Clinical fidelity of large language models in acute myeloid leukemia: the impact of guideline prompts.

Meenal Gehlawat, MD · Maria Alejandra Molina-Rodriguez, MD · Ayush Sharma · Saksham Gakhar, PhD · Dani Castillo, MD

Desert Valley Hospital, Victorville, California, USA

Introduction

- Large language models (LLMs) generate human-like medical responses and are increasingly used to support clinical decision-making.
- Acute myeloid leukemia (AML) is a **molecularly complex, rapidly evolving** disease in which classification systems, risk models and targeted therapies change faster than model training corpora.
- Models trained on static data are vulnerable to **knowledge drift** — fluent, confident answers anchored to superseded guidelines.
- LLM accuracy across the AML care pathway has **not been systematically evaluated**.

In this work, we aimed to

(1) measure the accuracy of five leading LLMs across the AML care pathway **(Figs. 1, 2)**, and

(2) determine whether **mandating reference to current NCCN and ESMO guidelines within the prompt** improves clinical fidelity, and in which domains **(Figs. 3, 4)**.

Methods

1 50 standardized clinical vignettes

Eight domains spanning the AML care pathway:

Diagnosis & classification	Cytogenetic & risk stratification
Pretreatment evaluation	Targeted therapy selection
Minimal residual disease	APL management
Adverse effects	Supportive care

2 Five leading LLMs tested

- ChatGPT GPT-5.5
- DeepSeek V3
- Meta AI Muse Spark
- Claude Sonnet 4.5
- Gemini 3 Pro

3 Two prompt conditions, identical questions

ARM A · BASELINE	ARM B · GUIDELINE-AUGMENTED
Clinical question presented alone, with no external reference specified.	Identical question plus an explicit instruction to reference current **NCCN** and **ESMO** guidelines.

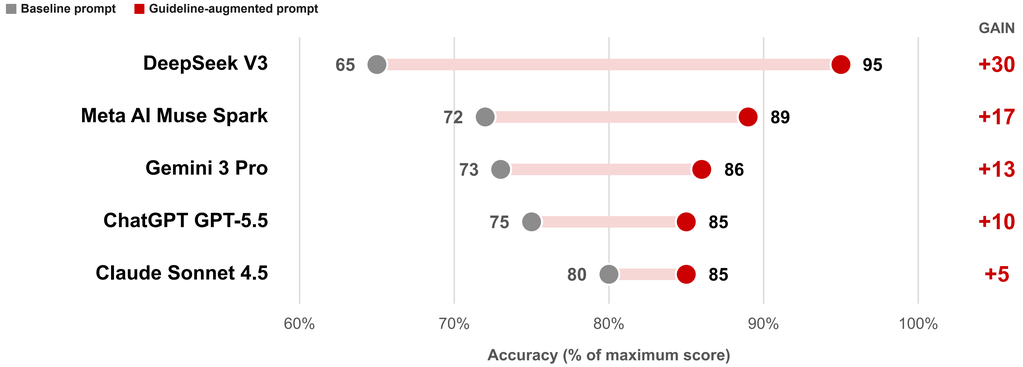
4 Blinded expert scoring

Each of the 500 responses graded against a standardized answer key:

0	1	2
Incorrect	Partially correct	Correct

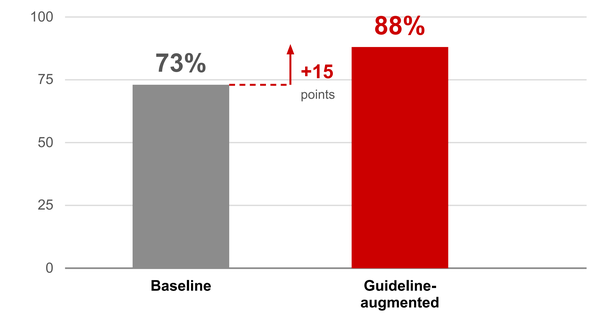
50 vignettes × 5 models × 2 prompt conditions = 500 scored responses. Maximum score 100 per model per arm (50 items × 2 points); accuracy is reported as the percentage of that maximum. Reviewers were blinded to model identity and prompt condition.

Fig. 1: Accuracy of each model under baseline and guideline-augmented prompting. Models ordered by baseline accuracy.



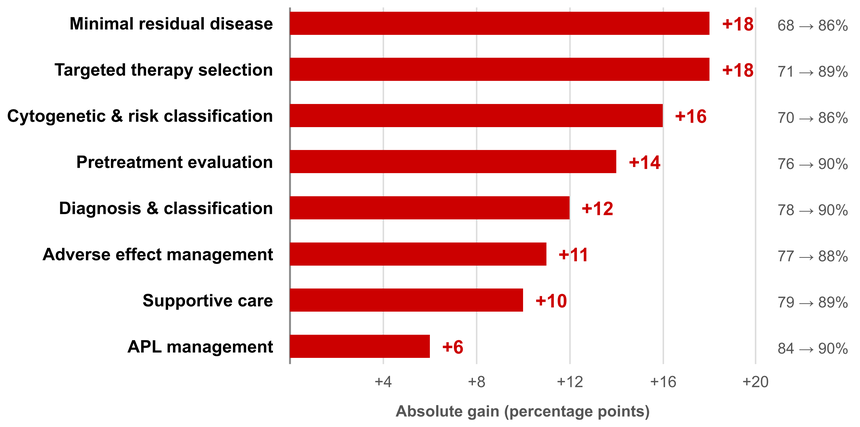
Every model improved. Claude Sonnet 4.5 had the highest baseline accuracy (80%); DeepSeek V3 the lowest (65%) and the largest gain.

Fig. 2: Mean accuracy across all five models, by prompt condition.



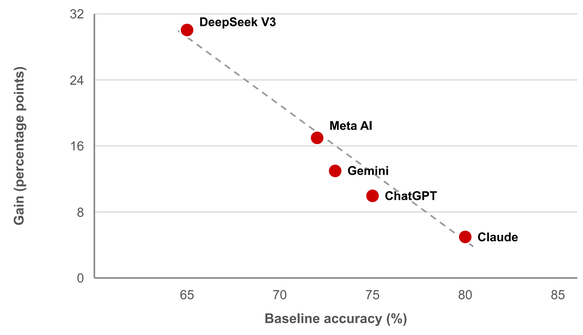
Mean accuracy rose from 73% to 88% of the maximum attainable score.

Fig. 3: Accuracy gain from guideline augmentation by AML domain, pooled across all five models. Figures at right give baseline → guideline-augmented accuracy.



Gains were largest where reasoning depends on current thresholds and drug approvals, and smallest in APL, where baseline performance was already high and the guidance is long-established.

Fig. 4: Gain from guideline augmentation against baseline accuracy, by model.



Gain was inversely related to baseline accuracy: the weaker the model, the more guideline prompting recovered.

Results and Discussion

GUIDELINE PROMPTING IMPROVED EVERY MODEL

- Mean accuracy rose from **73% to 88%** of the maximum score, a **+15 percentage-point** absolute gain. All five models improved; none degraded.
- The between-model range narrowed from **15 to 10 points**, so prompt construction mattered more than model choice.
- Residual errors clustered in **risk stratification and MRD threshold interpretation** — the domains that most directly drive transplant and treatment decisions.

KNOWLEDGE DRIFT IN ALL FIVE MODELS

- Under baseline prompting, models applied **superseded classification and risk systems**, delivered **fluently and without hedging** — no signal to the reader that the reference was outdated.
- Naming the guideline corrected much, but not all, of this drift.
- Ask the model to **state the edition and year** it is applying; drift is otherwise invisible in fluent output.

CONCLUSIONS

- LLMs hold impressive medical knowledge in AML but remain susceptible to **hallucination and knowledge drift**, supporting a shift from **knowledge retrieval to guideline adherence**.
- **Mandatory guideline integration** is simple, model-agnostic and zero-cost, and should be the cornerstone of safe LLM use in oncology.
- LLM output should be treated as **decision support requiring expert verification**, never as an autonomous clinical recommendation.

LIMITATIONS AND FUTURE DIRECTIONS

- Responses were scored by a **single expert reviewer per model**; the absence of dual scoring precludes inter-rater reliability estimates. Vignettes do not reproduce bedside ambiguity, and results reflect specific model versions at one time point.
- Future work should validate **prospectively in clinical settings** and test **automated guideline updating**.

Society of Hematologic Oncology (SOHO) Annual Meeting

Correspondence: Meenal Gehlawat, MD · Department of Internal Medicine, Desert Valley Hospital, Victorville, California

The authors report **no relevant financial disclosures**. No patient data were used in this study.