Fine-tuned LLMs Hit Perturbation - Validation set
LLMs for Drug-Drug Interaction Prediction

October 16, 2025

1 Empirical Design

Regarding prompt hit perturbations, we tested robustness to semantically neutral prompt changes for the two best fine-tuned models by F1 score on the validation set, using four variants (P1-P4):

- P1 reorders drug fields to genes→organisms→SMILES, testing dependence on positional encoding
- P2 paraphrases the instruction line while preserving semantic content (”classify whether their administration causes” → ”determine if administering these drugs results in”), testing lexical sensitivity
- P3 switches to a list format with Drug A/Drug B labels, testing structural robustness
- P4 uses a compact horizontal format with pipe-separated fields, testing spacing/delimiters and reduced texts.

Full prompts are provided in our online repository. For these new experiments, we used the same metrics (accuracy, sensitivity, and F1) and calculated the agreement percentage with the previously obtained results.

Results showed minimal variation (GPT-4o: ±0.6% accuracy, ±1.5% sensitivity; Phi-3.5: ±0.3% accuracy, ±0.6% sensitivity) and high agreement with the original predictions (95.0–96.1%). Critically, the baseline models themselves achieve high performance (GPT-4o: 0.926 accuracy; Phi-3.5: 0.913 accuracy), so 95%+ agreement indicates that perturbations induce minimal behavioral changes. Aggregate metrics (accuracy, F1) remained constant (SD < 1.5%) across perturbations, indicating that surface-form modifications do not jeopardize the learned DDI patterns in the models. Table 1 summarizes these results.

Table 1: Prompt perturbation robustness on the validation set for fine-tuned GPT-4o and Phi-3.5 2.7B. Agreement is versus the baseline prompt; N is the number of differing predictions out of 1,090 total instances.

Model	**Perturb.**	**Acc**	**Sens**	**F1**	**Agreement**	**N.**
GPT-4o	P1	0.922	0.923	0.922	96.1%	42
	P2	0.916	0.943	0.918	95.3%	51
	P3	0.909	0.908	0.909	95.6%	48
	P4	0.908	0.906	0.908	95.3%	51
	AVG	**0.914**	**0.920**	**0.914**	**95.6%**	**48**
	SD	**0.006**	**0.015**	**0.006**		
Phi-3.5 2.7B	P1	0.919	0.974	0.923	95.3%	51
	P2	0.917	0.971	0.922	95.7%	47
	P3	0.917	0.971	0.922	95.7%	47
	P4	0.911	0.985	0.917	95.0%	54
	AVG	**0.916**	**0.975**	**0.921**	**95.4%**	**49.75**
	SD	**0.003**	**0.006**	**0.002**		