

Amadeus AI: Benchmarking and Fairness in Agentic Systems for Travel

Conference Summary

SophI.A Summit 2025, November 19–21, Sophia Antipolis

Abstract. Eoin Thomas from Amadeus presented an approach to deploying AI in the travel industry that centers on realistic evaluation and fairness. By adapting TAU-bench to airline scenarios, the team tested agentic systems under realistic policy constraints and improved their behavior through targeted model and tool interventions. A fairness assessment revealed demographic disparities and inspired TravelBBQ, a travel-focused bias benchmark that helped reduce those disparities. Together, these efforts demonstrate pragmatic steps toward responsible AI in production.

Keywords: Agentic AI · Benchmarking · Fairness · Travel Technology · Large Language Models

1 Introduction

The travel industry poses specific challenges: millions of transactions, strict accuracy requirements, and varied airline policies. Amadeus developed methods to build production-ready AI systems that address these constraints while aiming for fairness and reliability. Their work spans operations research, machine learning, and generative AI, applied across the traveler journey from discovery and booking to on-trip assistance and post-trip feedback. Two themes stand out: domain-specific benchmarking for realistic validation, and fairness testing that catches biases general benchmarks miss.

2 Methods

TAU-bench for Airlines: The team adapted TAU-bench [1] to evaluate agentic AI in airline scenarios. The benchmark covers core functions (user details, reservation management, flight search), policy constraints like non-modifiable basic-economy fares, and varied user personas. Systematic optimization—improved models, tool sharpening, condensed policies, and refined prompts—raised GPT-4o’s success rate and reduced costs.

Fairness Evaluation: Adding demographic attributes revealed disparities: non-binary users received lower rewards, younger travelers fared worse, and female users triggered more tokens. In response, they adapted the BBQ benchmark [2] into TravelBBQ for travel contexts and reduced measured bias with domain-aware prompts.

Table 1. Benchmark comparison across models

Benchmark Category	Claude Sonnet 4.5	Claude Opus 4.1	GPT-5	Gemini 2.5 Pro
SWE-bench	77.2%	57%	74.5%	67.2%
TAU-bench	70.0%	63%	62.6%	-
OSWorld	61.4%	44.4%	-	-

3 Results and Limitations

Benchmarking revealed clear performance differences across models, with one model emerging as the strongest and others trailing by varying margins. Cost performance tradeoffs are important: smaller, specialized variants look attractive, while larger models provide incremental reliability gains.

Fairness testing identified measurable disparities that TravelBBQ reduced when using domain-aware prompts. Current fairness checks cover only age and gender, leaving out other critical dimensions such as ethnicity and disability. Finally, synthetic data helps scale evaluation but may introduce distribution-shift risks.

4 Discussion

Open questions remain. What engineering changes (for example, caching or model specialization) will improve cost and latency without sacrificing fairness? And how can multi-agent strategies be used to move these systems toward production-grade reliability? These are practical research directions for future work.

5 Conclusion

Deploying production AI systems in the travel industry requires more than general-purpose models. It demands domain-specific benchmarking, systematic optimization, and fairness-aware evaluation at every stage. The Amadeus methodology demonstrates how quantifying both performance and bias can guide responsible development in any high-stakes domain serving diverse user populations. By making their tools publicly available [3] and building on reproducible benchmarks [1,2], this work contributes to the broader advancement of responsible AI deployment.

References

- Yao, S., et al.: τ -bench: A Benchmark for Tool-Agent-User Interaction. arXiv:2406.12045 (2024)
- Parrish, A., et al.: BBQ: A Hand-Built Bias Benchmark for QA. arXiv:2110.08193 (2022)
- Amadeus IT Group: Travel-specific PIIs Pseudonymization. <https://github.com/AmadeusITGroup/Travel-specific-PIIs-pseudonymization> (2024)