Rigorous Evaluation of Large Language Models for Malaria Drug Discovery: Trade-offs in Performance, Scale, and Resource Utility

Marvellous O. Ajala[1] , **Zainab Ashimiyu-Abdusalam**[1] , **Comfort Adesina**[1]
[1]Magami Open Sciences Initiative
{marvellous, zainab, comfort}@magamios.org

Abstract

We introduce Malaria-Instruct, a curated instruction-following dataset derived from the ChEMBL Legacy Malaria corpus for Malaria virtual screening, and conduct a systematic evaluation of five open-source LLMs; Gemma-2 2B/9B, TxGemma-2B/9B, and LlaSMol-Mistral-7B, on a rigorous out-of-distribution data split. Performance was benchmarked against classical ML models (Random Forest, XGBoost) and frontier proprietary models (Gemini 2.5, OpenAI o3) under few-shot conditions. Fine-tuned LLMs substantially outperformed all baselines: TxGemma-9B achieved the highest ROC-AUC (0.731 ± 0.005) and LlaSMol-Mistral-7B the best enrichment factor (EF@1% $\approx$ 4.99). Domain-specific fine-tuning proved categorically indispensable with TxGemma-9B collapsing from ROC-AUC 0.731 to 0.499, under its best few-shot condition, and neither Gemini 2.5 (ROC-AUC $\approx$ 0.53) nor o3 (ROC-AUC $\approx$ 0.59) achieved reliable discrimination without fine-tuning. Biomedical pretraining conferred a measurable advantage at equivalent scale, while chemistry-aware pretraining yielded superior prospective enrichment. Fine-tuned open-source LLMs represent a compelling, resource-efficient paradigm for antimalarial VS, outperforming both classical pipelines and proprietary reasoning models under structurally challenging conditions.

1 Introduction

1.1 The Global Burden of Malaria and Advent of Virtual Screening

Malaria remains one of the most consequential infectious diseases confronting global public health, with an estimated 263 million cases and 597,000 deaths recorded globally in 2023 [WHO, 2024] and approximately 94% of all cases and 95% of deaths occurred in the WHO African Region. These figures underscore the profound human cost of the disease and the persistent structural inequities that concentrate its burden in low-resource settings. The emergence of partial artemisinin resistance (the current frontline standard of care following the progressive failure of chloroquine and sulfadoxine–pyrimethamine) in Africa [WHO, 2024; Lancet Microbe, 2025], following its spread from Southeast Asia, urgently motivates the identification of structurally novel chemotypes. Ligand-based virtual screening computationally ranks candidate molecules by predicted bioactivity, evaluated here primarily by enrichment factor at 1% (EF@1%), the operationally critical metric governing hit rate in experimental screening. For antimalarials specifically, the central challenge is that effective VS must generalise to structurally novel chemotypes beyond the historically narrow scaffold space of known actives, precisely the out-of-distribution condition this study is designed to test.

1.2 The Rise of Large Language Models in Molecular Biology and Cheminformatics

The representation of molecular structures as SMILES strings [Weininger, 1988],that are amenable to sequence modelling, catalysed a proliferation of chemistry-aware LLMs [Chithrananda *et al.*, 2020; Fabian *et al.*, 2020; Yu *et al.*, 2024]. More recent developments, including the SMolInstruct-trained LlaSMol family [Yu *et al.*, 2024], the therapeutics-focused TxGemma series [Wang *et al.*, 2025], and the predecessor Tx-LLM [Zambrano Chaves *et al.*, 2024], have produced instruction-following models that accept molecular queries in natural language, respond with predicted properties, and exhibit strong transfer across diverse chemical tasks. The emergence of frontier proprietary LLMs with broad reasoning capabilities, notably OpenAI's o3 and Google's Gemini 2.5, raises a further question of practical significance: can general-purpose reasoning models leverage their extensive scientific pretraining to perform useful bioactivity prediction in a few-shot setting, without domain-specific fine-tuning?

1.3 Fine-Tuning vs. Few-Shot Prompting: A Critical Distinction for Specialised Scientific Tasks

A foundational distinction in the deployment of LLMs for specialised scientific tasks concerns the mechanism by which task-specific knowledge is conveyed to the model. Parameter-efficient fine-tuning (QLoRA) internalises structure-activity

relationships at the gradient level in a way that ICL cannot [Hu *et al.*, 2022; Brown *et al.*, 2020], making domain-specific adaptation a testable necessity rather than a stylistic choice. Bioactivity prediction is not primarily a reasoning task amenable to pattern retrieval from a few examples as it requires the model to encode fine-grained structure-activity relationships that may not be recoverable from the statistical regularities present in even a handful of SMILES-label pairs. The hypothesis that domain-specific fine-tuning is necessary for reliable bioactivity prediction, rather than optional, is a central empirical claim of this work.

1.4 The Generalisation Problem: Why Scaffold Dissimilarity Splitting Matters

A recent study by [Meidı *et al.*, 2024] indicates that many splitting techniques provides negligible discriminative properties and that only dissimilarity-based and clustering-based splitting methods provide a meaningful test of out-of-distribution generalisation. In this study, we adopt the Lo-Hi dissimilarity splitting framework as the partitioning standard, enforcing strong Tanimoto dissimilarity train–test and train–validation, applied independently per assay. This design choice is deliberate and consequential: it ensures that reported performance metrics reflect true generalisation capacity rather than scaffold memorisation, and provides a substantially more demanding and ecologically valid benchmark than random or scaffold splitting would permit.

1.5 Research Contribution

Our primary contributions in this work are: (1) We introduce **Malaria-Instruct**, a novel instruction-following dataset derived from the ChEMBL Legacy Malaria corpus, curated with rigorous deduplication, assay harmonisation, and expert-informed contextualisation. (2) We establish a rigorous benchmark using Lo-Hi dissimilarity-based splitting that enforces meaningful out-of-distribution generalisation conditions, providing systematic head-to-head comparison of fine-tuned open-source LLMs against frontier proprietary models and classical cheminformatics baselines where non previously existed. (3) We provide a comprehensive evaluation of five open-source LLMs across fine-tuning and few-shot paradigms, with systematic characterisation of resource requirements to guide deployment decisions in low-resource research settings prevalent in Africa where these researches are often undertaken.

2 Related Works

2.1 ChEMBL and Curated Antimalarial Bioactivity Databases

Malaria-Instruct is built on ChEMBL [Mendez *et al.*, 2019; Zdrazil *et al.*, 2024], (details provided in appendix A) while addressing the while addressing the key curation challenges that exist in it: class imbalance, assay heterogeneity (across different organisms, time points, and measurement modalities), and duplicate bioactivity records (from the same molecule being tested across multiple experiments) persists. We address these challenges through a multi-stage curation procedure involving assay-level deduplication, harmonisation of 48-hour and 96-hour readouts, and ECFP-based negative sample augmentation for confirmatory assays lacking sufficient inactive observations.

2.2 Instruction Tuning and Dataset Curation for Scientific LLMs

Unlike prior general-purpose molecular instruction datasets [Fang *et al.*, 2023; Yu *et al.*, 2024; Cao *et al.*, 2024] discussed in appendix B, Malaria-Instruct is the first instruction-tuning dataset specifically constructed for antimalarial virtual screening, with assay-level contextualisation providing information about Plasmodium strain, assay duration, and mechanistic context. This contextualisation is specifically designed to support TxGemma's [Wang *et al.*, 2025] prompt format, which conditions predictions on additional biological context beyond the molecular structure alone. The dataset's construction incorporates per-assay, per-split negative sample augmentation.

2.3 Chemistry-Aware and Biomedically Specialised Language Models

The application of language modelling to molecular SMILES strings has a well-established lineage including ChemBERTa [Chithrananda *et al.*, 2020] and MolBERT [Fabian *et al.*, 2020] The most directly relevant prior work to this study is LlaSMol models [Yu *et al.*, 2024], which were produced by fine-tuning four open-source base models (Galactica, Llama 2, Code Llama, and Mistral) on SMolInstruct using LoRA. The Mistral-based variant, LlaSMol-Mistral, was identified as the best-performing chemistry LLM, outperforming GPT-4 and Claude Opus 3 by substantial margins on standard chemistry benchmarks at the time of publication. LlaSMol's chemistry-specific pretraining provides a principled inductive bias for molecular property prediction that general-purpose models lack, making it a particularly informative comparator in our evaluation. Domain-specific pretraining in therapeutic AI, grounded in the recognition that biomedical language: spanning clinical ontologies, protein nomenclature, molecular property distributions, and assay-specific terminology, constitutes a distinct subdomain that general-purpose models underrepresent relative to its downstream task density. TxGemma[Wang *et al.*, 2025] extends Tx-LLM [Zambrano Chaves *et al.*, 2024] achieving near-state-of-the-art performance on 43 of 66 Therapeutics Data Commons (TDC) tasks [Huang *et al.*, 2021], details provided in appendix C. The biomedical specialisation of TxGemma makes it a compelling candidate for antimalarial bioactivity prediction, but the extent to which its pretraining generalises to the specific structural and biological features of the Plasmodium screening context remains an open empirical question. While general-purpose LLMs like GPT-4 and the reasoning-focused o-series show promise in diverse scientific tasks, they consistently underperform compared to specialized models in predicting molecular bioactivity. This suggests that the "reasoning-by-analogy" used by these models cannot capture the high-dimensional, idiosyncratic structural patterns required for accurate drug-activity interpolation.

2.4 Data Splitting Strategies & Evaluation Frameworks in Cheminformatics

The choice of dataset splitting methodology is among the most consequential methodological decisions in molecular ML benchmarking, yet it has received insufficient attention in many published evaluations. Random splitting, the default in general ML frameworks, produces inflated performance estimates for molecular models due to the high structural similarity of molecules within the same dataset, effectively allowing models to interpolate between highly similar training and test compounds. Scaffold splitting [Bemis and Murcko, 1996] provides a more structurally aware partition but does not explicitly control the degree of inter-partition Tanimoto similarity, leaving open the possibility of near-analogues spanning the split boundary. The Lo-Hi framework [Steshin, 2023] operationalises dissimilarity-based splitting as a graph-theoretic optimisation problem: constructing a maximum independent set partition such that no two molecules from different partitions share a Tanimoto similarity above the specified threshold. Meidi *et al.,* (2024) provided a systematic comparison of splitting strategies across multiple bioactivity benchmarks, demonstrating that only dissimilarity-based and clustering-based splits (group structurally similar compounds and assign whole clusters to the same partition) provide a reliable test of out-of-distribution generalisation, with random and scaffold splits substantially overestimating practical model performance. Thus, the Lo-Hi splitting framework is adopted in this study as the standard for all partitioning operations, applied independently per assay to respect the distinct chemical space coverage of each screening dataset. Also, the choice of evaluation metric profoundly shapes the conclusions drawn from VS model benchmarks. In appendix D, we provide an extensive discussion on the different considerations for evaluation metrics. However, from the perspective of prospective VS deployment, we believe the Enrichment Factor (EF) is arguably the most operationally relevant metric. EF@1% quantifies the fold-enrichment of true actives in the top 1% of model-ranked compounds relative to the expected random baseline, directly measuring the practical efficiency of a VS campaign. A model with EF@1% of 5.0 concentrates five times as many true actives in the top 1% of its predictions as would be expected by chance, a difference that translates directly into reduced experimental costs and accelerated hit identification. The tension between ROC-AUC optimisation and EF optimisation, which arise from different parts of the score distribution, is a substantive methodological issue explored in the Discussion.

3 Methodology

3.1 Dataset

Dataset Curation

We introduce **Malaria-Instruct**, a novel instruction-following dataset for molecular bioactivity prediction constructed from the ChEMBL Legacy Malaria corpus [Mendez *et al.*, 2019; Zdrazil *et al.*, 2024]. [1]

Assay selection and deduplication. Only potency and IC50 assay entries were retained; all other bioactivity modalities were excluded. Because biological assays are routinely conducted in technical replicates, duplicate records were resolved at the assay level: for each unique molecule–assay pair, all replicate readings were aggregated and the datapoint was assigned a positive label if all replicates were concordantly active, a negative label if all replicates were concordantly inactive, and removed entirely if replicates returned conflicting classifications. This conservative conflict-removal policy prioritises label precision over dataset size, accepting a reduction in the number of usable datapoints to avoid the introduction of ambiguous supervision signals.

Temporal harmonisation. Antimalarial assays are standardly designed for 48-hour or 96-hour incubation periods, with intermediate 24-hour readings sometimes recorded. For assays with 24-, 48-, and 96-hour readings, labels were harmonised using the same concordance rule; conflicting time-point readings were discarded.

Assay contextualisation. TxGemma's prompt format conditions predictions on structured biological context beyond the molecular SMILES string alone, including information about assay type, biological target, and experimental conditions [Zambrano Chaves *et al.*, 2024] [Wang *et al.*, 2025]. To support this, an additional contextualisation column was constructed for each assay entry, encoding the nature of the assay, the specific *Plasmodium* strain under investigation, and distinguishing experimental characteristics. This contextualisation was produced through a combination of domain expert annotation and structured elicitation from state-of-the-art LLMs operating under expert-validated templates, ensuring biological accuracy while maintaining scalability across the full assay inventory. [2].

Negative sample augmentation. Confirmatory assays, designed as follow-up screens to validate hits from primary HTS campaigns, contain exclusively or predominantly active compounds by construction, providing no inactive observations from which a decision boundary can be learned. For such assays, negative sample augmentation was performed on a per-assay, per-split after the train-test-validation partitioning. ECFP4 fingerprints (radius 2, 2048 bits) were computed for molecules from previously discarded assays and dimensionality-reduced via PCA to the most informative components, retaining those explaining the greatest variance subject to a maximum of 100 dimensions or the number of available samples, whichever was smaller. The resulting embeddings were further projected to two dimensions via t-SNE using the Euclidean metric. Candidate negatives were then selected from the spatial region circumscribed by the convex envelope of confirmed positive samples in the t-SNE embedding, ensuring that augmented negatives occupy chemically plausible proximity to the active scaffold space rather than being drawn from remote or trivially dissimilar regions of

[1] The Malaria-Instruct dataset is publicly available on Zenodo at https://zenodo.org/records/19222923; the evaluation and fine-tuning code is available on GitHub at https://github.com/Magami-Open-Sciences-Initiative/LLMS-for-Malaria.

[2] This contextualisation was not added to the prompts for LlaSmol sticking with the formatting style used to train the model, leveraging only assay level details

chemical space. This approach was taken as it currently encapsulates how chemists currently computationally explore chemical space. Augmentation was performed to restore the positive-to-negative ratio observed in the broader MalariaInstruct corpus. [3]

Molecular standardisation. All retained molecules were standardised using the following sequential operations: charge neutralisation, parent fragment extraction for multicomponent mixtures, normalisation of unusual valence states and functional group representations, and canonical tautomer selection. Standardisation was applied uniformly across all splits to ensure that molecular identity comparisons and fingerprint computations reflect a consistent chemical representation.

Data Splitting

Lo-Hi splitting [Steshin, 2023] was adopted following [Meidı *et al.*, 2024], who demonstrate only dissimilarity-based splits provide genuine out-of-distribution evaluation. Partitioning was performed on ECFP4 fingerprints (radius 2, 2048 bits), selected for maximum substructural expressivity. Pairwise Tanimoto similarities were computed across the full molecular inventory, and partitions were constructed such that the maximum Tanimoto similarity between any train–test molecule pair did not exceed 0.4, and the maximum similarity between any train–validation pair did not exceed 0.55. These thresholds enforce a structural dissimilarity condition that approximates the novelty conditions encountered in real-world hit identification campaigns. Splitting was performed independently for each assay, preserving the structural diversity constraints within each biological context rather than applying a global partition that could allow assay-level leakage. For fine-tuned model experiments, an independent Lo-Hi resplit was drawn for each replicate, ensuring that performance standard deviations reflect genuine partition-induced variance rather than repeated evaluation on an identical structural partition.

Data Preparation

Fine-tuning format. Training instances for all fine-tuned LLMs were formatted following the data preparation protocol introduced by Tx-LLM [Zambrano Chaves *et al.*, 2024]. The resulting training corpus contains a mixture of 70% few-shot-formatted examples, with the number of in-context examples per instance drawn uniformly from $\{2, 3, 4, 5\}$ and 30% zero-shot examples. For few-shot training instances, both the in-context demonstration molecules and the query molecule for which a prediction is requested were drawn exclusively from the training partition, ensuring no validation or test information was accessible during training. Test and validation sets were formatted exclusively as zero-shot instances for final evaluation.

Few-shot evaluation format. For few-shot evaluation of unfinetuned models, in-context demonstration examples were drawn from either the training or test partitions, while the query molecule for which a prediction is requested was drawn strictly from the validation set. For each shot-count condition (3-shot, 4-shot, 5-shot), the dataset was independently resplit prior to constructing the few-shot evaluation instances, such that each shot-count condition operates on a distinct structural partition. The inferential consequences of this design for variance interpretation are discussed in §4.6.

[3] Augmented negatives were constrained to the pharmacophoric neighbourhood of training actives to ensure decision-boundary relevance.

3.2 Modelling

Model evaluation was structured around two complementary paradigms: supervised fine-tuning and few-shot in-context learning, applied to model classes spanning classical machine learning baselines, general-purpose open-source LLMs, and domain-specialised open-source LLMs, with additional few-shot evaluation of frontier closed-source models.

Fine-Tuning

Classical baselines. Random Forest [Breiman, 2001] and XGBoost [Chen and Guestrin, 2016] were trained on 2048-bit Morgan fingerprints with radius 2, computed as binary numpy arrays for each molecule. Both models were trained with default hyperparameter configurations and a fixed random state of 2024 and experiments conducted across five independent replicates with resplitting at each replicate. Justification for the choice of baseline provided in appendix E.

LLM fine-tuning. Three open-source model families were selected for parameter-efficient fine-tuning: Gemma-2 [Riviere *et al.*, 2024], TxGemma [Wang *et al.*, 2025], and LlaSMol-Mistral [Yu *et al.*, 2024]. Gemma-2 was included as a general-purpose reference to quantify the performance gap attributable to domain-specific pretraining. TxGemma represents the current state of the art among open-source therapeutic LLMs, built on the Gemma-2 architecture and pretrained on an instruction-tuned variant of the TDC benchmark [Huang *et al.*, 2021]. LlaSMol-Mistral, based on the Mistral-7B backbone, was selected as the best-performing variant from the LlaSMol model family, itself fine-tuned on the SMolInstruct dataset [Yu *et al.*, 2024]. Both 2B and 9B parameter variants were evaluated for Gemma-2 and TxGemma to characterise the performance–resource tradeoff across the small-to-medium scale range most relevant to resource-constrained research environments. Training parameters are defined in appendix F. All LLM experiments were conducted in duplicate with independent Lo-Hi resplitting at each replicate. Training used the training partition; evaluation during training monitored the test partition. Final performance metrics are reported exclusively on the validation set. Validation predictions were generated via vLLM to reduce inference latency.

Compute environment. All training was conducted on Google Colab Pro+ (paid tier); compute requirements are reported in Table 4

3.3 Few-Shot Evaluation

Open-source models. The in-context learning capability of all unfinetuned open-source models was evaluated under 3-shot, 4-shot, and 5-shot conditions. At each shot count, n molecule–activity demonstration pairs were prepended to the query prompt prior to soliciting a prediction for the target

molecule. See §4.6 and appendix G for inferential implications of per conditioning splitting.

Closed-source models. Gemini 2.5 and OpenAI o3 were evaluated under identical 3-, 4-, and 5-shot conditions via their respective public APIs. Fine-tuning was not pursued for closed-source models, as API-based fine-tuning is inconsistent with the resource-constrained, open-infrastructure research environment this study targets. To emulate the low-resource constraint predominant with researchers carrying out similar research in global south associated with API token costs, evaluation was conducted on a subsample of 500 molecules per condition[4]. All closed-source evaluations were conducted in duplicate at each shot-count condition, with mean and standard deviation reported across replicates.

Evaluation

All models were evaluated on the held-out validation set, which was retained as zero-shot examples across both fine-tuned and few-shot experimental conditions, ensuring that no validation molecule or its structural neighbours appeared in any prompt context seen during training or few-shot construction. Model outputs were parsed to extract binary activity predictions, and probabilistic scores were derived from output token logits where available, or from the rank ordering of predicted class labels otherwise. Performance was assessed across five complementary metrics selected to capture distinct and non-redundant aspects of model behaviour under class imbalance. Accuracy and precision are reported for completeness but are not treated as primary metrics, given their susceptibility to inflation by majority-class prediction under the positive-to-negative imbalance characteristic of antimalarial screening data. The Matthews Correlation Coefficient (MCC) serves as the primary balanced classification metric: unlike F1, MCC incorporates all four cells of the confusion matrix and remains informative under severe class imbalance, producing a score of zero for a classifier that predicts the majority class unconditionally [Chicco and Jurman, 2020]. ROC-AUC is reported as the standard global discrimination metric, corresponding to the probability that a randomly drawn active compound is ranked above a randomly drawn inactive, and is insensitive to the choice of classification threshold. The Enrichment Factor at the 1% level (EF@1%) is designated the primary operational metric for virtual screening evaluation. EF@1% quantifies the fold-enrichment of confirmed actives within the top 1% of model-ranked compounds relative to the expected rate under random selection, and directly indexes the practical efficiency of a prospective screening campaign, governing the number of experimental assays required per confirmed hit. A model achieving EF@1% of 5.0 concentrates five times the expected number of true actives in the top-ranked fraction, reducing experimental cost per hit by the same factor. Where ROC-AUC and EF@1% rankings diverge across models, EF@1% is treated as the operationally authoritative criterion for model selection. For fine-tuned models, all metrics are reported as mean ± standard deviation across the two independent experimental replicates. For few-shot open-source models, metrics are reported per shot count across the independently re-split evaluation conditions. For closed-source models, evaluated on a statistically powered subsample of 500 molecules, exceeding the minimum sample size of 326 derived via Finite Population Correction at 95% confidence and 5% margin of error, metrics are reported as mean ± standard deviation across duplicate runs at each shot count. Inference time per molecule and GPU memory requirements during both training and inference are additionally recorded for all models to support resource-constrained deployment decisions, and are summarised alongside performance metrics in Table 1.

[4]a sample size was chosen to exceed the minimum of 326 derived via Finite Population Correction at 95% confidence and 5% margin of error, providing statistically adequate coverage of the validation set distribution

4 Results

We provide the mean value of the results for the trained classical models, finetuned LLMs and the best value for the ICL of the closed source LLMs and the best values highlighted. TxGemma 9B gave the best ROC_AUC value (0.7315 ± 0.0053) while LlaSmol-Mistral had the best EF1% (4.9865 ± 0.0048) and MCC (0.56415 ± 0.0257) while XGBoost gave the best inference time of 0.00004s per sample. In Table 2, we present the result for the few-shot learning of the open source models. Except LlaSmol with accuracies within the range of 0.7952 0.8020 for 3 5 shots, all other models have accuracies less than 0.3 (Gemma 9B), 0.08 (TxGemma 2B) and 0.008 (TxGemma 9B). LlaSmol also had the highest MCC (0.0721 at 4 shots) while Gemma 9B had the highest AUC (0.5845 at 3 shots) and TxGemma had the highest EF1% (1.8160 at 3 shots). In Table 3, we present the result (mean) for the few-shot learning of the closed source models, with OpenAI o3 giving the higher AUC values (highest - 4 shots 0.5907) while Gemini 2.5 had the higher EF1% (highest - 4 shots, 3.788). For the MCC value, while o3 on average had higher values compared to Gemini, Gemini at 4 shots had the highest MCC (0.1689). In table 4, we present the compute requirement for training and inference. During training, the 2B models were trained on 16GB GPU while for inference, all models required atleast 24GB GPU for inference. Also, Gemma 2B had the best inference time (s) per molecule of 0.004s after finetuning while TxGemma 2B had the best inference time (s) per sample (0.0142s) during few-shoting. This value is further explored in Figure 1 and 2 where the inference time (s) per molecule is plotted against EF1% and AUC.

5 Discussion

The results of this study yield several interconnected findings that collectively illuminate the conditions under which LLMs can contribute meaningfully to antimalarial virtual screening, the boundaries of their utility, and the trade-offs between performance, domain specialisation, and computational resource requirements. We discuss these findings in turn, situating them within the broader context of molecular ML methodology.

Table 1: General Performance results of the experiments.

Category	Task Type	Model	Size	ROC_AUC	MCC	EF (1%)	Inf. Time (s)
Classical	Finetuned	RandomForest	–	0.6828	0.2581	2.5790	0.0004
	Finetuned	XGBoost	–	0.6652	0.2164	1.7908	**0.00004**
Open Source LLMs	Finetuned	Gemma 2	2B	0.7063	0.47735	3.66045	0.0040
	Finetuned	TxGemma 2	2B	0.6955	0.53275	4.6646	0.0047
	Finetuned	LlaSmol - Mistral	7B	0.70225	**0.56415**	**4.9865**	0.08015
	Finetuned	Gemma 2	9B	0.6383	0.4408	4.6347	0.0304
	Finetuned	TxGemma 2	9B	**0.73155**	0.5539	4.23025	0.03705
Closed Source LLM	Fewshot (Best)	OpenAI-o3	–	0.5284	0.1689	3.7879	–
		Gemini 2.5	–	0.5907	0.1584	1.64635	–

Table 2: Evaluation Metrics of the Fewshot Open Source Models

Model	Shots	Acc	AUC	MCC	EF
LlaSMol 7B	3	0.7952	0.5140	0.0545	1.5424
	4	0.8011	0.5196	**0.0721**	1.6980
	5	**0.8020**	0.5128	0.0455	1.4416
Gemma 2B	3	0.0061	0.5025	-0.0026	0.2971
	4	0.0059	0.5029	0.0053	0.7198
	5	0.0009	0.4993	0.0038	1.2486
TxGemma 2B	3	0.075	0.5363	-0.0258	0.8036
	4	0.0698	0.5268	-0.0221	0.8042
	5	0.0915	0.5535	-0.0156	0.669
Gemma 9B	3	0.2930	**0.5845**	0.0653	1.1100
	4	0.2622	0.5571	0.0475	1.1095
	5	0.2850	0.5697	0.0525	1.0761
TxGemma 9B	3	0.0073	0.4892	0.0290	**1.8160**
	4	0.0028	0.4991	0.0029	1.1037
	5	0.0023	0.4984	0.0059	1.2532

Table 3: Evaluation Metrics of the Fewshot Closed Source Models[cite: 39].

Model	Shots	ROC_AUC	MCC	EF (1%)
Gemini 2.5	3	0.5182	0.0945	2.342
	4	0.5284	**0.1689**	**3.788**
	5	0.5175	0.0883	2.254
OpenAI o3	3	0.5751	0.1285	1.411
	4	**0.5907**	0.1573	1.551
	5	0.5867	0.1583	1.646

Table 4: Compute requirement table of the finetuning and fewshoting task[cite: 101].

		Finetuning			Fewshot (Best)		
Size	Model	Train GPU	Inf. Time	EF	Inf. GPU	Inf. Time	EF
2B	Gemma 2	16GB	0.0040	3.660	24GB	0.0457	0.720
2B	TxGemma	16GB	0.0047	4.665	24GB	0.0142	0.804
7B	LlaSmol	24GB	0.0802	4.987	24GB	0.2337	1.698
9B	Gemma 2	24GB	0.0304	4.635	24GB	0.4895	1.110
9B	TxGemma	24GB	0.0371	4.230	24GB	0.0873	1.816

5.1 Fine-Tuning is Non-Negotiable: The Collapse of Few-Shot Performance Across All Model Classes

The most consequential finding of this study is the universal failure of few-shot in-context learning for antimalarial bioactivity prediction, observed across all model classes, including frontier proprietary models with reasoning capabilities. Notably, TxGemma-9B, which achieves the highest ROC-AUC (0.731) among fine-tuned models, collapses to ROC-AUC $\approx$ 0.499 under 4-shot prompting (as well as $<$0.01 accuracy), representing essentially random discrimination. This dramatic inversion cannot be attributed to insufficient model capacity or general reasoning ability: the same model, when fine-tuned with gradient-level domain adaptation, achieves strong discriminative performance. The performance of closed-source frontier models further underscores this conclusion. Gemini 2.5, evaluated under 3-shot conditions on a statistically powered subsample of 500 molecules, achieves ROC-AUC $\approx$ 0.52 indistinguishable from random. OpenAI o3, despite its advanced chain-of-thought reasoning architecture and strong general scientific performance, reaches only ROC-AUC $\approx$ 0.59 at best. These results are striking because both models represent the state of the art in general-purpose AI reasoning, with documented capabilities in mathematics, chemistry knowledge, and multi-step logical inference. Their failure on bioactivity prediction is not a failure of intelligence, it is a failure of the task-fit between reasoning-by-analogy and the empirical, high-dimensional structure-activity landscape that bioactivity prediction demands at the inference level. The frontier LLMs evaluated here have no mechanism for internalising the specific binding-activity relationships of antimalarial targets through a few SMILES-label examples. The gradients produced during fine-tuning provide the mechanism by which these specific dependencies are internalised into model parameters; in-context learning cannot substitute for this process regardless of the base model's general capability. More broadly, these results reinforce the finding from LlaSMol [Yu *et al.*, 2024] and prior ICL benchmarks [Guo *et al.*, 2023].

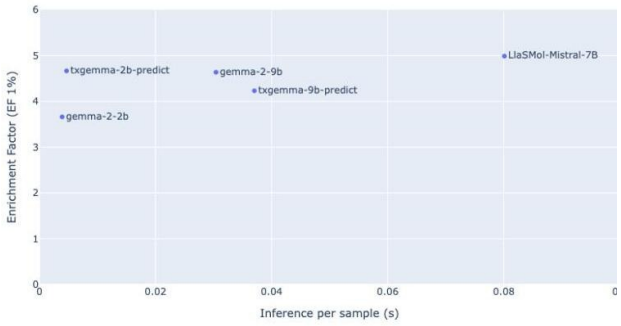

Figure 1: Plot of the Enrichment Factor 1% against inference time (s) per molecule.

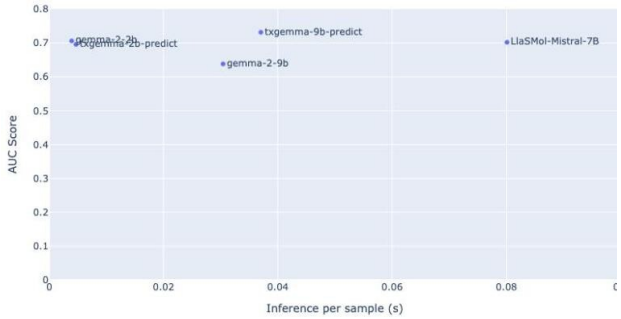

Figure 2: Plot of the AUC Score against inference time (s) per molecule

5.2 Does Biomedical Pretraining Confer a Genuine Advantage? Dissecting TxGemma's ROC-AUC Lead

Among fine-tuned models, TxGemma-9B achieves the highest ROC-AUC (0.731 ± 0.005), outperforming both the Random Forest baseline (0.683) and the general-purpose Gemma-2-9B (0.638). The within-scale comparison between TxGemma-9B (ROC-AUC: 0.731) and Gemma-2-9B (ROC-AUC: 0.638) models identical in architecture but differing in the domain of their pretraining fine-tune, provides the cleanest experimental window onto this question. The 9.3-percentage-point ROC-AUC advantage of TxGemma-9B over its general-purpose counterpart at equivalent scale is consistent with the hypothesis that biomedical pretraining confers a genuine and transferable advantage for molecular bioactivity classification. We interpret this advantage as arising from TxGemma's extensive exposure to diverse molecular property prediction tasks during TDC fine-tuning, which likely shapes the model's internal representations of SMILES tokens towards biologically grounded features, including patterns associated with pharmacophoric activity, target-binding motifs, and ADMET-related structural characteristics. Interestingly, the same comparison at the 2B scale reveals a much narrower performance gap: TxGemma-2B achieves ROC-AUC 0.695 versus Gemma-2-2B's 0.706, with Gemma-2-2B marginally superior. The 2B scale comparison complicates the domain-pretraining narrative in a way that warrants explicit acknowledgement. At 9B parameters, TxGemma outperforms Gemma-2 by 9.3 ROC-AUC percentage points, a gap consistent with biomedical pretraining providing a meaningful representational prior. At 2B parameters, this advantage inverts: Gemma-2-2B achieves ROC-AUC 0.706 against TxGemma-2B's 0.695. We interpret this reversal as a capacity-dependent interaction: at smaller parameter scales, the gradient signal from Malaria-Instruct fine-tuning is sufficient to overcome the head-start conferred by TDC pretraining, whereas at 9B the richer representational capacity of the larger model allows TxGemma's biomedical priors to be more effectively leveraged during task-specific adaptation. However, these findings require more extensive evaluation in subsequent works. The biomedical specialisation of TxGemma confers a measurable advantage specifically at larger model scales, consistent with the hypothesis that the benefit of domain pretraining scales with the model's capacity to utilise the additional representational priors it provides. This interaction between model scale and domain pretraining has important practical implications: research groups with access only to smaller models may find that the additional overhead of deploying specialised therapeutic LLMs yields limited marginal benefit over general-purpose alternatives at comparable parameter counts.

5.3 Chemistry-Aware Pretraining and Enrichment: Interpreting LlaSMol's Superior Enrichment Factor

While TxGemma-9B achieves the highest ROC-AUC, LlaSMol-Mistral attains the best enrichment factor at the 1% level (EF@1% $\approx$ 4.99 ± 0.005), substantially exceeding all other models — including TxGemma-9B (EF $\approx$ 4.23), Random Forest (EF $\approx$ 2.58), and XGBoost (EF $\approx$ 1.79). LlaSMol-Mistral also achieves the highest MCC (0.564 ± 0.026), suggesting that its decision boundary is better calibrated for the class-imbalance conditions characteristic of antimalarial screening data. These results indicate that chemistry-aware pretraining on SMolInstruct encodes structural pharmacophoric features that are particularly well-suited to prospective hit enrichment, even though LlaSMol-Mistral's global discriminative performance (ROC-AUC: 0.702) is not the highest in the cohort. A model can achieve high EF@1% through excellent concentration of actives in the extreme tail of its score distribution, even if its overall ranking is not perfect and LlaSMol-Mistral appears to achieve precisely this: exceptional precision at high confidence thresholds, reflecting its chemistry-aware pretraining's ability to assign high scores specifically to structurally credible antimalarial scaffolds. We hypothesise that this enrichment advantage arises from LlaSMol's training on SMolInstruct's diverse chemistry tasks, which include molecular property prediction, reaction prediction, and molecular name conversion. This multi-task molecular training likely equips LlaSMol with a richer internal vocabulary of substructure-property associations, including the recognition of pharmacophoric features associated with antiplasmodial activity, that is directly leveraged during Malaria-Instruct fine-tuning. By contrast, TxGemma's TDC pretraining is broader in biological scope (spanning cell lines, proteins, and clinical outcomes) but may be less densely populated with the struc-

tural chemistry examples needed for fine-grained pharmacophoric enrichment. The coexistence of TxGemma-9B's ROC-AUC lead and LlaSMol-Mistral's EF@1% lead is not a contradiction to be resolved but a model-selection signal to be acted upon. For researchers reporting to a benchmark leaderboard or comparing models on global discriminative capacity, TxGemma-9B is the recommended choice. For researchers deploying a model in a prospective VS campaign, where the only compounds that reach experimental assay are those in the top-ranked fraction, LlaSMol-Mistral is the operationally superior model: its EF@1% of 4.99 concentrates nearly five true actives for every one expected by chance, outperforming TxGemma-9B's 4.23 by a margin that translates directly into measurable reductions in experimental cost per confirmed hit.

5.4 The ROC-AUC vs. Enrichment Factor Tension: What Are We Actually Optimising For in Virtual Screening?

The discordance between ROC-AUC and EF@1% rankings in our results raises a fundamental question about model selection for VS applications: which metric should govern the choice of model for practical deployment? We argue that EF@1% is the operationally primary metric for VS, and that optimisation for ROC-AUC alone may lead to suboptimal model selection in resource-constrained screening campaigns. In a practical VS workflow, the output of a computational model is used to rank a chemical library of potentially millions of compounds, from which a small fraction, typically 1% or less, is selected for experimental synthesis and assay. The productivity of the screening campaign is determined almost entirely by the hit rate in this top-ranked fraction. A model that achieves ROC-AUC 0.731 but EF@1% of 4.23 (TxGemma-9B) produces a smaller fraction of confirmed actives per experimental unit cost than a model with ROC-AUC 0.702 but EF@1% of 4.99 (LlaSMol-Mistral), even though its global discriminative performance appears superior. From the perspective of return on investment in experimental screening, LlaSMol-Mistral is the preferred model, a conclusion that would be obscured by relying exclusively on ROC-AUC. This consideration has direct implications for the design of training objectives in molecular ML. Many existing bioactivity prediction models are trained with cross-entropy loss, which optimises a global classification objective and implicitly targets ROC-AUC-adjacent performance.

5.5 Classical Models Remain Competitive Baselines But at What Cost?

The Random Forest baseline achieves a ROC-AUC of 0.683 and EF@1% of 2.58, which while significantly below the performance of the best fine-tuned LLMs, constitutes a strong and highly resource-efficient result. For large-scale VS campaigns screening libraries of tens of millions of compounds, this difference in inference latency is non-trivial: LlaSMol-Mistral would require approximately 93 GPU-days to screen a 100-million-compound library, whereas Random Forest would accomplish the same task in under an hour on a single CPU core. It should also be noted that the performance advantage of fine-tuned LLMs over classical models, is measured under strict scaffold-dissimilarity conditions. Under random splitting, the gap between classical and LLM-based models is likely to be substantially reduced, as both model classes benefit from the ability to interpolate between structurally similar training and test molecules.

5.6 Reproducibility and Variance: What the Experimental Design Reveals

The experimental designs employed for fine-tuned and few-shot models are not parallel replication schemes, they are structurally distinct and answer different inferential questions. Interpreting the variance patterns in each case requires that distinction to be made explicit. For fine-tuned models, all experiments were conducted in duplicate with fully independent Lo-Hi resplitting of the Malaria-Instruct corpus for each replicate. Each duplicate therefore constitutes a genuinely independent experimental unit: a different dissimilarity-enforcing partition, a different training trajectory, and a different sequence of weight updates. The resulting standard deviations capture the *compound variance* of both partition-induced distributional shift and training stochasticity simultaneously. That the observed standard deviations remain small across all fine-tuned models: TxGemma-9B: ROC-AUC ±0.005; LlaSMol-Mistral: EF@1% ±0.005, is therefore a materially strong stability claim. It asserts that fine-tuned model performance is consistent regardless of which valid Lo-Hi partition is drawn from the corpus, a property that would not be guaranteed given the structural constraints of dissimilarity splitting and the relatively small size of individual assay-level subsets. This stability simultaneously validates the models and the benchmark: Malaria-Instruct produces evaluation conditions that are reproducible under the most demanding splitting conditions the cheminformatics literature currently prescribes. For few-shot models, the design is structurally different in a way that strengthens, rather than weakens, the null-result argument. The dataset was independently resplit before constructing each shot-count condition. This means variation observed across shot counts reflects the *joint effect* of shot count and partition change simultaneously, the two sources of variance are intentionally confounded within the shot-count comparison. The inferential consequences of this design and high variance of closed-source models are further discussed in appendix G and H respectively.

5.7 Computational Accessibility and Practical Deployment Considerations

A defining feature of this study is its focus on resource-constrained computational environments. All fine-tuned models in this study required either a 16GB or 24GB GPU for training and a minimum of 24 GB GPU for inference, corresponding to current mid-range GPU configurations (e.g., NVIDIA A100 40GB, available through Google Colab Pro+). Inference time per molecule ranges from 0.004 s (Gemma-2-2B) to 0.080 s (LlaSMol-Mistral), corresponding to throughputs of approximately 12,500 and 750 molecules per minute, respectively. These figures define the practical boundaries of VS campaign scale for each model: at 750 molecules per minute, LlaSMol-Mistral could screen a 100,000-compound

library in approximately 2.2 hours on a single GPU, a feasible timeline for academic VS campaigns. The scatter plot of inference time per molecule versus ROC-AUC and EF@1% (Figure 2) reveals an important efficiency frontier: Gemma-2-2B and TxGemma-2B occupy the optimal quadrant of high performance and low inference latency, processing approximately 4–7 ms per molecule while achieving ROC-AUC $>$ 0.695. These models represent the most resource-efficient operating point for VS applications where throughput is a primary constraint. LlaSMol-Mistral, despite its superior enrichment performance, requires approximately 20-fold more inference time, a trade-off that is acceptable for smaller library screening but may be prohibitive at scale. A significant caveat is that the inference time figures reported in Table 5 should be interpreted in the context of the inference framework employed. All reported per-molecule inference times were obtained using vLLM [Kwon *et al.*, 2023] as inference under standard PyTorch, the default framework for model deployment, is substantially slower by one to two orders of magnitude for all LLMs evaluated with extended analysis of why provided in appendix H. The QLora parameter-efficient fine-tuning approach (4-bit quantization, rank 8) adopted in this study is instrumental in achieving these resource efficiency figures. By reducing the effective memory footprint of 7–9B parameter models to the 24 GB GPU RAM regime, QLora enables the fine-tuning and deployment of models that would otherwise require 40–80 GB GPU configurations, extending their accessibility to a substantially broader research community. These results demonstrate that instruction-tuned open-source LLMs fine-tuned via QLoRA represent a practically achievable, resource-efficient alternative to classical cheminformatics pipelines for antimalarial VS in low-resource settings. This characterisation directly addresses the GPU-constrained reality of researchers in malaria-endemic settings, for whom free-tier platforms (Colab, Kaggle) are often the only available compute.

5.8 Limitations

Several limitations of this study should be acknowledged. First, Malaria-Instruct is derived exclusively from the ChEMBL Legacy Malaria corpus, which while comprehensive, captures only chemotypes that have historically been screened against Plasmodium, leaving dark chemical space, novel structural classes not previously investigated, entirely unexplored. The generalisation performance reported here therefore characterises structural novelty within the known antimalarial chemical universe, not the full scope of potential drug-like space. Second, all molecular representations in this study are SMILES-based, treating molecules as linear strings without explicit encoding of three-dimensional conformational information, stereochemical preferences, or target-binding geometries. Multi-modal representations that integrate SMILES with 3D structural features, molecular graphs, or protein pocket descriptors may offer substantially improved predictive performance, particularly for target-specific activity prediction where the binding site geometry is the primary determinant of activity. Third, the evaluation presented here is exclusively in silico: no experimental validation of model-predicted hits has been conducted. The practical utility of these models for prospective VS ultimately rests on the chemical and biological validity of their top-ranked predictions, which requires wet-lab confirmation. Prospective experimental validation of model-predicted antimalarial candidates represents the essential next step from this work. Lastly, the dataset reflects heterogeneity in ChEMBL assay provenance: different assays employ different Plasmodium strains (P. falciparum 3D7, Dd2, K1), measurement timepoints (48h vs. 96h), and readout modalities. While assay-level splitting and contextualisation partially address this heterogeneity, residual inter-assay variability may introduce label noise that differentially affects model classes.

6 Conclusion and Future Directions

The central empirical finding of this work is unambiguous: domain-specific fine-tuning is a categorical prerequisite for reliable antimalarial bioactivity prediction with LLMs. The 23-percentage-point ROC-AUC collapse observed in TxGemma-9B from 0.731 under fine-tuning to 0.499 under few-shot in-context learning best describes this with the pattern consistent across all model classes, including frontier proprietary models. Among fine-tuned models, biomedical and chemistry-aware pretraining both confer measurable advantages, but in ways that are metric-dependent and scale-sensitive. TxGemma-9B achieves the highest global discriminative performance (ROC-AUC: 0.731 ± 0.005), with its advantage over general-purpose Gemma-2-9B concentrated at the 9B parameter scale, suggesting a capacity-dependent interaction between domain pretraining and task-specific adaptation that diminishes at smaller model sizes while LlaSMol-Mistral, achieves the highest enrichment factor (EF@1%: 4.99 ± 0.005) and MCC (0.564 ± 0.026), outperforming all models on the metrics most directly relevant to prospective screening campaign productivity. Classical machine learning baselines also remain competitive reference points with significant throughput advantage over fine-tuned LLMs spans three to four orders of magnitude, and for VS campaigns operating at library scales of tens of millions of compounds, suggesting a tiered strategy combining classical pre-filtering with LLM re-ranking of shortlisted candidates may represent the optimal resource allocation. The inference time and GPU memory characterisation provided in this study, offers the quantitative foundation needed to make these deployment decisions on an evidence-based rather than intuitive basis. Several promising directions also emerge from this work. However, the most important is the prospective experimental validation of top-ranked model predictions, synthesising and screening the highest-confidence predicted actives from diverse structural clusters, represents the ultimate test of VS model utility. Such validation would provide both scientific evidence of practical impact and training data for subsequent model improvement cycles, closing the loop between computational prediction and experimental drug discovery.

References

[Bemis and Murcko, 1996] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. Molec-

ular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

[Breiman, 2001] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[Cao *et al.*, 2024] H. Cao, Z. Liu, X. Lu, Y. Yao, and Y. Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery, 2024.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[Chicco and Jurman, 2020] Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13, 2020.

[Chithrananda *et al.*, 2020] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[Fabian *et al.*, 2020] Benedek Fabian, Thomas Edlich, Hannes Gaspar, et al. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

[Fang *et al.*, 2023] Yin Fang, Y. Liang, N. Zhang, et al. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

[Guo *et al.*, 2023] T. Guo, K. Guo, Nan B., et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks, 2023.

[Hu *et al.*, 2022] Edward J. Hu, Yali Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685.

[Huang *et al.*, 2021] Kexin Huang, Tianfan Fu, Wenhao Gao, et al. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021.

[Kwon *et al.*, 2023] W. Kwon, Z. Li, S. Zhuang, et al. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023)*, pages 611–626, 2023.

[Lancet Microbe, 2025] Lancet Microbe. WHO world malaria report 2024. *The Lancet Microbe*, 6(2), 2025.

[Meidı *et al.*, 2024] S. Meidı, G. P. Wellawatte, and P. Gkeka. On the importance of dataset design: Better splitting for realistic evaluation of machine learning in molecular science. *arXiv preprint arXiv:2408.17052*, 2024.

[Mendez *et al.*, 2019] David Mendez, Anna Gaulton, A. Paula Bento, et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[Praski *et al.*, 2026] M. Praski, J. Adamczyk, and W. Czech. Benchmarking pretrained molecular embedding models for molecular representation learning, 2026.

[Riviere *et al.*, 2024] M. Riviere, S. Pathak, P.G Sessa, et al. Gemma 2: Improving open language models at a practical size, 2024.

[Steshin, 2023] S. Steshin. Lo-Hi: Practical ML drug discovery benchmark. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

[Wang *et al.*, 2025] E. Wang, S. Schmidgall, Paul F. Jaeger, et al. Txgemma: Efficient and agentic llms for therapeutics, 2025.

[Wei *et al.*, 2021] Jason Wei, Maarten Bosma, Vincent Y. Zhao, et al. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[Weininger, 1988] David Weininger. SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[WHO, 2024] WHO. *World Malaria Report 2024*. WHO Press, Geneva, 2024.

[Yu *et al.*, 2024] B. Yu, F. N. Baker, Z. Chen, X. Ning, and H. Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *Proceedings of the Conference on Language Modeling (COLM 2024)*, 2024. arXiv:2402.09391.

[Zambrano Chaves *et al.*, 2024] J. M. Zambrano Chaves, E. Wang, T. Tu, et al. Tx-LLM: A large language model for therapeutics. *arXiv preprint arXiv:2406.06316*, 2024.

[Zdrazil *et al.*, 2024] B. Zdrazil, E. Felix, F. Hunter, et al. The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 2024.

Appendix

A. ChEMBL Dataset

ChEMBL is a manually curated, large-scale open-access bioactivity database that integrates compound-target inter-

action data extracted from the primary medicinal chemistry literature, depositor submissions, and partner databases [Mendez *et al.*, 2019][Zdrazil *et al.*, 2024]. At present, ChEMBL contains bioactivity data for over two million distinct compounds, making it the most comprehensive public resource for QSAR modelling and computational drug discovery. Its ChEMBL Legacy Malaria dataset, compiled in partnership with the Medicines for Malaria Venture and aggregating decades of antimalarial screening data, constitutes one of the richest freely available sources of antimalarial bioactivity data, spanning multiple assay formats, Plasmodium strains, and time periods.

B. Instruction Tuning and Dataset Curation for Scientific LLMs

The instruction tuning paradigm, in which language models are fine-tuned on curated collections of (instruction, response) pairs to improve instruction-following fidelity and task generalisation, has fundamentally transformed the practical utility of large language models [Wei *et al.*, 2021][Ouyang *et al.*, 2022]. In the domain of scientific LLMs, instruction tuning has been applied to produce models capable of responding to molecular queries in natural language, with the instruction format serving as a compositional scaffold that enables multi-task learning across diverse property prediction objectives. Prior benchmark datasets for molecular instruction tuning include Mol-Instructions [Fang *et al.*, 2023] and SMolInstruct [Yu *et al.*, 2024]. Mol-Instructions spans molecule-text translation, property prediction, and molecular design tasks, while SMolInstruct extends coverage to 14 distinct chemistry tasks with rigorous quality control. Models trained on InstructMol [Cao *et al.*, 2024] integrate molecular graph representations with language instructions, leveraging structural inductive biases that pure SMILES-text models lack. The Tx-LLM and TxGemma data preparation pipeline [Zambrano Chaves *et al.*, 2024], [Wang *et al.*, 2025] further extends instruction tuning to multi-modal therapeutic tasks that integrate molecular, protein, cell line, and disease ontology information within a single prompt framework.

C. Biomedical Domain Specific Pretrained Models

Tx-LLM [Zambrano Chaves *et al.*, 2024], the direct predecessor to TxGemma, was a generalist therapeutic LLM fine-tuned from PaLM-2 on 709 datasets targeting 66 tasks spanning the drug discovery pipeline. Tx-LLM demonstrated that a single set of model weights could simultaneously encode knowledge about small molecules, proteins, nucleic acids, cell lines, and diseases, achieving near-state-of-the-art performance on 43 of 66 Therapeutics Data Commons (TDC) tasks [Huang *et al.*, 2021]. TxGemma [Wang *et al.*, 2025] improves on Tx-LLM by adopting the Gemma-2 architecture and retraining on the full TDC benchmark suite of 66 therapeutic tasks using 7 million curated examples. TxGemma-Predict, the predictive variant is available in 2B, 9B, and 27B parameter configurations, all of which have been benchmarked against specialist models on the TDC suite. Notably, TxGemma-9B-Predict has been validated to improve over Tx-LLM on 45 of 66 TDC tasks and to match or exceed best-in-class specialist model performance on 50 tasks

D. Evaluation considerations

The decision of which evaluation metric is chosen shapes the conclusions drawn from VS model benchmarks. Accuracy, the fraction of correctly classified instances, is a misleading primary metric under class imbalance: a model that always predicts the majority class (inactive) can achieve greater than 90% accuracy while providing no practical utility for VS. ROC-AUC provides a global measure of discriminative capacity that is insensitive to class imbalance and corresponds to the probability that a randomly drawn active is ranked above a randomly drawn inactive, making it a standard and interpretable benchmark metric. Matthews Correlation Coefficient (MCC) offers a balanced single-number summary that incorporates all four cells of the confusion matrix and is considered more informative than F1 under severe class imbalance [Chicco and Jurman, 2020].

E. Need for Strong Baseline

Praski *et al.* (2026) cautions against the assumption that architectural complexity necessarily translates to improved generalisation. In a systematic comparison of deep learning approaches to fingerprint-based baselines, they found that complex deep learning approaches offer negligible or non-significant improvements to fingerprint-based baselines in many bioactivity prediction settings, particularly under realistic data-splitting conditions. This necessitated our inclusion of Random Forest and XGBoost as primary classical baselines, not merely as weak foils, but as genuinely competitive reference points that LLMs must measurably surpass to justify their substantially greater computational cost.

F. Training Parameters

All LLMs were fine-tuned using QLoRA [Dettmers *et al.*, 2023] with 4-bit quantisation, LoRA rank 8, and the following target modules: q_proj, o_proj, k_proj, v_proj, gate_proj, up_proj, and down_proj. Training was conducted for one epoch over the full training partition. Optimisation used the paged AdamW 8-bit optimiser with a learning rate of 2×10^{-4}, weight decay of 0.001, and a warmup of 2 steps. Batch configuration was set to a train batch size of 2 with gradient accumulation of 2, or a train batch size of 4 with gradient accumulation of 1, depending on available GPU memory.

G. Inference Consequence of Few-shots resplitting

The confounding impart of the dataset being independently resplit before constructing each shot-count condition is analytically. While the variation observed across shot counts reflects the *joint effect* of shot count and partition change simultaneously, it answers the question that if in-context learning were genuinely capable of extracting structural bioactivity signals, performance should either be consistently good or improve monotonically as shot count increases *even across different partitions*, because each additional SMILES-label example should convey incrementally useful structure-activity information irrespective of the evaluation neighbourhood. The absence of any such monotonic improvement and

in several cases outright performance degradation from 3-shot to 5-shot cannot be attributed to a fixed partition artefact, because the partition changes with each condition. The near-random ICL performance is therefore robust to both variables simultaneously, substantially strengthening the conclusion that the failure is intrinsic to the task-model mismatch rather than an accident of a particular structural partition.

H. Closed Source Few-Shot Variance

The high variance observed in closed-source model performance under few-shot conditions carries a distinct and more pointed interpretation. Gemini 2.5 achieves EF@1% ranging from 2.34 to 3.79 across duplicate evaluations at 4-shot (±2.499), despite the shot count and partition being held constant within each duplicate pair with only the identity of the specific few-shot examples varying. This sensitivity to example identity is the behavioural signature of a model without genuine task grounding: a model that has internalised the structural determinants of antimalarial bioactivity would produce consistent rankings regardless of which specific SMILES-label pairs appear in the prompt. A model responding sensitively to the idiosyncratic features of individual example molecules is instead retrieving surface-level patterns from its prompt context. The high variance in closed-source ICL performance is therefore not methodological noise to be controlled away, it is itself evidence of the fundamental instability of in-context bioactivity prediction, and should be interpreted as such.

I. Inference Engine for LLMs: vLLM vs PyTorch Native

The source of this gap is between pytorch native implementation and vLLM is architectural rather than hardware-dependent. Autoregressive generation in standard PyTorch allocates GPU memory for the key-value cache on a per-sequence basis without batching or memory reuse across requests, producing significant GPU idle time between token generation steps even for short outputs. Because activity label prediction requires generating only one or two tokens per molecule, the ratio of cache initialisation overhead to productive computation is particularly unfavourable, amplifying the inefficiency relative to longer-form generation tasks where the fixed overhead is amortised across many tokens. vLLM's PagedAttention mechanism eliminates this fragmentation by managing the KV cache in fixed non-contiguous memory blocks, enabling continuous batching across variable-length requests and achieving near-complete GPU memory utilisation. The implication for practitioners is significant: the inference times reported here are achievable only with vLLM or a comparable optimised inference framework. Researchers deploying fine-tuned LLMs for VS using standard PyTorch should anticipate inference throughputs one to two orders of magnitude below those reported in Table 5, which would render library-scale screening with LlaSMol-Mistral or TxGemma-9B computationally infeasible without dedicated inference optimisation. Reporting inference times without specifying the inference framework, a common omission in the LLM benchmarking literature, therefore produces figures that are not reproducible in standard deployment contexts and may substantially overstate the practical accessibility of LLM-based VS.