Can Large Language Models Predict Prognosis in Nasopharyngeal Carcinoma?

Anonymous ACL submission

Abstract

Nasopharyngeal carcinoma prognosis prediction aims to estimate future clinical outcomes, such as recurrence or death, from patient clinical and imaging information to support risk stratification and clinical decision-making. Existing survival modeling approaches typically rely on predefined structured feature spaces, which makes it difficult to integrate heterogeneous data such as clinical narratives and often limits generalization across institutions. In this work, we propose CSR-Agent, a reasoning framework that reformulates interval-based survival prediction as a structured multi-stage reasoning process using LLMs. Clinical variables, radiomic features and unstructured clinical information are converted into unified textual representations, and the model performs a two-stage inference procedure consisting of clinical analysis followed by constrained probability generation with temporal consistency. To support privacy-preserving deployment, we further develop a lightweight open-source model via reasoning trace distillation, transferring the reasoning capability of a closed-source LLM to a locally deployable student model. Experiments on four independent multi-center NPC cohorts show that CSR-Agent consistently outperforms traditional survival models and strong LLM baselines, achieving improved predictive accuracy, cross-cohort generalization, and inference stability. The distilled lightweight model maintains competitive performance while enabling efficient local inference.

1 Introduction

Nasopharyngeal carcinoma (NPC) prognosis prediction aims to estimate a patient's future clinical outcomes using information available prior to the start of follow-up modeling (Sun et al., 2014). The input typically consists of baseline patient data, including demographic characteristics, clinical variables, imaging findings, and treatment-related information. Given these inputs, the goal is to estimate the probability that a clinically relevant event, such as death, distant metastasis, or disease recurrence, will occur within a predefined time horizon. These predictions enable risk stratification and support clinical decision-making and follow-up planning (Chen et al., 2016; Liu, 2024).

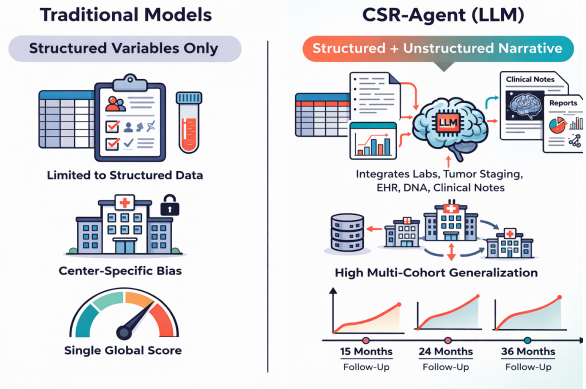

Figure 1: Comparison between the traditional method and the proposed CSR-Agent for survival prediction.

Current approaches for NPC prognosis prediction mainly rely on statistical survival models and deep learning methods. Traditional survival models estimate patient risk from statistical associations between clinical variables and time-to-event outcomes (Cox, 1972; Yu et al., 2011; Ishwaran et al., 2008). More recent studies apply deep learning techniques to capture nonlinear relationships among clinical variables, imaging features, and treatment information (Lee et al., 2018; Lv et al., 2023; Kou et al., 2025). Although these approaches achieve strong predictive performance, they remain largely data-driven and face several limitations. First, many existing methods mainly rely on structured tabular data, making it difficult to incorporate other clinical information such as textual medical records or medical images (e.g., CT or MRI). Second, models trained on data from a single institution often capture site-specific patterns, which limits their generalization to external centers with different patient populations or data distribu-

tions (Koh et al., 2021; Zeiberg et al., 2019; Lv et al., 2023; Kou et al., 2025). Third, many current frameworks focus on producing a single global risk score, which restricts their ability to model how patient risk evolves over time or to provide flexible survival estimates at multiple time horizons (Singer and Willett, 1993; Kvamme and Borgan, 2021).

Motivated by these challenges, we explore the use of LLMs to develop a new framework for NPC prognosis prediction. Leveraging their pretrained knowledge and reasoning capabilities (Wei et al., 2022; Zhou et al., 2022), LLMs can integrate diverse clinical information, including structured patient indicators as well as unstructured textual clinical narratives (Hegselmann et al., 2023; Singhal et al., 2025). Building on these advantages, we propose CSR-Agent, an LLM-based framework for interval-based survival prediction. Instead of predicting a single global risk score, CSR-Agent estimates event probabilities at multiple predefined time horizons and aggregates reasoning outputs to improve prediction stability. This design enables flexible, time-aware, and robust individualized risk estimation across diverse clinical settings. However, although this method provide strong reasoning capabilities, uploading medical records to external servers raises significant privacy risks. To enable privacy-preserving deployment, we further construct a lightweight open-source model that supports local inference. Specifically, patient records are first de-identified to remove personally identifiable information. The anonymized data are then used with a powerful closed-source LLM to generate reasoning traces, which are subsequently distilled to train the lightweight model.

To evaluate the effectiveness of the proposed framework, we construct a comprehensive evaluation benchmark using one publicly available dataset and additional clinical data collected from three independent medical centers. Extensive experiments demonstrate that large language models can effectively perform NPC prognosis prediction by reasoning over heterogeneous clinical information. Compared with directly applying LLMs for prognosis prediction, CSR-Agent consistently achieves significantly better predictive performance, highlighting the importance of structured reasoning over clinical factors. Furthermore, the closed-source LLM-driven version of CSR-Agent achieves state-of-the-art performance without task-specific training and shows strong generalization across all four medical centers in our benchmark. Moreover, the distilled lightweight model maintains competitive performance compared to the closed-source LLM-driven CSR-Agent pipeline, while enabling efficient local deployment in privacy-sensitive clinical environments.

Our main contributions are summarized as follows: (1) We observe that LLMs can effectively perform NPC prognosis prediction through reasoning, and that incorporating unstructured clinical information further improves prediction performance. (2) We develop a lightweight open-source model that enables accurate local risk stratification while preserving patient data privacy. (3) We conduct extensive experiments on one public dataset and three real-world clinical cohorts, demonstrating the effectiveness and robustness of CSR-Agent.

2 Related Work

2.1 Reasoning and Distillation

In recent years, LLMs have demonstrated strong capabilities in textual reasoning tasks (Mitra et al., 2024; Wang et al., 2023). Previous work shows that prompt design and structured reasoning strategies can improve controllability and coherence (Wei et al., 2022; Zhou et al., 2022; Sun et al., 2025; Li et al., 2025). However, most studies focus on free-form textual reasoning (Turpin et al., 2023; Paul et al., 2024). For structured prediction tasks involving numerical estimation and temporal constraints, model outputs remain unstable and difficult to evaluate. As LLMs scale, inference and deployment costs increase (Wan et al., 2023; Narayanan et al., 2021; Wang et al., 2024; Zhou et al., 2024). Post-training and knowledge distillation are widely used to adapt large models and transfer capabilities to smaller ones (Ouyang et al., 2022; Sun et al., 2019; Li et al., 2021). Recent work further applies distillation to improve stability and consistency in complex reasoning tasks (Hsieh et al., 2023; Shridhar et al., 2023; Magister et al., 2023; Ho et al., 2023), providing a promising direction for preserving advanced reasoning capabilities under limited model capacity.

2.2 Discretized Prediction and Generalization

In time-related structured prediction tasks, interval-based modeling is widely adopted (Li et al., 2016; Yu et al., 2011; Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2021; Wiegrebe et al., 2024), where the continuous time axis is discretized into predefined intervals to enable predictions at multi-

ple time points and enforce temporal consistency. Prognostic prediction further relies on heterogeneous features across multiple dimensions, including clinical variables, image-derived features, and their temporal dynamics (Kou et al., 2025; Lv et al., 2023; Wang et al., 2025; Xu et al., 2023; Lv et al., 2025). The combination of temporal constraints and multi-dimensional features makes these tasks particularly sensitive to distribution shifts and modeling inconsistencies. Multi-cohort evaluation provides a realistic setting for assessing generalization under distribution shifts, as models often exhibit degraded stability and consistency across centers or real-world environments (Koh et al., 2021; Zeiberg et al., 2019; Collins et al., 2015; Kou et al., 2025).Recent work also explores transforming numerical features into semantically structured textual representations to better support downstream prediction (Hegselmann et al., 2023; Chen et al., 2023; Singhal et al., 2025). Existing work studies reasoning, feature modeling, and generalization largely in isolation. We address this gap through a unified framework for discretized structured prediction under multi-center evaluation.

3 Task Formulation

Let $\mathcal{D} = \{(x_i, T_i, \delta_i)\}_{i=1}^{N}$ denote a cohort of N patients. For each patient i, $x_i \in \mathbb{R}^d$ represents the baseline clinical information, including structured variables (e.g., demographic characteristics and laboratory measurements) and optional unstructured information such as clinical narratives or imaging reports. T_i denotes the observed time-to-event, and $\delta_i \in \{0, 1\}$ indicates whether the event is observed ($\delta_i = 1$) or censored ($\delta_i = 0$).

The goal of prognosis prediction is to estimate the probability that a clinically relevant event occurs within a predefined time horizon. Formally, we define the survival function

$$S(t \mid x_i) = P(T_i > t \mid x_i), \tag{1}$$

which represents the probability that patient i remains event-free beyond time t. Equivalently, the cumulative event probability can be written as

$$F(t \mid x_i) = P(T_i \leq t \mid x_i) = 1 - S(t \mid x_i). \tag{2}$$

Instead of predicting survival continuously over time, we consider a discrete set of predefined time horizons $\{t_1, t_2, \ldots, t_K\}$. The model predicts a sequence of survival probabilities

$$\{S(t_1 \mid x_i), S(t_2 \mid x_i), \ldots, S(t_K \mid x_i)\}, \tag{3}$$

subject to the monotonicity constraint

$$S(t_1 \mid x_i) \geq S(t_2 \mid x_i) \geq \cdots \geq S(t_K \mid x_i), \tag{4}$$

reflecting the fact that survival probability decreases over time.

From a modeling perspective, the model outputs survival probabilities at multiple time horizons. To obtain a single risk estimate for each patient, we define a risk score based on the predicted event probability at a specific time horizon t^*:

$$r_i = F(t^* \mid x_i) = 1 - S(t^* \mid x_i), \tag{5}$$

where r_i denotes the predicted risk score for patient i.

For discrimination evaluation, the predicted risk scores $\{r_i\}_{i=1}^{N}$ can be used to rank patients, corresponding to metrics such as the C-index and time-dependent AUC. For risk stratification, a predefined threshold τ is applied to divide patients into different risk groups:

$$g_i = \begin{cases} \text{High-risk}, & r_i \geq \tau \\ \text{Low-risk}, & r_i < \tau \end{cases} \tag{6}$$

where τ is typically determined by cohort-level statistics such as the median or a predefined quantile of the predicted risk distribution.

This formulation enables individualized risk estimation while supporting clinically meaningful population-level risk stratification.

4 CSR-Agent

We propose CSR-Agent, a framework that leverages LLMs to integrate heterogeneous clinical information and perform structured prognostic reasoning for interval-based survival prediction. CSR-Agent addresses two key challenges in NPC prognosis prediction: incorporating heterogeneous clinical inputs, including structured variables and unstructured medical data, and performing clinically grounded reasoning to estimate time-dependent survival probabilities.

4.1 Unified Representation

Clinical prognosis prediction relies on diverse sources of patient information, including structured clinical variables and unstructured textual medical data. To enable LLM-based reasoning over such heterogeneous inputs, we convert all clinical

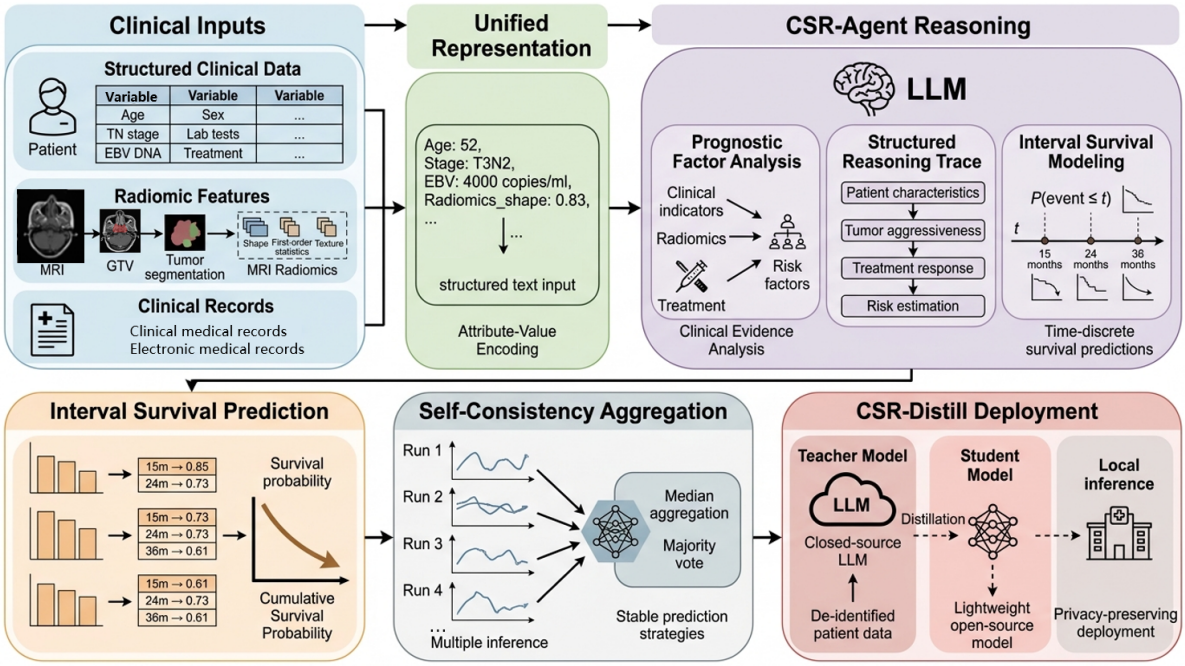

Figure 2: Overview of NPC-Agent pipeline for interval-based NPC prognosis prediction.

information into a unified textual representation. Structured clinical variables include demographic features, laboratory measurements, TN staging, and EBV DNA levels, which are represented as semantically labeled attribute–value pairs with explicit units. Radiomic features extracted from MRI images, such as shape descriptors and first-order statistics, are represented using the same textual encoding scheme, where each feature is expressed as a field–value pair whose field name explicitly encodes the feature identity, anatomical region, and temporal status.

Input Construction We consider two categories of input features: radiomic features derived from the gross tumor volume (GTV), and patients' clinical features. From patients' CET1-W image, we extract first-order statistical and shape-based radiomic features from the GTV. Post-treatment features are first selected using univariate Cox regression (Suresh et al., 2021) and the maximum relevance minimum redundancy (mRMR) criterion (Peng et al., 2005), and the corresponding features are extracted from pre-treatment images to form paired representations that characterize treatment-related imaging changes. Clinical features undergo necessary value transformations(e.g.,log transformation, median imputation.) and are selected using the same Cox regression and mRMR procedure, yielding a subset of clinical features. We convert the selected radiomic and clinical features into textual representations in "attribute-value" format and organize them into multiple semantically coherent input modules following a predefined structure and ordering. After obtaining the semantic representations of patient features, we further append task instructions and local prompts for interval-based survival prediction. The local prompt explicitly separates analysis and prediction into two consecutive subtasks, thereby guiding the LLMs to generate cumulative survival probability predictions at multiple predefined time points while ensuring temporal monotonicity.

4.2 NPC-Oriented Reasoning

Building on the unified clinical representation, CSR-Agent reformulates prognosis prediction as a structured multi-stage reasoning process rather than a direct end-to-end regression task. The model analyzes patient characteristics to derive survival probability estimates over time through clinically grounded reasoning.

First, the model examines clinical indicators and imaging-derived features to identify potential prognostic factors related to disease progression and treatment response. This process produces an intermediate reasoning trace that summarizes the key clinical evidence supporting the prediction.

Second, we adopt an interval-based survival formulation to enable temporally consistent predictions. Based on clinical studies indicating that most NPC recurrences occur within 12–18 months and

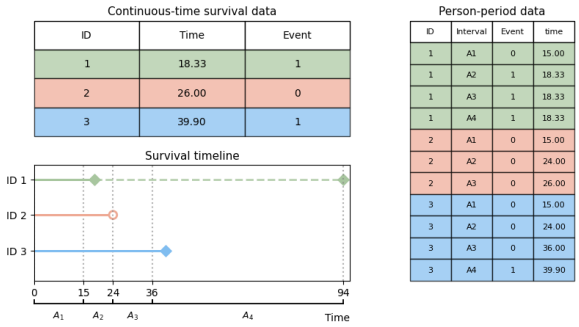

Figure 3: Person–period representation for discrete-time survival modeling

that 24-month progression-free survival (PFS24) is commonly used for risk assessment (Sun et al., 2014; Maklad et al., 2016; Chen et al., 2016; Liu, 2024), the time axis is discretized into clinically meaningful horizons at 15, 24, and 36 months, together with the end of follow-up. Continuous survival times are transformed into interval-level representations using a person–period expansion. Once an event occurs, its status is propagated to all subsequent intervals until the end of follow-up, which is consistent with the definition of cumulative event probability in survival analysis (Kleinbaum and Klein, 2012; Collett, 2015).

Third, conditioned on the reasoning trace, the model generates cumulative survival probability predictions for the predefined time points while enforcing a temporal monotonicity constraint, ensuring that survival probabilities decrease over time. This constraint reflects the inherent monotonic property of survival functions in survival analysis.

Inference Stabilization Finally, to improve inference stability, we adopt a self-consistency strategy. Multiple independent reasoning runs are performed for each patient, and the predictions are aggregated by taking the median of continuous probability estimates and majority voting for risk stratification decisions (Wang et al., 2022). This aggregation reduces stochastic variability and improves prediction robustness across cohorts.

4.3 CSR-Distill

To enable privacy-preserving deployment in clinical environments, we train a lightweight open-source model that reproduces the reasoning behavior of the closed-source LLM-driven CSR-Agent. The model is trained using distillation signals generated by the teacher model (Hinton et al., 2015; Romero, 2014). Instead of learning only from final survival probabilities, the student model is trained on structured reasoning traces produced during clinical analysis and probability generation. These traces capture how the teacher model interprets clinical evidence and derives survival predictions across multiple time horizons. By learning from these reasoning traces, the lightweight model can approximate the reasoning process of the larger model while maintaining the temporal consistency and output structure required for interval-based survival prediction.

5 Dataset for NPC Agent

We conduct NPC prognostic prediction on multiple mutually independent cohorts collected from different cohorts. Based on the available clinical follow-up annotations, disease-free survival (DFS) is used as the primary endpoint for the private multi-center cohorts, while progression-free survival (PFS) is adopted as the endpoint for the public datasets. DFS measures time to recurrence or death after treatment, whereas PFS measures time to disease progression or death (Food et al., 2018). Although the two endpoints are not identical in definition, both fall under the category of time-to-event survival prediction; therefore, we evaluate them within a unified analytical framework.

Training Data We primarily collect data from Cohort A, which serves as the main source for model training. This cohort includes 816 patients with structured clinical variables, multi-sequence MRI scans acquired pre- and post-treatment, and disease-free survival (DFS) annotations. All private data were approved by the institutional ethics committee. From the 816 patients in Cohort A, 616 patients are randomly sampled to construct the supervised training set for baseline models. No data from other centers or the public cohort are used for training.

Test Data For in-domain evaluation, we constructed a balanced test set by randomly sampling 200 instances (100 positive and 100 negative) from Cohort A, the remaining 616 patients in Cohort A were utilized for model training. To ensure rigorous evaluation, we maintained a strict separation with no overlap between the training and testing sets. To assess cross-center generalization under distribution shift, we further evaluate the trained models on two additional independent private cohorts (Cohort B: 96; Cohort C: 218) and one public cohort (Cohort D: 277), for each external cohort,

Method	Cohort A		Cohort B		Cohort C		Cohort D		**Avg**	
	C-index	AUC	C-index	AUC	C-index	AUC	C-index	AUC	C-index	AUC
Based on Traditional Methods										
DeepHit((Lee et al., 2018)	0.5001	0.5000	-	-	-	-	-	-	-	-
MTLR((Yu et al., 2011)	0.5857	0.6328	-	-	-	-	-	-	-	-
Cox((Cox, 1972)	0.6192	0.6660	-	-	-	-	-	-	-	-
RSF((Ishwaran et al., 2008)	0.6265	0.6842	-	-	-	-	-	-	-	-
Based on General LLMs										
Phi-4-mini-Instruct (Abouelenin et al., 2025)	0.5818	0.5875	0.5569	0.6044	**0.5918**	**0.6219**	0.5974	0.6290	0.5820	0.6107
Llama-3.3-8B-Instruct (Grattafiori et al., 2024)	0.5184	0.5222	0.4557	0.4600	0.5711	0.5916	0.5285	0.5003	0.5184	0.5185
Qwen2.5-7B-Instruct (Yang et al., 2024)	0.5370	0.5530	0.5199	0.5911	0.5300	0.5219	0.5178	0.5257	0.5264	0.5479
Mixtral-7B-Instruct (Jiang et al., 2024)	0.5596	0.6013	0.5459	0.5733	0.5845	0.6129	0.5670	0.5957	0.5643	0.5958
Qwen2.5-14B-Instruct (Yang et al., 2024)	0.5818	0.5875	0.5569	0.6044	**0.5918**	**0.6219**	0.5974	0.6290	0.5820	0.6107
Gemini3-Flash (Team et al., 2023)	0.5973	0.6523	**0.7077**	**0.7511**	0.5506	0.5884	0.5932	0.6044	0.6122	0.6491
Qwen3.5-27B (Qwen Team, 2026)	0.5676	0.6026	0.7031	0.7222	0.5535	0.5486	0.6125	0.6483	0.6092	0.6304
Based on CSR-Agent										
DeepSeek-V3.2 (Liu et al., 2025)	0.5744	0.5966	0.6912	0.6806	0.5408	0.5178	0.5891	0.6000	0.5989	0.5988
CSR-Agent	**0.7051**	**0.7860**	0.6877	0.7111	0.5560	0.5533	0.5590	0.5903	**0.6270**	**0.6602**
Based on CSR-Distill										
DeepSeek-R1-Distill-Qwen2.5-14B (Guo et al., 2025)	0.6031	0.6402	0.6985	0.7222	0.5284	0.5456	0.5456	0.5667	0.5939	0.6187
CSR-Distill-Agent	0.6308	0.6911	0.6490	0.6518	0.5638	0.5690	**0.6415**	**0.6823**	0.6213	0.6486

Table 1: Performance comparison across four clinical cohorts. **Bold** denotes the best performance; underlined denotes the second-best.

we constructed a balanced test set by including all available positive samples and randomly subsampling an equivalent number of negative samples to maximize the usable data size. This resulted in test sets of $n = 30$ for Cohort B, $n = 174$ for Cohort C, and $n = 86$ for Cohort D. Unlike our private cohorts, Cohort D is limited to pre-treatment multi-sequence MRI scans and a sparse set of clinical variables.

Evaluation Metrics Prognostic models are commonly evaluated along two dimensions: calibration, which assesses the agreement between predicted and observed event probabilities,and discrimination, which measures the ability to distinguish between high-risk and low-risk individuals (Steyerberg et al., 2010). In this work, we focus on evaluating predictive performance using several standard metrics, including the C-index, AUC. The C-index measures the concordance between predicted risk rankings and observed survival outcomes. Its value ranges from 0 to 1, with higher values indicating better risk ranking ability.

6 Experiments

In this section, we present the baselines, experimental setup, implementation details, main results, and ablation studies.

6.1 Experiment Setup

We compare CSR-Agent and its distilled variant against three groups of baselines: traditional survival models, general-purpose instruction-tuned LLMs, and backbone-matched LLM baselines. Traditional methods serve as clinical references, general-purpose LLMs assess whether off-the-shelf instruction-following models can perform the task directly, and backbone-matched baselines disentangle the effect of the proposed CSR framework from improvements due to stronger underlying models.

6.2 Implementation Details

We train CSR-Distill-Agent via supervised knowledge distillation. DeepSeek-V3.2 Thinking (Liu et al., 2025) is used as the teacher model to generate supervision signals. The student model is initialized from DeepSeek-R1-Distill-Qwen2.5-14B (Guo et al., 2025) and fine-tuned using supervised fine-tuning (SFT) with LoRA for parameter-efficient adaptation (Hu et al., 2022). The model is trained for five epochs during the distillation stage. The training pipeline is implemented using LlamaFactory (Zheng et al., 2024). All experiments are conducted under the PyTorch framework and executed on NVIDIA GPUs. Full distillation details are provided in Appendix.

6.3 Main Results

Table 1 presents the performance comparison across four clinical cohorts. The results highlight several key findings.

First, directly applying general-purpose LLMs to prognosis prediction yields limited performance. Although some models such as Gemini3-Flash achieve relatively strong results on individual cohorts, their performance is inconsistent across

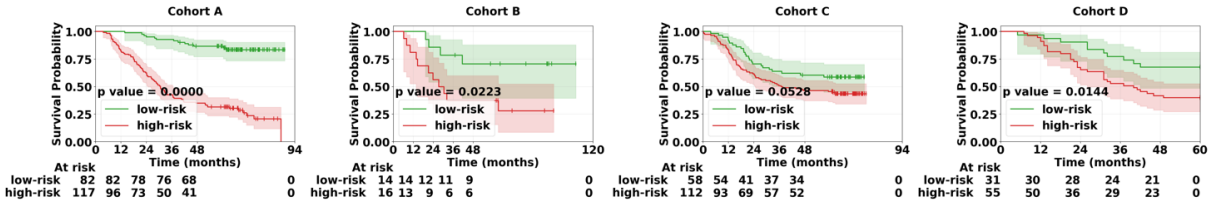

Figure 4: Kaplan–Meier survival curves of high- and low-risk groups stratified by CSR-Agent across multiple cohorts.

Shot	CSR	Cohort A		Cohort B		Cohort C		Cohort D		Avg.	
		C-index	AUC	C-index	AUC	C-index	AUC	C-index	AUC	C-index	AUC
Based on CSR-Distill-Agent											
Zero-shot	-	61.74	65.65	47.71	46.89	50.38	47.08	56.59	58.60	54.10	54.56
Zero-shot	✓	59.12	63.67	60.00	60.04	54.04	54.53	56.84	61.17	57.50	59.85
One-shot	-	60.39	64.39	45.69	42.05	54.02	54.96	59.15	59.98	54.81	55.35
One-shot	✓	59.41	64.01	**67.54**	**73.33**	55.61	56.80	59.13	61.93	60.42	64.02
Few-shots	-	62.11	64.88	43.59	38.10	54.01	52.94	59.09	60.81	54.70	54.18
Few-shots	✓	**63.08**	**69.11**	64.90	65.18	**56.38**	**56.90**	**64.15**	**68.23**	**62.13**	**64.86**

Table 2: Ablation of shot settings with and without the CSR framework across four cohorts (C-index and AUC, %).

datasets and generally inferior to specialized prediction frameworks. This indicates that naive prompting of LLMs is insufficient for reliable clinical risk prediction.

Second, CSR-Agent consistently outperforms direct LLM-based prediction across multiple evaluation settings. By introducing structured clinical reasoning and temporal survival constraints, CSR-Agent substantially improves predictive performance. For example, CSR-Agent achieves the best overall performance on Cohort A with a C-index of 0.7051 and an AUC of 0.7860, significantly surpassing all baseline methods. Averaged across cohorts, CSR-Agent also achieves the highest overall performance.

Third, the closed-source LLM-driven CSR-Agent demonstrates strong generalization ability across independent medical centers. Notably, this version of CSR-Agent achieves competitive or SOTA performance without task-specific training, while maintaining stable performance across all four cohorts. In contrast, traditional deep learning baselines are trained only on Cohort A, due to the limited feature spaceand fail to generalize to external datasets.

Finally, the distilled lightweight model maintains competitive performance while enabling privacy-preserving deployment. This demonstrates that the reasoning capability learned by the large model can be effectively transferred to a smaller open-source model, enabling accurate local inference in privacy-sensitive clinical environments.

Survival Stratification Figure 4 presents the Kaplan–Meier survival curves for the predicted high- and low-risk groups across one main-center cohort and three external cohorts. The clearest separation is observed in Cohort A ($p = 0.0000$), indicating strong stratification in the primary cohort. Among the external cohorts, B and D remain significant ($p = 0.0223$ and $p = 0.0144$), while C shows only marginal separation ($p = 0.0528$), possibly due to stronger cohort-specific heterogeneity or more pronounced feature shift relative to the main-center data. These results indicate that the proposed method achieves strong in-center performance and retains reasonable generalization ability in most external cohorts.

6.4 Ablation Study

Table 2 reports the ablation results of the CSR framework under different shot settings (Zero-shot, One-shot, and Few-shots) based on CSR-Distill-Agent. The results reveal two key observations regarding the effectiveness of our method.

Effect of CSR Across almost all configurations and cohorts, incorporating the CSR framework (w/ CSR) consistently outperforms the corresponding variants without it (w/o CSR). For example, under the few-shots setting, the CSR framework improves the average C-index from 54.70% to 62.13%, and the average AUC from 54.18% to 64.86%. This result indicates that directly prompting the model to predict survival outcomes is insufficient, and that imposing structured clinical reasoning is essential for accurate prognosis prediction.

Method	Cohort A	Cohort B	Cohort C	Cohort D
Based on Qwen2.5-14B				
	52.27	68.10	50.37	49.88
+ Distill	63.32	61.16	55.86	52.53
Based on DeepSeek-R1-Distill-Qwen2.5-14B				
	58.69	69.94	53.16	55.61
+ Distill	64.02	72.22	54.56	56.67

Table 3: Ablation Study on Knowledge Distillation (AUC, %).

In-Context Learning An additional observation concerns the interaction between the CSR framework and the number of in-context examples. Without the CSR framework, increasing the number of examples (from zero-shot to few-shots) fails to improve overall performance and may even lead to substantial degradation on external cohorts. For instance, the AUC for Cohort B drops to 38.10% under the w/o CSR few-shots setting. In contrast, when the CSR framework is applied, the model is able to effectively leverage in-context examples. The average performance steadily improves as the number of shots increases: the average AUC rises from 59.85% (zero-shot) to 64.02% (one-shot), and further reaches 64.86% under the few-shots setting. This finding suggests that the structured reasoning steps defined in the CSR framework play a critical role in enabling the model to learn from and generalize across few-shot demonstrations.

Effect of Distillation Table 3 presents the ablation results of the knowledge distillation procedure across four cohorts. Overall, distillation consistently improves predictive performance for both backbone models. For the Qwen2.5-14B backbone, applying the distillation procedure leads to noticeable improvements across all cohorts. The AUC on Cohort A increases from 52.27 to 58.31, while smaller but consistent gains are observed on Cohorts B, C, and D. These results indicate that transferring reasoning signals from the teacher model can effectively enhance the predictive capability of a weaker backbone model. A similar trend is observed for the stronger DeepSeek-R1-Distill-Qwen2.5-14B backbone. After distillation, the model achieves improved AUC on all cohorts, reaching 64.02, 72.22, 54.56, and 56.67 on Cohorts A–D, respectively. Compared with the Qwen2.5-14B backbone, the stronger initialization already provides better baseline performance, and distillation further refines its predictive ability. These results demonstrate that the proposed distillation process can effectively transfer prognostic reasoning knowledge to smaller models, consistently improving survival prediction performance across different model backbones and clinical cohorts.

7 Further Analysis

We further analyze the inference stability of the proposed framework. Since LLM-based predictions may exhibit stochastic variability across runs, we investigate aggregation strategies designed to stabilize model outputs.

Median Aggregation As shown in Table 4 in the Appendix,, the patient-level median aggregation outperforms the worst single-run performance across all cohorts, indicating improved robustness to stochastic inference variability. Across the three private cohorts (A–C), the aggregated results remain close to the best run, suggesting that stability is achieved without sacrificing predictive performance. Notably, on the external Cohort D, the median aggregation even surpasses the best single-run result for CSR-Distill-Agent, further demonstrating its effectiveness in mitigating prediction noise and improving overall ranking consistency.

Voting Aggregation As shown in Table 7 in the Appendix, majority voting yields limited and unstable gains for large-capacity models, while consistently improving performance for smaller distilled models. This suggests that its benefits are more pronounced when prediction variability is high, where voting helps mitigate stochastic fluctuations.

8 Conclusion

We propose the CSR framework for NPC survival prediction. Extensive evaluations on multi-center cohorts demonstrate that CSR achieves improved predictive performance and stronger cross-cohort generalization. The framework reformulates interval-based survival prediction as a constrained multi-stage reasoning task, combining structured semantic inputs with an explicit analysis–prediction pipeline to enforce temporal consistency under discrete-time modeling. Our results highlight the potential of applying general-purpose LLMs within a constrained and structured reasoning paradigm for complex clinical prognostic tasks. We hope that CSR can serve as a foundation for developing more robust, interpretable, and temporally consistent intelligent systems for real-world medical decision support.

9 Limitations

Our study has several limitations. First, the model is developed and validated using retrospective data, and its performance in prospective real-world clinical settings remains to be established. Second, the current model is trained with a fixed architecture and does not incorporate explicit mechanisms to handle feature inconsistency or distribution shift across cohorts, which may limit its generalization under substantial heterogeneity. Future work may explore shift-aware modeling or domain adaptation strategies, as well as reinforcement learning–based adaptive optimization with reward signals tied to cross-cohort generalization performance, to improve robustness across heterogeneous cohorts.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Kuan-Yu Chen, Ping-Han Chiang, Hsin-Rung Chou, Ting-Wei Chen, and Tien-Hao Chang. 2023. Trompt: Towards a better deep neural network for tabular data. *arXiv preprint arXiv:2305.18446*.

W.-H. Chen and 1 others. 2016. Prognostic value of plasma epstein–barr virus dna for local and regionally advanced nasopharyngeal carcinoma treated with cisplatin-based concurrent chemoradiotherapy in intensity-modulated radiotherapy era. *Medicine*, 95(5):e2642.

David Collett. 2015. *Modelling Survival Data in Medical Research*, 3rd edition. Chapman and Hall/CRC.

Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman, and Karel G.M. Moons. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod). *Circulation*, 131(2):211–219.

David R Cox. 1972. Regression models and life-tables. *Journal of the royal statistical society: Series B (methodological)*, 34(2):187–202.

Food, Drug Administration, and 1 others. 2018. Clinical trial endpoints for the approval of cancer drugs and biologics. *Guidance for industry*.

Michael F Gensheimer and Balasubramanian Narasimhan. 2019. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pages 5549–5581. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14852–14882.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. Random survival forests. *The Annals of Applied Statistics*, 2:841–860.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

David G. Kleinbaum and Mitchel Klein. 2012. *Survival Analysis: A Self-Learning Text*, 3rd edition. Springer.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, and 1 others. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.

Jia Kou, Jun-Yi Peng, Wen-Bing Lv, Chen-Fei Wu, Zi-Hang Chen, Guan-Qun Zhou, Ya-Qin Wang, Li Lin, Li-Jun Lu, and Ying Sun. 2025. A serial mri–based deep learning model to predict survival in patients with locoregionally advanced nasopharyngeal carcinoma. *Radiology: Artificial Intelligence*, 7(2):e230544.

Håvard Kvamme and Ørnulf Borgan. 2021. Continuous and discrete-time survival prediction with neural networks. *Lifetime data analysis*, 27(4):710–736.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. Dynamic knowledge distillation for pre-trained language models. *arXiv preprint arXiv:2109.11295*.

Yan Li, Jie Wang, Jieping Ye, and Chandan K Reddy. 2016. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1715–1724.

Yiqi Li, Yusheng Liao, Zhe Chen, Yanfeng Wang, and Yu Wang. 2025. Dice: Structured reasoning in llms through slm-guided chain-of-thought correction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6962–6977.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Y. Liu. 2024. Subsequent survival and loss of lifetime for patients with progression-free 24 months after treatment in nasopharyngeal carcinoma: A comprehensive nationwide population-based analysis. *International Journal of Radiation Oncology*Biology*Physics*, 120(2, Supplement):e767. ASTRO 2024: 66th Annual Meeting.

Wenbing Lv, Junyi Peng, Jiaping Hu, Yijun Lu, Zidong Zhou, Hui Xu, Kongzai Xing, Xiaodong Zhang, and Lijun Lu. 2025. Lmsst-gcn: Longitudinal mri sub-structural texture guided graph convolution network for improved progression prediction of knee osteoarthritis. *Computer Methods and Programs in Biomedicine*, 261:108600.

Wenbing Lv, Zidong Zhou, Junyi Peng, Lihong Peng, Guoyu Lin, Huiqin Wu, Hui Xu, and Lijun Lu. 2023. Functional-structural sub-region graph convolutional network (fsgcn): application to the prognosis of head and neck cancer with pet/ct imaging. *Computer Methods and Programs in Biomedicine*, 230:107341.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: short papers)*, pages 1773–1781.

A. M. Maklad and 1 others. 2016. Patterns of failure and survival in patients with nasopharyngeal carcinoma treated with intensity-modulated radiation therapy in saudi arabia. *OncoTargets and Therapy*, 9:6561–6567.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, and 1 others. 2021. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–15.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Adriana Romero. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073.

Judith D Singer and John B Willett. 1993. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2):155–195.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

X. Sun and 1 others. 2014. Long-term outcomes of intensity-modulated radiotherapy for 868 patients with nasopharyngeal carcinoma: an analysis of survival and treatment toxicities. *Radiotherapy and Oncology*, 110(3):398–403.

Zhenjie Sun, Naihao Deng, Haofei Yu, and Jiaxuan You. 2025. Table as thought: Exploring structured thoughts in llm reasoning. *arXiv preprint arXiv:2501.02152*.

Kritika Suresh, Cameron Severn, and Debashis Ghosh. 2021. Survival prediction models: an introduction to discrete-time modeling. *International Statistical Review*, 89(2):335–357.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, and 1 others. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. 2024. Model compression and efficient inference for large language models: A survey. *arXiv preprint arXiv:2402.09748*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yali Wang, Rui Gao, Taoyu Zhang, Yuyu Liu, Yiran Kong, Lijun Lu, and Wenbing Lv. 2025. Dynamics of habitat contour complexity in pre-and post-treatment mri images for survival prediction of nasopharyngeal carcinoma. In *Second International Conference on Future of Medicine and Biological Information Engineering (MBIE 2025)*, volume 13999, pages 55–62. SPIE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. 2024. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65.

Hui Xu, Wenbing Lv, Hao Zhang, Qingyu Yuan, Quanshi Wang, Yuankui Wu, and Lijun Lu. 2023. Multimodality radiomics analysis based on [18f] fdg pet/ct imaging and multisequence mri: application to nasopharyngeal carcinoma prognosis. *European Radiology*, 33(10):6677–6688.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Chun-Nam Yu, Russell Greiner, Hsien-Chang Lin, and Vickie Baracos. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*.

Daniel Zeiberg, Tejas Prahlad, Brahmajee K Nallamothu, Theodore J Iwashyna, Jenna Wiens, and

Michael W Sjoding. 2019. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PloS one*, 14(3):e0214465.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, and 1 others. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.

A Appendix

A.1 Distillation Details

The distillation dataset is constructed from the training split of Cohort A, which is treated as the main-center cohort in our experiments. The split contains 616 patients in total, and all of them are included in the distillation stage without any additional sampling or filtering. For the **w/ CSR** setting, we use trace distillation, where the teacher's reasoning traces are used as supervision signals. For the **w/o CSR** setting, we use label distillation, where only the teacher's final prognosis predictions are used as supervision signals. The corresponding configurations are summarized in Table 5.

A.2 Inference Details

For inference, we use sampling-based decoding with temperature 0.7, top-p 0.8, top-k 20, min-p 0.1, and repetition penalty 1.1. The maximum number of generated tokens is set to 16,384.The corresponding configurations are summarized in Table 6.

A.3 Case Study

We present an end-to-end case study to illustrate the complete processing pipeline of a single patient under the CSR framework.

Structured clinical and radiomic features are first converted into a unified semantic input template (see Appendix, Fig. 5–6). The model then performs two-stage reasoning (analysis $\rightarrow$ prediction) to generate interval-based cumulative survival probabilities. To account for stochastic variability, five independent inference runs are conducted, and the outputs are aggregated via median estimation for continuous probabilities and majority voting for risk labels. The aggregated survival probabilities are used for quantitative evaluation, while the aggregated risk labels are applied to Kaplan–Meier stratification and log-rank survival analysis (example outputs are shown in Appendix, Fig. 7).

Method	Cohort A			Cohort B			Cohort C			Cohort D		
	Best	Worst	Median	Best	Worst	Median	Best	Worst	Median	Best	Worst	Median
CSR-Agent (Few-shots)	0.7064	0.5904	0.7051	0.7277	0.6600	0.6877	0.5655	0.5512	0.5560	0.5709	0.5421	0.5590
CSR-Distill-Agent (Few-shots)*	0.6331	0.5863	0.6308	0.7490	0.6196	0.6490	0.5753	0.5325	0.5638	0.6091	0.5681	**0.6415**

Table 4: Effect of median aggregation under five independent inference runs. We report the best and worst single-run C-index and compare them with the patient-level median aggregated result.

```
You are Dr. Chen, a board-certified clinical oncologist specializing in nasopharyngeal carcinoma (NPC) prognosis.
Task 1:
    Inside <think>...</think>, generate clear, concise, step-by-step clinical reasoning for the following patient.
    You must integrate *both clinical features and radiomic features* in your reasoning.
    Required focus:
            ...
    Output constraints:
            ...
Task 2:
     After completing the reasoning inside <think>, produce the final cumulative DFS probability for each follow-up
interval provided in the input, based solely on your reasoning.
            ...
     Output requirements:
             1. Output one <answer>...</answer> line per interval.
            ...

       ### Few-shot examples (paired by interval)
                Meaning: Case A = shorter DFS time (poorer outcome);
                         Case B = longer DFS time (better outcome).
        Case A: This 51.0-year-old female patient had a diagnosis ...
            <think>...</think>,
            <answer>...</answer>;

        Case B: This 67.0-year-old female patient had a diagnosis ...
            <think>...</think>,
            <answer>...</answer>;
            ...
       End of Example.
       Now apply the same reasoning process and output format to the next patient below.
       Begin processing the following patient data:
```

Figure 5: Prompt Block1 used in CSR-Agent.

Item	**Configuration**
Teacher model	DeepSeek-V3.2 Thinking, T=0.7
Distillation data	616 patients (Cohort A training set)
Student model	DeepSeek-R1-Distill-Qwen2.5-14B
LoRA	rank=8, alpha=16, all linear layers
Training	5 epochs, LR=1e-4, cosine decay, warmup=0.1
Optimization	BF16, DeepSpeed ZeRO-3
Hardware	2 × RTX 5090, 3 hours
Seed	42

Table 5: Distillation configuration for CSR-Distill-Agent.

Item	Configuration
Decoding strategy	Sampling-based decoding (`do_sample=True`)
Temperature	0.7
Top-p	0.8
Top-k	20
Min-p	0.1
Repetition penalty	1.1
Max new tokens	16,384

Table 6: Inference configuration used in CSR-Agent.

Method	Setting	Cohort A	Cohort B	Cohort C	Cohort D
CSR-Agent(Few-shots)	Inconsistency Rate	37.18	16.67	47.65	10.47
	Mean ACC (Inconsistent)	55.41	48.00	51.85	44.44
	Multi-Vote ACC (Inconsistent)	66.22 **(+10.81)**	40.00 (-8.00)	51.85 **(0.00)**	44.44 **(0.00)**
CSR-Distill-Agent*(Few-shots)	Inconsistency Rate	51.03	53.85	56.55	58.33
	Mean ACC (Inconsistent)	55.96	48.57	52.84	53.47
	Multi-Vote ACC (Inconsistent)	63.64 **(+7.68)**	50.00 **(+1.43)**	56.84 **(+4.00)**	61.22 **(+7.75)**

Table 7: Effect of majority voting under five independent inference runs, we report the five-run prediction inconsistency rate and compare the mean single-run classification ACC with the majority-vote ACC on inconsistent samples.

This 51.0-year-old female patient had a diagnosis of Nasopharyngeal carcinoma (NPC). The patient previously smoked cigarettes, and has never consumed alcohol. The patient has never had a family history of cancer ...
Building on this, the study also extracted pre- and post-treatment GTV (gross tumor volume) radiomic features to further characterize the tumor phenotype and assess treatment response. The patient's 10 groups of GTV radiomic features — including first-order statistical features and 3D shape-based features, each containing both pre-treatment and post-treatment values — are shown below:

Pre-original-shape-Least Axis Length-GTV: 25.7467 mm
Post-original-shape-Least Axis Length-GTV: 19.6157 mm

...
Based on the above clinical features and radiomic features from gross tumor volume, answer:
Please predict the cumulative Disease-Free Survival probability at 15 months.
Please predict the cumulative Disease-Free Survival probability at 24 months.
Please predict the cumulative Disease-Free Survival probability at 36 months.
Please predict the cumulative Disease-Free Survival probability at 94 months.

Figure 6: Prompt Block2 used in CSR-Agent.

<think>
1. Compare pre- vs post-treatment radiomic values: Entropy increased slightly (0.04→0.0602), uniformity decreased slightly (0.9892→0.9836). Shape features: Least Axis Length minimal decrease (21.72→21.54 mm),...
2. Interpret direction of change and Δ: The minimal volume reduction (<3%) and increase in some dimensions indicate a very poor treatment response – essentially stable disease. The slight changes in entropy/uniformity are not clinically meaningful. This radiomic pattern is strongly unfavorable.
3. Impact on DFS: - Positive clinical factors: young age (33), female, low-risk histology, normal HGB/ALB/LDH, undetectable EBV DNA pre- and post-induction (strong favorable). - Negative clinical factors: previous smoking/alcohol (mild negative), elevated CRP (13 mg/L), advanced stage T3N2, lack of concurrent chemotherapy, extremely low cumulative cisplatin dose (inadequate systemic therapy). - Radiomic factors: minimal shrinkage and increase in some axes → strong negative impact; entropy/uniformity changes negligible.
4. Conflicts & weighting: The favorable EBV DNA status is outweighed by the high-risk local factors (advanced stage, poor radiomic response, absence of concurrent chemo). Dominant negative factors are the radiomic evidence of treatment resistance and inadequate systemic therapy, which predict high risk of early recurrence.
5. Interval dependency: NPC recurrences peak within the first 2–3 years. Given the high early risk, cumulative DFS will decline steeply in the first 24–36 months, then plateau slowly after 5 years. Probabilities must satisfy 15m > 24m > 36m > 94m.
6. DFS trend estimation: The cumulative DFS probability is expected to be moderate at 15 months, with a substantial drop by 24–36 months, reaching a low plateau by 94 months. Summary: Integrating clinical and radiomic features, the patient has a moderate-to-high risk profile, leading to estimated cumulative DFS probabilities of 0.8000 at 15 months, 0.6500 at 24 months, 0.5500 at 36 months, and 0.4500 at 94 months.
</think>

<answer>"By 15 months, the patient's cumulative Disease-Free Survival probability is **0.8000**."</answer>
<answer>"By 24 months, the patient's cumulative Disease-Free Survival probability is **0.6500**."</answer>
<answer>"By 36 months, the patient's cumulative Disease-Free Survival probability is **0.5500**."</answer>
<answer>"By 94 months, the patient's cumulative Disease-Free Survival probability is **0.4500**."</answer>

Figure 7: CSR-Agent Output with reasoning and predictions.