Comparative Evaluation of Large Language Models (LLMs) for Guideline-Concordant Therapy Selection in Multiple Myeloma: A Cross-Sectional Assessment of ChatGPT, Claude, and Gemini

S. MAQBOOL[1], A. IHTESHAM[2], W. DITA[3], M. S. KHAN[4], T. D. AMADASUN[1], M. A. ABBAS[1], I. KHAN[1] and A. IFTIKHAR[5]

[1] HCA Healthcare/Centerpoint Medical Center, Independence, MO, USA; [2] Rawalpindi Medical University, Rawalpindi, Pakistan; [3] NYC H&H/Woodhull Medical Center, New York, NY, USA; [4] London North West University Healthcare NHS Trust, United Kingdom; [5] University of Arizona, Tucson, AZ, USA.

INTRODUCTION:

Treatment selection in multiple myeloma (MM) requires nuanced, individualized clinical decision-making based on disease stage, cytogenetic risk, transplant eligibility, prior lines of therapy, and rapidly evolving guideline recommendations. Large language models (LLMs) have emerged as potential clinical decision-support tools; however, their reliability in complex hematologic oncology scenarios remains insufficiently characterized. We conducted a comparative evaluation of three widely used LLMs for guideline-concordant treatment recommendation in MM.

METHOD:

This cross-sectional study compared ChatGPT, Claude, and Gemini on 40 standardized multiple myeloma vignettes (20 newly diagnosed, 20 relapsed/refractory), balanced by transplant eligibility and cytogenetic risk (standard- and high-risk, including del(17p), t(4;14)). Each model received identical structured prompts for individualized treatment recommendations based on current NCCN/ASH/ESMO guidance. Responses were benchmarked against NCCN-recommended regimens and classified as fully concordant, partially concordant, discordant, or unsafe, with response variability and error patterns assessed as secondary outcomes.

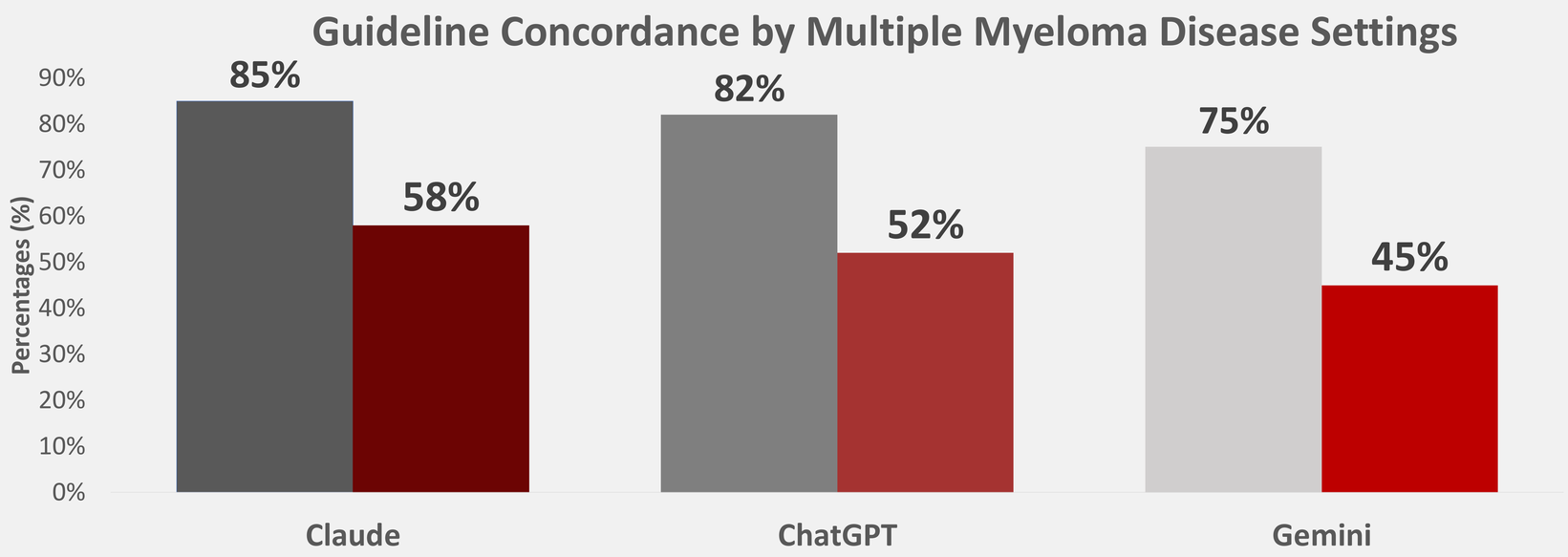
AI model	Overall guideline concordance
Claude	72.5%
ChatGPT	68.0%
Gemini	60.0%

RESULTS

Performance varied meaningfully across models. Accuracy deteriorated further in high-risk cytogenetic scenarios, with concordance decreasing to approximately 45–50% across all models, reflecting frequent omission of risk-adapted therapeutic intensification. Unsafe recommendations occurred in approximately 5% of outputs, primarily involving premature use of advanced cellular therapies or incomplete regimen selection. Response inconsistency on repeated prompting was observed in 12–18% of cases, with Gemini demonstrating the highest variability. Common error patterns included under-treatment of biologically aggressive disease and over-recommendation of novel therapies in earlier treatment settings.

CONCLUSION:

LLMs demonstrate moderate capability in generating guideline-concordant treatment recommendations for MM, with stronger performance in straightforward newly diagnosed cases but reduced reliability in complex, high-risk, and relapsed settings. While these tools show promise as adjunctive educational or decision-support resources, current performance limitations and safety concerns preclude independent clinical use without expert oversight.



Shahzaib Maqbool, hasanshahzaib299@gmail.com
PGY-2, Internal Medicine- HCA Center Medical Point, MO