

INTRODUCTION

Access to timely specialist input in cancer remains variable, particularly in community settings without routine tumor board support.

COMPLIMA was developed as a **multi-large language model (LLM) framework to generate structured, consensus-style treatment recommendations** designed to reflect U.S. community oncology practice patterns.

CONCLUSIONS

- **Multi-LLM consensus was consistent**, with 70% inter-model agreement and **85.2% concordance with expert-preferred management**.
- AI-AI and AI-expert **discordance was concentrated in clinically ambiguous scenarios**, including incomplete staging and therapeutic equipoise.
- COMPLIMA supports the **feasibility of multi-LLM consensus as an adjunctive decision-support** approach in lung cancer care.
- **Larger-cohort validation is warranted** to assess generalizability.

METHODS

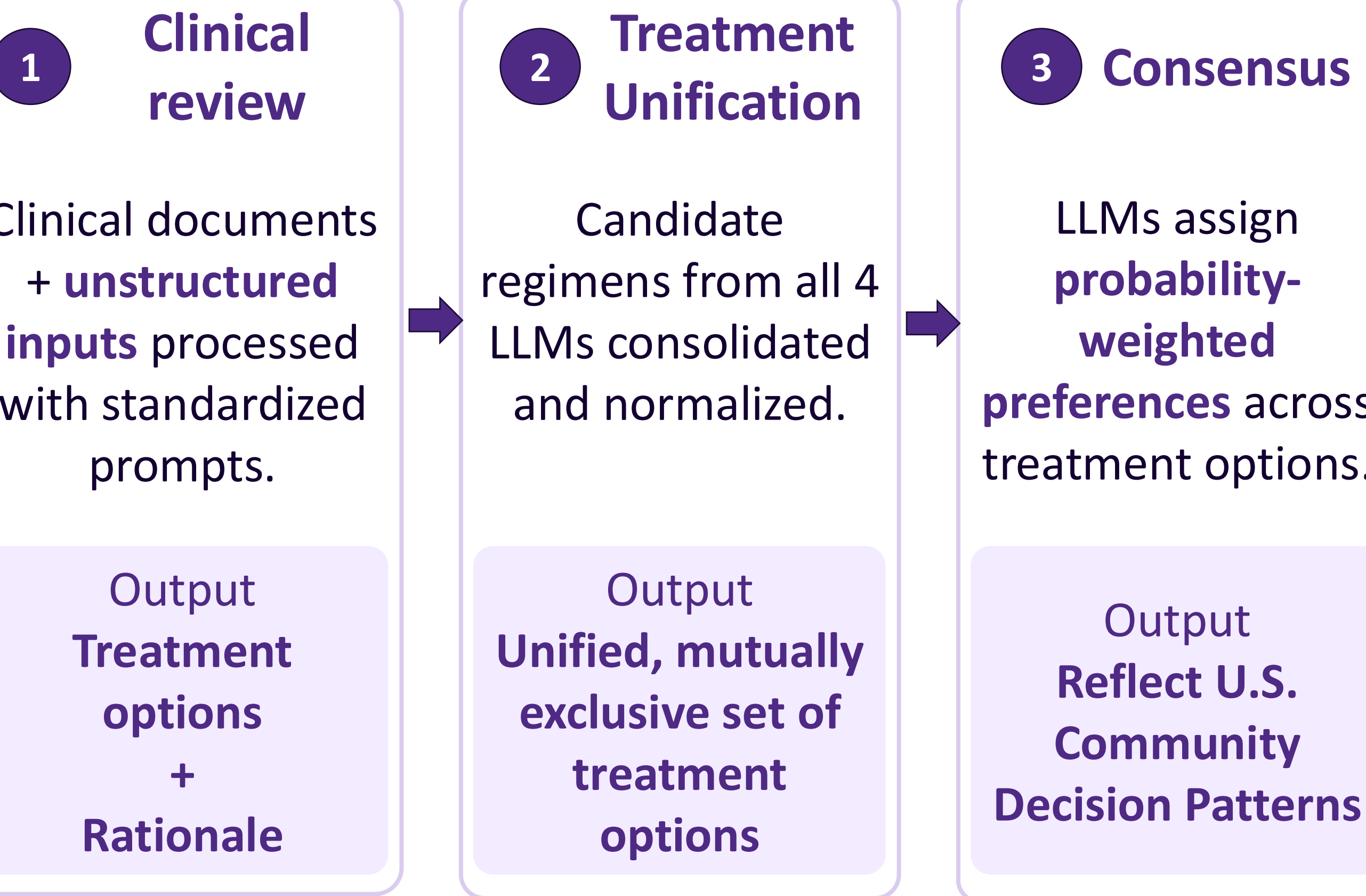
10 REAL-WORLD LUNG CANCER CASES

From practicing oncologists | Diverse disease stages
Variable biomarkers



COMPLIMA MULTI-LLM FRAMEWORK

4 LLMs Working in Parallel
ChatGPT-5.2, Grok-4, Gemini 3 Pro, Claude Sonnet 4.6

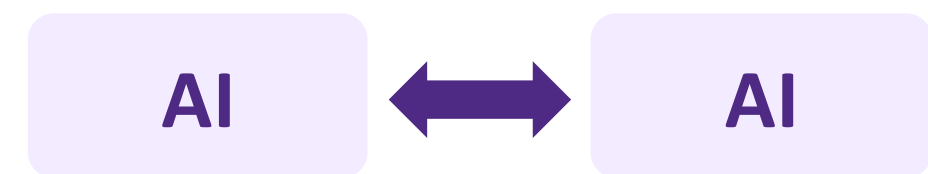


EXPERT ONCOLOGIST ASSESSMENT

3 expert oncologists reviewed treatment options and rationale

Treatment Concordance

Agreement of treatment preferences



Rationale Assessment

Experts scored the accuracy of each LLMs' rationale for their top recommendation.



Concordance and distributional similarity across models & experts

Kendall's τ & Jensen-Shannon (JS) divergence

ChatGPT-5.2, Gemini 3 Pro, Claude Sonnet 4.6, Grok-4 were used to run the analysis described in the abstract

RESULTS

KEY FINDINGS

Mean Kendall's τ
0.690

Inter-model ranking concordance

Mean Jensen-Shannon (JS) divergence
0.226

Inter-model distributional divergence

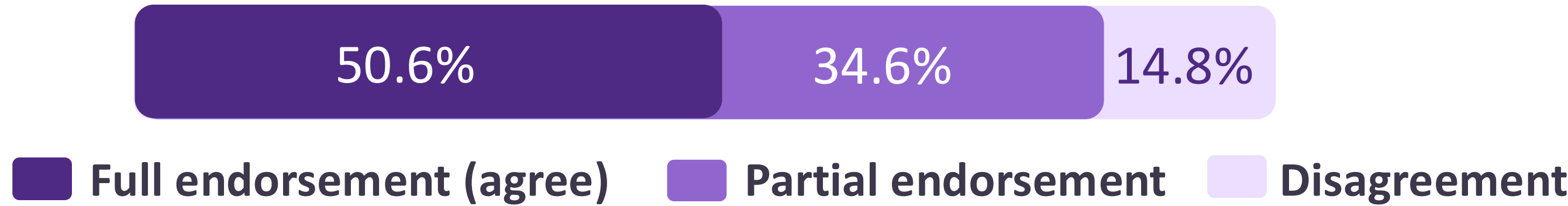
7/10 (70%)

Cases with the same top-ranked treatment across all 4 LLMs

Partial divergence in 3 cases was associated with incomplete staging or known areas of therapeutic variability (e.g., ALK inhibitor selection, treatment intensification in co-mutant EGFR disease).

EXPERT ENDORSEMENT OF AI RATIONALES

85.2% of AI-generated rationales received full or partial expert endorsement



54 AI-generated rationales reviewed by oncologists (162 total assessments)

AI-EXPERT RATIONALE CONCORDANCE

LLM	Mean τ	Mean JS	Top Match
ChatGPT-5.2	0.357	0.706	3/10
Claude Sonnet 4.6	0.258	0.700	3/10
Gemini 3 Pro	0.326	0.622	5/10
Grok 4	0.255	0.717	4/10

Expert rationale scores (agree=2, partial=1, disagree=0) were aggregated across oncologists by treatment option and compared with LLM preference distributions.

Higher τ and lower JS indicate greater concordance.

Top Match: cases where the LLM top-ranked option matched the highest expert-scored option.