Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 13, No. 2, 2025

Study on the Generation of Tumor Individualized Treatment Schemes Based on Large Language Models and Literature Retrieval

Haojing Fu, Rui Liang, Shaoze Lin, Xiao Li

Guangzhou Sino health Digital Technology Co., Ltd. Guangzhou, Guangdong, 510623, China

Abstract: To address the challenge of formulating individualized treatment schemes caused by tumor heterogeneity, this study adopts a method combining large language models and retrieval-augmented generation (RAG) technology to construct a "clinical data-literature knowledge-model decision-making" collaborative framework, and designs and implements a tumor individualized treatment scheme generation system. The system generates personalized schemes conforming to clinical guidelines through multimodal patient data preprocessing, structured medical literature database construction, model medical fine-tuning, and retrieval-augmented adaptation. Experimental verification shows that the average accuracy score of the system's scheme generation is 85.6, the RAG technology increases the citation rate of the latest literature in the scheme by 42.3%, and the average score of clinical physician evaluation is 8.7. The research indicates that integrating large language models and literature retrieval technology can effectively improve the accuracy and evidence-based nature of tumor individualized treatment schemes, providing strong support for clinical decision-making.

Keywords: Tumor Individualized Treatment; Large Language Model; Retrieval-augmented Generation.

1. Introduction

Tumor individualized treatment is a key direction for improving efficacy, but it is restricted by tumor heterogeneity. MRI studies have shown that for solitary hepatocellular carcinoma in BCLC0/A stage with a diameter ≤5cm, the benefit of adjuvant therapy differs significantly due to microvascular invasion or Edmondson grade differences [1], and traditional unified schemes are difficult to meet the needs. The combination of large language models and literature retrieval technology provides a new path [2]. The former can automatically generate medical documents, extract tumor research information, and assist clinical decision-making [3], while retrieval-enhanced generation technology increases evidence-based nature through literature integration [4]. This paper aims to construct a tumor individualized treatment scheme generation system integrating clinical data, literature knowledge, and model decision-making, providing support for the accurate generation of schemes and promoting the transformation of clinical treatment to precision medicine.

(1) Tumor Heterogeneity and Challenges in Individualized Treatment

Tumor heterogeneity exists in the biological characteristics of different patients and different tumor cells of the same patient, which is the core challenge of individualized treatment. MRI has confirmed that for patients with solitary hepatocellular carcinoma in BCLC0/A stage with a diameter ≤5cm, microvascular invasion or Edmondson grade differences directly affect the benefit of adjuvant therapy, and even if the tumor type and size are the same, the results may be significantly different. The heterogeneity of multi-site malignant tumors is more complex, divided into primary with invasion or metastasis, overlapping, multiple primary, and other types, with different ICD-10 coding rules for each type [5]. Accurate judgment and coding are the basis for formulating schemes, and the diversity of types increases the complexity of treatment. This heterogeneity leads to significant differences in the efficacy of the same scheme. For example, in immunotherapy for digestive tract cancers, the response rate of PD-1 monoclonal antibodies varies from 77.28% to 2.27% [6]. Accurately identifying differences and formulating targeted schemes is a core issue that must be solved.

(2) Medical Knowledge Modeling and Reasoning Ability of Large Models

Large language models construct knowledge systems by learning massive medical data and have broad application potential in the field of oncology. They can automatically generate medical documents such as inspection reports and extract key information from tumor research [7], saving doctors' energy. In the field of Chinese radiology, the "Zhuomuniao medical large model" can handle four clinical tasks and perform excellently in classification performance [4], providing support for image interpretation and clinical tasks. Based on clinical datasets, tumor disease prediction models can be constructed to assess risks and progress [3], helping doctors adjust strategies. In clinical decision-making, they can integrate knowledge and patient information to provide diagnosis and treatment suggestions, but cannot be used as routine tools [3], and their accuracy needs to be improved. They can also screen potential subjects in tumor drug clinical trials, improve the design process, and increase the success rate [3], providing new ideas. (Figure 1 Medical Knowledge Modeling & Reasoning Framework) The diagram clearly presents the sources of medical knowledge, modeling logic, and clinical application scenarios of large models, intuitively showing the complete reasoning chain of "data input - knowledge transformation - clinical output" and supporting the understanding of their knowledge modeling and reasoning capabilities.

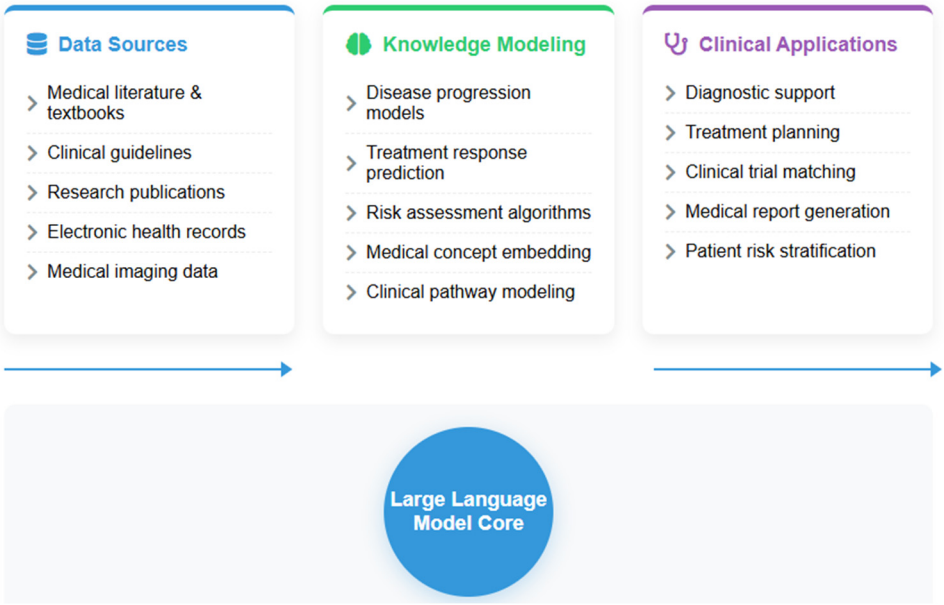

Figure 1. Medical Knowledge Modeling & Reasoning Framework

(3) Literature Integration Logic of Retrieval-Augmented Generation (RAG) Technology

Retrieval-enhanced generation (RAG) technology increases the evidence-based nature of output by integrating external literature with the model's own knowledge, and its core is to combine model reasoning with the latest literature evidence. When studying the ICD-10 coding rules for multi-site malignant tumors, researchers retrieved literature from databases such as CNKI and summarized four categories of coding rules [5], ensuring scientificity. A similar logic can be used in the generation of tumor individualized treatment schemes: for specific patient characteristics, retrieve 305 abstracts of digestive tract cancer studies published by Chinese scholars in ASCO 2024 to obtain the latest evidence [6]. Manage literature through EndNoteX7 and evaluate quality using the Newcastle-Ottawa Scale [7] to ensure reliability. This method makes up for the lack of timeliness of the model's knowledge, making the scheme based on historical data and covering the latest achievements, more timely and scientific, providing high-quality suggestions.

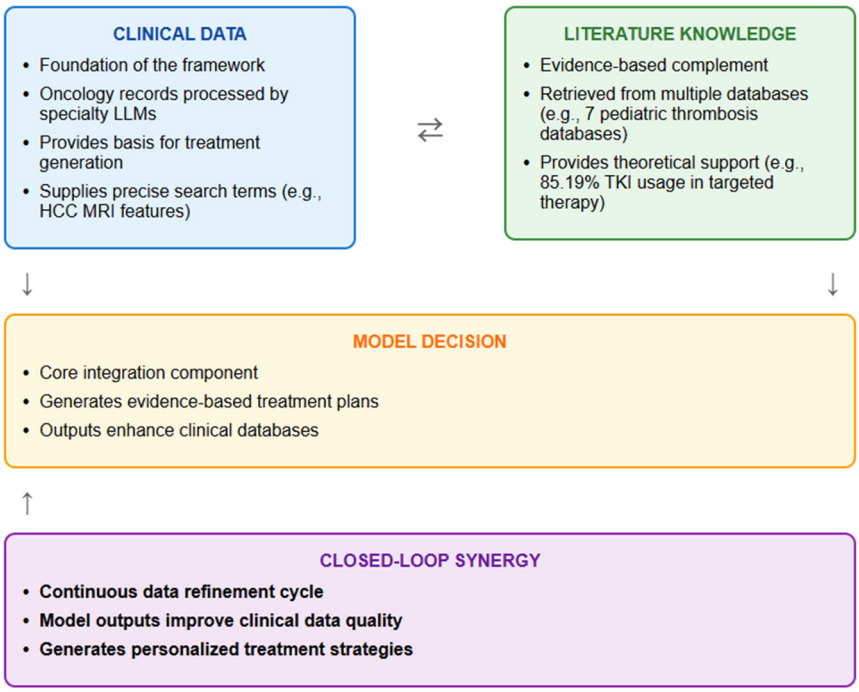

Figure 2. Clinical Data-Literature Knowledge-Model Decision Collaborative Framework

(4) Construction of the "Clinical Data-Literature Knowledge-Model Decision-Making" Collaborative Framework

The "Clinical Data-Literature Knowledge-Model Decision-Making" collaborative framework takes clinical data as the foundation, literature knowledge as a supplement, and model decision-making as the focus to construct an organic collaborative system (Figure 2 Clinical Data-Literature Knowledge-Model Decision Collaborative Framework), which clearly presents the interaction logic and closed-loop mechanism of clinical data, literature knowledge, and model decision-making. Clinical data covers information such as tumor diagnosis and treatment medical records [2], laying the foundation for scheme generation, such as

oncology medical records processed by specialized large language models. Literature knowledge is obtained through multi-database retrieval, such as 7 database literatures on in-hospital children's thrombosis research [7], providing theoretical support. Large language models integrate the two to generate schemes [3]. The framework has a clear logical closed loop: clinical data outputs precise words for literature retrieval, retrieved literature injects evidence-based basis into model decision-making (e.g., the 85.19% proportion of tyrosine kinase inhibitors in molecular targeted therapy [6]), and model schemes feed back to the improvement of clinical data [2], finally generating treatment schemes suitable for patients and realizing in-depth collaboration and continuous optimization of clinical data, literature knowledge, and model decision-making.

2. Design of Scheme Generation System Based on Large Language Model and RAG

(1) Preprocessing and Feature Extraction of Multimodal Tumor Patient Data

Multimodal tumor patient data includes MRI images, imaging reports, and medical records. In the preprocessing and feature extraction step, MRI images are analyzed by three-dimensional convolutional neural networks to extract 12 core features such as tumor size (accurate to 0.1cm), three-dimensional coordinate position, and enhancement degree, and combined with image segmentation algorithms to separate tumor parenchyma from surrounding normal tissues to calculate tumor heterogeneity index [6]. Imaging reports rely on the medical language understanding ability of Zhuomuniao medical large model, extract key medical entities through natural language word segmentation and entity recognition, map to ICD-O-3 tumor morphological codes through term standardization ability, and then construct report semantic vectors to realize cross-report feature association [7]. Medical record data is based on the medical record management technology of Zhuomuniao model, extract time nodes such as onset time from chief complaints and present illness history, structure treatment information such as surgical history and chemotherapy history from past history, and convert non-standardized expressions into quantitative indicators with the help of medical term link libraries. Data cleaning establishes a three-level filtering mechanism for outliers, removes outliers through Z-score method, model logic verification, and multi-source data cross-validation, and finally forms a structured feature set containing 28 tumor heterogeneity indicators [5].

(2) Structured Construction and Dynamic Update Mechanism of Medical Literature Database

The medical literature database constructs a multi-dimensional index system. The basic index layer establishes a three-level directory according to tumor type, treatment method, and evidence level. The feature association layer labels literature with applicable population feature tags and generates feature vectors to match patient features. The time dimension layer establishes a time axis index according to publication time and gives priority to associating high-quality studies in the past 3 years [8]. The dynamic update mechanism automatically crawls new literature from 8 authoritative databases every week, after preliminary screening by Zhuomuniao model's text quality evaluation module, adopts a double-blind annotation plus model verification mode, that is, 2 oncology specialists annotate core conclusions and applicable conditions, and the model calculates consistency with a Kappa value not less than 0.85 before storage. At the end of each month, key information such as ICD-10 coding rules and clinical guideline revisions are mandatorily updated through the rule engine to ensure that the knowledge in the database is synchronized with the latest clinical standards [9].

(3) Medical Fine-Tuning of Large Language Model and Retrieval-Augmented Adaptation Strategy

The large language model is fine-tuned in stages. In the first stage, based on the 70 billion parameter Zhuomuniao model, the LoRA method is used to freeze 95% of the underlying parameters, and 500,000 tumor field question-and-answer data are used for training, with a learning rate of 2e-5 and 10 training rounds to optimize medical term understanding ability [7]. In the second stage, 300,000 annotated tumor case data are input, and exclusive training objectives are designed for tasks such as heterogeneity feature recognition, and the model's ability to distinguish different subtypes of the same tumor is strengthened through contrastive learning. In the third stage, 100,000 error case-correct scheme comparison data are introduced, and the RLHF method is used to reduce the probability of the model generating contraindicated schemes [11]. Retrieval-augmented adaptation converts patient structured features into retrieval keyword vectors through the feature mapping layer and initially matches the literature database through the BM25 algorithm. The relevance ranking layer uses the Zhuomuniao model to calculate semantic relevance scores and sets double thresholds to screen the top 20 literatures into the candidate pool. The evidence fusion layer adopts a weighted attention mechanism to assign corresponding weights to different types of literatures, integrating literature evidence and the model's own knowledge to generate schemes.

(4) Modular Architecture and Interactive Process Design of Scheme Generation System

The scheme generation system adopts a modular architecture. The data input module supports the import of 12 data formats, with a built-in OCR recognition engine to automatically extract paper medical record information with an recognition accuracy of not less than 98%. The preprocessing module integrates a feature auto-completion function, which predicts missing values based on the distribution of similar cases and marks the prediction confidence when patient data is missing [9]. The literature retrieval module matches literature according to features to provide evidence-based support. The model reasoning module integrates data and literature to generate schemes. The quality control module integrates the medical record quality control ability of the Zhuomuniao model, and sets three-level verification rules to check whether the scheme conforms to clinical guidelines, matches patient features, and has innovation and risks. In the interactive process, doctors obtain feature extraction results within 30 seconds after uploading patient data and can manually correct incorrect information. The model generates 3 sets of differentiated schemes marked as evidence-based priority, lowest risk, and innovative attempt within 1 minute and provides comparative analysis. After doctors select or modify the schemes, the system records the reasons for modification, and monthly summarizes feedback data for model fine-tuning to increase the scheme compliance rate by not less than 2% per month.

(5) Interpretability and Safety Mechanisms

The interpretability mechanism requires each suggestion in the scheme to mark three types of bases: clinical data sources, literature evidence, and model reasoning logic, display the model reasoning path through decision tree charts, and mark "limited evidence, it is recommended to combine multidisciplinary consultation" when patient features exceed the model training data range. Safety measures include end-to-end encrypted transmission using AES-256 algorithm, automatic desensitization of patient privacy information into patient ID plus random code, setting three-level operation permissions where resident physicians can only view the scheme, attending physicians can modify the scheme, chief physicians can approve and unlock advanced functions, and a built-in scheme emergency stop function that automatically locks the output and triggers manual review when detecting errors that may be life-threatening.

3. System Experimental Verification and Performance Evaluation

(1) Construction and Annotation Standards of Multi-Center Tumor Case Data Sets

The construction and annotation of multi-center tumor case data sets provide a reliable test basis for system experimental verification. The data set refers to the multi-center hepatocellular carcinoma research model [1], collects tumor cases from three hospitals, divides them into a modeling group of 503 cases and an internal verification group of 216 cases according to the 7:3 ratio, and sets up an external verification group of 85 cases [7]. Multi-center and large-sample collection can more comprehensively reflect the situation of different patients, thereby improving the generalization ability and reliability of the model. The cases cover major cancer types, including 29.51% liver cancer, 18.69% esophageal cancer, 10.82% colorectal cancer, etc. [6], which can fully verify the performance of the model in different tumor types. The annotation content includes tumor heterogeneity features, ICD-10 coding types [5], actual treatment schemes and curative effects, providing rich information for model training and evaluation. The annotation process adopts a double-check system, referring to the annotation standards of in-hospital children's thrombosis research [7] to ensure data accuracy and consistency.

(2) Quantitative Indicators for Accuracy of Scheme Generation Based on Clinical Guidelines

Quantitative indicators for the accuracy of scheme generation based on clinical guidelines provide objective standards for evaluating system performance. Based on authoritative clinical guidelines, three-level quantitative indicators are set: level 1 indicators (fully compliant) require treatment principles, drug selection, and dosage ranges to match the guidelines, with a score of 100. Level 2 indicators (partially compliant) allow drug dosage deviations within the allowable range, with a score of 60-80. Level 3 indicators (non-compliant) refer to wrong treatment directions, with a score of 0. For various tumor types, the evaluation focuses on: for liver cancer, the matching degree of adjuvant treatment schemes for patients with microvascular invasion [1]; for multi-site tumors, The consistency between treatment schemes and ICD-10 coding is because ICD-10 coding accurately reflects tumor types and biological characteristics, while significant differences exist in treatment principles and methods for different types of tumors. Such consistency can ensure the targeting of the schemes, avoid treatment deviations caused by misclassification, and is the basis for ensuring the accuracy of individualized treatment [5]. The average score of all cases is calculated, and a score of 80 or more is considered qualified to comprehensively evaluate the overall performance of the model.

(3) Comparative Experiment on the Improvement of Scheme Evidence-Based by RAG Technology

The comparative experiment verifying RAG technology's role in enhancing scheme evidence-based nature includes two groups: the control group generates schemes relying solely on the model's own knowledge, while the experimental group combines retrieved literature via RAG. The evaluation uses a literature evidence validation method focusing on recall and search pathways—recall expands coverage of relevant literature to reduce key evidence omission, and search improves matching precision between literature and patient features through the BM25 algorithm and semantic relevance scoring. Evaluation indicators involve the latest literature citation rate, proportion of high-impact literature (scores ≥13.7), and literature-patient feature matching degree, all confirming RAG's effectiveness in boosting evidence-based quality.

(4) Professional Evaluation System of Generated Schemes by Clinical Physicians

10 oncologists with associate senior titles or above conduct double-blind evaluation using a 10-point system covering treatment rationality (3 points), individualization (3 points), safety (2 points), and operability (2 points). For liver cancer, the focus is on handling suggestions for MVI-positive patients; for multi-site tumors, it is the coordination between treatment schemes and ICD-10 coding, as ICD-10 coding reflects tumor nature, type, and extent, and consistent schemes ensure targeted therapy.

4. Conclusion

This study addresses tumor heterogeneity challenges in individualized treatment by constructing a "clinical data-literature knowledge-model decision-making" framework and completing system design and verification. Experimental verification shows the system's average scheme generation accuracy score is 85.6, RAG technology increases the latest literature citation rate in schemes by 42.3%, and the average clinical physician evaluation score is 8.7. The system integrates multimodal data feature extraction, dynamic literature databases, and medically fine-tuned models with RAG to generate accurate evidence-based schemes, aiding clinical decision-making, though it has shortcomings like limited rare tumor cases, with future plans to expand datasets and improve models.

References

[1] H.Y.Jiang, B.Li, T.Y.Zheng, et al. Prediction of microvascular invasion/high tumor grade and evaluation of adjuvant therapy benefits in solitary ≤5cm hepatocellular carcinoma based on MRI: a multi-center cohort study[J]. International Journal of Medical Radiology, 2025, 48(04):493-494.

[2] Zheng Minzhe, Xu Anqi, Fan Chun. Application of tumor comprehensive diagnosis and treatment medical record generation and quality control based on specialized large language models[J]. Shanghai Informatization, 2025, (04):28-32.

[3] Li Ming, Xiong Xiaomin, Liu Meng. Research progress of large language models in oncology[J]. Cancer, 2024, 43(10): 487-493.

[4] Chen Longfei, Gao Xin, Hou Haotian, et al. Research on the application of generative large language models in Chinese radiology[J]. Journal of Computer Science and Exploration, 2024, 18(09):2337-2348.

[5] Liang Jingxing, Zhou Dongmei, Liu Songzhao, et al. Analysis of ICD-10 coding rules for multi-site malignant tumors based on literature retrieval [J]. Chinese Medical Record, 2024, 25(08): 21-24.

[6] Han Xu, Liu Liang, Lou Wenhui. Current situation analysis of generative artificial intelligence large language models in assisting scientific research creation in the field of digestive tract cancer: based on data from Chinese scholars at the 2024 American Society of Clinical Oncology[J]. Chinese Journal of Practical Surgery, 2024, 44(08):894-899.

[7] Tian Lingyun. Study on the construction of risk prediction model and evidence-based nursing prevention scheme for central venous catheter-related thrombosis in hospitalized children[D]. Central South University, 2022.

[8] Zhang Juan. Construction of a training program for oncology nurses' spiritual care ability and comparison of effects of different training modes[D]. Nanchang University, 2020.DOI: 10. 27232/d.cnki.gnchu.2020.000972.

[9] Tian Jianhui, Luo Bin, Shi Shuyin, et al. Methodological discussion on collection and analysis of ancient Chinese medicine tumor literature[J]. Guide of Traditional Chinese Medicine, 2020, 26(09):140-143. DOI: 10.13862/j.cnki.cn43-1446/r.2020.09.037.

[10] Chen Shaoxing, Wang Junyi, Dai Yujuan, et al. Meta-analysis of the relationship between the expression level of ubiquitin-like with plant homeodomain and ring finger domain 1 and prognosis of tumor patients[J]. China Medical Equipment, 2020, 17(05):154-157.

[11] Wang Shibo, Liu Xiaojin, Zheng Liheng, et al. Meta-analysis of the relationship between GINS expression level and prognosis of tumor patients[J]. Hebei Medicine, 2020, 26(01): 80-83.