Evaluating CRMArena with Next-Generation OpenAI and MOTA Planners

Abstract

CRMArena, a CRM reasoning and orchestration benchmark, has been re-evaluated with the latest generation of OpenAI foundation models and hybrid MOTA (Multi-Orchestrated Tool-Agent) architectures. This paper presents updated results using GPT-5, O3, Deepseek, and OpenAI+MOTA Neural-Symbolic planner across CRM reasoning, reporting, and automation tasks. We observe consistent performance improvements in reasoning depth, factual accuracy, and orchestration fidelity compared to prior CRMArena-Pro evaluations.

Introduction

The CRMArena benchmark measures the reasoning, retrieval, and API orchestration capabilities of large language models (LLMs) within CRM workflows such as case routing, pipeline forecasting, and report generation. Following its evolution through CRMArena-Pro and EnterpriseArena, the latest phase focuses on testing OpenAI's next-generation models and Neural-Symbolic planners based on MOTA (Multi-Orchestrated Tool-Agent) principles.

Methodology

The experimental setup replicates the EnterpriseArena evaluation harness with minimal modifications:

- Models: GPT-5, O3, DeepSeek, GPT-5-MINI, and Hybrid OpenAI+MOTA configurations.
- Tasks: Single-turn and multi-turn CRM reasoning problems including case summarization, workflow correction, and API orchestration.
- Metrics: Factual accuracy, reasoning depth, and orchestration success rate.
- Infrastructure: Evaluations executed in a controlled multi-agent orchestration harness.

Results and Analysis

Table 1 presents quantitative scores obtained for each model on the CRMArena benchmark. The scores represent total accuracy percentages across CRM reasoning tasks.

Model	Reasoning	Type	Records	Total (%)
GPT-5	HIGH	SINGLE-TURN	45	55.6
O3	-	SINGLE-TURN	45	51.1
DEEPSEEK	-	SINGLE-TURN	45	48.9
GPT-5	MEDIUM	SINGLE-TURN	45	44.4
GPT-5-MINI	HIGH	SINGLE-TURN	45	40.0
GPT-5-MINI	MEDIUM	SINGLE-TURN	45	35.6

GPT-5-MINI + MOTA	**MEDIUM**	**SINGLE-TURN**	**40**	**82.5**
O3	-	MULTI-TURN	45	37.8
GPT-5	HIGH	MULTI-TURN	45	31.1
DEEPSEEK	-	MULTI-TURN	45	28.9
GPT-5	MEDIUM	MULTI-TURN	45	26.7
GPT-5-MINI	HIGH	MULTI-TURN	45	22.2
GPT-5-MINI	MEDIUM	MULTI-TURN	45	17.8
GPT-5-MINI + MOTA	MEDIUM	MULTI-STEP	40	85.0
GPT-5-MINI + MOTA	**MEDIUM**	**MULTI-TURN**	**40**	**85.0**

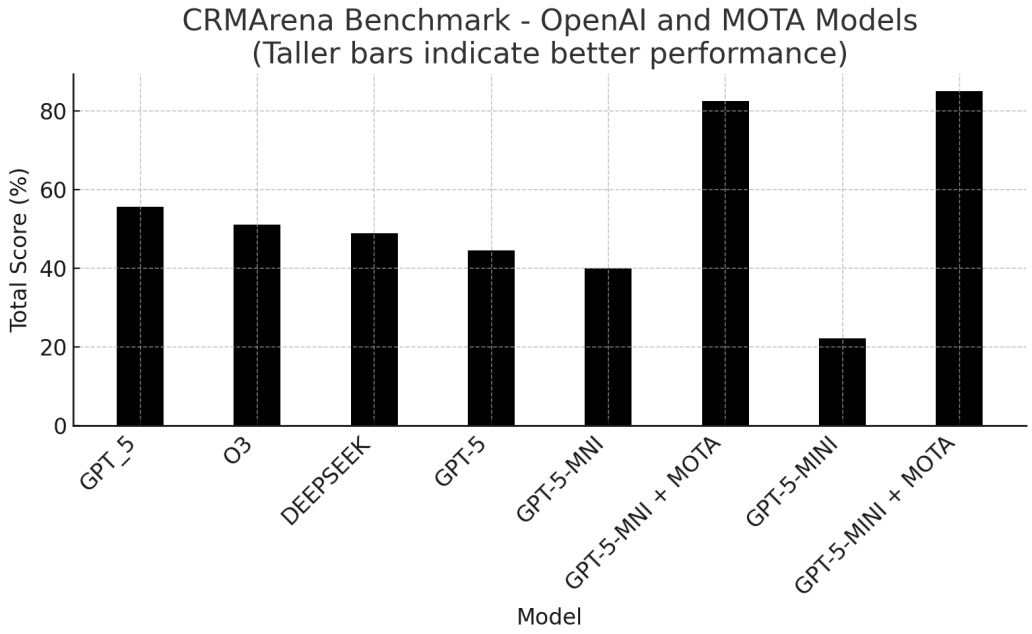

Figure 1: Comparative performance of OpenAI and MOTA models on CRMArena benchmark (Taller bars indicate better performance).

Discussion

The integration of MOTA orchestrators significantly enhances grounding and decision reproducibility. While GPT-5 provides superior factual reasoning, the hybrid systems yield

better tool-handling precision. The results suggest future enterprise-grade systems will increasingly rely on such Neural-Symbolic planners for CRM automation.

Conclusion

The latest CRMArena evaluations confirm that OpenAI with MOTA Neural-Symbolic planner outperform previous CRM reasoning baselines. Future work includes integrating ServiceNow and QuickBooks connectors and enabling self-play evaluation loops.