Multimodal Survival Analysis with Locally Deployable Large Language Models

Moritz Gögl
University of Oxford
moritz.gogl@keble.ox.ac.uk

Peter Watkinson
University of Oxford
peter.watkinson@ndcn.ox.ac.uk

Christopher Yau
University of Oxford
christopher.yau@wrh.ox.ac.uk

Abstract

We study multimodal survival analysis integrating clinical text, tabular covariates, and genomic profiles using locally deployable LLMs. As many institutions face tight computational and privacy constraints, this setting motivates lightweight, on-premises models. Our approach jointly estimates calibrated survival probabilities and generates concise, evidence-grounded prognosis text via teacher–student distillation and principled multimodal fusion. On a TCGA cohort, it outperforms baselines, avoids reliance on cloud services and associated privacy concerns, and reduces the risk of hallucinated or miscalibrated estimates from base LLMs.

1 Introduction

Survival analysis estimates the probability of an event over time and is central to medical decision-making (e.g., forecasting mortality or disease progression). Classical models operate on structured covariates (e.g., age, sex, genomic data), while clinical practice also generates rich unstructured data, such as clinical reports. Recent large language models (LLMs) can reason over such text and produce human-readable assessments, but cloud-hosted models raise privacy concerns and heavyweight local deployments are impractical for many institutions [15]. Moreover, base LLMs are not calibrated for survival prediction as they are not trained on raw survival data; they typically recall published summary statistics–and may hallucinate–rather than producing data-grounded estimates [22].

We present a unified, locally deployable multimodal survival framework that pairs a compact causal LLM with structured covariates and gene expression. Our model jointly produces calibrated survival curves and concise prognosis explanations. A teacher–student pipeline first queries a large teacher LLM for numeric survival probabilities at fixed horizons and a brief assessment; the student then learns from both the teacher's verbalized reasoning and the observed survival outcomes. The architecture supports either a discrete-time hazards model or a Cox proportional hazards (CoxPH) [2] model, and fuses modalities by concatenation or via separate gated heads. Compared to prior multimodal survival models [23, 17], our approach couples survival estimation with concise explanations while remaining lightweight and locally deployable.

Contributions. Our work makes three contributions: (1) a calibrated locally deployable multimodal survival framework that couples a compact causal LLM (1.5B parameters) with covariates and gene expression, supporting both discrete and CoxPH heads with flexible fusion; (2) a teacher–student fine-tuning scheme with a single forward-pass text objective that distills numeric survival probabilities and rationales; and (3) an empirical evaluation on a TCGA cohort [21] showing improved performance over baselines, alongside concise, verbalized prognosis explanations.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

2 Related Work

Methods in survival analysis range from parametric and semiparametric models to nonparametric estimators, with modern ensemble and deep-learning approaches that learn flexible risk and event-time structures. Cox models [2] remain standard but assume a log-linear covariate effect and proportional hazards, while Random Survival Forests [8] relax such assumptions. Deep learning introduced end-to-end models that infer complex survival patterns: DeepSurv [11] replaces the Cox linear predictor with a neural network, DeepHit [14] models discrete hazards and competing risks, and Neural Survival Clustering [10] learns mixture structure in survival data. Leveraging text for survival prediction emerged with BERTSurv [23], which uses transformer embeddings of clinical notes to improve concordance over tabular baselines, employing a BERT-based backbone [4] (e.g., ClinicalBERT [7]). A recent survey by Jeanselme et al. [9] reviews language-model strategies for survival, covering direct prompting, feature extraction, and fine-tuning pipelines, and highlights open issues around censoring and evaluation protocols. Foundation models have also been explored for multimodal survival: Song et al. [17] show that zero-shot embeddings from foundation models can be combined with classical survival models to yield gains over unimodal baselines, and discuss risks of hallucination in text summarization. Complementary work applies LLMs directly to pathology reports for cancer type, stage, and prognosis assessment [16], focusing on text-only predictions without calibrated survival curves. Moreover, our framework draws motivation from the approach introduced in [18], which combines calibrated hidden-state and verbalized signals in the context of guided deferral systems.

3 Methods

3.1 Problem Setup and Overview

We consider right-censored survival data consisting of triples $x_i = (x_i^{\text{path}}, x_i^{\text{cov}}, x_i^{\text{ge}})$ and outcomes (t_i, e_i) for samples $i = 1, \ldots, N$, where t_i denotes the observed follow-up time and $e_i \in \{0, 1\}$ is the event indicator ($e_i = 1$ indicating death, $e_i = 0$ indicating censoring). Here, x_i^{path} is a free-text pathology report; $x_i^{\text{cov}} \in \mathbb{R}^{d_c}$ are tabular covariates (e.g., age, sex, cancer type); and $x_i^{\text{ge}} \in \mathbb{R}^{d_g}$ is a high-dimensional gene-expression vector. For the text channel, we form a combined input x_i^{text} by appending a formatted patient-info string to the report. We estimate the conditional survival distribution $S(t \mid x)$ in two complementary ways (Fig. 1). First, a hidden-state pathway encodes the text with a compact causal LLM whose representation feeds a survival head (either discrete-time hazards or CoxPH). Tabular covariates and gene-expression latents are fused via early concatenation or late fusion with gated heads. Second, a verbalized pathway has the LLM generate an explicit survival probability together with a concise rationale, which we map to a full survival curve.

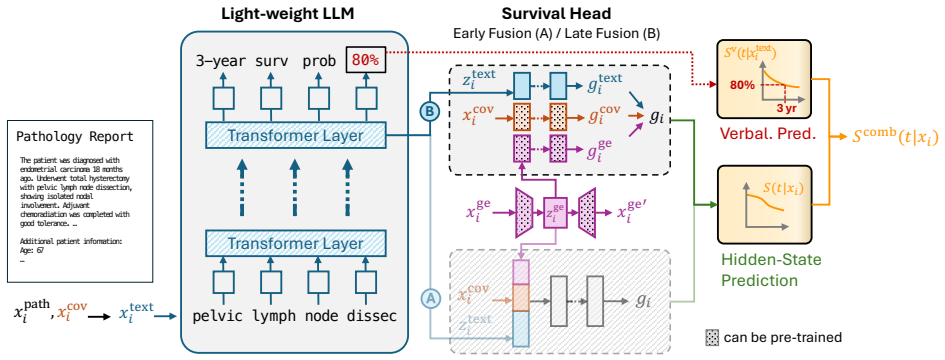

Figure 1: Overview of the proposed framework. A compact causal LLM encodes the input text into hidden embeddings used by a survival model (here, CoxPH) and produces verbalized survival estimates with explanatory text. Covariates and gene expression are fused either early (A) or late (B).

3.2 Hidden-State Survival Predictions

We use hidden representations from a compact causal LLM together with covariates and gene expression to produce well-calibrated survival estimates:

Text encoder and pooling. A causal LLM encodes the tokenized pathology report into hidden states $H \in \mathbb{R}^{L\times d}$. We form a fixed-size embedding z^{text} via simple self-attention pooling, which aggregates token representations by similarity and averages them into a sequence-level vector:

$$A = \text{softmax}(HH^\top), \quad \tilde{H} = AH, \quad z^{\text{text}} = \frac{1}{L}\sum_{\ell=1}^{L} \tilde{H}_{\ell:}. \tag{1}$$

Gene-expression autoencoder. High-dimensional gene expression (GE) is compressed with an autoencoder $\text{Dec}(\text{Enc}(x_i^{\text{ge}}))$, yielding a latent $z_i^{\text{ge}} = \text{Enc}(x_i^{\text{ge}})$. The reconstruction objective $\mathcal{L}_{\text{AE}} = \mathbb{E}_i[\|\text{Dec}(\text{Enc}(x_i^{\text{ge}})) - x_i^{\text{ge}}\|_2^2]/d_{\text{g}}$ regularizes the latent while preserving survival-relevant signal [19].

Survival models. We support two survival network heads: (i) a *discrete-time hazards* model [6, 12], which outputs logits $o \in \mathbb{R}^B$ over B time bins (hazards $h_b = \sigma(o_b)$, survival $S(t_b) = \prod_{k\le b}(1-h_k)$) and is trained with a masked Bernoulli objective; and (ii) *CoxPH* [11], which outputs a log-risk score g and is optimized via the negative partial log-likelihood. The corresponding training objectives are

$$\mathcal{L}_{\text{disc}} = \frac{\sum_{i,b} a_{ib}\,\text{BCE}(h_{ib}, y_{ib})}{\sum_{i,b} a_{ib}} \qquad \mathcal{L}_{\text{cox}} = -\frac{1}{\sum_i e_i}\sum_{i:e_i=1}\Big(g_i - \log\sum_{j:t_j\ge t_i} e^{g_j}\Big). \tag{2}$$

Here, $y_{ib} = \mathbb{1}\{e_i = 1,\ t_{b-1} < t_i \le t_b\}$ indicates an event for individual i in bin b, and $a_{ib} = \mathbb{1}\{\text{at risk at the start of } b\}$ masks at-risk samples that have not been censored or died by time bin b.

Fusion strategies. We consider two regimes to integrate text, covariates, and gene expression [5]:

(A) Early fusion concatenates $z = [z^{\text{text}}; x^{\text{cov}}; z^{\text{ge}}]$ and feeds a single head f, enabling rich cross-modal interactions at the cost of tighter coupling.

(B) Late fusion learns modality-specific heads $f^{\text{text}}, f^{\text{cov}}, f^{\text{ge}}$ and combines their outputs with learned gates $\gamma^{\text{text}}, \gamma^{\text{cov}}, \gamma^{\text{ge}}$ (see Appendix C.2). This enables separate *pre-training* of f^{cov} and f^{ge} without being constrained by the larger memory footprint of end-to-end LLM fine-tuning.

3.3 Verbalized Survival Prediction and Assessment

Inspired by an approach introduced for instruction-tuned LLMs in guided deferral systems [18], we additionally use the generative capabilities of the same compact causal LLM to produce a concise prognosis explanation and an explicit 3-year survival probability statement.

Teacher–student distillation. As illustrated in Fig. 2, we first query a larger *teacher* LLM (here: DeepSeek-R1 Distill Qwen-32B [3]) offline with two prompts: (1) a sequence of numeric-only instructions to return survival probabilities at 1/3/5 years conditioned on x^{text}; and (2) an explanation prompt conditioned on x^{text} and the rounded 3-year survival probability, `TEACHER_PROB`, computed from an exponential fit of the extracted numeric predictions (see Appendix C.1). The generated teacher explanation and probability estimate are then used to construct the target sequence: `[TEACHER_EXPLANATION] «VPROB»\n\n The estimated 3-year survival probability is: [TEACHER_PROB]%.«END_VPROB»`, which is learned by the *student* model (here: DeepSeek-R1 Distill Qwen-1.5B [3]) during training, conditioned on x^{text}.

For our use case, the verbalized prediction sentence–delimited by `«VPROB»` and `«END_VPROB»`–and, within it, the numeric probability, `TEACHER_PROB`, are the most critical parts of the assessment. We therefore upweight the cross-entropy on tokens in this span and on the numeric substring using weights w and w^{num}, yielding the following loss:

$$\mathcal{L}_{\text{text}} \;=\; \mathcal{L}_{\text{text}}^{\text{full}} \;+\; (w-1)\,\mathcal{L}_{\text{text}}^{\text{vprob}} \;+\; (w^{\text{num}}-1)\,\mathcal{L}_{\text{text}}^{\text{num}}. \tag{3}$$

Calibration correction By default, all samples contribute to the text loss. However, since teacher predictions are obtained by prompting an LLM rather than learned directly from observed survival outcomes, they may be miscalibrated. Following a similar idea to [18], we optionally mask out text loss contributions for samples whose teacher 3-year survival estimate, `TEACHER_PROB`, is inconsistent with the observed status at the assessment horizon: (i) the event occurred before 3 years yet the teacher assigns a high survival probability (>50%), or (ii) the individual is known to be alive/at-risk at 3 years yet the teacher assigns a low survival probability (<50%). All samples remain in the dataset

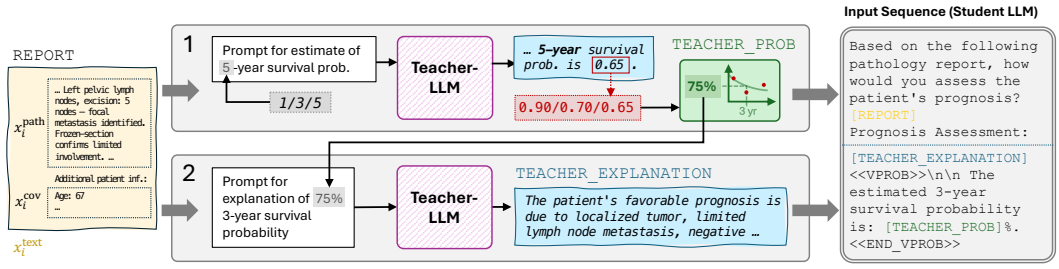

Figure 2: Teacher pipeline for constructing the student target. (1) Numeric prompting at 1/3/5 years from the input text; extract survival probabilities, and fit a parametric curve (exponential by default) to obtain the 3-year value. (2) Explanation prompting conditioned on the text and the rounded 3-year percentage; the explanation plus a marked probability sentence form the student's training target.

and fully contribute to survival objectives; only language-model targets are masked to reduce the influence of potentially miscalibrated supervision in text generation.

3.4 Objective and Optimization

We jointly optimize the survival model, gene-expression reconstruction, and text generation, reweighting the latter two with coefficients α and β. The total loss is then given by

$$\mathcal{L} = \mathcal{L}_{\text{surv}} + \alpha\,\mathcal{L}_{\text{AE}} + \beta\,\mathcal{L}_{\text{text}}, \tag{4}$$

where $\mathcal{L}_{\text{surv}} \in \{\mathcal{L}_{\text{disc}}, \mathcal{L}_{\text{cox}}\}$ depending on the selected survival model. At test time, we compute the survival function $S(t|x_i)$ (probability of surviving beyond t) from the survival-head outputs, and fit an exponential curve $S^{\text{v}}(t|x_i^{\text{text}})$ from the verbalized 3-year survival probability. We then form a convex combination $S^{\text{comb}}(t|x_i) = (1-\lambda)\,S(t|x_i) + \lambda\,S^{\text{v}}(t|x_i^{\text{text}})$, with $\lambda \in [0,1]$ selected on the validation set to maximize concordance [18].[1]

4 Experiments

4.1 Experimental Setup

We evaluated on a TCGA-derived cohort [21] of 8,902 samples comprising pathology reports, tabular covariates, and gene expression; details are provided in Appendix B. Across both survival models (discrete and CoxPH), we ablate fusion strategy (early vs. late), optional pre-training of covariate- and gene-expression-specific survival heads under increased batch size, calibration correction (CC), and a variant conditioned only on x^{text}. Baselines include BERTSurv [23] and unimodal experiments on covariates and gene expression under both survival models. We additionally report our teacher's verbalized prediction performance, deriving survival functions via exponential fitting. We report time-dependent concordance (C^{td}) and integrated Brier score (IBS); definitions are given in Appendix C.3.

4.2 Results

Table 1 summarizes performance across configurations; our standard setting (late fusion, no pre-training, no calibration correction) is highlighted in gray. Across settings, hidden-state and combined predictions outperformed baselines, confirming gains from multimodal fusion; hidden-state predictions were consistently stronger than verbalized ones, and blending the two yielded modest, consistent improvements in C^{td} and IBS. The teacher achieved substantially better verbalized performance, reflecting its larger model capacity. Moreover, its prediction was obtained via three separate prompts, making it more robust. Late fusion generally improved hidden-state and combined performance; a variant conditioned only on x^{text} underperformed, consistent with the strong predictive signal in gene expression observed in baselines. Calibration correction left CoxPH unchanged but markedly improved the discrete model's performance, yielding the best overall discrimination. Pre-training modality-specific heads with larger batches to overcome memory constraints under end-to-end LLM

[1] If no probability can be extracted, we set the combined prediction to the hidden-state curve $S^{\text{comb}}(t|x_i) = S(t|x_i)$, and mean-impute $S^{\text{v}}(t|x_i^{\text{text}}) = \overline{S^{\text{v}}(t|x^{\text{text}})}$ across test samples for verbalized-only evaluation.

Table 1: Performance across configurations and baselines.

Model/Configuration								Hidden-state		Verbalized		Combined	
Name	Text*	Cov	GE	Survival	Fusion	Pretrain	CC	$C^{\text{td}}\uparrow$	IBS$\downarrow$	$C^{\text{td}}\uparrow$	IBS$\downarrow$	$C^{\text{td}}\uparrow$	IBS$\downarrow$
BERTSurv	✓	✓	✓	CoxPH	Early	—	—	0.691	0.150	—	—	—	—
Cov-only	✗	✓	✗	Discrete	—	—	—	0.665	0.149	—	—	—	—
	✗	✓	✗	CoxPH	—	—	—	0.668	0.149	—	—	—	—
GE-only	✗	✗	✓	Discrete	—	—	—	0.734	0.139	—	—	—	—
	✗	✗	✓	CoxPH	—	—	—	0.751	0.135	—	—	—	—
Teacher	✓	✗	✗	—	—	—	—	—	—	0.746	0.141	—	—
Ours	✓	✓	✓	Discrete	Late	✗	✗	0.765	0.138	0.626	0.164	0.766	0.135
	✓	✓	✓	Discrete	Late	✓	✗	0.740	0.149	0.628	0.165	0.745	0.145
	✓	✓	✓	Discrete	Late	✗	✓	**0.774**	0.135	0.613	0.167	**0.778**	0.130
	✓	✗	✗	Discrete	—	✗	✗	0.673	0.152	0.637	**0.159**	0.697	0.144
	✓	✓	✓	Discrete	Early	✗	✗	0.741	0.141	**0.648**	0.166	0.744	0.139
	✓	✓	✓	CoxPH	Late	✗	✗	0.773	**0.126**	0.575	0.165	0.775	**0.126**
	✓	✓	✓	CoxPH	Late	✓	✗	0.743	0.144	0.625	0.165	0.746	0.141
	✓	✓	✓	CoxPH	Late	✗	✓	0.770	0.127	0.600	0.164	0.773	0.128
	✓	✗	✗	CoxPH	—	✗	✗	0.729	0.141	0.621	0.163	0.733	0.140
	✓	✓	✓	CoxPH	Early	✗	✗	0.753	0.135	0.611	0.168	0.755	0.134

*For teacher and student models, this comprises x^{path} and verbalized x^{cov}.

training yielded no improvement. Among baselines, CoxPH on gene expression performed best and exceeded BERTSurv, even though BERTSurv leverages information from all modalities.

Qualitative analysis. We show a positive example of a generated assessment from our model (discrete-time survival model, standard configuration) alongside an excerpt of the corresponding raw pathology report (Fig. 3). Consistent evidence spans are highlighted in the same color. Beyond such positive cases, we observe that very long reports can challenge the model: textual coherence may drop and the stated 3-year probability may occasionally be omitted. Additional negative and edge-case examples illustrating these patterns are presented in Appendix A.

Raw Pathology Report (excerpt).

```
[...] The tumor cells. demonstrate only
mild atypia. No mitotic figures are
seen [...]. There is no microvascular
proliferation or necrosis. [...]
Scattered MIB-1 reactive cells are
present [...][;] a labeling index of 2.5%
is calculated, consistent with the
low grade. histologic features.
Addendum Diagnosis: Well differentiated
astrocytoma (who grade II). [...]
```

Generated Assessment (Student LLM).

```
The favorable prognosis is due to the
well-differentiated nature of the tumor
(grade II), the absence of mitotic
activity, and the lack of microvascular
or necrotic features, indicating a
low-grade glioma.

The estimated 3-year survival probability
is: 90%.
```

Figure 3: Qualitative example: original report (left) and generated assessment (right). Consistent evidence spans are highlighted in the same color.

5 Conclusion

We presented a unified, locally deployable multimodal survival framework that fuses unstructured text (e.g., clinical reports), tabular covariates, and gene expression on a compact LLM backbone with comparatively low hardware requirements. The approach jointly learns calibrated survival estimates and concise, verbalized assessments via a teacher–student scheme. We explored key design choices, including different underlying survival models and fusion strategies. Ongoing work will condition the language model itself on genomics to surface patient-specific, verbalized risk factors (e.g., salient genes or mutations), enabling more trustworthy human–AI collaboration in clinical workflows.

Acknowledgments

MG is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). CY is supported by a UKRI Turing AI Acceleration Fellowship (EP/V023233/1) and received additional funding through EPSRC grant EP/Y018192/1. The results shown here are in whole or part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`.

References

[1] N. E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45, Apr. 1975.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2):187–202, Jan. 1972.

[3] DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[5] K. Gadzicki, R. Khamsehashari, and C. Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, 2020.

[6] M. F. Gensheimer and B. Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, Jan. 2019.

[7] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019.

[8] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), Sept. 2008.

[9] V. Jeanselme, N. Agarwal, and C. Wang. Review and evaluation of language models for survival analysis. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[10] V. Jeanselme, B. Tom, and J. Barrett. Neural survival clustering: Non-parametric mixture of neural networks for survival clustering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 92–102. PMLR, 07–08 Apr 2022.

[11] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), Feb. 2018.

[12] H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, Oct. 2021.

[13] H. Kvamme, Ørnulf Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

[14] C. Lee, W. Zame, J. Yoon, and M. Van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[15] P. Mistry. Infrastructure for innovation: getting the NHS and social care ready for AI | The King's Fund. `https://www.kingsfund.org.uk/insight-and-analysis/long-reads/infrastructure-nhs-social-care-ai`, 2025. [Accessed 04-09-2025].

[16] R. Saluja, J. Rosenthal, A. Windon, Y. Artzi, D. J. Pisapia, B. L. Liechty, and M. R. Sabuncu. Cancer type, stage and prognosis assessment from pathology reports using llms. *Scientific Reports*, 15(1), July 2025.

[17] S. Song, M. Borjigin-Wang, I. Madejski, and R. L. Grossman. Multimodal survival modeling in the age of foundation models, 2025.

[18] J. Strong, Q. Men, and J. A. Noble. Trustworthy and practical ai for healthcare: A guided deferral system with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28413–28421, Apr. 2025.

[19] L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Medical Informatics and Decision Making*, 20(1), Sept. 2020.

[20] W. Weibull. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18(3):293–297, Sept. 1951.

[21] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, Sept. 2013.

[22] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Y. Zhao, Q. Hong, X. Zhang, Y. Deng, Y. Wang, and L. Petzold. Bertsurv: Bert-based survival models for predicting outcomes of trauma patients, 2021.

Appendix

A Additional Qualitative Analysis of Generated Assessments

While the main paper presents a positive example of a generated assessment (see Fig. 3), here we include additional negative examples to illustrate common failure modes. We find that the model struggles particularly with longer pathology reports, where fluency can degrade and the numeric prediction may become less reliable. In Fig. A.1, the explanation remains sensible and grounded in the report but the 3-year probability is over-optimistic. Fig. A.2 shows two negative cases with poor English fluency; the right example additionally fails to produce an explicit verbalized probability.

Raw Pathology Report (excerpt).

Laterality:Mucosa is ulcerated and necrotic. [...] Specimen type: Gastrectomy. Tumor site: Antrum. Tumor size: 5 x3x1.5cm. Tumor features: Ulcerated. Histologic type: Adenocarcinoma, signet ring type. Histologic grade: Poorly differentiated. Tumor extent: Serosa (visceral peritoneum). Lymph nodes: 11/11 positive for metastasis (Regional 11/11). [...]

Generated Assessment (Student LLM).

The prognosis is poor due to the presence of poorly differentiated tumor cells, which are aggressive and prone to metastasis. Additionally, the large tumor size (5 cm) and ulceration further contribute to a lower survival probability.

The estimated 3-year survival probability is: 65%

Figure A.1: Additional qualitative example: original report (left) and student-generated assessment (right). Consistent evidence spans are highlighted with the same color; missing but clinically relevant information is framed in orange. The explanation is reasonable and faithful to the report, but the 3-year survival probability (65%) is clearly over-optimistic.

Generated Assessment (Student LLM).

DE serous carcinoma.

The patient's age is 50 years old.

The estimated 3-year survival probability is: The prognosis is poor due to the presence of metastatic disease in multiple sites, including the uterus, fallopian tubes, peritoneum, and lymph nodes, which are all indicators of advanced disease and reduced survival rates.

The 3年 survival probability estimate is: approximately 40%.

Generated Assessment (Student LLM).

right side of the serosa, there is a serosal margin of 1 cm, which is smooth, and there is no evidence of serosal invasion.

The favorable prognostic factors include: low-grade serous adenocarcinoma, absence of myo-invasion, no lymph node metastasis, and no distant metastasis.

Figure A.2: Two negative examples of student-generated assessments. Both outputs exhibit degraded English fluency; the right example further fails to provide an explicit verbalized probability.

B Data and Preprocessing

The Cancer Genome Atlas (TCGA) [21] is a large multi-cancer resource providing harmonized clinical, pathology, and genomics data. We assembled a cohort pairing survival outcomes with three inputs per sample: an unstructured pathology report, structured covariates, and gene-expression profiles. Survival times were administratively censored after 5 years. Covariates were processed following a data preparation strategy inspired by [9].[2] We included age at initial pathologic diagnosis,

[2] https://github.com/Jeanselme/LLM-For-Survival-Analysis

sex, race, AJCC pathologic tumor stage (I, II, III), and cancer type. Age and stage were min–max scaled. Sex and race were encoded as binary indicators. Cancer type was grouped into families (gastrointestinal, gynecological, genitourinary, respiratory, skin, brain, and other) and represented via one-hot indicators. Gene-expression profiles comprised 20,531 genes; within-gene missing entries were imputed to 0. Samples with other missing critical information were excluded (survival outcome, clinical report, gene expression, or covariates). The remaining 8,902 samples were split into training/validation/test sets in proportions 70/10/20%.

C Implementation Details

C.1 Details on Teacher Pipeline and Student Input Sequence Construction

We construct the student's training target with a two-step teacher pipeline (Fig. 2). First, a larger teacher LLM is queried in three prompts with numeric-only instructions to return survival probabilities at 1, 3, and 5 years (same horizons as in [9]) from the input text consisting of the pathology report with a short patient-information snippet. We extract the numeric probabilities using regular expressions, interpolate any missing probabilities (if all three probabilities are missing, we impute them using the corresponding means), and then fit a simple parametric survival curve. This multi-horizon fitting grounds the 3-year estimate in three independent numeric prompts rather than a single query, improving robustness to prompt variance. We explored Weibull [20], log-logistic, spline, and exponential curves and found that a simple exponential fit performed best in terms of IBS and C^{td}. We therefore used the exponential model throughout to estimate the 3-year survival probability for the second step, in which the teacher is prompted to generate a concise prognosis explanation conditioned on the same text and the 3-year survival percentage (rounded to the nearest 5%). The student model learns a target sequence that concatenates the explanation with a marked sentence verbalizing the 3-year probability, delimited by «VPROB» and «END_VPROB», enabling span- and number-weighted language-model losses during training.

C.2 Late Fusion Blending

We present the precise late fusion blending used for discrete-time and CoxPH heads. For discrete heads, the learnable modality gates are per–time-bin vectors $\gamma^{\text{text}}, \gamma^{\text{cov}}, \gamma^{\text{ge}} \in [0,1]^B$; for CoxPH, the modality gates are scalars $\gamma^{\text{text}}, \gamma^{\text{cov}}, \gamma^{\text{ge}} \in [0,1]$. The discrete logits o and CoxPH scores g are then given by

$$\begin{aligned} o &= (1-\gamma^{\text{ge}})\left[(1-\gamma^{\text{cov}})\, o^{\text{text}} + \gamma^{\text{cov}}\, o^{\text{cov}}\right] + \gamma^{\text{ge}}\, o^{\text{ge}}, \\ g &= (1-\gamma^{\text{ge}})\left[(1-\gamma^{\text{cov}})\, g^{\text{text}} + \gamma^{\text{cov}}\, g^{\text{cov}}\right] + \gamma^{\text{ge}}\, g^{\text{ge}}. \end{aligned} \tag{5}$$

C.3 Metrics Computation

We report time-dependent concordance C^{td} and integrated Brier score (IBS) calculated as

$$C^{\text{td}} = \mathbb{P}\Big(\hat{S}(t_i \mid x_i) < \hat{S}(t_i \mid x_j) \;\Big|\; t_i < t_j,\ e_i = 1\Big) \tag{6}$$

$$\text{IBS} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\hat{S}(t \mid x_i)^2\, \mathbb{1}\{t_i \le t,\ e_i = 1\}}{\hat{G}(t_i)} + \frac{\big(1 - \hat{S}(t \mid x_i)\big)^2\, \mathbb{1}\{t_i > t\}}{\hat{G}(t)} \right] dt, \tag{7}$$

with $t_{\max} = \max_i t_i$. $\hat{G}(\cdot)$ is the Kaplan–Meier estimate of the censoring survival function used for inverse-probability-of-censoring weighting (IPCW) [13].

C.4 Model Configurations, Hyperparameters, and Computational Details

The configurations and hyperparameter settings of all models are provided in Table C.1. Specifically, we implement both survival heads and the autoencoder as MLPs. We use distinct learning rates across parameters in different model components (LLM, survival head, autoencoder, and gating parameters). In particular, the LLM's learning rate is set lower to avoid overwriting pre-trained knowledge. To reduce memory consumption, we freeze all but the last 18 layers of the student LLM during fine-tuning, truncate pathology reports to 820 tokens, and use a small batch size.

Table C.1: Hyperparameters of student model, teacher model, and baselines.

Model	Component	Parameter	Value
Student Model	LLM	LLM Name	DeepSeek-R1 Distill Qwen-1.5B
		Freezing	All but last 18 transformer layers
		Precision	bfloat16
		Attention dropout	0.1
		Learning rate	5e-5
	Autoencoder	Latent dim	128
		Encoder layers	[4096, 2048, 1024, 512, 256]
		Activation	ReLU
		Dropout	0.3
		α (CoxPH / Discrete)	1e-8 / 1e-9
		Learning rate	1e-3
	Survival Model	Layers	[100,100,100]
		Activation	ReLU
		Dropout	0.3
		Time bins B (Discrete)	30
		β (CoxPH / Discrete)	5.0 / 1.0
		Learning rate	1e-3 (gates 1e-4)
	Optimization	Batch size	16 (512 under pre-training)
		Epochs / Patience	30 / 5 (1000 / 5 under pre-training)
		Optimizer	AdamW (weight decay 0.01)
		Span weights	sentence w=2.0; number w^{num}=5.0
	Generation Parameters	Temperature	0.3
		Min / max # of new tokens	50 / 350
		# Beams	3
		Top-k	20
		Top-p	0.9
		Repetition penalty	1.3
		No-repeat-ngram-size	3
Teacher Model	Generation Parameters	LLM Name	DeepSeek-R1 Distill Qwen-32B
		Temperature	0.1
		Max # of new tokens (Num. / Expl. prompt)	10 / 300
BERTSurv	LLM	LLM Name	ClinicalBERT
	Survival Model	Type	CoxPH
		Layers	[100,100,100]
		Activation	SELU
		Dropout	0.1
	Optimization	Batch size	24
		Epochs / Patience	30 / 5
		LR	1e-2
		Optimizer	Adam (weight decay 0.01)
Cov-only / GE-only	Survival Model	AS IN STUDENT MODEL	
	Optimization	Batch size	512
		Epochs / Patience	1000 / 5
		LR	1e-3
		Optimizer	AdamW (weight decay 0.01)

For baselines, because no public BERTSurv implementation is available, we re-implemented it in-house following [23] and applied the reported hyperparameters. As BERTSurv does not directly handle gene-expression inputs, we trained the same gene-expression autoencoder offline to obtain fixed latents, which we then concatenated to the ClinicalBERT text embedding (together with other tabular covariates). For the covariates-only and gene-expression-only baselines, we reused the same survival-model configuration as in our framework to ensure comparability. In CoxPH settings, we compute the Breslow baseline hazards [1] on the combined training and validation sets and use them for test-time evaluation.

All experiments were performed on a computer cluster with one NVIDIA A100 (80 GB) GPU, six CPU cores, and 20 GB system RAM per task. At inference, the student model's resource footprint is substantially lower; evaluation can run on smaller GPUs or even CPU-only, supporting local deployability.