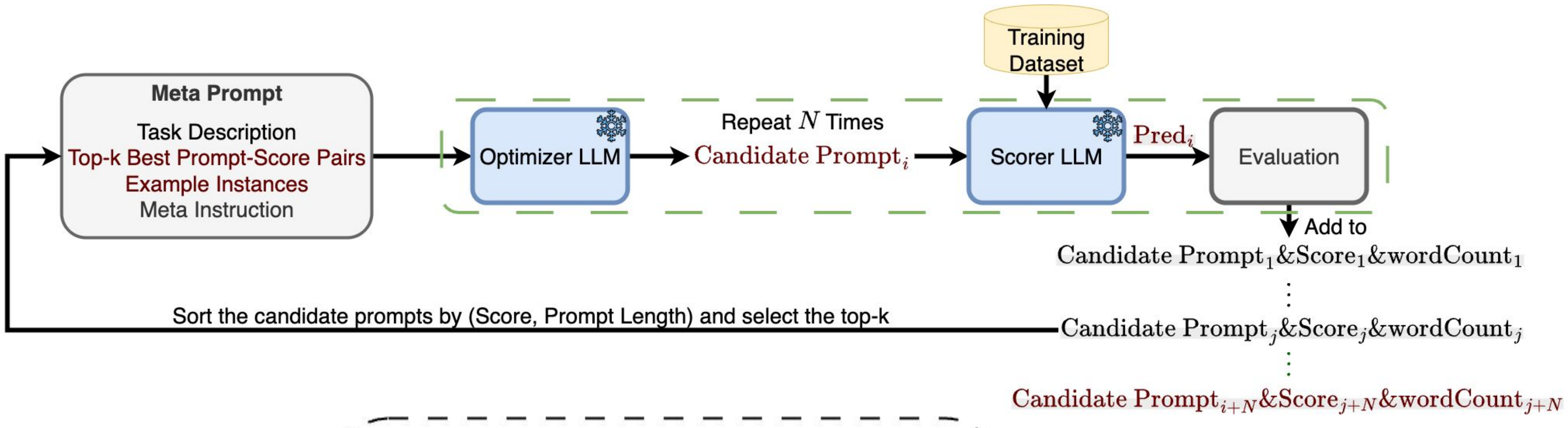
[1]Santa Clara University, [2]Princeton University, [3]University of Maryland, [4]Navan Inc., [5]Columbia University

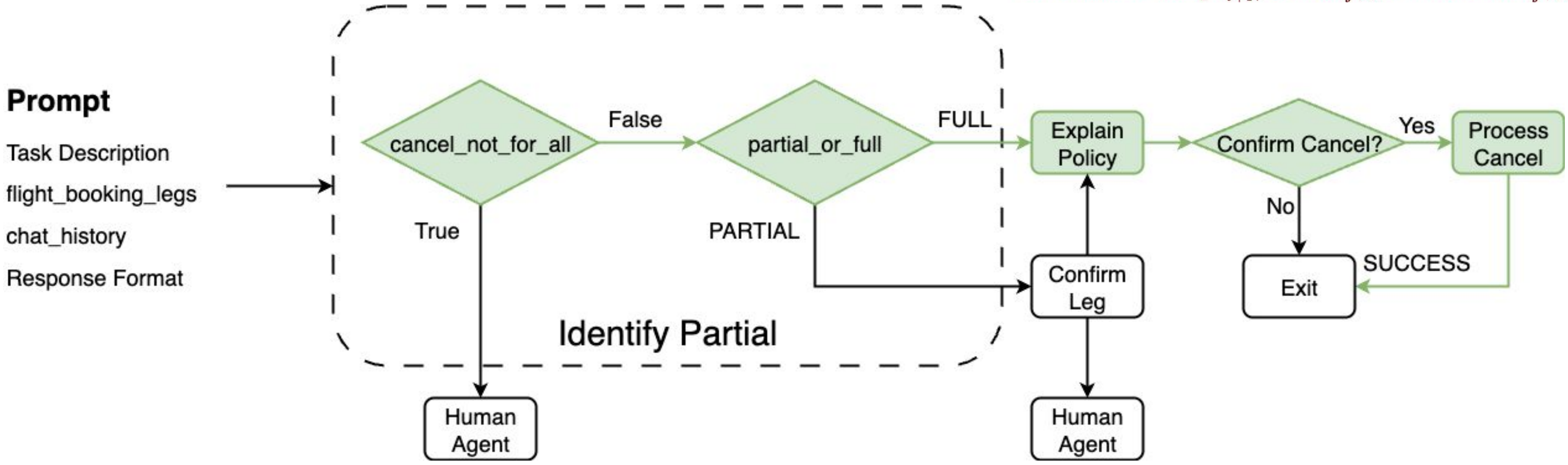
*This work was done during the internship at Navan. Correspondence: zpeng@scu.edu

NAVAN

BayLearn
Bay Area Machine Learning Symposium

Santa Clara Engineering





Introduction / Motivation

LLM-driven agents are increasingly deployed in production settings, yet their workflows often depend on static, human-written prompts that fail to guarantee optimal performance across scale, cost, and accuracy requirements. In this work, we present a data-driven framework for optimizing prompts in deployed agentic AI systems using the OPRO (Optimization by Prompting) methodology. We study Ava, a production AI assistant handling high-volume airline cancellation requests, where classification errors in intent detection (e.g., distinguishing partial vs. full cancellations) can significantly impact customer experience and operational costs. By mining labels directly from user engagement data, we construct balanced datasets and apply iterative OPRO-based prompt refinement with LLMs serving both as scorers and optimizers. Our experiments demonstrate that OPRO significantly improves adjusted balanced accuracy while reducing API cost and latency, outperforming baseline prompts across multiple model families.

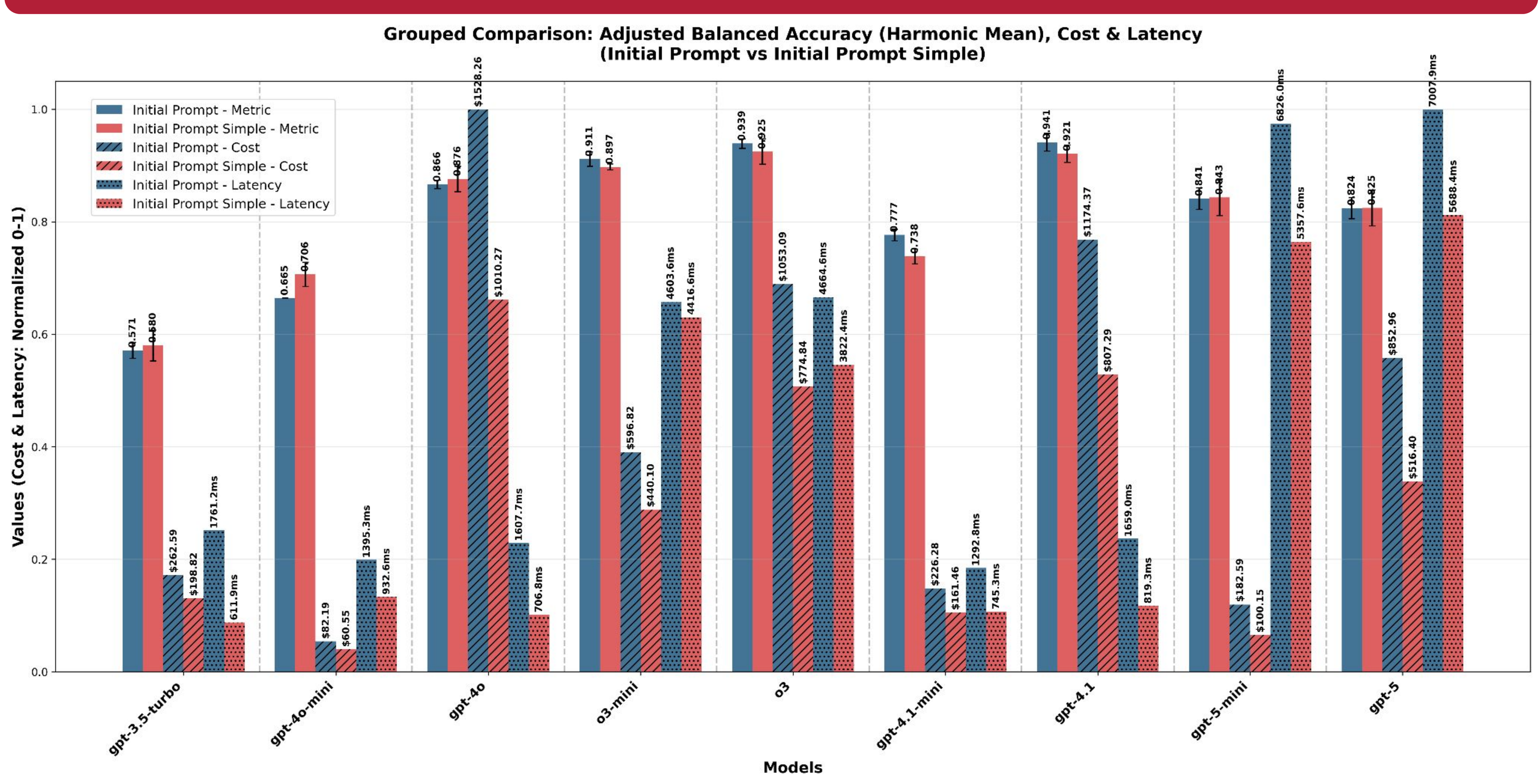
Method

Cancel_not_for_all: “True” when the cancellation is for all the passengers, otherwise “False”. Partial_or_full: “FULL” when the cancellation is for all the booked flights, otherwise “PARTIAL”. As shown in the Green colored path, the user will be routed to different branchs based on the two predicted labels and thus we can annotate the two labels by tracing how historical conversations were routed through the workflow. With the collected data, OPRO is utilized for optimizing the human-written prompt.

	Training Set (208 sessions)		Test Set (90 sessions)	
cancel_not_for_all	Count	%	Count	%
false	148	71.2%	65	72.2%
true	60	28.8%	25	27.8%
partial_or_full	Count	%	Count	%
full	100	48.1%	42	46.7%
null	60	28.8%	25	27.8%
partial	48	23.1%	23	25.6%

Table 1: Distribution of labels in training and test sets.

Results



Methods	cancel_not_for_all	partial_or_full	Avg	# In	# Out	Prompt Length	Cost
Without Reasoning							
4o-mini	62.52 ± 2.69	81.24 ± 1.90	70.64 ± 2.14	379.64	6.00	258	60.55
4o-mini-OPRO	87.75 ± 1.80	73.91 ± 0.00	80.23 ± 0.75	187.64	6.00	66	31.75
4.1-mini	73.60 ± 1.96	74.14 ± 2.13	73.84 ± 1.32	379.64	6.00	258	161.46
4.1-mini-OPRO	97.23 ± 0.62	78.26 ± 0.00	86.72 ± 0.24	216.64	6.00	95	96.26
4.1	93.60 ± 1.96	90.66 ± 2.13	92.09 ± 1.56	379.64	6.00	258	807.29
4.1-OPRO	1.00 ± 0.00	88.48 ± 4.48	93.83 ± 2.59	221.64	6.00	100	491.29
With Reasoning							
4o-mini	56.92 ± 0.00	79.81 ± 0.00	66.45 ± 0.00	424.64	30.83	301	82.19
4o-mini-OPRO	97.54 ± 0.75	73.91 ± 0.00	84.10 ± 0.28	201.64	31.33	78	49.05
4.1-mini	78.40 ± 1.96	76.98 ± 1.74	77.65 ± 1.03	424.64	35.27	301	226.28
4.1-mini-OPRO	93.85 ± 0.75	81.24 ± 1.17	87.08 ± 0.52	191.64	33.78	68	130.71
4.1	94.40 ± 1.96	93.91 ± 2.13	94.14 ± 1.57	424.64	40.64	301	1174.37
4.1-OPRO	97.75 ± 2.26	93.91 ± 3.48	95.73 ± 1.66	215.64	40.73	92	757.14

Table 1: Main results. OPRO variants are highlighted in gray.

Conclusion

By carefully analyzing how customers are directed to different branches, weak labels can be mined from the historical data and thus OPRO can be conducted to improve the performance including cost, and adjust balanced accuracy. Even with the strong base model GPT-4.1, we improve the adjust balanced accuracy from 94.14 to 95.73 while reducing the cost from 1174.37 to 757.17 ($ per 1M LLM calls).