Few-shot Perturbations
LLMs for Drug-Drug Interaction Prediction

October 16, 2025

1 Empirical Design

we conducted a sensitivity analysis on the two best-performing models (Claude 3.5 Sonnet and GPT-4o) using similarity-based example selection on the 1,090-instance validation set. We focused sensitivity analyses on the best-performing models to demonstrate that even top performers are sensitive to prompt structure, avoid overwhelming the paper with redundant analyses (18 models x 4 perturbations would generate many supplementary tables without additional scientific insight), and establish an upper bound on few-shot robustness before pivoting to fine-tuning.

Specifically, we applied a semantically equivalent prompt perturbation that:
(1) reordered per-drug fields to genes→organisms→SMILES, (2) paraphrased the system instruction (e.g., "You are an expert of drug-drug interaction" → "You are a drug-drug interaction specialist"), and (3) reversed the order of the 10 retrieved examples. We then compared the results using the same metrics and reporting the Δ with the previously obtained results.

Table 1: RQ2: Few-shot similarity-based robustness on the validation set under a combined prompt perturbation (reworded system instruction, reordered drug fields, reversed example order). Deltas (Perturbed−Baseline) are shown in parentheses.

Model	**Setting**	**Acc**	**Sens**	**F1**
Claude 3.5 Sonnet	Baseline	0.8376	0.8422	0.8384
	Perturbed	0.7917 (-0.0459)	0.6899 (-0.1523)	0.7681 (-0.0702)
GPT-4o	Baseline	0.7917	0.8404	0.8014
	Perturbed	0.7688 (-0.0229)	0.7450 (-0.0954)	0.7632 (-0.0382)

The results of the sensitive analysis (see Table 1) showed significant performance degradation (Sonnet: -4.59% accuracy, -15.23% sensitivity; GPT-4o: -2.29% accuracy, -9.54% sensitivity), confirming that few-shot learning is extremely sensitive to prompt structure variations—a known limitation that has been documented in previous work [3, 2, 1].

References

[1] Qingxiu Dong et al. "A Survey on In-context Learning". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 1107–1128.

[2] Nelson F Liu et al. "Lost in the Middle: How Language Models Use Long Contexts". In: *Transactions of the Association for Computational Linguistics* 11 (2024), pp. 157–173.

[3] Sheng Lu et al. "Are Emergent Abilities in Large Language Models just In-Context Learning?" In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 5098–5139.