Onco-Logic-Bench: Evaluating Role-Prompted LLMs on Long-Horizon Oncology Trajectories

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as oncology assistants to summarize multi-year charts, check regimens, and "audit" clinical decisions. We introduce **Onco-Logic-Bench**, a de-identified benchmark built from real oncology records to study how such role-prompted LLMs handle long-horizon trajectories. Onco-Logic-Bench contains 68 snippets from 33 patient trajectories (Onco-Logic-68) and defines six logical dimensions: temporal alignment (T1), template-conflict detection (C2), numerical consistency (N3), causal reasoning (C4), cross-document inference (I5), and anatomical/terminology precision (P6). We evaluate two Chinese-developed, general-purpose instruction-tuned LLMs, Qwen-3-32B and DeepSeek-R1-Distill-Llama-70B, prompted as virtual oncology assistants, using an LLM-as-a-Judge protocol with Gemini-3-Flash. The judge provides 1–10 scores and three error flags: fidelity errors, prior-overriding-evidence "professional hallucinations," and rationalization. Both models are near-perfect on temporal and template-conflict dimensions, yet fail systematically on numerical and precision reasoning and show non-trivial professional hallucinations under high average scores in inference. These findings reveal a *fidelity–intuition paradox* and caution against using current LLMs as logical auditors for oncology regimens.

1 Introduction

Large language models (LLMs) are rapidly entering medical and oncology workflows, often through simple role prompts that ask a general-purpose model to act as a "tumor board assistant" or "oncology attending" (Singhal et al., 2023; Nori et al., 2023; Yang et al., 2023; Tu et al., 2024; Li et al., 2024; Thirunavukarasu et al., 2023). In daily use, such role-prompted LLMs are asked to summarize years of notes, reconcile imaging and pathology reports, check chemotherapy and radiotherapy regimens, and even "audit" whether a patient has been treated according to guidelines.

These uses are high-stakes. Oncology trajectories span years and dozens of documents, mixing progress notes, imaging reports, radiotherapy plans, and discharge summaries. Logical consistency requires aligning timelines, anatomical sites, dosages, and response assessments across heterogeneous documentation, including residual copy–paste errors and outdated templates (Meystre et al., 2008; Pampari et al., 2018; Stewart et al., 2022; Wang et al., 2023). At the same time, LLMs are prone to hallucinations—fluent but unsupported content (Ji et al., 2023; McKenna et al., 2023; Luo et al., 2024; Wadhwa et al., 2024). In oncology, strong domain priors can produce *professional hallucinations*: clinically plausible narratives that quietly overwrite conflicting or missing chart evidence instead of flagging it.

Existing work on medical and clinical LLMs mostly evaluates short-form question answering or single-visit summaries (Jin et al., 2019; Singhal et al., 2023; Yang et al., 2023; Thirunavukarasu et al., 2023; Mullery et al., 2024; Zheng et al., 2023; Uapinyoying et al., 2024). Traditional clinical NLP benchmarks emphasize concept extraction or single-note understanding (Meystre et al., 2008; Pampari et al., 2018), while long-context benchmarks study generic reasoning over extended inputs (Yoon et al., 2023; Chiang et al., 2024; Sun et al., 2023). None of these settings directly capture the logical pressures of *decadal* oncology trajectories with real-world documentation noise and explicit evidence conflicts.

In this work, we study how off-the-shelf Chinese-developed, general-purpose LLMs behave when they are *prompted as virtual oncology assistants* on long-horizon cancer trajectories. We introduce **Onco-Logic-Bench**, a de-identified benchmark derived from real-world on-

cology records, and an evaluation protocol that makes professional hallucinations quantitatively measurable. For brevity, we refer to these role-prompted models as "clinical LLMs" in the remainder of this paper, even though they are not domain-finetuned.

Our contributions are three-fold:

- **Onco-Logic-Bench.** We curate **Onco-Logic-68**, a de-identified corpus of 68 snippets from 33 real oncology trajectories, and construct six prototypical multi-document cases (Patients A–F) that instantiate six logical dimensions for long-horizon oncology reasoning: temporal alignment (T1), template-conflict detection (C2), numerical consistency (N3), causal reasoning (C4), cross-document inference (I5), and anatomical/terminology precision (P6).

- **Evaluation protocol for professional hallucinations.** We design an LLM-as-a-Judge pipeline where Gemini-3-Flash scores role-prompted LLMs and annotates three error types: fidelity errors (E_F), prior-overriding-evidence "professional hallucinations" (E_{POE}), and rationalization (E_R), turning these behaviors into measurable quantities on de-identified, multi-year oncology trajectories.

- **Empirical study of role-prompted LLMs.** We evaluate two Chinese-developed, general-purpose instruction-tuned LLMs, Qwen-3-32B and DeepSeek-R1-Distill-Llama-70B, as athlete models on all six dimensions. Both models are near-perfect on temporal and template-conflict dimensions, yet fail systematically on numerical and precision reasoning and exhibit non-trivial professional hallucinations under high average scores in cross-document inference, revealing a *fidelity–intuition paradox* for long-horizon oncology logic.

2 Onco-Logic-Bench Dataset

2.1 Corpus

Onco-Logic-Bench is built from a de-identified oncology corpus that we denote as **Onco-Logic-68**. The source consists of Chinese clinical narratives from a national oncology quality-control project, covering patients with lung cancer, esophageal cancer, lymphoma, and nasopharyngeal carcinoma. We manually select 68 high-density snippets that contain key diagnostic, treatment, and follow-up information and group them into 33 patient-level trajectories. Among these, 12 trajectories contain multiple notes spanning 5–11 years, while the remaining 21 trajectories contain a single note and are not used in the benchmark tasks.

Each snippet is assigned an integer identifier (1–68) that preserves temporal order within a trajectory. Multi-document trajectories are given synthetic labels (Patients A–L); single-snippet trajectories receive generic labels (*Patient_Single_XX*) for potential future extensions.

2.2 Logical dimensions

From the 12 multi-document trajectories, we select six prototypical cases (Patients A–F) and use them to define six logical dimensions for long-horizon oncology reasoning:

- **T1-Temporal:** long-horizon temporal alignment of staging, treatment, and response (lung cancer, Patient A).

- **C2-Conflict:** detection of documentation conflicts caused by re-used templates (esophageal neuroendocrine carcinoma, Patient B).

- **N3-Numerical:** numerical consistency of cycles and dosages in chemotherapy regimens (lymphoma, Patient C).

- **C4-Causal:** causal reasoning about treatment changes and adverse events (esophageal squamous cell carcinoma, Patient D).

- **I5-Inference:** cross-document inference about distant metastasis from partially explicit evidence (nasopharyngeal carcinoma with bone involvement, Patient E).

- **P6-Precision:** anatomical and terminology precision when aligning MRI reports with clinical documentation (nasopharyngeal carcinoma, Patient F).

All subsequent experiments are conducted on these six multi-document trajectories, which serve as focused stress tests for long-horizon oncology logic.

2.3 De-identification

All snippets are automatically de-identified before any interaction with LLMs. Personal names, hospital names, and geographic locations are replaced with placeholders (e.g., `[PATIENT]`, `[HOSPITAL]`, `[LOCATION]`), ID numbers and contact information are removed, and years are normalized to a generic form (`20XX`) while preserving month and day. Clinician signatures are mapped to generic staff markers. This procedure preserves the temporal and logical structure of trajectories while eliminating direct identifiers; athlete and judge models only see the de-identified text.

3 Models and Evaluation Protocol

3.1 Athlete models and prompting

We evaluate two Chinese-developed, general-purpose instruction-tuned LLMs as *athlete models*: **Qwen-3-32B** and **DeepSeek-R1-Distill-Llama-70B**. Both are widely used assistant models with strong Chinese capabilities and no additional clinical fine-tuning. For each logical dimension (T1–P6), we take its prototypical trajectory (Patients A–F), sort the de-identified snippets by identifier (1–68), and concatenate them into a multi-document context. The user prompt appends the Chinese question for that dimension.

The system prompt positions the model as a rigorous oncology chart auditor and explicitly instructs it to prioritize documented evidence over generic medical priors, and to reason inside a `<think>` tag before producing a final answer, following chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022). We query each athlete model three times per task with the same prompts and a low sampling temperature and later aggregate scores and error rates over runs.

3.2 Judge and metrics

We adopt an LLM-as-a-Judge setup (Zheng et al., 2023; Uapinyoying et al., 2024) with **Gemini-3-Flash** as an external judge. The judge receives the de-identified multi-document context, the athlete answer, a short gold answer, and a brief rubric for the target dimension, and returns a JSON object containing a 1–10 score and three binary error flags: a fidelity error E_F (explicit contradiction of chart facts), a prior-overriding-evidence error E_{POE} (professional hallucination, where generic clinical priors overwrite or ignore chart evidence), and a rationalization flag E_R (invented assumptions used to justify an inconsistent answer). We use the score and the three error flags for each (dimension, athlete, run) triple as the basic units for the analysis in Section 4.

4 Results

4.1 Overall performance

Table 1 summarizes performance by logical dimension for the two athlete models. Scores are averaged over three runs per task; error rates denote the fraction of runs where the judge flags a prior-overriding-evidence error E_{POE} (professional hallucination). Aggregated over all six dimensions, Qwen-3-32B achieves an average score of 8.7/10 with E_{POE} in 11% of runs, while DeepSeek-R1-Distill-Llama-70B averages 7.6/10 with professional hallucinations in 17% of runs.

Both models are near-perfect on **T1** (temporal alignment) and **C2** (template conflicts), reliably recovering long-horizon staging and treatment timelines and detecting obvious copy–paste inconsistencies. Performance diverges on more demanding dimensions. On **N3** (numerical consistency), Qwen-3-32B reaches 6.7/10 on average with professional hallucinations in 33% of runs, while DeepSeek-R1 collapses to 2.3/10 with E_{POE} in all runs, indicating systematic failure to track a small set of critical numbers in a short lymphoma regimen. On **C4** (causal reasoning), both models remain strong (8.0–9.0/10, no E_{POE}). On **I5** (cross-document inference), DeepSeek-R1 attains 10/10 on all runs with no errors, whereas Qwen-3-32B is unstable: two runs achieve perfect scores but one run is severely penalized, yielding a 7.3/10 mean and 33% E_{POE}. On **P6** (precision), Qwen-3-32B again scores 10/10 with no errors, while DeepSeek-R1 drops to 5.3/10 and exhibits fidelity errors in two of three runs due to laterality and anatomical misassignments.

4.2 Error profiles and fidelity–intuition paradox

Looking across dimensions, the error profiles of the two models differ markedly. Qwen-3-32B concentrates its errors on **N3** and **I5**: on N3 it alternates between near-correct and numerically inconsistent summaries, and on I5 it mixes two perfect inferences with one run that ignores clearly documented radiotherapy details and is flagged for both fidelity and professional hallucination. In contrast,

Dim	Score ↑		E_{POE} rate ↓	
	Qwen	DeepSeek	Qwen	DeepSeek
T1	10.0	10.0	0%	0%
C2	10.0	9.0	0%	0%
N3	6.7	2.3	33%	100%
C4	8.0	9.0	0%	0%
I5	7.3	10.0	33%	0%
P6	10.0	5.3	0%	0%

Table 1: Performance by logical dimension. Scores are averaged over three runs per athlete and dimension. E_{POE} denotes the fraction of runs flagged by the judge as prior-overriding-evidence professional hallucinations.

DeepSeek-R1 shows a broader spread of fidelity errors. On **N3**, all runs combine E_F, E_{POE} and E_R despite fluent explanations, suggesting that the model defaults to generic chemotherapy patterns and then justifies them instead of reconciling the two snippets. On **P6**, two low-scoring runs swap laterality or mis-anchor a key lymph node to the wrong anatomical region.

These patterns illustrate a *fidelity–intuition paradox*: long-context clinical LLMs can appear highly capable on average—aligning decadal timelines, detecting template conflicts, and producing coherent narratives—yet remain brittle on numerically and inferentially demanding oncology logic. For deployment, relying on a single role-prompted LLM as a logical auditor of regimens or staging is therefore risky even when average scores look strong. Systems built on such models may need to surface E_F/E_{POE}-like signals, aggregate over multiple runs, or combine LLM intuition with symbolic and numerical checks before emitting high-stakes oncology recommendations.

5 Discussion and Limitations

Our evaluation shows that scalar scores alone are a poor proxy for clinically meaningful reliability. A single dimension-level mean (e.g., 7.3/10 for Qwen-3-32B on I5) aggregates runs that are indistinguishable from expert behavior together with runs that contradict clear chart facts. The error triad (E_F, E_{POE}, E_R) helps unpack this gap: E_F captures hard contradictions with the record, E_{POE} highlights professional hallucinations where priors override evidence, and E_R marks attempts to rationalize these mistakes. In practice, we rarely observe E_{POE} in isolation; it typically co-occurs with E_F and E_R when a model both ignores chart evidence and then retrofits a plausible explanation. Monitoring a small set of discrete error signals is therefore more informative than inspecting scores alone.

Our study has several limitations. First, we rely on a single external judge (Gemini-3-Flash), which may introduce its own biases and occasional misjudgments. We mitigate this with de-identified inputs, concise rubrics focused on factual alignment, and a constrained JSON output schema, but our results should still be interpreted as an LLM-vs-LLM evaluation rather than a substitute for physician grading. Second, Onco-Logic-Bench currently covers six multi-document trajectories from one national oncology quality-control setting and a small set of logical dimensions (T1–P6). It is a targeted stress test, not a comprehensive coverage of oncology practice; rare tumor types, multimodal evidence and complex comorbidities remain out of scope.

Taken together, our findings illustrate a *fidelity–intuition paradox* for role-prompted general-purpose LLMs in oncology. The same models that align decadal timelines and detect template conflicts can still fail on numerically and inferentially demanding decisions, sometimes with confident professional hallucinations. This suggests that treating such LLMs as drop-in "logic engines" for regimen or staging audits is premature. More robust systems will likely need to combine LLM intuition with explicit symbolic and numerical checks, aggregate over multiple independent samples, and expose fidelity-oriented signals such as E_F and E_{POE} to human users. Onco-Logic-Bench offers a small but realistic testbed for developing and evaluating these hybrid designs.

Ethics Statement. All trajectories in Onco-Logic-Bench are fully de-identified before any interaction with language models, with direct identifiers, institutions and locations replaced by placeholders and dates partially masked while preserving temporal order. The dataset is used only to study model behavior and is never linked back to clinical care.

Data and Code Availability. Upon acceptance, we will release the evaluation code at an anonymized repository. The de-identified corpus and task specifications will be available for non-commercial research under a controlled-access agreement that forbids re-identification or redistribution.

References

Wei-Lin Chiang and 1 others. 2024. Can large language models be used for medical evaluation? *NEJM AI*.

Ziwei Ji and 1 others. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55.

Qiao Jin and 1 others. 2019. Pubmedqa: A dataset for biomedical word-of-mouth question answering. In *Proceedings of EMNLP*.

Takeshi Kojima and 1 others. 2022. Large language models are zero-shot reasoners. *NeurIPS*.

James Li and 1 others. 2024. Longitudinal clinical reasoning: Challenges for large language models. In *Proceedings of EMNLP 2024*.

Yifei Luo and 1 others. 2024. Knowledge conflicts in large language models: A survey. In *Proceedings of ACL 2024*.

Nick McKenna and 1 others. 2023. Sources of hallucination in llms: Prior knowledge vs. contextual evidence. *Proceedings of EMNLP 2023*.

Stephane M Meystre and 1 others. 2008. Automatic de-identification of medical records. *Journal of the American Medical Informatics Association*.

S. Mullery and 1 others. 2024. Evaluating temporal reasoning in clinical narratives. *Journal of Biomedical Informatics*.

Harsha Nori and 1 others. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Anusri Pampari and 1 others. 2018. emrqa: A large-scale corpus for question answering on electronic medical records. *Proceedings of EMNLP*.

Karan Singhal and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Robert Stewart and 1 others. 2022. Extracting cancer phenotypes from clinical narratives. *Lancet Digital Health*.

M. Sun and 1 others. 2023. Reasoning on patient trajectories: A new benchmark for longitudinal ehr.

Arun James Thirunavukarasu and 1 others. 2023. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

X. Tu and 1 others. 2024. Towards generalist biomedical ai: Evaluation of llms on clinical logic. *BioNLP @ ACL 2024*.

P. Uapinyoying and 1 others. 2024. Rigorous evaluation of medical llms on safety and reasoning. *JAMA Internal Medicine*.

S. Wadhwa and 1 others. 2024. Faithful medical reasoning in long-form documentation. *ACL Findings 2024*.

X. Wang and 1 others. 2023. Clinical concept extraction using llms. *NPJ Digital Medicine*.

Jason Wei and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Guangyu Yang and 1 others. 2023. Clinicalgpt: Optimizing llm for clinical decisions. *arXiv preprint arXiv:2306.09968*.

Jaehong Yoon and 1 others. 2023. Longbench: A comprehensive benchmark for long-context llms. *arXiv preprint arXiv:2308.14508*.

Lianmin Zheng and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.