

Online Resource 1

Implementation and validation details for Graph-Augmented LLMs with Risk-Adjusted Safety Filtering for Personalized Advice and Medication Recommendation

This supplementary file provides only the implementation and validation details needed to support the main LETTER. It is intentionally limited to selected experimental settings, online inference profiling, routing/advice-consistency control, and safety-threshold selection.

S1 Experimental setup and selected hyperparameters

The system was implemented in PyTorch 2.0.0 with Python 3.8.16 and run on a Xeon Silver 4110 CPU, 264 GB RAM, and an NVIDIA RTX 2080 Ti GPU. The random seed was fixed to 1203. Dataset splits follow the Letter: MIMIC-IV uses the CogNet temporal split, while Diagnose-me uses stratified train/validation/test splits (77.8%/11.0%/11.2%). All medication-recommendation baselines were re-evaluated under the same MIMIC-IV split and metrics rather than taken from the original publications. MIMIC-IV evaluates medication recommendation, whereas Diagnose-me evaluates routing and lifestyle-advice support. Confidence intervals were computed by 1,000-iteration bootstrap resampling on the held-out test set. Selected settings are summarized in Table S1.

Table S1: Selected implementation settings.

Setting	Value
Framework	PyTorch 2.0.0
Python version	3.8.16
Hardware	Xeon Silver 4110 + RTX 2080 Ti
Random seed	1203
<i>Encoder-decoder and graph module</i>	
Embedding dimension d	64
Beam width	4
Maximum generation length	45
Batch size	16
Training epochs	200
Optimizer	Adam
Learning rate η	0.005
Loss weights $(\lambda_1, \lambda_2, \lambda_3)$	(0.3, 0.4, 0.3)
Severity-scaling parameter β	2.0
Ensemble weights (ω_1, ω_2)	(0.6, 0.4)
Safety threshold τ_e	0.15
<i>LLaMA-2 routing/advice module</i>	
Base model	LLaMA-2-13B-chat
Optimizer	AdamW
Learning rate schedule	Cosine, 2,000-step warmup
Weight decay	0.1
Gradient clipping	1.0
Nucleus sampling p	0.95
Temperature τ	0.9
Repetition penalty	1.1
Fine-tuning conversations	179,900

All encoder-decoder hyperparameters were selected on validation data. The embedding dimension was searched over $\{32, 64, 128, 256\}$, the learning rate over $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$, the severity-scaling parameter over $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$, and the beam width over $\{1, 2, 4, 8, 16\}$. Loss weights were selected from 15 combinations with $\lambda_1 + \lambda_2 + \lambda_3 = 1$ in increments of 0.1. The safety threshold τ_e was swept over $\{0.05, 0.10, 0.15, 0.20, 0.25\}$ on a physician-annotated validation set ($N = 100$); full sweep results are reported in Section S5.

S2 Online inference cost and deployment profile

Table S2 separates one-time cached preparation from per-case online inference. The measurements were obtained on the experimental hardware described in Section S1 and are intended to characterize implementation cost rather than clinical deployment performance.

Table S2: Inference profile of LLM-AMR. Offline components are cached, while online components are executed per case or conditionally invoked when advice is generated.

Component	Execution	GPU	Latency	Peak mem.	Throughput
EHR/DDI/GMR graph construction	Offline cached	Opt.	~ 4.2 min	~ 1.8 GB	One-time
Propensity-based risk estimation	Offline cached	No	~ 2.1 min	~ 0.6 GB	One-time
Conversational routing	Online per case	Rec.	~ 180 ms	~ 14 GB ^a	~ 5.5 cases/s
Lifestyle-advice generation, when invoked	Online conditional	Rec.	Output-length dep. ^b	~ 14 GB ^a	Output-length dep.
Adaptive graph fusion + medication scoring	Online med. case	Rec.	~ 38 ms	~ 1.2 GB	~ 26 cases/s
Risk lookup + safety filtering	Online med. case	No	~ 4 ms	~ 0.3 GB	~ 250 cases/s
Medication-rec. path	Online per case	Rec.	~ 224 ms^c	$\sim 14\text{--}15$ GB	~ 4.5 cases/s
Advice/both path	Online conditional	Rec.	~ 224 ms + gen.^{b,c}	$\sim 14\text{--}15$ GB	Output-length dep.

^aMemory for conversational routing and advice generation mainly reflects loading the LLaMA-2-13B-chat backbone; the classifier head adds negligible overhead.

^bAdvice-generation latency depends on output length and decoding settings.

^cThe medication-recommendation path includes routing, graph fusion, medication scoring, and risk filtering.

Note: Offline latency is measured once over the training split, after which graphs and risk tables are cached. Online latency is the mean wall-clock time per case under controlled implementation profiling. Peak memory denotes the maximum observed footprint during each step, not an additive cost across independently executed modules. These measurements are not clinical deployment guarantees.

For operational maintenance, the modules can be monitored and recalibrated separately through periodic audits or drift-triggered updates. Routing drift can be tracked through confidence scores, manual-review rates, and audited routing samples. The EHR and GMR graphs can be refreshed when local prescribing or demographic patterns shift, while DDI and contraindication resources should be updated when external drug knowledge changes. Propensity-based risk tables should be recalibrated when coding practices, patient mix, or medication distributions differ substantially from the training period.

S3 Routing labels and advice-consistency control

The conversational module performs three-way routing with labels *advice-only*, *medication-only*, and *both*. Labels were derived from Diagnose-me annotations, with spaCy NER and an RxNorm drug lexicon (8,000+ entries) used to identify medication mentions during label construction.

- **advice-only**: contains lifestyle or non-pharmacological advice phrases but no medication names,
- **medication-only**: contains medication entities but no advice phrases,
- **both**: contains both advice and medication content.

To further justify the choice of the conversational backbone, we compared LLaMA-2-13B-chat with several lighter alternatives on the same Diagnose-me split. As shown in Table S3, LLaMA-2-13B-chat achieved the strongest routing performance, with 0.913 accuracy and 0.911 Macro-F1. LLaMA-2-7B-chat performed closest to the 13B model, which is expected because it belongs to the same conversational model family but has fewer parameters. ClinicalBERT and BioClinicalBERT provided stronger performance than DistilBERT among the encoder-only baselines, reflecting the benefit of clinical-domain pretraining; however, these BERT-based models support only routing and do not provide the same generative conversational capacity needed for lifestyle-advice support. DistilBERT is the smallest alternative and offers a low-cost routing option, but it showed the largest performance trade-off. Therefore, LLaMA-2-13B-chat was selected as the shared conversational backbone in LLM-AMR because it provides the best routing performance while also supporting non-medicinal lifestyle-advice generation. Smaller encoders remain possible replacements when only task routing is required and deployment resources are limited.

Table S3: Comparison with lighter conversational routing backbones on the same Diagnose-me split. The table reports the role evaluated in this study, not the full capability of each model.

Model	Parameters	Evaluated role	Routing Acc.	Macro-F1
DistilBERT	66M	Routing only	0.804	0.799
ClinicalBERT	110M	Routing only	0.831	0.827
BioClinicalBERT	110M	Routing only	0.843	0.839
LLaMA-2-7B-chat	7B	Routing only	0.871	0.868
LLaMA-2-13B-chat	13B	Routing + lifestyle-advice support	0.913	0.911

Low-confidence routing cases are deferred to manual review when the confidence threshold τ_c is not met:

$$\max_y P(y \mid x, H) < \tau_c = 0.70.$$

For advice-oriented cases, the LLM is restricted to non-pharmacological lifestyle guidance, while medication decisions remain delegated to the structured recommender. When the predicted task is *both*, generated advice is checked against the medication output using a contradiction-control mechanism:

$$P(m_i \mid O_A, x, H) = P(m_i \mid x, H) (1 - \lambda \mathbb{I}(m_i, O_A)), \quad (1)$$

where $\mathbb{I}(m_i, O_A) \in \{0, 1\}$ is a conflict indicator based on SBERT cosine similarity, with threshold 0.75 and penalty weight $\lambda = 0.6$ selected on validation cases to balance missed conflicts against false conflict alarms. Sensitivity checks around these values mainly affected borderline advice-medication pairs, which were flagged for manual review rather than treated as definitive contradictions. Conflicting advice is suppressed or flagged for review. This control is intended to reduce unsafe or contradictory free-text advice rather than guarantee complete elimination of hallucination.

S4 Ablation of graph weighting schemes

To clarify the contribution of the learnable graph design, we compared the full model with fixed and partial graph-weighting variants on the MIMIC-IV test split. As shown in Table S4, the full

learnable design improves the main recommendation-quality metrics, indicating that adaptive weighting across graph sources is more effective than fixed or single-source weighting schemes.

Table S4: Ablation of graph weighting schemes on the MIMIC-IV test split. The full learnable design improves the main recommendation-quality metrics compared with fixed and partial weighting variants.

Model Variant	Jaccard	F1	PRAUC	DDI	MAP
Fixed Uniform ($w = 1/ M $)	0.587	0.741	0.738	0.073	0.785
Fixed Linear Depth ($w = d/d_{\max}$)	0.596	0.753	0.749	0.078	0.793
Attention-based (EHR only)	0.609	0.768	0.758	0.081	0.801
Exponential Severity (DDI only)	0.615	0.774	0.763	0.084	0.806
Multi-Dimensional GMR	0.618	0.777	0.766	0.085	0.808
Full Learnable (Ours)	0.629	0.787	0.775	0.097	0.815

GMR graph construction

The GMR graph uses structured demographic fields from the EHR, mainly sex and age group, together with historical medication records to estimate demographic medication-use and restriction patterns. These signals are used as soft graph weights in the medication representation and are not treated as complete contraindication rules. More specific factors, such as pregnancy, pediatric or elderly-specific warnings, and organ-function restrictions, are reflected only when they are available in structured covariates or external contraindication resources used by the safety-filtering stage.

S5 Safety-threshold selection

The risk-adjusted safety module filters medications using the thresholded safe set

$$\mathcal{M}_s(d_j) = \{m_i \in \mathcal{M} : R(m_i, d_j) < \tau_e\} \setminus \mathcal{M}_c(d_j),$$

where $R(m_i, d_j)$ is the ensemble risk score defined in Letter Eq. 5, and $\mathcal{M}_c(d_j)$ denotes medications contraindicated for diagnosis d_j according to FDA drug labels and the AHFS DI database.

The operating threshold τ_e was selected on a physician-annotated validation set ($N = 100$) by sweeping $\tau_e \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ (see also Section S1). Figure S1 shows the trade-off among F1 score, safe-medication retention, and unsafe-combination exclusion across all evaluated thresholds.

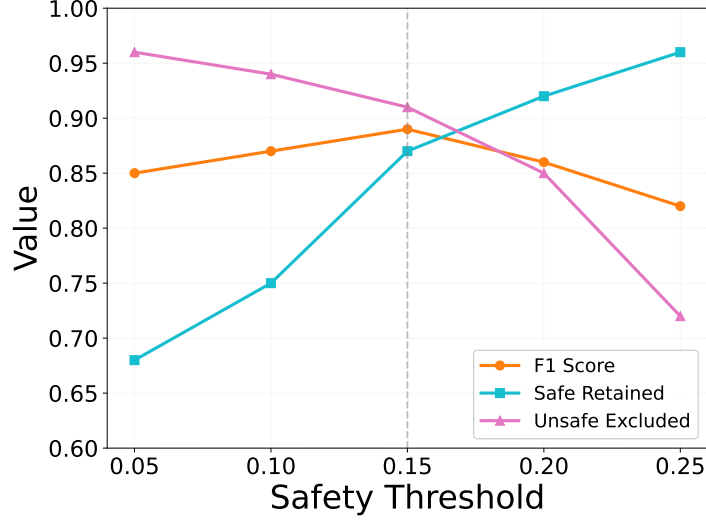


Figure S1: Sweep of the physician-annotated validation set with respect to safety. The selected threshold $\tau_e = 0.15$ provides the best F1 trade-off between safe-medication retention and unsafe-combination exclusion.

The selected threshold was $\tau_e = 0.15$, which achieved the best F1 trade-off while retaining 87% of safe medications and excluding 91% of unsafe combinations. Approximate 95% confidence intervals at this operating point were 79–93% for safe-medication retention and 84–96% for unsafe-combination exclusion. Stricter thresholds exclude more unsafe combinations but also remove more safe medications, whereas looser thresholds retain more safe medications but allow more unsafe combinations through.

Because the annotated validation set is modest in size, these rates should be interpreted as operating-point estimates rather than population-level safety guarantees. Also, the threshold could be influenced by the amount of noise in the individual medication–diagnosis label (annotation noise) and by the variation in prescription practices, coding patterns, and the review of contraindications from one site to another. Thus, the value of τ_e should be adapted if the framework is deployed in clinical environments where the distributions of EHRs are significantly different and/or local EHR review policies are different. The scores obtained should be treated as pragmatic observations of safety within the observed EHR covariates and not as causal certainties.