Rebuttal PDF

DICTrank accuracy (%)	Ternary (n=1181)	Ternary on Predicted No/Most (n)	Binary on Ground Truth No/Most (n)
Full pipeline	84.6	92.5 (761)	**93.0** (603)
Keyword summary prompt	88.1	90.4 (924)	93.7 (603)
No CoT	77.8	67.4 (629)	92.5 (603)
GPT-3.5	77.6	77.8 (855)	88.4 (603)
RAG fragment context			94.2 (584)
askFDALabel (previous SOTA)			**77.7** (584)
DILIrank accuracy (%)	Ternary (n=819)	Ternary on Predicted No/Most (n=525)	Binary on Ground Truth No/Most (n=363)
Full pipeline	81.1	85.0	86.2
DIRIL accuracy (%)	Ternary (n=269)	Ternary on Predicted No/Most (n=177)	Binary on Ground Truth No/Most (n=269)
Full pipeline	71.3	76.8	72.9

Table 2: Validation results across cardiotoxicity (DICTrank), liver toxicity (DILIrank), and renal toxicity (DIRIL) comparing our predictions to expert ratings from the FDA. We show accuracy using our ternary prompt on the full dataset. We also show accuracy after filtering to only include predicted “No" or "Most” drugs to allow the model to set aside borderline cases. Finally, we show accuracy using our binary prompt after filtering to only include ground-truth “No” or “Most” drugs, allowing for apples-to-apples comparisons with askFDALabel. Our full pipeline on DICTrank significantly outperforms the previous state-of-the-art (askFDALabel). In our ablations, adding additional cardiotoxicity keywords into our summary prompt had an uneven effect on accuracy. Removing the Chain-of-Thought step and moving to GPT-3.5 consistently hurt accuracy on DICTrank. Running our binary prompt on just the fragments of the drug label returned by the FDA’s RAG system slightly outperforms using the full drug label, perhaps by limiting extraneous information. Because we only have the FDA’s RAG fragments for the ground-truth No/Most subset of DICTrank, we cannot compare to results on the full dataset. Finally, we achieve similarly high accuracy on DILIrank, but our predictions perform worse on DIRIL perhaps due to a differing methodology.