

## A APPENDIX

### A.1 DRUG RANKING RESULTS OF DEFENSE

Figure 3 presents four representative defense settings for the drug ranking task under adversarial poisoning:

(a) **Poisoned ranking results without defense:** The majority of points lie below the diagonal, indicating that most targeted drugs are promoted after poisoning, consistent with prior findings Yang et al. (2024a).

(b) **Link faithfulness defender (Medium level) Yang et al. (2024a):** Moderate filtering partially mitigates ranking distortion, with some recovery for top-ranked drugs, but many mid- and low-ranked drugs remain affected.

(c) **Link faithfulness defender (High level) Yang et al. (2024a):** Aggressive filtering strongly suppresses poisoning effects, but over-defends by removing legitimate high-ranking drugs, leading to the loss of valuable candidates.

(d) **Our GNN-based reconstruction:** By reweighting relations and pruning only low-confidence edges, our method maintains top-ranked drugs while resisting adversarial promotion, achieving a balanced trade-off between robustness and completeness.

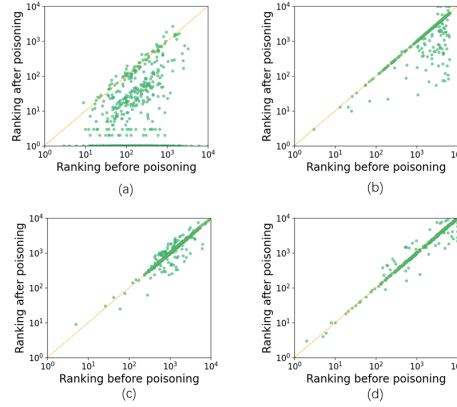


Figure 3: Drug ranking evaluation under poisoning attacks using different defense methods. (a) No defense. (b) Medium-level defense. (c) High-level defense. (d) Our GNN-based KG reconstruction.

### A.2 ABLATION EXPERIMENT

To dissect the contributions of key components in our structure-aware KG reconstruction framework and validate their necessity, we conduct systematic ablation experiments. All evaluations are performed under Scorpis attacks (summary injection), using BioGPT, LLaMA2, and Meditron as base models with PrimeKG, and results are averaged over 5 independent runs on PubMedQA and MedQA benchmarks.

#### A.2.1 EFFECT OF EDGE PRUNING THRESHOLD $\tau$

The threshold  $\tau$  in topology refinement (Eq. 7) determines which edges are retained in the reconstructed graph, balancing between filtering adversarial links and preserving valid medical relations. We test  $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$  and compare performance metrics in Table 5 and 6.

-  $\tau = 0.05$ : Retains excessive low-confidence edges, including a large number of adversarial injections. This leads to degraded precision and F1 scores (BioGPT: Precision=0.872, F1=0.865 on PubMedQA) due to noisy propagation in multi-hop reasoning.

-  $\tau = 0.10$ : Still retains some low-confidence adversarial edges, resulting in suboptimal performance (LLaMA2: Accuracy=0.864, Recall=0.850 on PubMedQA).

- $\tau = 0.15$ : Achieves optimal balance, preserving high-confidence clinical relations (e.g., primary treatment links with  $\hat{\alpha}_{ij} \geq 0.6$ ) while filtering most adversarial edges. This configuration yields the highest scores across all metrics (BioGPT: Accuracy=0.907, Precision=0.926, Recall=0.915, F1=0.921 on PubMedQA), consistent with our main results.
- $\tau = 0.20$ : Moderately over-prunes edges, removing some valid low-weight relations, which reduces recall (Meditron: Recall=0.874 on PubMedQA).
- $\tau = 0.25$ : Further increases pruning stringency, causing noticeable drops in recall and F1 (LLaMA2: Recall=0.862, F1=0.883 on PubMedQA) due to loss of legitimate medical connections.

PubMedQA $\tau$	BioGPT				LLaMA2				Meditron			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
0.05	0.862	0.872	0.855	0.865	0.851	0.863	0.840	0.852	0.847	0.859	0.836	0.848
0.10	0.876	0.889	0.865	0.882	0.864	0.881	0.850	0.871	0.859	0.872	0.845	0.865
0.15	<b>0.907</b>	<b>0.926</b>	<b>0.915</b>	<b>0.921</b>	<b>0.901</b>	<b>0.918</b>	<b>0.907</b>	<b>0.912</b>	<b>0.895</b>	<b>0.912</b>	<b>0.901</b>	<b>0.906</b>
0.20	0.892	0.910	0.885	0.895	0.882	0.904	0.879	0.883	0.877	0.896	0.868	0.877
0.25	0.881	0.899	0.870	0.885	0.873	0.892	0.862	0.883	0.866	0.881	0.855	0.868

Table 5: Performance metrics under varying edge pruning threshold  $\tau$  on PubMedQA.

MedQA $\tau$	BioGPT				LLaMA2				Meditron			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
0.05	0.855	0.863	0.847	0.855	0.843	0.856	0.832	0.844	0.839	0.851	0.828	0.840
0.10	0.870	0.881	0.862	0.871	0.858	0.872	0.846	0.859	0.853	0.865	0.841	0.853
0.15	<b>0.916</b>	<b>0.905</b>	<b>0.924</b>	<b>0.917</b>	<b>0.907</b>	<b>0.919</b>	<b>0.917</b>	<b>0.913</b>	<b>0.911</b>	<b>0.910</b>	<b>0.905</b>	<b>0.907</b>
0.20	0.894	0.892	0.899	0.891	0.885	0.905	0.892	0.889	0.886	0.894	0.882	0.888
0.25	0.880	0.884	0.876	0.880	0.871	0.887	0.865	0.876	0.868	0.876	0.859	0.867

Table 6: Performance metrics under varying edge pruning threshold  $\tau$  on MedQA.

### A.2.2 CONTRIBUTION OF ROBUSTNESS ENHANCEMENT MODULES

In Table 7 and 8, we ablate two core modules: Adversarial Anomaly Detection(AAD) (Eq.6) and Drug Ranking Consistency(DRC) (Eq.7), evaluating their individual and combined impacts.

- Full Model: Achieves the highest performance across all metrics (BioGPT: Accuracy=0.907, Precision=0.926, Recall=0.915, F1=0.921 on PubMedQA) by leveraging both modules synergistically.
- Without AAD: Removes the mechanism to flag suspicious high-risk relations, leading to decreased precision and F1 (LLaMA2: Precision=0.875, F1=0.875 on PubMedQA).
- Without DRC: Disables enforcement of clinical relevance ranking, causing reduced recall (Meditron: Recall=0.880 on PubMedQA).
- Without Both Modules: Results in cumulative performance degradation, with the lowest scores across all metrics (BioGPT: Accuracy=0.870, Precision=0.862, Recall=0.858, F1=0.870 on PubMedQA).

### A.2.3 IMPACT OF LOSS LOSS FUNCTION COEFFICIENTS

Our dual dual-objective loss formulation  $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{struct}} + \lambda_2 \mathcal{L}_{\text{adv}}$  orchestrates a critical tradeoff between two competing objectives:

1. Structural fidelity ( $\mathcal{L}_{\text{struct}}$ ): Maintaining topological alignment with the clean medical knowledge graph to preserve evidence-based relationships;
2. Adversarial resilience ( $\mathcal{L}_{\text{adv}}$ ): Suppressing suppressing suppression of adversarial edges injected via attacks like Scorpius.

To optimize this balance, we conducted exhaustive experiments with  $\lambda_1 \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$  (where  $\lambda_2 = 1 - \lambda_1$ ), evaluating BioGPT on both PubMedQA and MedQA. Representative results for BioGPT are visualized in Figure 4, with consistent trends observed across all models.

1. **Optimal Balance at  $\lambda_1 = 0.5$ :** All metrics peak at this configuration, demonstrating synergistic alignment of structural preservation and adversarial filtering. For BioGPT:

- PubMedQA: Accuracy = 0.907, Precision = 0.926, Recall = 0.915, F1 = 0.921
- MedQA: Accuracy = 0.916, Precision = 0.905, Recall = 0.924, F1 = 0.917

This balance is clinically critical—preserving high-confidence treatment pathways (e.g., FDA-approved drug-indication pairs) while eliminating spurious contraindications injected by adversaries.

2. **Risk of Over-Suppression ( $\lambda_1 \leq 0.3$ ):** Overweighting  $\mathcal{L}_{adv}$  leads to aggressive pruning that inadvertently removes valid low-weight medical relationships (e.g., off-label uses with emerging evidence). This causes:

- Significant recall degradation (BioGPT on PubMedQA: Recall = 0.840 at  $\lambda_1 = 0.0$ )
- Compromised clinical completeness, as critical differential diagnosis pathways are truncated

3. **Risk of Under-Suppression ( $\lambda_1 \geq 0.7$ ):** Overweighting  $\mathcal{L}_{struct}$  preserves adversarial edges (e.g., Scorpius-injected drug-disease links), corrupting inference enough.

- Precision erosion (BioGPT on PubMedQA: Precision = 0.862 at  $\lambda_1 = 1.0$ )
- Clinically hazardous recommendations, including contraindicated drug combinations

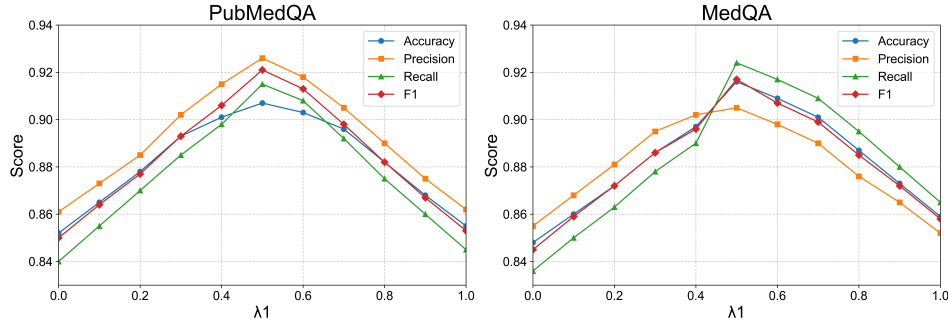


Figure 4: BioGPT performance metrics across  $\lambda_1$  values. Metrics include Accuracy (blue), Precision (orange), Recall (green), and F1 (red).

PubMedQA Configuration	BioGPT				LLaMA2				Meditron			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
W/O AAD + W/O DRC	0.870	0.862	0.858	0.870	0.855	0.854	0.846	0.855	0.863	0.860	0.852	0.863
W/O AAD	0.889	0.901	0.892	0.889	0.880	0.875	0.888	0.875	0.875	0.882	0.880	0.874
W/O DRC	0.896	0.915	0.885	0.896	0.889	0.902	0.880	0.889	0.880	0.890	0.880	0.880
Full Model	<b>0.907</b>	<b>0.926</b>	<b>0.915</b>	<b>0.921</b>	<b>0.901</b>	<b>0.918</b>	<b>0.907</b>	<b>0.912</b>	<b>0.895</b>	<b>0.912</b>	<b>0.901</b>	<b>0.906</b>

Table 7: Performance metrics of robustness module ablation on PubMedQA.

MedQA Configuration	BioGPT				LLaMA2				Meditron			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
W/O AAD + W/O DRC	0.866	0.865	0.860	0.866	0.869	0.870	0.862	0.869	0.850	0.852	0.848	0.850
W/O AAD	0.886	0.892	0.890	0.886	0.882	0.880	0.885	0.882	0.876	0.878	0.875	0.876
W/O DRC	0.892	0.895	0.898	0.892	0.888	0.901	0.889	0.888	0.882	0.885	0.880	0.882
Full Model	<b>0.916</b>	<b>0.905</b>	<b>0.924</b>	<b>0.917</b>	<b>0.907</b>	<b>0.919</b>	<b>0.917</b>	<b>0.913</b>	<b>0.911</b>	<b>0.910</b>	<b>0.905</b>	<b>0.907</b>

Table 8: Performance metrics of robustness module ablation on MedQA.

4. **Dataset-Specific Consistency:** MedQA (structured clinical questions) consistently outperforms PubMedQA (unstructured literature-derived queries) across all  $\lambda_1$  values, with a 1.2 - 2.3% F1

---

702 gap. This highlights the importance of balanced loss weighting for unstructured medical text,  
703 where adversarial signals are more subtly embedded.  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755