Analysis of Phi3.5 2.7B fine-tuned on 2,000 samples
LLMs for Drug-Drug Interaction Prediction

October 16, 2025

1 Empirical Analysis

To directly address whether the DDI-specific task requires more training data than suggested by general NLP studies, we conducted an additional experiment with Phi-3.5 2.7B – the best-performing model in terms of sensitivity. We trained Phi-3.5 2.7B using 2,000 training new samples (doubling our original size) – using the same methodology for the original 1,000 training samples, and the same hyperparameters as reported in the manuscript, but for 5 epochs – and tested it on the external datasets. The results showed minimal improvement over the 1,000-sample model: the average accuracy increased from 0.881 to 0.885; the average sensitivity decreased from 0.978 to 0.974; and the average F1 score improved from 0.900 to 0.902. When weighted by dataset size, the 2,000-sample model achieved an accuracy of 0.921 versus 0.919 for the 1,000-sample model, a sensitivity of 0.991 versus 0.985, and an F1 score of 0.926 versus 0.924.

While references [1, 3, 4, 2] address general NLP tasks, our contribution demonstrates empirically that the efficiency of fine-tuning with limited examples extends to specialized biomedical prediction tasks. These above-mentioned results suggest that performance plateaus around 1,000 examples for this specific DDI prediction task with fine-tuned LLMs, supporting our original design choice.

Table 1: Performance comparison of Phi-3.5 2.7B trained with 2,000 samples on external datasets. Between brackets is reported the Δ between the metric values computed for the new fine-tuned version and the 1,000-sample fine-tuned one.

Dataset	**Acc. (Δ)**	**Sens. (Δ)**	**F1 (Δ)**
CredibleMeds	1.000 (–)	1.000 (–)	1.000 (–)
Corpus 2013	0.891 (+0.016)	0.938 (–)	0,896 (+0.013)
Corpus 2013	0.899 (-0.020)	0.946 (-0.054)	0.903 (-0.022)
French Ref.	0.922 (+0.012)	0.997 (+0.025)	0.927 (+0,012)
HEP	0.925 (+0.001)	0.995 (-0.005)	0.930 (–)
HIV	0.921 (-0.006)	0.995 (-0,005)	0.926 (-0,006)
KEGG	0.921 (-0.001)	0.992(+0.001)	0.926 (-0.001)
NDF-RT	0.933 (-0.025)	0.958 (-0.042)	0.934 (-0.025)
NLM Corpus	0.833 (+0.056)	0.889 (–)	0.842 (+0.042)
Onc Non-Int.	0.921 (+0.002)	1.000 (–)	0.927 (+0.001)
OSCAR	0.911 (+0.010)	0.958 (+0.017)	0.915 (+0.010)
PK Corpus	0.500 (–)	1.000 (–)	0.667 (–)
WorldVista	0.929 (+0.005)	0.997 (+0.011)	0.934 (+0.005)
AVG	**0.885 (+0.004)**	**0.974 (-0.004)**	**0.902 (+0.002)**
Weighted AVG	**0.921 (+0.002)**	**0,991 (+0.006)**	**0,926 (+0.002)**

References

[1] Chunting Zhou et al. "Lima: Less is more for alignment". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 55006–55021.

[2] Scott Barnett et al. "Fine-tuning or fine-failing? debunking performance myths in large language models". In: *arXiv preprint arXiv:2406.11201* (2024).

[3] Michael Oliver and Guan Wang. "Crafting efficient fine-tuning strategies for large language models". In: *arXiv preprint arXiv:2407.13906* (2024).

[4] Xiaoyong Zhao et al. "Research on Fine-Tuning Optimization Strategies for Large Language Models in Tabular Data Processing". In: *Biomimetics* 9.11 (2024), p. 708.