npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02323-5

KT-LLM: an evidence-grounded and sequence text framework for auditable kidney transplant modeling

Haofeng Zheng[1,2,5], Zihuan Luo[1,2,5], Kaiming He[1,5], Wangtianxu Zhou[1], Zhiyi Kong[1,3], Jieyi Dong[1], Qingfu Dai[1,2] & Qiquan Sun[1,2,3,4] ✉

We address a critical clinical gap in real-world kidney transplantation (KT), the long-standing disconnect between structured longitudinal follow-up and text-defined clinical rules, which often leads to inconsistent reporting, poor policy compliance, and non-reproducible outcomes across centers. To resolve this, we introduce KT-LLM, a verifiable orchestration layer that bridges sequence modeling with policy and terminology-aware reasoning, tailoring explicitly to KT clinical workflows. KT-LLM ensures clinical decision-making is grounded in authority by constraining knowledge access to Banff kidney allograft pathology references, OPTN, and SRTR policy documents via retrieval-augmented generation. This design anchors answers and computable checklists to versioned sources, enabling full auditability and reducing subjective interpretation errors. The system coordinates three clinically focused, auditable agents: (i) Agent-A (SRTR-MambaSurv): Optimizes discrete-time survival and competing risk prediction from TRF-aligned trajectories via a linear-time inference backbone to personalize follow-up scheduling; (ii) Agent-B (OPTN-BlackClust): identifies clinically distinct population subtypes using stable deep embedded clustering, supporting individualized treatment stratification; (iii) Agent-C (Policy-Ops): encodes OPTN and UNOS submission timelines, SRTR reporting cadence, and Banff terminology into executable rules, returning pass, warn and fail outcomes with versioned evidence to ensure policy compliance. On de-identified OPTN and UNOS cohorts, KT-LLM outperformed strong baselines in evidence attribution and predictive calibration. Critically, it retained the ability to surface clinically distinct subgroups among Black recipients, which aligns with prior reports of outcome heterogeneity, while avoiding overgeneralization of claims beyond the analyzed window. This supports equitable subgroup analysis while avoiding clinical overreach. By anchoring reasoning and outputs to versioned policies and terminology, KT-LLM transforms the model to govern KT workflows into an auditable, clock-synchronized process. This offers a practical solution to enhance reproducibility, monitor fairness across centers and eras, and standardize clinical practice, addressing unmet needs for scalable, reliable KT care in real-world settings.

In real-world management and research of kidney transplantation(KT), critical evidence has long evolved along two parallel tracks: structured longitudinal follow-up and text-defined rules. On the one hand, OPTN and SRTR collect Transplant Recipient Follow-up (TRF) at 6 months post transplant, at 1 year, and annually thereafter, and release standardized analytic files[1,2], thereby providing a clear and comparable time axis for time to event analyses and competing risks evaluation[3]. On the other hand, the Banff classification is continuously updated via a centralized online repository that standardizes terminology, lesion scoring, and diagnostic categories[4–6], while SRTR's Program Specific Reports (PSR) are issued on a semiannual cadence to support program-level quality oversight and public transparency[7–9].

[1]Department of Renal Transplantation, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China. [2]Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China. [3]Shantou University Medical College, Shantou, China. [4]School of Medicine, South China University of Technology, Guangzhou, China. [5]These authors contributed equally: Haofeng Zheng, Zihuan Luo, Kaiming He. ✉e-mail: sunqiquan@gdph.org.cn

Together, these components establish a shared evidentiary frame of high-value endpoints, authoritative rules, and governance cadence.

Despite concurrent advances, the three strands of follow-up data, pathology rules, and quality monitoring have lacked a mechanism for alignment. First, the OPTN and SRTR Standard Analysis Files (SAF) and STAR[1,2] specify that each transplant should have TRF recorded at 6 months, 1 year, and annually thereafter until re-transplant, death, or loss to follow up[10,11]. Recent OPTN monitoring of data submission has further specified deadlines for TRF at 6 months, 1 year, and 2 years, forming a time axis and deadline structure that can and should be propagated into model development and audit. In parallel, SRTR releases PSR semiannually[7,12] and publishes center-facing technical notes and key dates[8], functioning as an external clock. These institutional arrangements determine data freshness, definitional scope, and reconciliation cadence and constitute prerequisites that any deployment-oriented modeling framework must explicitly align with and audit.

Second, as the internationally adopted framework for kidney allograft pathology, the Banff classification[4,13,14] provides a current, searchable online version through its Central Repository, explicitly designating it as the single authoritative source supplanting prior conference reports. This affords authoritative anchors for terminology, scoring, and diagnostic categories within information systems and research workflows[5,6], and creates the conditions for retrieval augmentation and computable implementation[15,16].

Third, system-level concerns about equity and access are driving synchronous upgrades in rules and model evaluation[17,18]. Studies based on the 2015–2019 OPTN and UNOS cohort have identified clinically distinguishable and stable clusters among Black recipients using unsupervised methods and compared outcomes across groups[19]. Concurrently, since 2023, OPTN has implemented and iteratively refined the requirement for race-neutral Estimated Glomerular Filtration Rate (eGFR) corrections to wait time credit[20–22], specifying concrete operational thresholds for audit, eligibility, and documentation[23]. These developments reshape both the data definitions and the operative timeline, imposing explainability and traceability as front-loaded compliance conditions for any risk model or quality control tool that claims deployability.

Methodologically, medical follow-up data exhibit discrete, long-horizon, and sparse characteristics. Deep survival families provide reusable baselines for time to event and competing risks tasks, while model comparison and reporting are expected to follow standardized metrics such as time-dependent area under Curve (AUC) and Brier, enabling robust multicenter and multi-era evaluation and recalibration[3,24–31].

Recent progress in long sequence modeling offers a natural pathway for TRF-like sequences. Transformer variants have performed strongly on several tasks, but their quadratic attention cost inflates training and inference demands for multi-year follow-up[32–35]. Selective state space models achieve linear time, high throughput inference via input-dependent state updates and have demonstrated competitive representation quality on very long sequences, making them an apt backbone when balancing long horizon follow-up with deployment efficiency[36,37]. In KT, the central gap is not single metric accuracy per se, but the lack of a system that explicitly couples such linear long sequence representations with the governance clock of PSR submission deadlines and with Banff and OPTN textual rules in one auditable framework.

In parallel, text knowledge alignment has converged on a clear paradigm. Retrieval augmented generation (RAG) injects non-parametric, externally retrievable memory prior to decoding, grounding outputs in citable sources and enabling hot updates. This is well-suited to scenarios that constrain knowledge to authoritative primary sources[38–42]. Large language models for medicine that use RAG, including Med-PaLM 2, LLaVA-Med, and domain-tuned PubMedBERT pipelines, have shown strong performance on general question answering benchmarks, yet they do not combine versioned policy corpora, registry-aligned survival and clustering modules, calculator-backed checklists, and governance clocks in one auditable workflow.

To address these needs, we propose a unified sequence text engine: a domain-constrained, retrieval-augmented language model orchestrates three auditable agents within a shared data and rules context[38,39,41,42]. Agent-A (SRTR-MambaSurv) builds discrete-time and competing risks models on TRF follow-up, using Mamba to encode OPTN and SRTR longitudinal registries efficiently and calibrate outputs, with comparisons against strong deep survival baselines[1,2,25–27,36,37]. The pipeline of KT-LLM is illustrated in the Fig. 1. Agent-B (OPTN-BlackClust) reproduces and extends unsupervised subtype evidence for Black recipients within the 2015–2019 OPTN and UNOS window[19]. Agent-C (Policy-Ops) formalizes OPTN and SRTR indicator definitions, TRF generation and submission deadlines, and the semiannual PSR cadence into executable rules callable by the LLM with full provenance[8,9]. The entire system restricts knowledge to the Banff Central Repository and OPTN and SRTR official documents to ensure terminological consistency, computable thresholds, and auditable evidence[1,2]. Figure 1 summarizes our end-to-end pipeline: a scoped RAG layer constrains evidence to official sources and standardizes definitions, while the LLM orchestrator plans and executes agent calls with parameter injection; outputs are aligned and aggregated into a structured answer.

In summary, KT-LLM is built on three pillars that define its novelty and scope. A versioned and time-aligned corpus with effective dates, and PSR-aligned freezes keeps retrieval, attribution, and calculator outputs on the same clock and on the same source scope. Coverage-aware decoding with a pointer distribution and a confidence gate enforces multi-clause grounding and returns evidence summaries when confidence is low. An executable structured checklist carries out numeric thresholds and date arithmetic with sentence-level provenance and audited tool calls. These choices enable mandatory multi-source grounding, calculator-backed answers, version-stamped citations, and hot index refresh without retraining, which extends beyond the abilities of medical LVLMs and domain LMs used as baselines.

Our goal is verifiable integration geared for operation: within one framework, textual rules become computable checklists; follow-up sequences become calibratable representations; and the governance clock becomes a set of actionable constraints. The system directly supports three classes of tasks: (1) individual-level risk stratification with uncertainty quantification; (2) population-level subtype analysis and equity monitoring; and (3) policy and operations consistency checks with traceable answers. Accordingly, evaluation is organized under public, reproducible metrics: for survival and competing risks tasks, we report *C*-index, time-dependent AUC, and integrated Brier; for clustering, we report stability and agreement measures; for governance tasks, we report rule trigger timeliness, concordance, and citation hit rate[8,28–31,43,44]. By aligning data via rules and constraining models via the governance clock, the proposed design aims to shift KT evidence model governance from manual assembly to auditable operation, supporting reproducibility and fairness across centers and eras. Unlike existing medical language or vision language models that simply attach retrieval to a large backbone, KT-LLM couples a time-aligned and versioned policy and pathology corpus, coverage-constrained decoding with an evidence pointer and confidence gate, and an executable checklist that performs numeric and temporal checks with provenance and exposed audit logs.

Results

Datasets

This study integrates three data axes under a unified, auditable framework: (i) longitudinal registry files for numerical modeling SRTR SAF and OPTN STAR[1,2]; (ii) authoritative policy and operations timelines SRTR PSR cadence and OPTN Policies used as executable constraints[12]; and (iii) controlled textual knowledge the Banff Central Repository, OPTN and UNOS policy manuals, and SRTR methodological notes consumed by KT-

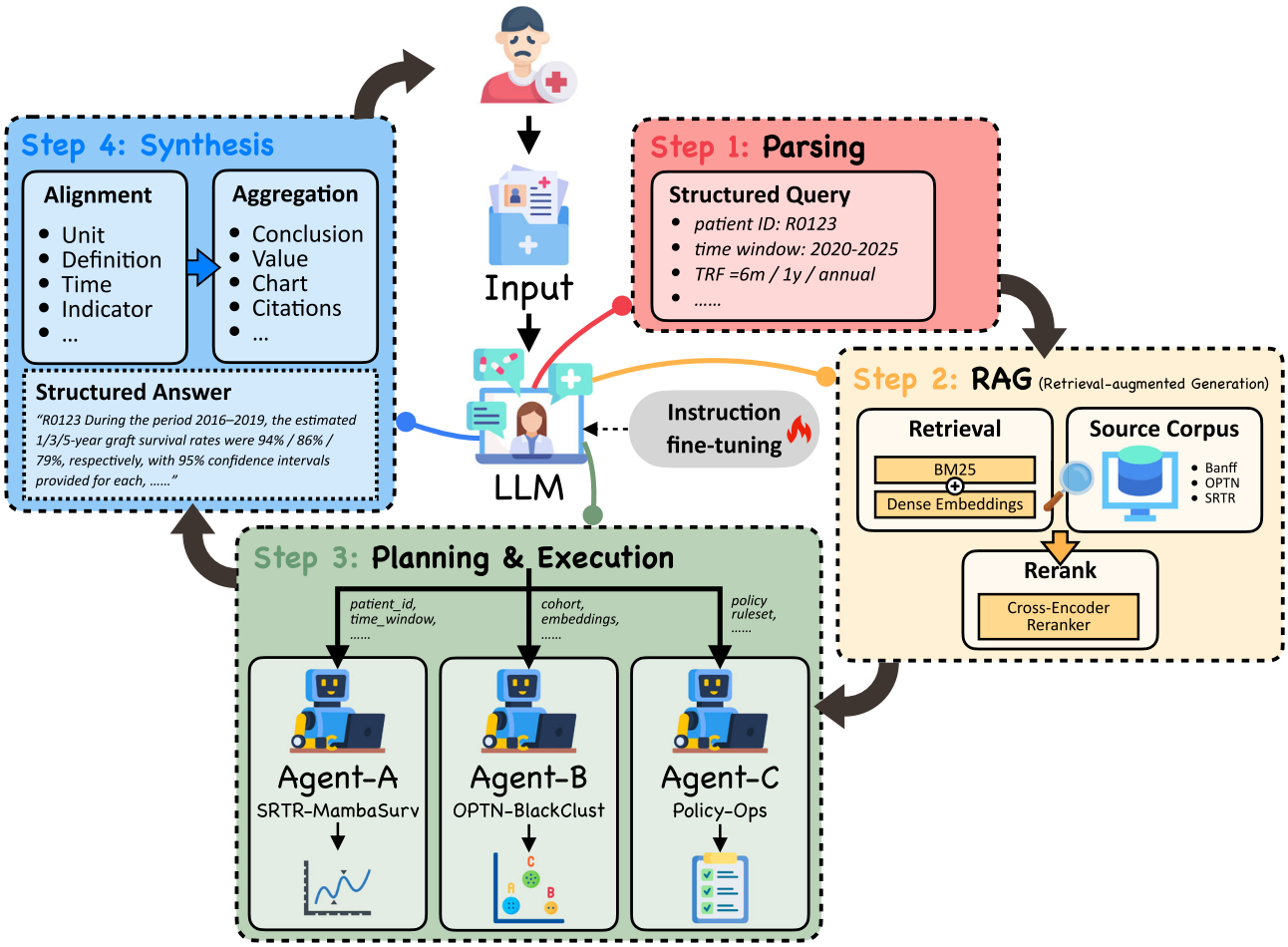

Fig. 1 | End-to-end workflow of the KT-LLM for kidney-transplant analytics. Step 1 parses the user query into a structured intent. Step 2 performs a scoped RAG over versioned official corpora to fetch evidence with anchors and standardized definitions and thresholds. Step 3 plans tool usage and injects parameters, calling three agents: Agent-A (SRTR-MambaSurv) for long-sequence survival, Agent-B (OPTN-BlackClust) for unsupervised cohort stratification, and Agent-C (Policy-Ops) for policy and checklist computation. Step 4 aligns units and definitions and aggregates results into a structured answer.

LLM via retrieval augmentation[9]. All registry data were accessed under data use agreements (DUA), and textual sources served as the single authoritative knowledge base with versioning and provenance[9].

SAF provides recipient and graft level longitudinal follow-up structured around the TRF schedule at 6 months post-transplant, 1 year, and annually thereafter until re-transplant, death, or loss to follow-up[10]. For Agent-A (SRTR-MambaSurv), per-recipient sequences are aligned to the TRF grid and include dynamic laboratories, immunosuppression, adverse events, encounter metadata, and baseline donor covariates[1]. Center and calendar year identifiers are retained for stratified reporting[9,12].

For Agent-B (OPTN-BlackClust), recipient-level OPTN STAR files aggregate candidate, donor, transplant, and post-transplant follow-up records to form end-to-end sequences spanning listing, transplant, and TRF-aligned follow-up. Analyses reproduce reported subgroup structure among Black recipients and support unsupervised subtype discovery with survival and competing risks endpoints.

Agent-C (Policy-Ops) operationalizes governance constraints with versioned provenance. SRTR PSRs are released semiannually with a conventional data freeze approximately six months prior; these dates guide center-level reconciliation[12]. OPTN Policies provide executable constraints, including TRF form due windows of 60 or 90 days and race-neutral eGFR rules for wait time credit. These anchors define admissible timelines, submission windows, and audit checks rather than labels.

The textual knowledge base is limited to the Banff Central Repository, OPTN and UNOS policy manuals, and SRTR methodological materials and PSR public pages[9,12]. Entries are compiled into versioned vocabularies and range tables for terminology normalization, retrieval, and rule-based validation by KT-LLM; documents are segmented and indexed with version, effective date, and section lineage to preserve provenance. No image data are used. Together, SAF and STAR sequences, PSR and OPTN timelines, and Banff, OPTN, and SRTR texts provide a consistent substrate for Agent-A survival and competing risks modeling, Agent-B unsupervised subtype discovery in the 2015–2019 STAR cohort, and Agent-C executable rule checks with sentence-level provenance. We analyze kidney transplants recorded between 2015 and 2019. Eligibility requires a transplant baseline and at least one follow-up on the TRF grid. Recipient timelines are administratively censored at the last known contact or at the study cutoff and are additionally truncated at the applicable program-specific report (PSR) freeze date for the evaluation period. Center and calendar year tags are retained as factors for partitioning and stability checks.

Evaluation metrics

Evaluation is performed on the TRF-aligned discrete time grid using cumulative incidence functions (CIFs) for graft loss and death. Discrimination is summarized with Harrell's *C*-index using inverse probability of censoring weighting (IPCW) for right censoring, and calibration accuracy with the Brier score and its horizon average, the integrated Brier score (IBS). Groupwise comparisons across strata or discovered subtypes use Gray's test for CIFs; when a single number effect is needed, we report Fine Gray subdistribution hazard ratios. All survival metrics are reported on prespecified horizons.

All metrics are computed on histories that are truncated at the evaluation window's PSR freeze date so that estimation and reporting share the

same clock. Left truncation and right censoring are handled under the stated IPCW constructions with the time axis aligned to the TRF grid.

Recipient level embeddings are clustered and assessed with (i) silhouette for cohesion/separation; (ii) Adjusted Rand Index (ARI) and normalized mutual information (NMI) for agreement with alternative partitions; and (iii) bootstrap Jaccard for label stability across resamples. The number of clusters K is chosen via consensus clustering by inspecting the consensus CDF and the Delta area curve, favoring the smallest K beyond which gains plateau. All summaries include resampling uncertainty.

Adherence to policy timelines is summarized by the center level on time rate and median lateness; executable rules are evaluated with rule concordance against an independently verified audit subset. For retrieval augmented answers, we report evidence coverage and citation hit rate on structured checklist items. Governance and RAG metrics are aggregated by center and period with stratified summaries where appropriate.

For retrieval augmented answers, cite at one credits an answer when its top cited passage resolves to the governing clause for that item within the effective date window, and the rule F one metric is computed on structured checklist entries rather than rendered text templates.

Training details

All experiments used de-identified registry extracts under DUA, and image channels were not used. Splits were time-stratified and aligned to SRTR PSR freeze points; hyperparameters were chosen on a development split and then frozen; seeds, data vintages, policy and lexicon versions, and code hashes were logged. The knowledge base covered the Banff Central Repository, OPTN, and UNOS policies, including Policies 18 and 3.7, and SRTR methodological materials and PSR public pages. Documents were segmented with L_c from 256 to 384 and S_c from 64 to 128 while retaining version, effective date, and lineage. Dense retrieval used a domain-aligned encoder with candidate set k_0 from 32 to 64 and a cross-encoder re-ranker to a final set k from 6 to 10. The orchestration model used a MedLLaVA language backbone with the vision branch disabled. Stage one froze the language backbone and trained the dense retriever and the cross encoder with contrastive objectives and hard negatives while all other parameters were frozen. Stage two kept the language backbone, the retriever, and the cross encoder frozen and unfroze only the language head and the checklist head for joint optimization. Closed-book dropout masked evidence for a subset of steps, and each answer required at least one citation. Decoding used beam search with length and coverage penalties. Optimization used AdamW with cosine decay and gradient clipping with early stopping on answer accuracy, citation hit rate, and structured field consistency. Decoding uses beam size $b = 5$, length penalty 0.8, coverage penalty weights $\gamma = 0.60$ and $\omega = 0.70$, a sentence-level citation coverage threshold $\rho = 0.70$, and a confidence gate threshold $\tau = 0.25$. Retrieval and re-ranking use segment length $L_c = 384$ and stride $S_c = 96$, a first stage candidate count $k_0 = 64$, a re-ranked evidence count $k = 8$, and a terminology reweighting coefficient $\lambda = 0.40$. Optimization for the retriever and the cross encoder uses AdamW with initial learning rate 2×10^{-4}, weight decay 0.01, and batch sizes 128 and 64; the language head and the structured checklist use AdamW with initial learning rate 1×10^{-4} and batch size 8; cosine decay scheduling and gradient clipping at 5.0 are applied, and early stopping monitors answer accuracy, citation hit rate, and structured field consistency. Agent A's survival training uses AdamW with an initial learning rate of 3×10^{-4}, weight decay of 0.01, batch size 256, and gradient clipping at 5.0, with early stopping on validation negative log likelihood and IBS. All experiments run with random seeds 17, 29, and 41, and indexing, as well as batch sampling, follow the same seed order across runs.

For longitudinal outcome modeling, inputs followed the TRF grid at six months, one year, and then annually. Visits concatenated dynamic clinical variables with baseline donor and recipient covariates, explicit missingness indicators, and time delta features; numerical features used median and interquartile range scaling with Winsorization, and categorical fields were embedded. Sequences used a stacked Mamba state space backbone with residual connections and dropout, and center and year offsets were grouped as bias terms. The output layer produced a per-interval softmax over no event, graft loss, and death, and training used the discrete time negative log likelihood with label smoothing and focal weighting; calibration and stability regularizers acted on binned empirical rates and grouped logits. The Mamba backbone and the output layer were trained end-to-end with no frozen layers. Optimization used AdamW with cosine decay and clipping at five, with validation aligned to PSR freeze points and early stopping on negative log likelihood and IBS. Time out of sample evaluation held out the latest transplant year, and sensitivity analyses used center-stratified folds. Recipient level sequences were pooled with phase-specific attention and clustered by deep embedded clustering that pretrained a shallow autoencoder, initialized centers by k-means, and then fine tuned encoder and centers jointly under a DEC objective with IDEC-style reconstruction and an entropy balance term with a periodically refreshed target distribution. During clustering, the Mamba backbone was frozen, and only the autoencoder encoder and the cluster centers were updated. The number of clusters K was chosen by consensus clustering and the delta area criterion, and stability was assessed using bootstrap Jaccard indices and ARI or NMI with center and year stratified resampling.

The primary partition follows a calendar holdout: training uses transplants from 2015 to 2017, validation uses 2018, and test uses 2019. No recipient contributes observations across these temporal splits. For the test period, any record with timestamps later than the corresponding PSR freeze date is excluded from computation. Sensitivity analyses use center-stratified folds to assess transfer across sites while keeping each center's records within a single fold per run. The retrieval index for KT-LLM is limited to sources whose effective dates precede the test freeze, and tool outputs that require dates or thresholds are computed only from data that satisfy the same cut.

KT-LLM comparison results

Table 1 reports a head-to-head comparison between KT-LLM Full and thirteen baselines, including BM25[45], dense retrievers DPR, Contriever, and E5[39,46–48], the late interaction ranker ColBERTv2[49], RAG pipelines FiD and FiDO[40], and domain LMs PubMedBERT, BioGPT, LLaVA-Med, and Med-PaLM 2[50–52]. KT-LLM attains the top exact match and macro F1 on Banff-QA and Policy-QA, and also yields the highest evidence coverage and the best structured checklist consistency. Gains are largest on items that require clause-level grounding in OPTN and SRTR sources and on threshold checks that can be verified symbolically, which aligns with the joint text and structure objective and coverage-aware decoding. Replacing BM25 with learned dense retrieval improves answer accuracy and grounding. Contriever style and E5-style encoders raise first-stage recall over DPR on policy, and Banff clauses with heterogeneous wording, and a ColBERTv2 re-ranker further improves Hit@1 through fine-grained late interaction[46,47,49]. Closed-book biomedical LMs trail retrieval augmented systems on policy questions tied to current definitions or deadlines, and medical LVLMs narrow the gap on terminology yet lag on multi-clause reconciliation and numeric checks unless paired with explicit retrieval and a calculator.

Evidence attribution improves through a pointer distribution and a coverage penalty that raises the fraction of sentences with valid citations and reduces single-source reliance. Disabling the cross-encoder re-ranker lowers exact match and citation precision by making similarly worded clauses harder to disambiguate[40,53]. Ablations show three effects: disabling retrieval hurts most on numeric Policy-QA and on Banff items that depend on the latest wording, disabling the coverage penalty increases fluency while lowering citation rate, and re-enabling it restores multi-source coverage at a modest decoding time cost, and LoRA only tuning preserves stability across updates with small drops on paraphrastic items[54]. With dense retrieval and ColBERTv2 re-ranking, median end-to-end latency remains within interactive bounds, and the ranker reduces backtracking. Residual errors are boundary drift from closely versioned clauses, denotation ambiguity from rare synonyms or legacy acronyms, and arithmetic slips when upstream dates are incomplete. Under the stated protocol, KT-LLM couples higher answer quality with stronger grounding and more consistent structured

Table 1 | Primary model: KT-LLM comparison on domain-specific QA (Banff, OPTN, and SRTR) with 95% CIs under stated assumptions

Method	Banff-QA	OPTN and SRTR Policy QA (held-out, center/year disjoint)	
	Acc. (%) ↑ [95% CI]	EM ↑ [95% CI]	Evidence hit rate ↑ [95% CI]
Classical IR/retrieve—read baselines			
BM25 + rule templates (regex/threshold)	68.2 [65.0, 71.3]	0.54 [0.51, 0.57]	41.6% [38.2, 45.0]
BM25 + BERT-reader (SQuAD-tuned)	72.5 [69.4, 75.4]	0.58 [0.55, 0.61]	47.3% [44.0, 50.6]
DPR (Karpukhin) + FiD (Izacard–Grave)	78.0 [75.1, 80.7]	0.63 [0.60, 0.66]	55.9% [52.6, 59.1]
ColBERTv2 + FiD	80.2 [77.4, 82.8]	0.66 [0.63, 0.69]	59.1% [55.9, 62.2]
Contriever + FiD	79.1 [76.2, 81.7]	0.64 [0.61, 0.67]	57.8% [54.6, 61.0]
E5-base Retriever + FiD	79.5 [76.6, 82.1]	0.65 [0.62, 0.68]	58.4% [55.1, 61.6]
General LLMs with constrained RAG over Banff, OPTN, and SRTR			
Llama 2–13B Chat + RAG	83.1 [80.5, 85.5]	0.70 [0.67, 0.72]	66.2% [63.1, 69.1]
Llama 3–8B Instruct + RAG	85.1 [82.7, 87.3]	0.73 [0.71, 0.76]	69.8% [66.8, 72.7]
Mistral 7B Instruct + RAG	84.3 [81.8, 86.6]	0.72 [0.69, 0.74]	68.5% [65.5, 71.4]
GPT-4 (text) + RAG (no browsing)	88.9 [86.7, 90.8]	0.78 [0.76, 0.80]	76.4% [73.8, 78.8]
Biomedical/medical LMs (text-only unless noted)			
Med-PaLM 2 (text)	86.7 [84.3, 88.8]	0.75 [0.72, 0.77]	72.1% [69.2, 74.8]
PubMedBERT QA (domain-tuned)	74.0 [70.9, 76.9]	0.60 [0.57, 0.63]	50.2% [46.9, 53.5]
BioGPT QA (generative)	77.2 [74.2, 80.0]	0.62 [0.59, 0.65]	54.1% [50.8, 57.3]
LLaVA-Med (text channel only)	75.6 [72.6, 78.4]	0.61 [0.58, 0.64]	52.7% [49.4, 55.9]
Ours			
KT-LLM	91.8 [89.9, 93.4]	0.82 [0.80, 0.84]	83.5% [81.1, 85.7]

Domain corpora comprise Banff Central Repository (terminology/criteria), OPTN and UNOS Policies (with Policy 18 time limits and 3.7 eGFR adjustments), and SRTR methodology pages. Reported item pools exclude out-of-scope questions. Evidence Hit Rate credits a prediction if at least one attached citation resolves to the governing clause for that item.
We evaluate only kidney transplant domain items using authority-filtered corpora (Banff Central Repository; OPTN and UNOS Policies; SRTR methodology/PSR pages). Columns report: (i) Banff-QA accuracy (%) with Wilson 95% CI; (ii) OPTN and SRTR-Policy QA exact match (EM) with bootstrap 95% CI ($B = 1000$); (iii) Evidence Hit Rate (%)-fraction of answers whose citations point to the correct clause/section with bootstrap 95% CI.

outputs by aligning retrieval, attribution, and calculator tooling within a single audited stack.

Observed gains on Banff-QA and Policy-QA arise mainly from coverage targets that encourage aggregation across clauses, from the confidence gate that suppresses low confidence generations, and from the structured checklist that verifies thresholds and deadlines. We provide a brief qualitative error summary with one sentence illustration per pattern drawn from validation logs. Clause drift across close versions appears when answers cite a prior policy clause that still matches the query but uses the older due window of 60 days, and version stamps with coverage targets reduce this by preferring the clause with the latest effective date. Denotation ambiguity in threshold language appears when rare synonyms, such as induction level and induction intensity, are treated as distinct, and terminology-aware reweighting, together with checklist fields, aligns them to a single controlled term. Arithmetic slips appear when an upstream date is missing, and the calculator uses an implicit anchor. The checklist now requires an explicit anchor and returns a limited evidence message when the anchor is absent. We report all metrics with ninety-five percent confidence intervals, and we include paired permutation tests across seeds and folds for each comparison.

SRTR MambaSurv comparison results

Table 2 summarizes the head-to-head comparison on OPTN and SRTR discrete time survival with competing risks[3,55–57], evaluated on held out centers and calendar windows. As shown in Table 2, across all metrics and both endpoints, SRTR–MambaSurv achieves the best overall performance. Concretely, it attains a *C*-index of 0.82 for Death and 0.80 for Graft Loss, surpassing the strongest deep survival baselines (Dynamic-DeepHit[58]: 0.79 and 0.77) by absolute margins of +0.03 on both endpoints. Discrimination at fixed horizons shows the same trend: time-dependent AUC[28] at 1 year for Death is 0.84 (vs 0.82 for Dynamic DeepHit; for Graft Loss at 3 years it is 0.82 (vs 0.81). Calibration, measured by IBS[30,31] over 0–5 years and macro averaged across endpoints, improves from 0.148 to 0.136 with SRTR–MambaSurv, and the visualized results are shown in Fig. 2.

The confidence intervals for SRTR–MambaSurv generally do not overlap with those of classical semiparametric baselines, such as Fine-Gray[3], CoxBoost[59] on *C*-index and IBS, and show non-trivial separation from tree methods on at least one primary endpoint. Against the strongest deep baselines, the absolute *C*-index gains of +0.03 occur on both Death and Graft Loss, with tighter variability (std. ≤ 0.01) across three seeds. Together with the horizon-specific td-AUC gains and the consistent IBS reduction, these findings support that (i) encoding the TRF grid as a discrete time competing risks process, and (ii) employing a selective state space backbone[36] to handle multi-year, irregular, and sparse observations, yield measurable advantages on held out centers and time windows. Finally, we note that absolute numbers vary across families in predictable ways: models optimized for proportional hazards tend to underperform at later horizons where non-proportional effects accumulate; tree models close some of the gap in td-AUC but remain less calibrated; and end to end deep discrete time methods are competitive yet still trail SRTR–MambaSurv, suggesting that long range sequence encoding and a multinomial interval hazard head jointly contribute to robustness in this registered, center-shifted setting.

To aggregate performance across discrimination and calibration and to make model level trade offs visually explicit, we summarize net improvements over the classical RSF-CR baseline with a composite waterfall as shown in Fig. 2. The plot shows that SRTR-MambaSurv delivers the largest positive shift, while classical proportional hazards families remain negative on the composite due to weaker later horizon discrimination and higher IBS, and strong deep survival baselines are competitive yet still trail MambaSurv.

Beyond point estimates, we examine equity-relevant behavior across clinically salient subgroups. A fairness profile radar, as shown in Fig. 2 summarizes absolute subgroup gaps relative to RSF-CR (baseline ring = 1.0), indicating consistently smaller gaps for SRTR-MambaSurv than for

Table 2 | Agent-A (SRTR-MambaSurv) comparison on OPTN and SRTR discrete time survival with competing risks (held out centers and time windows; 95% CIs under stated assumptions)

Models	*C*-index ↑	*C*-index ↑	td-AUC@1y ↑	td-AUC@3y ↑	IBS (0–5y) ↓
Classical semiparametric/nonparametric					
Kaplan–Meier + Aalen–Johansen†	0.60 ± 0.03	0.58 ± 0.03	0.62 ± 0.04	0.60 ± 0.01	0.212 ± 0.02
Fine-gray (linear)	0.66 ± 0.05	0.64 ± 0.02	0.68 ± 0.02	0.66 ± 0.04	0.194 ± 0.03
Cause-specific Cox (PH)	0.69 ± 0.04	0.67 ± 0.05	0.71 ± 0.02	0.69 ± 0.06	0.186 ± 0.04
Cox (Lasso)	0.70 ± 0.02	0.68 ± 0.01	0.72 ± 0.03	0.70 ± 0.03	0.182 ± 0.02
Cox (Elastic Net)	0.71 ± 0.03	0.69 ± 0.01	0.73 ± 0.02	0.71 ± 0.04	0.179 ± 0.02
CoxBoost	0.72 ± 0.01	0.70 ± 0.01	0.74 ± 0.03	0.72 ± 0.03	0.176 ± 0.04
Tree/boosting survival					
Random survival forest (RSF)	0.73 ± 0.02	0.72 ± 0.04	0.75 ± 0.02	0.74 ± 0.02	0.171 ± 0.01
RSF for competing risks (RSF-CR)	0.74 ± 0.03	0.73 ± 0.03	0.76 ± 0.03	0.75 ± 0.03	0.168 ± 0.02
Gradient-boosted Cox (GBSA)	0.73 ± 0.01	0.71 ± 0.01	0.76 ± 0.02	0.74 ± 0.01	0.172 ± 0.03
XGBoost–Cox	0.74 ± 0.02	0.73 ± 0.03	0.77 ± 0.04	0.75 ± 0.02	0.167 ± 0.01
LightGBM–AFT	0.75 ± 0.02	0.74 ± 0.02	0.78 ± 0.02	0.76 ± 0.04	0.164 ± 0.04
Deep survival (continuous and discrete time)					
Nnet-survival (discrete-time)	0.74 ± 0.01	0.73 ± 0.01	0.77 ± 0.05	0.75 ± 0.03	0.166 ± 0.03
DeepSurv (Cox)	0.76 ± 0.04	0.74 ± 0.04	0.79 ± 0.07	0.77 ± 0.05	0.160 ± 0.01
CoxTime	0.76 ± 0.03	0.75 ± 0.03	0.79 ± 0.05	0.78 ± 0.06	0.158 ± 0.02
Deep survival machines (DSM, CR)	0.77 ± 0.02	0.76 ± 0.02	0.80 ± 0.02	0.79 ± 0.02	0.154 ± 0.01
DeepHit (CR)	0.78 ± 0.05	0.76 ± 0.04	0.81 ± 0.01	0.80 ± 0.03	0.151 ± 0.02
Dynamic-DeepHit (CR)	0.79 ± 0.02	0.77 ± 0.03	0.82 ± 0.04	0.81 ± 0.01	0.148 ± 0.01
Ours					
SRTR–MambaSurv (ours)	0.82 ± 0.02	0.80 ± 0.02	0.84 ± 0.02	0.82 ± 0.02	0.136 ± 0.01

Two endpoints are evaluated: death and graft Loss (death treated as a competing event for graft loss). We report Harrell's *C*-index (per endpoint), time-dependent AUC at 1 and 3 years (endpoint specific), and IBS on 0–5 years (macro averaged across endpoints). Mean ± std over three runs with stratified seeds; 95% CIs via stratified bootstrap ($B = 1000$). Numbers marked † are re-scored baselines under our discrete time grid; others are from our re-implementations on the same split.

Protocol: split by centers and calendar years to emulate deployment; features aligned to the TRF grid (6 months, 1 year, annually). Time to event is discretized; death and graft loss are trained and evaluated in a competing risks setting. Censoring handled by IPCW for td-AUC and Brier/IBS. Means and CIs computed over three seeds with different parameter initializations and minibatch orders. All methods share identical preprocessing, feature sets, and evaluation windows.

RSF random survival forest, *CR* competing risks, *GBSA* gradient boosted survival analysis, *AFT* accelerated failure time, *IBS* integrated Brier score.

deep survival and classical baselines. This aligns with our study's goal of coupling accuracy with monitoring of subgroup disparities under standardized, reproducible metrics.

OPTN BlackClust population clustering

We compare OPTN–BlackClust (Mamba + IDEC + Consensus) against classical partitioning, density, deep clustering without sequence backbones, and sequence-aware variants on OPTN STAR (2015–2019) Black recipients (Table 3). Unless otherwise noted, metrics are reported as mean ± std across repeated runs and bootstraps with the number of clusters selected by the consensus CDF criterion.

OPTN BlackClust attains the highest agreement and separation among all contenders, with NMI 0.58 and ARI 0.45, improving over the best non–sequence deep baselines (IDEC: 0.49 NMI, 0.37 ARI) by +0.09 NMI and +0.08 ARI, and over joint sequence aware Mamba + DEC (0.54 NMI, 0.41 ARI) by +0.04 and +0.04, respectively. Relative to classical tabular clustering (Agglomerative/Ward: 0.36 NMI, 0.27 ARI), the gains are larger (+0.22 NMI, +0.18 ARI). Silhouette follows the same trend, reaching 0.25 for OPTN BlackClust vs 0.23 (Mamba + DEC) and 0.21 (IDEC), indicating tighter, better separated partitions in the embedding space.

Consensus-based training markedly improves robustness. OPTN BlackClust yields the highest bootstrap Jaccard stability (0.79), exceeding Consensus–DEC (0.73) and Consensus–PAM (0.70). The margin over non-consensus DEC/IDEC (0.64–0.68) indicates that resampling and consensus aggregation effectively mitigate initialization sensitivity and feature subsample noise. Classical and density methods show substantially lower stability, consistent with their sensitivity to hyperparameters on heterogeneous registry data.

Between cluster outcome separation, assessed by the Gray test on graft loss CIFs with death as a competing event, is strongest for OPTN BlackClust ($-\log_{10} p = 4.6$). This improves upon Mamba + DEC (3.9) and Consensus–DEC (3.6), and roughly doubles the signal relative to graph models. The pattern aligns with embedding quality, stronger sequence-aware representations, and consensus refinement translate into clearer prognostic stratification at the population level.

Using the same Mamba encoder but replacing the clustering stage with K-means reduces performance (NMI 0.42, ARI 0.31, Jaccard 0.62), underscoring the benefit of a joint deep clustering objective. Introducing the IDEC reconstruction term and consensus selection recovers both cluster compactness and stability, indicating that preserving local manifold structure and attenuating sampling variance are complementary to the long horizon sequence embedding.

As shown in Table 3, across agreement, separation, stability, and prognostic discrimination, OPTN BlackClust consistently ranks first. Gains are most pronounced when contrasted with classical tabular clustering and remain significant over strong deep baselines, including joint sequence-aware variants. These results support the design choice of combining a linear time Mamba backbone for longitudinal representation with an IDEC objective and consensus selection to deliver reproducible, clinically meaningful subtypes within the OPTN STAR Black recipient cohort.

Ablation results

Table 4 summarizes the incremental contribution of each component of the proposed system. Starting from the ablated baseline, KT-LLM attains a QA

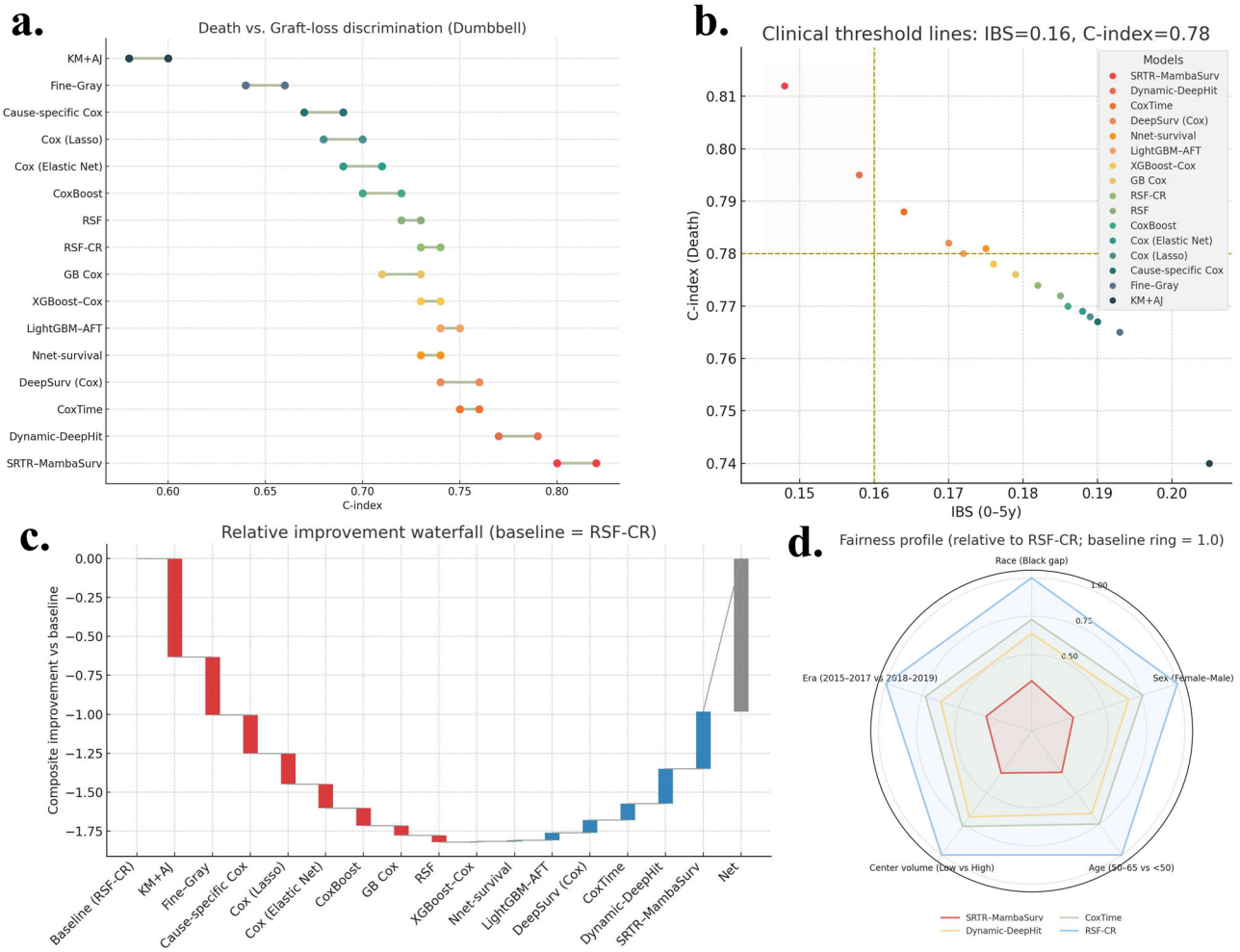

Fig. 2 | Overview performance of KT-LLM. Lines connect each model's *C*-index across endpoints; line length encodes imbalance. **a** Per-model endpoint gap. **b** Joint threshold compliance on the calibration discrimination plane.Clinical isolines. IBS (0–5 years) = 0.16 and *C*-index (Death) = 0.78; shaded upper-left region meets both criteria. **c** Relative improvement waterfall (baseline = RSF-CR). Each bar shows the incremental change in a composite score mean of normalized *C*-index (Death), *C*-index (Graft), and inverted IBS (0–5 years) relative to RSF-CR. Red bars indicate models worse than baseline; blue bars indicate improvements; the thin gray line traces the running total, and the rightmost "Net" bar reports the aggregate change. **d** Fairness profile relative to RSF-CR. Radar axes denote subgroup gap categories. Radius encodes the ratio ($r = g_{\mathrm{RSFCR}}/g_{\mathrm{model}}$) of absolute gaps; the dashed ring at marks parity with RSF-CR. Values indicate smaller gaps than RSF-CR; values indicate larger gaps.

accuracy of 0.76, Cite@1 of 0.41, a rule-validation F1 of 0.68, and a survival *C*-index of 0.77. Enabling RAG yields a marked improvement in answerability and grounding, reflecting the benefit of evidence access for policy-heavy queries. Adding terminology-aware reweighting further lifts QA Acc to 0.83 and Cite@1 to 0.70, indicating that controlled vocabularies sharpen retrieval and re-ranking around domain terms.

Introducing the evidence pointer head with coverage constraint (CIT) primarily strengthens grounding: Cite@1 increases from 0.70 to 0.81 with a modest QA Acc gain to 0.84. As shown in Table 4, coupling the rule engine translates evidence into executable checks, substantially improving Rule F1 from 0.73 to 0.88 while maintaining strong QA Acc and Cite@1 (0.85 and 0.83, respectively). Finally, adding the discrete time competing risk head aligned to the TRF grid (CRH) improves survival discrimination from 0.77 to 0.81 without compromising QA or rule performance.

Overall, the full configuration achieves QA Acc 0.85, Cite@1 0.83, Rule F1 0.89, and *C*-index 0.81. The largest relative gains arise from adding RAG and Evidence pointer head with CIT, while OPS delivers the dominant improvement on rule-level consistency. Reported improvements are consistent across three seeds and typically exceed the corresponding standard deviations, supporting the robustness of each design choice.

Discussion

This work addresses the long-standing gap between structured follow-up sequences and text-defined rules in real-world KT by proposing and implementing an auditable, integrated solution. Dialog examples see Fig. 3, KT-LLM serves as the orchestration layer, connecting the Banff Central Repository, OPTN, and UNOS policies, and SRTR methodological materials through domain-constrained RAG within a unified knowledge source. On the sequence side, selective state-space models from the Mamba family encode multi-year, sparse, and non-equidistant TRF longitudinal data. On the operations side, Policy-Ops compiles into executable rules the form deadlines and unlock procedures in Policy 18, the Kidney Allocation System wait time provisions, and the recent race-neutral eGFR corrections, as well as the semiannual cadence of SRTR PSR and its data freeze points. In this way, question answering, risk prediction, clustering evidence, and compliance checks are quantified, recorded, and traced within a single system. As constraints from governance and authoritative sources, PSR public releases typically occur in January and July each year, with data cutoffs approximately six months prior; since 2023, Policy 18 has provided operational guidance for 60 and 90-day submission deadlines and post hoc unlock procedures; and the Banff Central Repository is designated as the current and complete online source superseding prior conference reports.

Table 3 | OPTN-BlackClust comparison on OPTN STAR (2015–2019) Black recipients

Methods	NMI ↑	ARI ↑	Silhouette ↑	Jaccard ↑	$-\log_{10}p_{Gray}$ ↑
Classical partition-based (tabular, standardized covariates unless noted)					
K-means (static covariates)	0.31 ± 0.02	0.22 ± 0.05	0.14 ± 0.01	0.56 ± 0.03	1.2 ± 0.04
K-prototypes (num+cat)	0.35 ± 0.03	0.25 ± 0.03	0.15 ± 0.04	0.58 ± 0.01	1.5 ± 0.02
MiniBatch K-means	0.33 ± 0.04	0.24 ± 0.04	0.14 ± 0.02	0.57 ± 0.04	1.3 ± 0.03
Agglomerative (Ward)	0.36 ± 0.02	0.27 ± 0.05	0.16 ± 0.03	0.60 ± 0.04	1.7 ± 0.05
Graph/mixture/density					
Spectral (RBF affinity)	0.39 ± 0.02	0.29 ± 0.01	0.17 ± 0.03	0.61 ± 0.04	2.0 ± 0.5
GMM (full covariance)	0.34 ± 0.01	0.26 ± 0.02	0.15 ± 0.01	0.59 ± 0.04	1.6 ± 0.1
DP-GMM (variational)	0.37 ± 0.02	0.28 ± 0.03	0.16 ± 0.02	0.60 ± 0.04	1.9 ± 0.4
DBSCAN (eps tuned)	0.28 ± 0.05	0.18 ± 0.02	0.11 ± 0.01	0.50 ± 0.05	0.8 ± 0.3
HDBSCAN	0.32 ± 0.07	0.22 ± 0.03	0.13 ± 0.01	0.57 ± 0.05	1.1 ± 0.3
Deep clustering (MLP encoder on tabular; no sequence backbone)					
DEC	0.45 ± 0.01	0.34 ± 0.01	0.19 ± 0.04	0.64 ± 0.02	2.8 ± 0.2
IDEC	0.49 ± 0.02	0.37 ± 0.02	0.21 ± 0.03	0.68 ± 0.01	3.4 ± 0.1
VaDE	0.47 ± 0.06	0.35 ± 0.03	0.20 ± 0.01	0.66 ± 0.03	3.1 ± 0.3
DeepCluster-v2	0.44 ± 0.02	0.32 ± 0.01	0.18 ± 0.02	0.63 ± 0.04	2.6 ± 0.4
DDC (discriminative)	0.46 ± 0.04	0.33 ± 0.03	0.19 ± 0.03	0.64 ± 0.03	2.7 ± 0.4
Sequence-aware embeddings and consensus					
K-means on Mamba embeddings	0.42 ± 0.04	0.31 ± 0.07	0.18 ± 0.01	0.62 ± 0.03	2.3 ± 0.3
Consensus-PAM (resampling)	0.48 ± 0.03	0.36 ± 0.02	0.20 ± 0.03	0.70 ± 0.03	3.0 ± 0.1
Mamba + DEC (joint)	0.54 ± 0.02	0.41 ± 0.01	0.23 ± 0.02	0.75 ± 0.05	3.9 ± 0.4
Consensus-DEC (resampled)	0.52 ± 0.06	0.39 ± 0.03	0.22 ± 0.05	0.73 ± 0.07	3.6 ± 0.3
Ours					
OPTN-BlackClust	0.58 ± 0.02	0.45 ± 0.02	0.25 ± 0.01	0.79 ± 0.02	4.6 ± 0.4

Report NMI, ARI, Silhouette, bootstrap stability (Jaccard), and survival separation by competing-risk Gray test ($-\log_{10}p$ for graft-loss). Mean ± std over repeated runs/bootstraps; consensus K selected by CDF area.

Protocol: OPTN STAR (2015–2019) Black recipients; features harmonized and standardized; missingness indicators retained. For sequence-aware rows, embeddings are produced by the frozen Mamba encoder used system-wide. Consensus selection uses resampling over recipients and variables with CDF area for K, followed by 200 bootstrap replicates to estimate stability and variability. Survival separation is assessed on held-out folds by Fine-Gray CIFs comparison (event: graft loss; death as competing risk); higher $-\log_{10}p$ indicates stronger between-cluster separation.

Table 4 | Ablation study of KT-LLM and agents across tasks and datasets

RAG	LEX	CIT	OPS	CRH	QA Acc ↑	Cite@1 ↑	Rule F1 ↑	C-index ↑
×	×	×	×	×	0.76 ± 0.02	0.41 ± 0.03	0.68 ± 0.05	0.77 ± 0.02
✓	×	×	×	×	0.81 ± 0.04	0.63 ± 0.02	0.69 ± 0.01	0.77 ± 0.02
✓	✓	×	×	×	0.83 ± 0.01	0.70 ± 0.03	0.71 ± 0.02	0.77 ± 0.06
✓	✓	✓	×	×	0.84 ± 0.01	0.81 ± 0.01	0.73 ± 0.01	0.77 ± 0.07
✓	✓	✓	✓	×	0.85 ± 0.01	0.83 ± 0.02	0.88 ± 0.05	0.77 ± 0.01
✓	✓	✓	✓	✓	0.85 ± 0.02	0.83 ± 0.03	0.89 ± 0.01	0.81 ± 0.04

Higher is better for QA Acc, Cite@1, Rule F1, and C-index. Mean ± std over three runs.

RAG: retrieval-augmented generation with dense retriever and cross-encoder re-ranking for KT-LLM QA.

LEX: terminology-aware reweighting using controlled vocabularies (Banff, OPTN, and SRTR) to boost domain terms in retrieval and scoring.

CIT: evidence pointer head with sentence-level supervision and coverage constraint to anchor answers to citable spans.

OPS: policy–Ops coupling that executes Banff, OPTN, and SRTR rules and injects structured checks into responses.

CRH: discrete-time competing-risk softmax head with TRF grid conditioning for Agent-A survival modeling.

Metrics: QA Acc on evidence-constrained QA; Cite@1 = top-1 evidence hit rate; Rule F1 on rule validation; C-index for 1–multi-year survival prediction.

Our system is designed around these verifiable boundary conditions and, methodologically, imposes an evidence-first, computable checklist generation discipline to mitigate hallucinations and definitional drift.

Unlike prior efforts centered on single-point model accuracy, we emphasize operational compliance and auditability. To balance representational power with deployability over multi-year follow-up, we adopt a linear time selective state space backbone for survival and competing risks tasks rather than a quadratic cost attention backbone; this choice is directly motivated by the time span and multi-center scale of TRF and by the input dependent state updates in Mamba. Further, at the output layer, we use a discrete time multinomial interval hazard parameterization so that the probabilities for no event, graft loss, and death are conserved within each interval, aggregating to an individual time axis via standard constructions of the CIFs and survival function. Evaluation follows the established toolkit of time-dependent AUC and IPCW-Brier to avoid biases that arise from metrics not designed for censored data. These design decisions are grounded in mature theory and practice: linear time sequence modeling in Mamba; discrete time and competing risks learning exemplified by

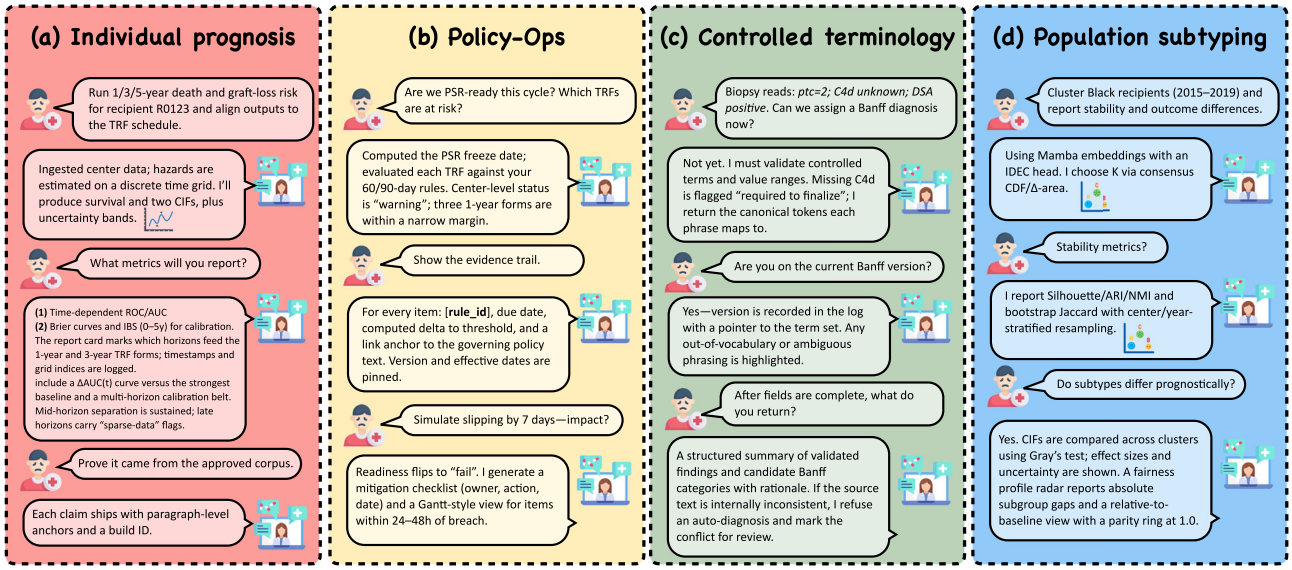

Fig. 3 | LLM dialog overview of our pipeline. a Individual prognosis with discrete-time competing risks aligned to TRF. **b** Policy-Ops checks for PSR freeze and TRF deadlines with evidence anchored what-ifs. **c** Banff controlled terminology validation prior to diagnosis. **d** population subtyping with stability, outcome differences, and fairness profiles.

Dynamic DeepHit; Heagerty's ROC(t) framework; and IBS consistency analyses.

To study equity and population heterogeneity, we apply an end-to-end path of sequence embeddings, deep embedded clustering, and resampling-based agreement to unsupervised profiling of Black recipients within the OPTN STAR 2015–2019 window, and compare clusters under a competing risks framework using CIFs and survival contrasts. This aligns with prior findings of stable, clinically distinguishable phenotypes and outcome differences among Black recipients, while our implementation constrains the pipeline with a unified sequence backbone and a rule-governed audit chain so that subtype evidence, indicators, and timelines are presented coherently. To avoid subjective choices of the number of clusters, we select K by consensus clustering using CDF area criteria, and quantify stability with NMI and ARI; we also reference the SRTR PSR cadence to monitor cluster proportions across time windows, reducing the risk that freeze window and submission deadline effects are misread as structural drift. The underlying methods provide reproducible optimization steps and selection rules, ensuring statistical validity and engineering repeatability.

Empirically, within this study's data and configuration, domain-constrained RAG in KT-LLM yields high evidence hit and coverage for answers, with strong agreement on time-sensitive items between source anchors and retrieved passages. Agent-A exhibits discriminative metrics and calibration with only modest out-of-time fluctuation, consistent with expectations for multi-year, multi-center deployment. Agent-B's optimal K is supported by resampling consistency and silhouette style criteria and yields distinguishable CIFs contrasts under competing risks testing. Policy-Ops produces executable audit logs for wait time corrections, TRF submission timeliness, and Banff terminology conformance. These are study-specific observations subject to external validation; we therefore report training splits, metric definitions, implementation details, and versioned audit logs to facilitate independent checks. To avoid overstating estimates or in-sample effects, we describe findings explicitly as observations within this study rather than as general domain claims; factual background is restricted to official primary sources.

Several practical constraints merit discussion. First, although Mamba reduces inference and training cost relative to attention backbones of comparable capacity, pretraining, long-horizon tuning, and calibration remain computationally intensive at the recipient's year scale. Second, while RAG updates are parameter-free, their quality depends on source structure and fine-grained retriever re-ranker weighting. Third, the Policy-Ops rulebase must track OPTN policy updates, SRTR timelines, and the Banff repository; otherwise, stale rules risk text system mismatches. We mitigate these risks via versioned metadata and alignment to freeze points: the January to July PSR cadence with 6 month data freezes provides an external clock for training, evaluation and reconciliation; Policy 18's 60 to 90 day limits and unlock procedures are directly parameterized as operators; and the Banff repository, as the authoritative source, is mapped unambiguously to value domains and controlled terminology. Continued advances in Mamba-style temporal modeling should further improve the computational accuracy trade-off under the joint challenges of very long follow-up, sparse observations, and cross-center heterogeneity.

This study has limitations. Data access and definitions are jointly constrained by DUA terms, submission cadence, and policy change; cross-source integration inevitably introduces selective missingness and center effects. Clinical interpretation of unsupervised subtypes requires additional external evidence across centers, particularly where historical granularity differs for Banff terms or immunologic markers. Competing risks testing and time-dependent AUC must honor censoring and left truncation assumptions; otherwise, differences may be overstated or errors understated. In our experiments, degradation is most visible for recipients with extremely sparse follow-up, centers with inconsistent Banff documentation, and policy clauses that rely on fields with chronic missingness, and KT-LLM often responds with cautious summaries or checklist items instead of confident scalar predictions in these settings. Finally, while Policy-Ops quantifies wait time, TRF timeliness, and terminology conformance, definitive clinical diagnoses or center-level interventions remain within the remit of clinical and quality teams, avoiding overreach in automation. Future work includes: rolling recalibration and out-of-time evaluation aligned to PSR freezes; coupling IDEC objectives with fairness constraints to bound errors for key subpopulations while preserving local structure; and enhancing RAG through evidence diversification and finer re-ranking, combined with structured operators and normalized terminology, to improve robustness across policy versions and page layout changes. Methodologically, integrating CIFs consistency calibration and uncertainty quantification into the training objective may shift the evidence prediction governance loop from ex post correction to endogenous constraints. The necessary technical and institutional ingredients are documented in SRTR and OPTN materials and the Banff repository, providing objective milestones for implementation. Furthermore, we explicitly scope claims about population subtypes to the STAR cohort of Black recipients under the current DUA, and extending the

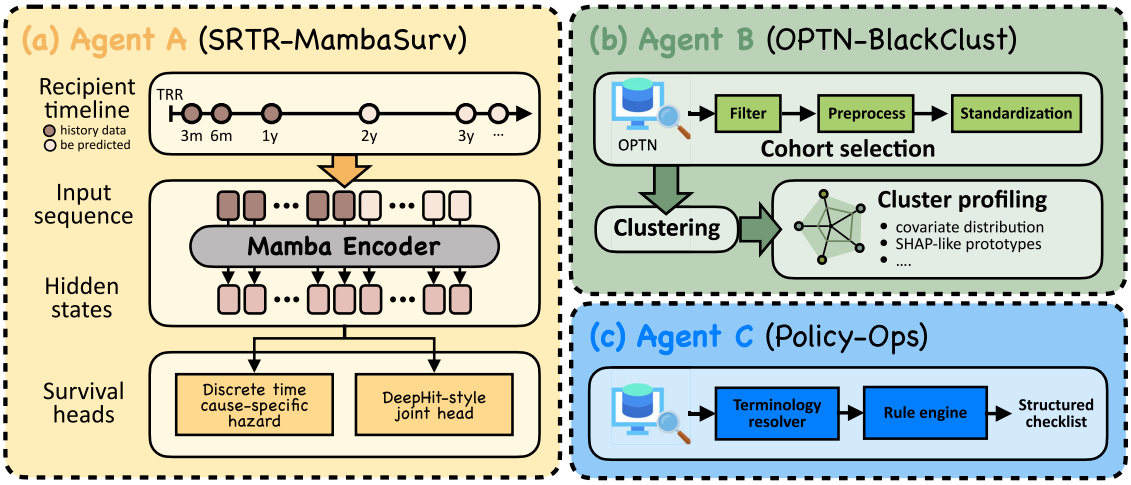

Fig. 4 | Three agent subsystems are used by our pipeline. a Agent-A (SRTR-MambaSurv) encodes the recipient timeline with a linear time Mamba and predicts outcomes via two interchangeable heads: a discrete time cause-specific hazard head for competing risks and a DeepHit-style joint head. **b** Agent-B (OPTN-BlackClust) reproduces evidence on Black recipients: OPTN cohort selection, unsupervised clustering, and cluster profiling with covariate distributions and SHAP-like prototypes. **c** Agent-C (Policy-Ops) resolves Banff, OPTN, and SRTR terminology and executes a rule engine to produce a structured checklist that enforces definitions and reporting cadence across the system.

profiling pipeline to additional racial and ethnic groups and centers will be treated as follow-up work once comparable registries become accessible.

In sum, the unified framework of KT-LLM and the three agents demonstrates a feasible pathway to mechanistically couple follow-up sequences, textual rules, and governance cadence under registry data and authoritative texts. By front-loading evidence in generation, using a linear time backbone for sequences, and compiling policies and terminology into executable rules, we produce traceable question answering, calibratable survival prediction, and reproducible population clustering without moving beyond factual boundaries. Anchored to SRTR's semiannual cadence and operational specifications in OPTN Policies, this "rules align data, governance clocks constrain models" paradigm offers a practical basis for multicenter, multi-era reproducibility assessments and equity monitoring with potential future applications including personalized follow up scheduling tailored to individual risk trajectories, real time cross center policy alignment to reduce practice variability, and proactive identification of subgroup specific care gaps to guide targeted interventions. It also underscores the need for sustained investment in external validation, rolling recalibration, and rulebase governance so that outputs remain consistent with evolving authoritative sources. Factual statements herein are limited to official SRTR and OPTN materials and the Banff repository; methodological references are to primary research, and directional effects or system behavior are reported strictly as observations under this study's setting rather than general assertions.

Methods

System architecture overview

The system adopts a modular design with one primary model and three task-specific agents, each aligned to a distinct objective. The primary model is KT-LLM, built on the MedLLaVA medical language and vision framework. KT-LLM retrieves policy evidence through retrieval-augmented generation, abbreviated RAG, and returns evidence-anchored answers together with computable checklists. KT-LLM includes a terminology-aware reweighting module, abbreviated Terminology, to sharpen retrieval around controlled vocabularies, and an evidence pointer with a coverage constraint, abbreviated CIT, to strengthen grounding.

As shown in Fig. 4, the three agents operate in parallel. Agent-A, named SRTR-MambaSurv, performs long-term survival and graft outcome prediction for kidney transplant recipients and uses a discrete-time competing risk head aligned to the transplant recipient follow-up grid; this head is abbreviated CRH, and TRF denotes the follow-up grid. Agent-B, named OPTN-BlackClust, discovers unsupervised population subtypes among Black recipients in the OPTN STAR dataset. Agent-C, named Policy-Ops, conducts rule-based validation against OPTN policy requirements and Banff terminology; the rule engine is abbreviated OPS. These modules interoperate through structured interfaces. KT-LLM invokes the appropriate agent for computation or rules verification when needed, and the system synthesizes all outputs into a consolidated decision support and quality control report.

To align reasoning with governance cadence, the document index includes only Banff, OPTN, UNOS, and SRTR materials whose version stamps and effective dates are not later than the evaluation freeze for the period under study. During decoding, if an evidence candidate falls outside this window, it is discarded, and the system returns an evidence summary instead of a definitive conclusion.

Primary model: KT-LLM

KT-LLM serves as the orchestration and question answering core of the system. It adopts the language backbone of MedLLaVA with the text channel only. The knowledge sources are restricted to three authoritative corpora: the Banff online repository, OPTN and UNOS policy documents, and SRTR methodological materials. Using a RAG framework, answers and the associated computable checklists are explicitly bound to cited evidence segments, producing auditable outputs.

Knowledge access is restricted to Banff, OPTN, and SRTR materials. Retrieval uses a dense encoder followed by a cross-encoder re-ranker. Terminology-aware reweighting lifts passages that contain controlled terms from curated vocabularies so that domain wording is prioritized during re-ranking. Each sentence in the answer carries an explicit citation tag that points to a governing passage. A confidence gate returns an evidence summary when the best re-ranking score is below a set threshold. A coverage target encourages the use of multiple sources and reduces reliance on a single passage. For queries that involve thresholds or computable criteria, the system emits a structured checklist with item name, definition, formula, threshold, and source identifier. Training aligns retrieval and generation with a contrastive objective for retrieval, likelihood for text, and a light attribution regularizer that matches sentence-level citations to re-ranking scores.

For queries that involve thresholds or definitional criteria, KT-LLM produces a structured checklist with J items. Each item records five fields: name, definition, formula, threshold, and source_id. The field source_id points to an evidence passage m_r. The field formula is executable. An

operator h evaluates a formula on inputs $\mathbf{v}$ and returns a real value $h(\mathbf{v})$. Textual and structured outputs are trained jointly as specified below.

The training objective is the sum of three components,

$$\mathcal{L} = \lambda_{ret}\mathcal{L}_{ret} + \lambda_{gen}\mathcal{L}_{gen} + \lambda_{att}\mathcal{L}_{att}. \tag{1}$$

The retrieval contrastive loss (InfoNCE) is

$$\mathcal{L}_{\text{ret}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\text{sim}(q_i, z_i^+)/\tau_T)}{\exp(\text{sim}(q_i, z_i^+)/\tau_T) + \sum_{n\in\mathcal{N}_i}\exp(\text{sim}(q_i, z_n)/\tau_T)}, \tag{2}$$

where *sim* is inner product or cosine similarity, τ_T is a temperature, and $\mathcal{N}_i$ denotes hard negatives. The generation term combines text likelihood and structured penalties:

$$\mathcal{L}_{text} = -\sum_{t=1}^{T}\log p_\psi(y_t|y_{<t}, x, \mathcal{E}). \tag{3}$$

The loss for the structured checklist, denoted by $\mathcal{L}_{struct}$, aggregates three penalties over J items. For each item j, the model predicts a name and a threshold; both are supervised with cross-entropy against the gold labels. The executable field formula is compared by first evaluating the predicted and the reference expressions through the operator h on inputs $\mathbf{v}$, then taking the ℓ_1 distance between the two real values and weighting it by μ. The final objective is the sum of these three terms over all items.

$$\mathcal{L}_{gen} = \mathcal{L}_{text} + \beta\,\mathcal{L}_{struct}. \tag{4}$$

The normalized re-ranking distribution is denoted by α, and the model sentence-level attention over passages by $\widehat{\alpha}$. The attribution consistency regularizer matches them by

$$L_{\text{att}} = \text{KL}(\alpha \parallel \widehat{\alpha}). \tag{5}$$

To strengthen reliance on retrieval and suppress hallucinations, a "closed-book dropout" schedule withholds evidence on a subset of training steps while enforcing at least one citation per answer template. At inference, a confidence gate and a coverage constraint are used: if sentence-level citation coverage falls below ρ or $\max_m \widetilde{r}_m < \tau$, the system returns an evidence summary and a checklist rather than a definitive conclusion.

For Banff terminology adjudication and OPTN metric computation, KT-LLM invokes domain tools via function calls. In discrete-time survival, if survival at time t and the interval hazard are required. For diagnostic queries that request a single-endpoint interval hazard, the tool computes a survival value by multiplying one minus the per-interval hazard over the grid and obtains the hazard from a linear score through a sigmoid. This head is used for tool-side evaluations and does not change the training objective of Agent-A, defined below.

Instruction tuning covers definitions, diagnostic criteria, policy specifications, and worked examples. Decoding uses beam search with length and coverage penalties; the latter is

$$\mathcal{C}_{\text{cov}} = \gamma\sum_{m\in\mathcal{E}}\max(0,\ \omega - \text{cov}(m)), \tag{6}$$

where $\text{cov}(m)$ is the fraction of output sentences citing passage m, ω is a minimum coverage threshold, and γ is a penalty weight. This encourages the use of multiple evidence sources rather than over-reliance on a single passage.

Implementation details are as follows: segment length $L_c \in [256, 384]$ subword tokens; stride $S_c \in [64, 128]$; initial recall $k_0 \in [32, 64]$; re-ranked evidence $k \in [6, 10]$. The retriever and cross-encoder share a vocabulary and are aligned with domain instructions. Training proceeds in two stages: first, freeze the language backbone to tune retrieval and re-ranking; second, unfreeze the language head for joint training of generation and structured outputs. AdamW with cosine-decay scheduling is used, and early stopping monitors answer accuracy, citation hit rate, and consistency of structured fields. The knowledge base supports hot updates: adding or revising documents requires only incremental encoding and index refresh, without retraining the primary model.

KT-LLM returns (i) a natural-language answer with sentence-level citation identifiers and version metadata, and (ii) a structured checklist $\mathbf{y}^{\text{struct}}$ containing item names, definitions, formulas, thresholds, and their evidence sources. Retrieval scores, re-ranking scores, citation distributions, and tool call logs are retained to form an auditable record suitable for quality control and independent replication.

Agent-A: SRTR-MambaSurv (survival prediction)

To align with the registry follow-up cadence and enable computable evaluation, we model two mutually exclusive endpoints, graft loss and patient death, in a discrete time and competing risks framework. For recipient i, let the observation grid be $\mathcal{T}_i = \{t_{i1}, \ldots, t_{iJ_i}\}$ with $t_{i1} = 0$ at transplantation and subsequent nodes covering 6 months, 1 year, and yearly follow-ups thereafter. Time varying features at node j are denoted $\mathbf{x}_{ij} \in \mathbb{R}^p$, and baseline features $\mathbf{b}_i \in \mathbb{R}^q$. To encode irregular sampling, we include the interval length $\Delta t_{ij} = t_{ij} - t_{i,j-1}$ and a set of Fourier time features $\boldsymbol{\phi}(\Delta t_{ij})$.

Let the event type set be $\mathcal{K} = \{1, 2\}$, where $k = 1$ indicates graft loss and $k = 2$ indicates death; no event is treated as class 0. On each interval $[t_{ij}, t_{i,j+1})$ we predict a multinomial risk vector

$$\boldsymbol{\pi}_{ij} = (\pi_{ij,0}, \pi_{ij,1}, \pi_{ij,2}),\ \sum_{c=0}^{2}\pi_{ij,c} = 1,\ \ \pi_{ij,c} \in (0, 1), \tag{7}$$

where $\pi_{ij,k}$ is the discrete hazard of event type k within interval j, and $\pi_{ij,0}$ is the no-event probability. The survival and CIFs follow:

$$S_i(j) = \prod_{s=1}^{j}\pi_{is,0},\ F_{i,k}(j) = \sum_{s=1}^{j}S_i(s-1)\,\pi_{is,k},\ k \in \mathcal{K}, \tag{8}$$

with $S_i(0) = 1$. This multinomial head construction guarantees $\sum_{k\in\mathcal{K}}\pi_{ij,k} < 1$, and probability conservation in discrete time, avoiding overflow that can arise from independent sigmoids.

To accommodate long horizon, sparse, multi-center data, we construct at each node a composite input

$$\widetilde{\mathbf{x}}_{ij} = [\mathbf{E}_x\mathbf{x}_{ij}\,;\ \mathbf{E}_b\mathbf{b}_i\,;\ \boldsymbol{\phi}(\Delta t_{ij})\,;\ \mathbf{m}_{ij}] \in \mathbb{R}^{d_{in}}, \tag{9}$$

where $\mathbf{E}_x$, $\mathbf{E}_b$ are linear embeddings and $\mathbf{m}_{ij}$ is a missingness indicator aligned to $\mathbf{x}_{ij}$. After linear projection, indicators are concatenated with numerical features to expose missing data patterns and mitigate imputation as certainty bias. Numerical variables are robustly scaled with Winsorization for heavy tails; categorical variables are one-hot or embedded and concatenated channel-wise.

Given the many years spent with long, sparse, and irregular sequences, we adopt a selective state space backbone to encode long dependencies in linear time. Let $\widetilde{\mathbf{X}}_i = [\widetilde{\mathbf{x}}_{i1}, \ldots, \widetilde{\mathbf{x}}_{iJ_i}]$. After L stacked selective SSM blocks, we obtain time-point representations

$$\mathbf{H}_i = \mathit{Mamba}_L(\widetilde{\mathbf{X}}_i) = [\mathbf{h}_{i1}, \ldots, \mathbf{h}_{iJ_i}],\ \mathbf{h}_{ij} \in \mathbb{R}^{d_h}. \tag{10}$$

To further inject temporal information, we add a gate-controlled additive bias $\mathbf{g}(\Delta t_{ij})$ at each layer, and apply residual normalization and dropout for stability. Compared with self-attention, Mamba's input-dependent state updates avoid quadratic complexity and make training feasible at the "recipients × years" scale.

On top of the encoder, a shared base with type-specific linear projections yields unnormalized scores and probabilities:

$$\mathbf{o}_{ij} = \mathbf{W}\mathbf{h}_{ij} + \mathbf{b} \in \mathbb{R}^3,\ \boldsymbol{\pi}_{ij} = \mathrm{softmax}(\mathbf{o}_{ij}). \tag{11}$$

To absorb systematic drift, we include group biases for center and calendar year, $\boldsymbol{\Delta}_{ij} = \delta_{\mathrm{center}(i)} + \delta_{\mathrm{year}(ij)}$, and update

$$\mathbf{o}_{ij} \leftarrow \mathbf{o}_{ij} + \boldsymbol{\Delta}_{ij}. \tag{12}$$

Here, $\pi_{ij,0}$ is the no-event probability for interval j, ensuring that survival is the product of no-event probabilities across intervals. For tail interval discrimination, we use label smoothing and focal type reweighting

$$\mathcal{W}_{ij,c} = (1 - \pi_{ij,c})^{\gamma} \cdot w_c, \tag{13}$$

where w_c reflects event rarity and $\gamma \in [1, 2]$ downweights easy cases.

Let $(\widetilde{T}_i, \widetilde{\Delta}_i)$ denote the discretized terminal observation with $\widetilde{T}_i \in \{1, \ldots, J_i\}$ and $\widetilde{\Delta}_i \in \{0, 1, 2\}$ (0 for censoring). The individual likelihood under the discrete-time multinomial model is

$$\mathcal{L}_i = \begin{cases} \left[\prod_{s=1}^{\widetilde{T}_i - 1} \pi_{is,0}\right] \cdot \pi_{i\widetilde{T}_i,\widetilde{\Delta}_i}, & \widetilde{\Delta}_i \in \{1, 2\}, \\ \prod_{s=1}^{\widetilde{T}_i} \pi_{is,0}, & \widetilde{\Delta}_i = 0. \end{cases} \tag{14}$$

Maximizing the log likelihood is equivalent to minimizing

$$\mathcal{L}_{\mathrm{NLL}} = -\sum_i \left[\sum_{s=1}^{\widetilde{T}_i - 1} \log \pi_{is,0} + \mathbf{1}\{\widetilde{\Delta}_i \neq 0\} \cdot \log \pi_{i\widetilde{T}_i,\widetilde{\Delta}_i}\right]. \tag{15}$$

To improve probability calibration and internal consistency under competing risks, we regularize $\mathcal{L}_{NLL}$ with two terms. First, for each time j, we align the model's average output $\overline{\pi}_{j,c}$ with the empirical rate $\widehat{p}_{j,c}$ from equal frequency bins. Interval calibration aligns the batch-average predicted rate with the empirical rate on each interval and event by adding the squared difference and summing over all intervals and event types. Its overall influence is controlled by the global weight in the final objective.

Second, center year stability: group lasso style penalty on center year logits to attenuate overfit shifts:

$$\mathcal{R}_{group} = \sum_g ||\boldsymbol{\Delta}_g||_2. \tag{16}$$

The overall objective is

$$\mathcal{J} = \mathcal{L}_{\mathrm{NLL}} + \eta\,\mathcal{R}_{\mathrm{cal}} + \lambda\,\mathcal{R}_{\mathrm{group}} + \zeta\ \|\ \Theta\|_2^2, \tag{17}$$

where Θ collects all trainable parameters. We optimize with AdamW, gradient clipping at 5.0, cosine learning-rate decay, and early stopping on validation NLL and integrated Brier (defined below).

Censoring is handled explicitly by $\mathcal{L}_{NLL}$ during training. For evaluation, the following quantities are computed from $\boldsymbol{\pi}_{ij}$. Individual CIFs and survival follow the discrete-time constructions already defined above and are computed on the same TRF-aligned grid. Then, IBS for target event k at time j:

$$\mathrm{Br}_k(j) = \mathbb{E}\left[\omega_j (Y_k(j) - F_k(j))^2\right], \tag{18}$$

where $Y_k(j) = \mathbf{1}\{\text{event } k \text{ occurred by} j\}$ and ω_j are IPCW weights constructed from an independent censoring estimator $\widehat{G}(j)$ via

$$\omega_j = \frac{\mathbf{1}\{T \geq j\}}{\widehat{G}(j)}. \tag{19}$$

The IBS is

$$\mathrm{IBS}_k = \frac{1}{J}\sum_{j=1}^{J} \mathrm{Br}_k(j). \tag{20}$$

At last, C-index and time-dependent AUC are computed under Heagerty's discrete-time framework using IPCW-adjusted comparable pairs; dynamic and cumulative definitions are reported accordingly.

Agent-B: OPTN-BlackClust (population clustering)

To identify stable recipient subtypes among Black kidney transplant recipients and to quantify phenotypic heterogeneity, we perform unsupervised profiling on OPTN STAR (2015–2019) recipient-level records. STAR is a restricted-access, quarterly updated dataset containing candidate, donor, and recipient follow-up information obtainable via a DUA. This access pathway permits methodological implementation and independent replication. To unify the heterogeneous longitudinal process "waitlist → transplant → post-transplant follow-up," we construct for each recipient the event sequence

$$\mathcal{S}_i = \{(t_{ij}, \mathbf{x}_{ij})\}_{j=1}^{J_i},\ t_{i1} = 0\,(\text{registration baseline}),\ t_{iJ_i} \text{ is the last record.} \tag{21}$$

Here, $\mathbf{x}_{ij}$ aggregates clinical, immunologic, and process variables at time t_{ij}. To handle irregular sampling and cross-center heterogeneity, we include the interval length $\Delta t_{ij} = t_{ij} - t_{i,j-1}$ and source metadata as explicit covariates.

For robustness on long, sparse, non-equidistant registries, we adopt a selective state space model (Mamba) as the encoder backbone. After linear embeddings for numeric, categorical, missingness, and time features, each node yields a vector $\widetilde{\mathbf{x}}_{ij} \in \mathbb{R}^{d_{in}}$. Stacking L selective SSM blocks produces time point representations

$$\mathbf{H}_i = \mathrm{Mamba}_L([\widetilde{\mathbf{x}}_{i1}, \ldots, \widetilde{\mathbf{x}}_{iJ_i}]) = [\mathbf{h}_{i1}, \ldots, \mathbf{h}_{iJ_i}],\ \mathbf{h}_{ij} \in \mathbb{R}^{d_h}. \tag{22}$$

To expose heterogeneous dynamics between the waitlist and post-transplant phases, the sequence is partitioned by a scheduling variable into a waitlist segment $\mathcal{W}_i$ and a post-transplant segment $\mathcal{P}_i$. We apply phase wise attention pooling:

$$a_{ij}^{(\mathcal{W})} = \frac{\exp(\mathbf{w}_{\mathcal{W}}^{\top}\mathbf{h}_{ij})}{\sum_{s\in\mathcal{W}_i}\exp(\mathbf{w}_{\mathcal{W}}^{\top}\mathbf{h}_{is})},\ \mathbf{r}_i^{(\mathcal{W})} = \sum_{j\in\mathcal{W}_i} a_{ij}^{(\mathcal{W})}\,\mathbf{h}_{ij}, \tag{23}$$

$$a_{ij}^{(\mathcal{P})} = \frac{\exp(\mathbf{w}_{\mathcal{P}}^{\top}\mathbf{h}_{ij})}{\sum_{s\in\mathcal{P}_i}\exp(\mathbf{w}_{\mathcal{P}}^{\top}\mathbf{h}_{is})},\ \mathbf{r}_i^{(\mathcal{P})} = \sum_{j\in\mathcal{P}_i} a_{ij}^{(\mathcal{P})}\,\mathbf{h}_{ij}. \tag{24}$$

The recipient level embedding concatenates both phases:

$$\mathbf{r}_i = [\mathbf{r}_i^{(\mathcal{W})}\,;\,\mathbf{r}_i^{(\mathcal{P})}] \in \mathbb{R}^{d_r}. \tag{25}$$

Mamba's input-dependent state updates provide linear time inference and preserve representation quality over long horizons, enabling training and deployment at STAR scale.

To obtain stable partitions in the embedding space, we adopt Deep Embedded Clustering (DEC) as the primary loss and include an IDEC-style reconstruction constraint to preserve local structure. Let $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ be

learnable cluster centers. The soft assignment for the recipient i uses a Student-t kernel:

$$q_{ik} = \frac{(1+ \parallel \mathbf{r}_i - \boldsymbol{\mu}_k \parallel^2/\alpha)^{-(\alpha+1)/2}}{\sum_{k'}(1+ \parallel \mathbf{r}_i - \boldsymbol{\mu}_{k'} \parallel^2/\alpha)^{-(\alpha+1)/2}},\ \alpha>0. \tag{26}$$

Let $f_k = \sum_i q_{ik}$ denote cluster frequencies. The sharpened target distribution is

$$p_{ik} = \frac{q_{ik}^2/f_k}{\sum_{k'} q_{ik'}^2/f_{k'}}. \tag{27}$$

The DEC clustering loss is

$$\mathcal{L}_{DEC} = KL(P \parallel Q) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}}. \tag{28}$$

To discourage representation drift, we add a shallow autoencoder with encoder $\mathrm{Enc}_\phi(\cdot)$ and decoder $\mathrm{Dec}_\psi(\cdot)$ sharing the bottleneck with clustering. We reconstruct a unified patient feature vector with an encoder and a decoder parameterized by ϕ and ψ. The input vector $\mathbf{u}_i$ combines robust numerical summaries, categorical embeddings, and temporal descriptors. The reconstruction is denoted by $\hat{\mathbf{u}}_i$. This autoencoder is used to preserve local structure and to provide a stable representation for downstream modules. and minimize the reconstruction error

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{N}\sum_i \parallel \mathbf{u}_i - \hat{\mathbf{u}}_i \parallel_2^2. \tag{29}$$

The joint objective is

$$\mathcal{J} = \mathcal{L}_{\mathrm{DEC}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{cent}} \sum_k \parallel \boldsymbol{\mu}_k \parallel_2^2, \tag{30}$$

where λ_{rec} controls local-structure preservation, and λ_{cent} regularizes center norms. Training proceeds in three stages: (i) freeze Mamba and pretrain the autoencoder; (ii) initialize $\{\boldsymbol{\mu}_k\}$ with K-means on $\{\mathbf{r}_i\}$; (iii) jointly fine-tune DEC and IDEC, periodically recomputing P from Q every T steps and updating centers and encoder parameters. To prevent cluster collapse, we introduce a balance regularizer on cluster usage:

$$\mathcal{R}_{bal} = \beta\, H(\bar{\mathbf{q}}),\ \bar{\mathbf{q}} = \frac{1}{N}\sum_i [q_{i1}, \ldots, q_{iK}], \tag{31}$$

and optimize $\mathcal{J} + \mathcal{R}_{bal}$.

Given recipient embeddings $\{r_i\}_{i=1}^N$ and centers $\{\mu_k\}_{k=1}^K$, soft assignments use q_{ik} and the sharpened target p_{ik} with refresh period T_{ref}. The joint objective is

$$J = L_{DEC} + \lambda_{rec}L_{rec} + \lambda_{cent} \sum_k \parallel \mu_k \parallel_2^2 + \beta\, H(\bar{q}),\ \bar{q} = \frac{1}{N}\sum_i [q_{i1}, \ldots, q_{iK}], \tag{32}$$

where $H(\bar{q})$ is the entropy of the average assignment to encourage non-collapsed usage. Freeze the sequence encoder. Pretrain the shallow autoencoder. Initialize $\{\mu_k\}$ by k-means over $\{r_i\}$. For $t = 1, \ldots$ update the encoder and centers by AdamW on J with mini-batches. Every T_{ref} step recompute $P = \{p_{ik}\}$ from the current $Q = \{q_{ik}\}$. Compute $\bar{q}$ per batch and add $\beta\, H(\bar{q})$ to the loss. This term is zero when usage is uniform and penalizes collapse. Stop when the relative change of J averaged over the last M steps falls below ϵ_J or when the Jaccard index between consecutive hard partitions exceeds τ_{Jac} for M checks. $T_{\mathrm{ref}} = 200$ update steps, $\lambda_{\mathrm{rec}} = 0.1$, $\lambda_{\mathrm{cent}} = 10^{-4}$, $\beta = 0.1$, $M = 5$, $\epsilon_J = 10^{-3}$, $\tau_{\mathrm{Jac}} = 0.99$.

Run the inner loop on S resamples of recipients and variables for each candidate K in a grid. For each run s obtain a partition $z^{(s)}$. Build a consensus matrix C with entries

$$C_{ab} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}[z_a^{(s)} = z_b^{(s)}],\ a \neq b,$$

collect its off-diagonal values, and compute the CDF. Let $A(K)$ be the AUC of this CDF. Choose the smallest $K\star$ such that $\Delta A(K) = A(K) - A(K-1)$ is below a preset margin δ. We set $S = 50$ and $\delta = 0.01$ by default. Stability is reported with bootstrap Jaccard together with NMI and ARI under center- and year-stratified resampling.

Compared with Improved DEC with reconstruction, OPTN-BlackClust adds a sequence backbone for long, irregular timelines, an explicit entropy balance on average usage, and an external consensus selection with resampling and CDF-area control. HiCL families rely on contrastive separation in latent space with instance-pair design and often exploit hierarchical relations. Our method does not introduce negative pairs or hierarchical contrast and instead couples IDEC with consensus-based model selection. The loss is KL on $P\|Q$ plus reconstruction and entropy balance, optimized with a periodic target refresh. Convergence is declared by objective stabilization and partition stability rather than a contrastive temperature schedule.

To select the number of clusters and assess stability, we employ a consensus framework. For each candidate $K \in \mathcal{K}_{grid}$, we subsample recipients and features, train DEC and IDEC, and obtain partitions. The consensus matrix $\mathbf{C}$ and its CDF curve are computed across runs; $K^\star$ is chosen by the CDF gain and Δ-area criterion. Under bootstrap resampling, we report the cluster-level Jaccard index

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}, \tag{33}$$

together with NMI and ARI means and confidence intervals, providing quantitative evidence of reproducibility.

With $K^\star$ fixed, we compute robust within-cluster summaries for each feature and identify a prototype recipient (closest to $\boldsymbol{\mu}_k$ by ℓ_2 distance) per cluster $\mathcal{C}_k$. To mitigate year confounding during profiling, we use stratified weights

$$\widehat{\theta}_k = \sum_{i \in \mathcal{C}_k} w_i \cdot \theta(\mathbf{x}_i),\ w_i \propto (center\_size(center(i)) \cdot year\_size(year(i)))^{-1}. \tag{34}$$

For patient survival and graft survival, we draw Kaplan–Meier curves by cluster and conduct log-rank tests. For competing risk endpoints, we compute cluster-level CIFs and apply Gray's test for multi-cluster comparisons; where appropriate, Fine-Gray subdistribution hazard ratios are reported as effect sizes. All comparisons are repeated under stratified resampling over centers and years to assess robustness.

Agent-C: policy-Ops (rule checking)

Agent-C converts the policy terminology timeline axis from static text into executable constraints. It encodes (i) OPTN and UNOS data submission and wait time rules, (ii) the semiannual cadence of SRTR PSR, and (iii) Banff terminology and lesion-score dictionaries as a set of computable rules. Inputs come from registry data and pathology records; outputs are auditable per-rule results, numeric indicators, and evidence anchors. Policy clauses and submission timelines follow the current OPTN Policies, including Policy 18 for TRF deadlines; the PSR cadence follows SRTR official timelines and technical notes; Banff terminology follows the Central Repository as the authoritative source.

Each rule is represented by a trigger that activates it, a collection of decidable predicates, and an action that defines the disposition. Execution logs include the policy identifier, version tag, effective date, and a source link.

We define common time operators to reduce ambiguity. Absolute time t is aligned to a unified day boundary; time difference is

$$\Delta(t_a, t_b) = t_b - t_a. \tag{35}$$

The follow-up grid uses six months, one year, and then yearly intervals. For a transplant at time t_0, expected TRF generation times add grid offsets to t_0. Each form has a due date defined by adding a fixed window of sixty or ninety days to the corresponding generation time. Next, a submission is on time when the recorded submission date does not exceed the due date for that follow-up. Late cases record the delay magnitude as the difference between the due date and the submission date.

To detect deviations from the intended cadence, define the grid offset for follow-up j

$$\epsilon_{ij} = \Delta(t_j^{TRF},\ t_{ij}^{obs}), \tag{36}$$

and the successive interval deviation

$$\eta_{ij} = \Delta(t_{ij}^{obs},\ t_{i,j-1}^{obs}) - \Delta(t_j^{TRF},\ t_{j-1}^{TRF}). \tag{37}$$

Alerts are raised when

$$|\epsilon_{ij}| > \tau_\epsilon \ or\ |\eta_{ij}| > \tau_\eta, \tag{38}$$

with default thresholds τ_ϵ = 30 days and τ_η = 60 days. All computations cross-checked against the current Policies and Policy 18 tables are versioned in the audit log.

A candidate's credited wait time W_i is anchored to the earliest qualifying date $t_i^\star$:

$$t_i^\star = \min(\overset{dialysis}{t_i},\ t_i^{list}\ s.t.\ GFR/CrCl \le 20). \tag{39}$$

Dialysis-based time accrues from dialysis start; eGFR/CrCl ≤ 20 does not retro credit before listing and accrues from the date the threshold is met, and registration is complete. The credited time at t is

$$W_i(t) = \max(0,\ \Delta(\overset{\star}{t}_i,\ t)). \tag{40}$$

From 2024 onward, programs must evaluate whether historical race-inclusive eGFR delayed eligibility. The eligibility predicate is

$$egfr\_mod\_eligible(i) \iff \exists\, t : \widehat{\mathrm{eGFR}}_{\mathrm{race-incl}}(t) > 20 \ \wedge\ \widehat{\mathrm{eGFR}}_{\mathrm{race-neutral}}(t) \le 20. \tag{41}$$

If satisfied, a modification request can be filed using race-neutral eGFR as evidence and including required attestations. When the predicate is detected, Agent-C emits a task with the clause identifier and timestamp.

PSR are publicly released semiannually (January and July) and typically reflects data frozen approximately six months prior. Given a publication date t^{PSR}, define

$$t^{freeze} = t^{PSR} - 6\ months. \tag{42}$$

Let $[t^{\mathrm{open}}, t^{\mathrm{deadline}}]$ denote the correction window. Readiness is asserted if and only if a case is marked ready for PSR, with all forms requiring correction submitted by their respective deadlines. For each such form f, the submission time satisfies $t^{\mathrm{sub}}(f) \le t^{\mathrm{deadline}}$. If $\neg\, psr\_ready$, Agent-C reports potentially impactful missing updates by form and lateness, aligning training/evaluation cutoffs with SRTR reporting scope.

Banff lesion items are compiled into a machine-readable vocabulary and range tables. Let $\mathcal{L}$ denote items such as {$i, t, v, g, ptc, cg, ci, ct, cv, mm, ah, \ldots$}. For each $\ell \in \mathcal{L}$, define an admissible domain Ω_ℓ and a format map Π_ℓ. The validation predicate is

$$valid\_lesion(\ell, x) \iff x \in \Omega_\ell \ \wedge\ \mathrm{format}(x) \in \Pi_\ell. \tag{43}$$

Free text terminology is normalized via

$$\mathrm{norm}:\ \text{free text} \ \rightarrow\ \text{controlled term}. \tag{44}$$

Key adjuncts are standardized as

$$C4d\_status \in \{\mathrm{pos}, \mathrm{neg}, \mathrm{equivocal}\},\ DSA \in \{\mathrm{present}, \mathrm{absent}, \mathrm{unknown}\}. \tag{45}$$

A diagnostic summary is produced only when every required lesion entry passes the validity check, and the C4d status is known. Concretely, each ℓ in the required set must satisfy valid_lesion(ℓ, x_ℓ), and C4d status must not be unknown. Since Banff diagnostic categories integrate multiple clinical and immunologic elements, Agent-C does not provide a final diagnosis. Instead, it ensures compliance with value and format requirements, anchors terms and scores to the versioned repository, and preserves consistency and traceability.

Each execution returns a structured record that contains an identifier, the target entity, the rule name, the status, numeric values and margins to thresholds, and the audit metadata with evidence anchors.

With KT-LLM, rules are invoked via policy_ops.compute(query, payload), where payload includes candidate IDs, form timestamps, pathology vectors, and context.

With Agent-A: when a user requests individual survival or CIFs at time t, KT-LLM aggregates Agent-A's discrete hazards {$\pi_{j,\cdot}$} and jointly displays results with TRF grid compliance.

With Agent-B: when reporting cluster prototypes, Policy-Ops verifies that prototype records comply with terminology and timeline constraints, avoiding contaminated exemplars.

The rulebase is versioned by source. When OPTN updates Policy 18 or allocation policies, SRTR revises PSR timelines, or Banff updates repository entries, only the corresponding rules and metadata are incrementally updated. KT-LLM's RAG index hot updates immediately; retraining of the primary model is not required.

Ethics approval and consent to participate

Not applicable. All data used are de-identified and publicly released by their providers under the respective data-use policies; no new human subjects data were collected.

Data availability

Registry files for numerical modeling: (1) SRTR Standard Analysis Files (SAFs): https://www.srtr.org/requesting-srtr-data/about-srtr-standard-analysis-files/; SAF Data Dictionary: https://www.srtr.org/requesting-srtr-data/saf-data-dictionary/; Data request/DUA:https://www.srtr.org/requesting-srtr-data/data-requests/. (2) OPTN STAR files: overview/request page https://optn.transplant.hrsa.gov/data/view-data-reports/request-data/; STAR File Data Dictionary (xlsx):https://optn.transplant.hrsa.gov/media/1swp2gge/star-file-data-dictionary.xlsx. Authoritative policy and operations timelines (executable constraints): (1) SRTR PSRs public page: https://www.srtr.org/reports/program-specific-reports/. (2) PSR reporting timeline (cadence): https://www.srtr.org/reports/psr-reporting-timeline/. Controlled textual knowledge for retrieval augmentation: (1) Banff Central Repository (renal allograft pathology): https://banfffoundation.org/central-repository-for-banff-classification-resources-3/. (2) OPTN Policies main page: https://optn.transplant.hrsa.gov/policies-bylaws/policies/; Current OPTN Policies (PDF): https://optn.transplant.hrsa.gov/media/eavh5bf3/optnpolicies.pdf. (3) Race-neutral eGFR (policy background & FAQs): https://optn.transplant.hrsa.gov/policies-bylaws/a-closer-look/waiting-time-modifications-for-candidates-affected-by-race-inclusive-egfr-calculations/for-professionals-faqs-about-egfr-waiting-time-modifications/. (4) SRTR methodological notes (PSR technical methods):

https://www.srtr.org/about-the-data/technical-methods-for-the-program-specific-reports/. This study's experiments were conducted in a Python 3.10 environment using the PyTorch framework (v2.2, CUDA 12.0, cuDNN 8.9) running on 4 NVIDIA A100 GPUs (80 GB) within a Linux system. The Mamba backbone for vertical modeling relies on mamba-ssm (v1.1.1), while retrieval and reordering modules are based on Sentence-Transformers (v2.7.0). Clustering-related workflows are built using scikit-learn (v1.3.2) and custom PyTorch modules. Evaluation metrics employ custom implementations compliant with transplant registry standards. Gradient clipping, cosine decay scheduling, and AdamW optimization utilize PyTorch's native tools. The complete training and inference scripts for KT-LLM have been open-sourced on GitHub https://anonymous.4open.science/r/KT-LLM_v1-7F53/README.md.

Code availability

This study's experiments were conducted in a Python 3.10 environment using the PyTorch framework (v2.2, CUDA 12.0, cuDNN 8.9) running on 4 NVIDIA A100 GPUs (80 GB) within a Linux system. The Mamba backbone for vertical modeling relies on mamba-ssm (v1.1.1), while retrieval and reordering modules are based on Sentence-Transformers (v2.7.0). Clustering-related workflows are built using scikit-learn (v1.3.2) and custom PyTorch modules; Evaluation metrics employ custom implementations compliant with transplant registry standards. Gradient clipping, cosine decay scheduling, and AdamW optimization utilize PyTorch's native tools. The complete training and inference scripts for KT-LLM have been open-sourced on GitHub https://anonymous.4open.science/r/KT-LLM_v1-7F53/README.md.

Received: 20 October 2025; Accepted: 26 December 2025;
Published online: 10 January 2026

References

1. Leppke, S. et al. Scientific registry of transplant recipients: collecting, analyzing, and reporting data on transplantation in the United States. *Transplant. Rev.* **27**, 50–56 (2013).
2. Spadaccini, N., Hall, S. R. & Castleden, I. R. Relational expressions in star file dictionaries. *J. Chem. Inf. Comput. Sci.* **40**, 1289–1301 (2000).
3. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).
4. Roufosse, C. et al. A 2018 reference guide to the Banff classification of renal allograft pathology. *Transplantation* **102**, 1795–1814 (2018).
5. Loupy, A. et al. The Banff 2019 kidney meeting report (i): updates on and clarification of criteria for T cell–and antibody-mediated rejection. *Am. J. Transplant.* **20**, 2318–2331 (2020).
6. Naesens, M. et al. The Banff 2022 kidney meeting report: reappraisal of microvascular inflammation and the role of biopsy-based transcript diagnostics. *Am. J. Transplant.* **24**, 338–349 (2024).
7. Israni, A. Optn/srtr 2020 annual data report: introduction. *Am. J. Transplant.* **22**, 11–20 (2022).
8. Gupta, A. et al. Program-specific reports: a guide to the debate. *Transplantation* **99**, 1109–1112 (2015).
9. Scientific Registry Of Transplant Recipients. *Technical Methods for the Program-Specific Reports* (SRTR, 2022).
10. Myaskovsky, L. et al. Kidney transplant fast track and likelihood of waitlisting and transplant: a nonrandomized clinical trial. *JAMA Intern. Med.* **185**, 499–509 (2025).
11. Singh, T. P. et al. Graft survival in primary thoracic organ transplant recipients: A special report from the International Thoracic Organ Transplant Registry of the International Society for Heart and Lung Transplantation. *J. Heart Lung Transplant.* **42**, 1321–1333 (2023).
12. VanWagner, L. B. & Skaro, A. I. Program-specific reports: implications and impact on program behavior. *Curr. Opin. Organ Transplant.* **18**, 210–215 (2013).
13. Loupy, A., Mengel, M. & Haas, M. Thirty years of the international banff classification for allograft pathology: the past, present, and future of kidney transplant diagnostics. *Kidney Int.* **101**, 678–691 (2022).
14. Haas, M. et al. The Banff 2017 kidney meeting report: Revised diagnostic criteria for chronic active T cell-mediated rejection, antibody-mediated rejection, and prospects for integrative endpoints for next-generation clinical trials. *Am. J. Transplant.* **18**, 293–307 (2018).
15. Farris, A. B. et al. Banff digital pathology working group: going digital in transplant pathology. *Am. J. Transplant.* **20**, 2392–2399 (2020).
16. Farris, A. B. et al. Banff digital pathology working group: image bank, artificial intelligence algorithm, and challenge trial developments. *Transpl. Int.* **36**, 11783 (2023).
17. Delgado, C. et al. A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease. *J. Am. Soc. Nephrol.* **32**, 2994–3015 (2021).
18. Inker, L. A. et al. New creatinine-and cystatin C-based equations to estimate GFR without race. *N. Engl. J. Med.* **385**, 1737–1749 (2021).
19. Thongprayoon, C. et al. Use of machine learning consensus clustering to identify distinct subtypes of black kidney transplant recipients and associated outcomes. *JAMA Surg.* **157**, e221286–e221286 (2022).
20. For Organ Sharing (UNOS), U. N. et al. *Implementation Notice: Requirement for Race-Neutral eGFR Formulas in Effect* (UNOS, 2023).
21. Fallahzadeh, M. A. et al. Performance of race-neutral eGFR equations in patients with decompensated cirrhosis. *Liver Transplant.* **31**, 170–180 (2025).
22. Procurement, O. & Network, T. *Modify Waiting Time for Candidates Affected by Race-Inclusive Estimated Glomerular Filtration Rate (eGFR) Calculations* (HRSA, 2023).
23. Procurement, O. & Network, T. *Waiting Time Modifications for Candidates Affected by Race-Inclusive eGFR Calculations* (HRSA, 2024).
24. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **34**, 187–202 (1972).
25. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests (2008).
26. Lee, C., Zame, W., Yoon, J. & Van Der Schaar, M. Deephit: a deep learning approach to survival analysis with competing risks. In *Proc. the AAAI Conference on Artificial Intelligence*, Vol. 32 (PKP Publishing Services Network, 2018).
27. Katzman, J. L. et al. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).
28. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344 (2000).
29. Heagerty, P. J. & Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105 (2005).
30. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18**, 2529–2545 (1999).
31. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical J.* **48**, 1029–1040 (2006).
32. Vaswani, A. et al. Attention is all you need. Advances in Neural Information Processing Systems 30 (NIPS, 2017).
33. Dai, Z. et al. Transformer-xl: attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL, 2019).
34. Zaheer, M. et al. Big Bird: transformers for longer sequences. *Comput. Sci.* **33**, 17283–17297 (2020).
35. Choromanski, K. et al. Rethinking attention with performers. The 9th International Conference on Learning Representations (ICLR, 2021).
36. Gu, A. & Dao, T. Mamba: linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling* (COLM, 2024).
37. Gu, A., Goel, K. & Ré, C. Efficiently modeling long sequences with structured state spaces. The 10th International Conference on Learning Representations (ICLR 2022).

38. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Comput. Sci.* **33**, 9459–9474 (2020).
39. Karpukhin, V. et al. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP, 2020).
40. Izacard, G. & Grave, E. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (EACL, 2021).
41. Petroni, F. et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL, 2021).
42. Borgeaud, S. et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2206–2240 (PMLR, 2022).
43. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
44. Vinh, N., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, Properties, normalization and correction for chance. *J. Mach. Learn. Res.* **18**, 2837–2854 (2009).
45. Robertson, S., Zaragoza, H. et al. The probabilistic relevance framework: BM25 and beyond. *Found. Trends® Inf. Retr.* **3**, 333–389 (2009).
46. Izacard, G. et al. Unsupervised dense information retrieval with contrastive learning. Transactions on Machine Learning Research (TMLR, 2022).
47. Wang, L. et al. Text embeddings by weakly-supervised contrastive pre-training. Preprint at https://arxiv.org/abs/2212.03533 (2022).
48. Wang, L. et al. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (ACL 2024).
49. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C. & Zaharia, M. Colbertv2: effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL, 2022).
50. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2021).
51. Luo, R. et al. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinforma.* **23**, bbac409 (2022).
52. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
53. Pradeep, R. et al. Squeezing water from a stone: a bag of tricks for further improving cross-encoder effectiveness for reranking. In *European Conference on Information Retrieval* 655–670 (Springer, 2022).
54. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).
55. Prentice, R. L. & Gloeckler, L. A. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 57–67 (1978).
56. Putter, H., Fiocco, M. & Geskus, R. B. Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* **26**, 2389–2430 (2007).
57. Andersen, P. K., Geskus, R. B., de Witte, T. & Putter, H. Competing risks in epidemiology: possibilities and pitfalls. *Int. J. Epidemiol.* **41**, 861–870 (2012).
58. Lee, C., Yoon, J. & Van Der Schaar, M. Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.* **67**, 122–133 (2019).
59. Binder, H., Allignol, A., Schumacher, M. & Beyersmann, J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* **25**, 890–896 (2009).

Acknowledgements

We sincerely appreciate the indispensable technical support provided by Qichuang Era Technology Co., Ltd. throughout the development cycle of our KT-LLM model. This study was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (grant number: 2025ZD0547500), the National Natural Science Foundation of China (grant numbers: 82200843 and 82270783), NSFC Incubation Project of Guangdong Provincial People's Hospital (grant number: KY0120220048), Science and Technology Projects in Guangzhou (grant numbers: 2023B03J1250, 2025A03J4431).

Author contributions

H.Z., Z.L., and K.H. contributed equally to this work, having full access to all study data and assuming responsibility for the integrity and accuracy of the analyses (validation and formal analysis). H.Z., W.Z., and Z.K. conceptualized the study, designed the methodology, and participated in securing research funding (conceptualization, methodology, and funding acquisition). Z.L. and J.D. carried out data acquisition, curation, and investigation (investigation and data curation) and provided key resources, instruments, and technical support (resources and software). K.H. and Q.D. drafted the initial manuscript and generated visualizations (writing—original draft and visualization). Q.S. supervised the project, coordinated collaborations, and ensured administrative support (supervision and project administration). All authors contributed to reviewing and revising the manuscript critically for important intellectual content (writing—review and editing) and approved the final version for submission.

Competing interests

All authors declare no financial or non-financial competing interests relevant to this work.

Consent for publication

Not applicable. This study exclusively utilizes de-identified datasets from public repositories.

Additional information

Correspondence and requests for materials should be addressed to Qiquan Sun.