# Fine-tuned LLMs Hit Perturbation - External datasets 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 external datasets, 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.

More in detail, we assessed prompt robustness on the largest external dataset (KEGG, with 13,262 instances) using the same four perturbation variants applied to the two best fine-tuned models (GPT-4o and Phi-3.5 2.7B) in terms of F1-score values. To this aim, we employed accuracy, sensitivity, and F1 score metrics, and computed the agreement percentage, and standard deviation with respect the original results reported in the manuscript.

Across perturbations (see Table 1), the best fine-tuned models in terms of F1-score (i.e., GPT-4o and Phi-3.5 2.7B) remained highly stable. Specifically, GPT-4o achieved accuracy ranging from 0.929 to 0.934 and F1 from 0.933 to 0.935, while Phi-3.5 2.7B achieved accuracy from 0.909 to 0.920 and F1 from 0.916 to 0.926. Averages and standard deviations were small, and the performance pattern reported in the paper held consistently: GPT-4o preserved the highest accuracy and F1 on KEGG, while Phi-3.5 maintained very high sensitivity (average 0.993) with competitive F1.

Table 1: Prompt perturbation robustness on KEGG for fine-tuned GPT-4o and Phi-3.5 2.7B. Each perturbation variant is evaluated on the complete KEGG test set.

| **Model** | **Perturbation** | **Accuracy** | **Sensitivity** | **F1** |
|---|---|---|---|---|
| GPT-4o | P1 | 0.9340 | 0.9548 | 0.9354 |
| | P2 | 0.9296 | 0.9787 | 0.9329 |
| | P3 | 0.9315 | 0.9508 | 0.9328 |
| | P4 | 0.9323 | 0.9609 | 0.9342 |
| | **AVG** | **0.9320** | **0.9610** | **0.9340** |
| | **SD** | **0.0020** | **0.0110** | **0.0010** |
| Phi-3.5 2.7B | P1 | 0.9204 | 0.9922 | 0.9258 |
| | P2 | 0.9204 | 0.9928 | 0.9257 |
| | P3 | 0.9184 | 0.9911 | 0.9239 |
| | P4 | 0.9085 | 0.9950 | 0.9158 |
| | **AVG** | **0.9170** | **0.9930** | **0.9230** |
| | **SD** | **0.0050** | **0.0010** | **0.0040** |