ONCODAILY MEDICAL JOURNAL

abstract

Analysis of Large Language Models in Clinical Workflow Management, Cost, and Productivity Performance Through Stepwise Uro-Oncology Scenarios

Mehmet Halici, Muhammed Emin Polat, Burak Ertan, Bahadır Koylu, Yusuf Kasap, Erkan Olcucuoglu, Alaattin Ozen, Fuzuli Tugrul, Ibrahim Cem Balci, Serkan Salturk, Huseyin Uvet

DOI: 10.69690/ODMJ-018-3101-6479

AMSTRO
Asia and Middle East Society of
Therapeutic Radiation and Oncology
Affiliated with ASTF

Asia and Middle East Society for Radiation Therapy and Oncology, 2026

ONCODAILY MEDICAL JOURNAL

abstract

Analysis of Large Language Models in Clinical Workflow Management, Cost, and Productivity Performance Through Stepwise Uro-Oncology Scenarios

Mehmet Halici, Muhammed Emin Polat, Burak Ertan, Bahadır Koylu, Yusuf Kasap, Erkan Olcucuoglu, Alaattin Ozen, Fuzuli Tugrul, Ibrahim Cem Balci, Serkan Salturk, Huseyin Uvet

DOI: 10.69690/ODMJ-018-3101-6479

AMSTRO
Asia and Middle East Society of
Therapeutic Radiation and Oncology
Affiliated with ASTF

Asia and Middle East Society for Radiation Therapy and Oncology, 2026

abstract

Analysis of Large Language Models in Clinical Workflow Management, Cost, and Productivity Performance Through Stepwise Uro-Oncology Scenarios

Authors: Mehmet Halici[1], Muhammed Emin Polat[2], Burak Ertan[3], Bahadır Koylu[4], Yusuf Kasap[2], Erkan Olcucuoglu[2], Alaattin Ozen[1], Fuzuli Tugrul[5], Ibrahim Cem Balci[6], Serkan Salturk[7], Huseyin Uvet[6]

Affiliation: [1]Radiation Oncology, Başakşehir Çam and Sakura City Hospital, Istanbul, Turkey
[2]Urology, Ankara Bilkent City Hospital, Ankara, Turkey
[3]Computer Engineering, Yıldız Technical University, Istanbul, Turkey
[4]Medical Oncology, Koç University, Faculty of Medicine, Istanbul, Turkey
[5]Radiation Oncology, Acıbadem Hospital, Eskisehir, Turkey
[6]Mechatronics Engineering, Yıldız Technical University, Istanbul, Turkey
[7]Electronics and Communication Engineering, Yıldız Technical University, Istanbul, Turkey

DOI: 10.69690/ODMJ-018-3101-6479

Introduction: This study compared the time-based cost-efficiency and generative efficiency of three large language models (LLMs), ChatGPT-5, Gemini 2.5, and Claude Opus 4.1, using comprehensive stepwise uro-oncologic scenarios, and evaluated the effect of optimized prompting on economic performance.

Methodology: Ten clinically realistic stepwise scenarios (five prostate and five bladder cancer cases) were developed by a radiation oncologist, a urologist, and a medical oncologist through expert consensus. Each scenario comprised three stages: diagnosis, treatment, and follow-up, with two open-ended questions per stage. One pilot case tested two prompting strategies: standard stepwise and Sequential Waterfall Prompting (SWP), where preceding steps were cumulatively appended. All scenarios were presented to each model through its APIs using the selected method. For each response, input, reasoning, and output tokens and response times were recorded. Total Cost (USD) was calculated per question using official token pricing. Models were compared for Total Cost, Response Time (s), Generative Efficiency (Output/Input), and Economic Efficiency (Cost/Time). Non-parametric Kruskal–Wallis and one-way ANOVA tests were applied in SPSS v20, comparing both models and scenario steps.

Results: Using SWP, a 17-fold cost reduction was achieved compared with standard stepwise prompting. The average cost per scenario was 0.98 USD, and the average total token usage was 35,124 tokens (input and output combined). Significant differences were found among models for cost, response time, and generative efficiency ($p < 0.001$). Gemini 2.5 achieved the lowest cost and fastest responses, while ChatGPT-5 demonstrated the highest Generative and Economic Efficiency, reflecting a balanced trade-off between performance and resource utilization. Claude Opus 4.1, despite its higher cost, remained competitive in response speed **(Figure 1)**. In the stepwise analysis, cost differences were not significant ($p = 0.066$), but the diagnostic stage showed shorter response

abstract

Analysis of Large Language Models in Clinical Workflow Management, Cost, and Productivity Performance Through Stepwise Uro-Oncology Scenarios

Authors: Mehmet Halici[1], Muhammed Emin Polat[2], Burak Ertan[3], Bahadır Koylu[4], Yusuf Kasap[2], Erkan Olcucuoglu[2], Alaattin Ozen[1], Fuzuli Tugrul[5], Ibrahim Cem Balci[6], Serkan Salturk[7], Huseyin Uvet[6]

Affiliation: [1]Radiation Oncology, Başakşehir Çam and Sakura City Hospital, Istanbul, Turkey
[2]Urology, Ankara Bilkent City Hospital, Ankara, Turkey
[3]Computer Engineering, Yıldız Technical University, Istanbul, Turkey
[4]Medical Oncology, Koç University, Faculty of Medicine, Istanbul, Turkey
[5]Radiation Oncology, Acıbadem Hospital, Eskisehir, Turkey
[6]Mechatronics Engineering, Yıldız Technical University, Istanbul, Turkey
[7]Electronics and Communication Engineering, Yıldız Technical University, Istanbul, Turkey

DOI: 10.69690/ODMJ-018-3101-6479

Introduction: This study compared the time-based cost-efficiency and generative efficiency of three large language models (LLMs), ChatGPT-5, Gemini 2.5, and Claude Opus 4.1, using comprehensive stepwise uro-oncologic scenarios, and evaluated the effect of optimized prompting on economic performance.

Methodology: Ten clinically realistic stepwise scenarios (five prostate and five bladder cancer cases) were developed by a radiation oncologist, a urologist, and a medical oncologist through expert consensus. Each scenario comprised three stages: diagnosis, treatment, and follow-up, with two open-ended questions per stage. One pilot case tested two prompting strategies: standard stepwise and Sequential Waterfall Prompting (SWP), where preceding steps were cumulatively appended. All scenarios were presented to each model through its APIs using the selected method. For each response, input, reasoning, and output tokens and response times were recorded. Total Cost (USD) was calculated per question using official token pricing. Models were compared for Total Cost, Response Time (s), Generative Efficiency (Output/Input), and Economic Efficiency (Cost/Time). Non-parametric Kruskal–Wallis and one-way ANOVA tests were applied in SPSS v20, comparing both models and scenario steps.

Results: Using SWP, a 17-fold cost reduction was achieved compared with standard stepwise prompting. The average cost per scenario was 0.98 USD, and the average total token usage was 35,124 tokens (input and output combined). Significant differences were found among models for cost, response time, and generative efficiency ($p < 0.001$). Gemini 2.5 achieved the lowest cost and fastest responses, while ChatGPT-5 demonstrated the highest Generative and Economic Efficiency, reflecting a balanced trade-off between performance and resource utilization. Claude Opus 4.1, despite its higher cost, remained competitive in response speed **(Figure 1)**. In the stepwise analysis, cost differences were not significant ($p = 0.066$), but the diagnostic stage showed shorter response

times ($p = 0.019$) and higher generative efficiency ($p = 0.021$) compared with the treatment and follow-up phases **(Table 1, see the appendix)**.

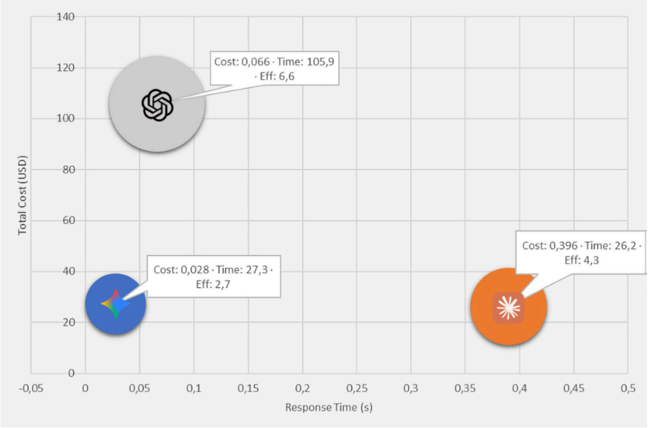

Figure 1: *Cost-time-generative efficiency map illustrating the multidimensional relationship among three LLMs*

Conclusion: This study offers a preliminary yet objective evaluation of LLM performance within structured clinical workflows and delineates model-specific strengths. Gemini 2.5 emerged as the most cost- and time-efficient model, whereas ChatGPT-5 achieved the most favorable overall performance, characterized by superior Generative and Economic Efficiency. The increased productivity observed during the diagnostic phase likely reflects its more focused analytical demands. These findings underscore that future LLM-based clinical decision-support systems should be optimized not only for cost and speed but also for clinical accuracy and explainability (1–3). Continued validation of clinical accuracy will be essential to determine the real-world applicability of these models in oncologic practice.

Conflict of interests: The authors declare no conflict of interests.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

times ($p = 0.019$) and higher generative efficiency ($p = 0.021$) compared with the treatment and follow-up phases **(Table 1, see the appendix)**.

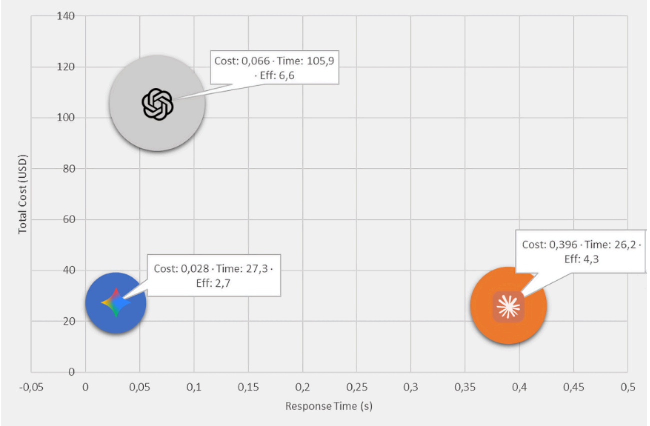

Figure 1: *Cost-time-generative efficiency map illustrating the multidimensional relationship among three LLMs*

Conclusion: This study offers a preliminary yet objective evaluation of LLM performance within structured clinical workflows and delineates model-specific strengths. Gemini 2.5 emerged as the most cost- and time-efficient model, whereas ChatGPT-5 achieved the most favorable overall performance, characterized by superior Generative and Economic Efficiency. The increased productivity observed during the diagnostic phase likely reflects its more focused analytical demands. These findings underscore that future LLM-based clinical decision-support systems should be optimized not only for cost and speed but also for clinical accuracy and explainability (1–3). Continued validation of clinical accuracy will be essential to determine the real-world applicability of these models in oncologic practice.

Conflict of interests: The authors declare no conflict of interests.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Appendix

Table 1: *Comparison of three LLMs and clinical steps across cost, time, and efficiency metrics*

Model Comparison	**Model**	**Mean ± SD**	**Mean Rank**	**p-value**	**Post-hoc Comparison**
Total Cost (USD)	GEMINI	0.0280 ± 0.0068	31.05	< 0.001	OPUS>GPT>GEMINI
	GPT	0.0663 ± 0.0163	89.95		
	OPUS	0.3958 ± 0.0130	150.5		
Response Time (s)	GEMINI	27.30 ± 6.30	64.8	< 0.001	GPT>GEMINI≈OPUS
	GPT	105.93 ± 42.64	150.43		
	OPUS	26.22 ± 6.43	56.27		
Generative Efficiency (Output/Input)	GEMINI	2.70 ± 0.96	40.8	< 0.001	GPT>OPUS>GEMINI
	GPT	6.64 ± 1.88	139.73		
	OPUS	4.31 ± 1.43	90.97		
Economic Efficiency (Cost/Time)	GEMINI	0.00102 ± 0.00067	90.13	< 0.001	OPUS>GEMINI>GPT
	GPT	0.00066 ± 0.00015	30.87		
	OPUS	0.01568 ± 0.00257	150.5		
Step Comparison	Step	**Mean ± SD**	**Mean Rank**	**p-value**	**Post-hoc Comparison**
Total Cost (USD)	Diagnosis	0.1543 ± 0.1675	77.72	0.066	No significant difference
	Treatment	0.1675 ± 0.1681	96.22		
	Follow-up	0.1684 ± 0.1650	97.57		
Response Time (s)	Diagnosis	41.89 ± 28.24	75.05	0.019	Treatment≈Follow-up>Diagnosis
	Treatment	56.12 ± 43.77	98.97		
	Follow-up	61.44 ± 56.81	97.48		
Generative Efficiency (Output/Input)	Diagnosis	5.04 ± 2.30	102.92	0.021	Diagnosis>Treatment>Follow-up
	Treatment	4.59 ± 2.12	91.97		
	Follow-up	4.03 ± 2.04	76.62		
Economic Efficiency (Cost/Time)	Diagnosis	0.00615 ± 0.00769	91.38	0.978	No significant difference
	Treatment	0.00547 ± 0.00676	89.4		
	Follow-up	0.00575 ± 0.00713	90.72		
Total Cost (USD)	Diagnosis	0.1543 ± 0.1675	77.72	0.066	No significant difference