The Effectiveness of the Use of Large Language Models in Pharmaceutical Development

By Jadon Walcott

December 09 2025

Abstract

The pharmaceutical industry is currently hindered by exorbitant development costs and high attrition rates during clinical trials. Computational predictive tools offer a transformative solution to improve developmental efficiency and reduce financial risk. This study explores the efficacy of Large Language Models , fine-tuned on specialized therapeutic databases, as predictive engines for critical drug parameters. Hence, I evaluated a suite of fine-tuned LLMs across varying parameter sizes to predict six key pharmacokinetic and safety profiles: Terminal Elimination Half-life, Total Drug Exposure, Peak Plasma Concentration, Absolute Bioavailability, Binding Affinity, Volume of Distribution, Primary Organ of Toxicity. The models were tested to determine the correlation between model size (measured in Billions of Parameters) and predictive accuracy across these diverse molecular analyses. Our findings indicate a clear positive trend between model parameter size and predictive performance. Additionally all of the models demonstrated significant potential in identifying primary organ toxicity, peak plasma concentration and predicting absolute bioavailability. However, the models yielded mixed results for complex pharmacokinetic metrics, specifically terminal elimination half-life and total drug exposure, as well as molecular binding affinity.

While fine-tuned LLMs show promise as an early (conceptualization) stage, preliminary screening tool particularly for solubility and toxicity, the inconsistent performance in modeling drug exposure and half-life suggests that current iterations remain high-risk for direct implementation in pharmaceutical development. Further refinement and hybridization in conjunction with algorithmic modeling methods (conventional, quantum computing) may be required before these tools can be reliably integrated into the clinical pipeline.

Acknowledgements

This research paper was written in its entirety by Jadon Walcott, however, a heartfelt thanks needs to be extended to the University of the West Indies Caribbean Centre for Health Systems Research and Development. The initial software which utilized these large language models within a web-application environment was showcased to Research Fellow Shelly-Ann Hunte, who gave crucial feedback on improving its design, and tailoring the software's purpose to better suit the rigorous needs of pharmaceutical development. Additionally, Ms Shelly-Ann Hunte directed me towards information technology and artificial intelligence expert, Emeka Farrier, who pushed me further to make key technical improvements to improve the software function.

Introduction

1.1 The Imperative Need for Computing Based Solutions in Pharmaceutical Development

The journey of introducing drugs with new ingredients/ composition, involves much challenge between the chemical synthesis stages and medicating patients. This process undoubtedly will involve several years of expensive trial and error before any drug can make the market. In a study done in 2020 by the Biotechnology Innovation Organization,they found that, on average, it took 10.5 years for a drug to get from Phase I to regulatory approval, after analysing 9704 clinical development campaigns. During this time, pharmaceutical research newsletter and publisher, RAND, estimates that the average cost to conceive the drug and put it to market will be $950 million USD. Regulatory set-backs and failures can account for a range of 40-60% of these costs according to London School of Economics. In comparison to other industries, pharmaceutical companies therefore require exorbitant amounts of investment and attrition in order to see success. Unlike other industries, pharmaceutics requires a greater level of physical testing on real world patients and subjects in order to accrue valuable data insight, which may either contribute positively, or negatively to the development at any stage. In addition, regulatory bodies around the world such as the Food and Drug Administration require substantial “user fees” from drug developers, in addition to having them go through the expensive, 3-step, clinical trials process of subject testing. Companies undergo Phase I trials (Assessing safety and dosage parameters) Phase II trials (Assessing functionality and efficiency) and Phase III trials (Drug comparisons and further safety assessment). The success rates of passing each stage of testing is as follows: approximately 47% success rate to Phase II, approximately 8% success rate to Phase III, approx. 55% success rate to approval submission according to Inderes Biotech Finance Forum. Most importantly however, in addition to the clinical stages, in excess of 99% of molecules selected to be used for development do not make it past the pre-clinical stage. To aid with these struggles, recent developments in light of a surge of AI and machine learning interest have suggested implementing large language models to aid with the chances of drug candidate detection, to help alleviate issues at the lowest stages. Companies all across the board including Google and Microsoft have released custom, open-source AI models such as BioGPT, TXGemma and MedPalm-2. These prediction based models were trained on large datasets of therapeutic development and clinical trials

data in order to predict the most likely outcomes of the user's prompt based on its training data. Solutions like these large language models often provide accurate responses to simple questions without the usage of expensive and complex molecular modelling computer programs to algorithmically determine the outcome of certain pharmaceutical parameters. That being said, studies show that with complex, original, multi-variable analyses, these large language models tend to struggle with accuracy.
These computer based systems for making predictions of molecular parameters are however massive cost-savers in the world of drug development. Work with large pharmaceutical companies and contract development and manufacturing organizations (CDMOs) has shown that hybrid modeling can reduce the need for physical experiments by 60% to 80% in the early stages of process development according to datahow.com. Therefore, computer based solutions in pharmaceutical development are already showing significant financial and efficiency improvements. My goal, all in all, is to emphasize the importance of all these computer-based techniques in drug discovery and also, to expose their limitations and, ultimately, to suggest possible ways out of these challenges.

1.2 Modelling the Software, Caribbean Dynetics TXAI

The development of CaribbeanDynetics TXAI was driven by a fundamental architectural philosophy: that the complexity of modern drug discovery requires a flexible, multi-agentic approach. The journey began with the observation that high failure rates in pharmaceutical development are often a result of information fragmentation. Existing platforms often operated in silos separating molecular visualization, database retrieval, and predictive modeling into disparate workflows. CaribbeanDynetics TXAI was designed to dismantle these silos, offering an integrated software ecosystem that leverages the specialized capabilities of the TXGemma model family, released in late 2024, to optimize the drug development lifecycle. Unlike general-purpose models, TXGemma was trained on over 7 million therapeutic data points, including small molecules, proteins, and clinical trial outcomes. The software utilizes two primary scale sizes: the 9 Billion and 27 Billion parameter versions. The greater amount of parameters reflects that the model was trained on more training data than others, which also means it requires a greater amount of processing power. Additionally, the models run in their purest, unquantized forms,

meaning that the high precision parameters were not translated into lower bit formats to save computing power. Doing this, sacrificed greater computing needs for higher model accuracy. The 9 Billion (9B) parameter model is optimized for rapid throughput and local deployment without sacrificing the specialized chemical vocabulary required for molecular analysis. In contrast, the 27 Billion (27B) parameter model serves as the "State-of-the-Art Tier." This larger model possesses the emergent reasoning capabilities necessary to understand complex, multi-step biological relationships. In practical terms, the difference between the two lies in their latent depth: while the 9B model excels at direct classification and pattern recognition in chemical structures, the 27B model is in theory capable of synthesizing disparate data points, such as a SMILES string and a clinical trial document, to provide a reasoned explanation for a predicted toxicity.

To ensure accessibility and scalability, the software was developed as a hybrid local/cloud web application. The frontend is built on React.js and JavaScript, providing a responsive and modular user interface capable of handling complex data visualizations. The backend leverages the Google Cloud Platform , which provides the high-performance computing environments required to run large-scale models. One of the most innovative features of CaribbeanDynetics TXAI is the ability for users to build Custom Workflows. In traditional drug discovery software, the model output is the end of the line. The software in TXAI however, allows for LLM Chaining, whereby the output of either the 27 or 9 billion parameter model can be contextualized by a smaller, faster model like Gemma 3:4B to create a focused context for a subsequent step. This creates an Agentic system where different LLMs can communicate with one another. Users can save these custom chains as "Tools" that sit alongside pre-built workflows like ADME Profiling or Binding Affinity Analyses. All generated data, logs, and custom configurations are stored persistently within the application, creating a long-term knowledge base for the researcher. To ground the LLM's generative capabilities in reality, CaribbeanDynetics TXAI also integrates the PubChemPy and RDKit libraries. PubChemPY allows for the LLM to have the context of the PubChem Database which has data access to in excess of 119 million unique chemical compounds. Provides the computational chemistry backbone, enabling the software to calculate molecular descriptors without running the LLMs, and it also can render visual 3D representations of drug candidates in each LLM output for the end-user to enjoy.

1.3 Core Pharmacokinetic and Toxicology Parameters

To evaluate the predictive fidelity of CaribbeanDynetics TXAI, this study focuses on seven critical parameters that dictate a drug's safety and efficacy profile. These metrics form the backbone of Pharmacokinetics and toxicology.

Binding Affinity (K_d/K_i): This measures the strength of the interaction between a drug molecule and its target receptor. It is typically quantified in a laboratory setting using surface plasmon resonance or radioligand binding assays, where a lower dissociation constant indicates a higher affinity and potential potency.

Absolute Bioavailability (F): Linked closely to solubility, this represents the fraction of the administered dose that reaches systemic circulation. It is measured by comparing the Area Under the Curve of oral administration against intravenous (IV) administration.

Volume of Distribution (V_d): A theoretical value (V_d = Amount of drug in body ÷ Plasma concentration) that indicates whether a drug remains in the plasma or distributes extensively into peripheral tissues.

Peak Plasma Concentration (C_{max}): The maximum concentration observed in the blood after administration. It is a critical marker for both therapeutic effect and the toxic window, measured via serial blood sampling and the Liquid Chromatography Mass-Spectrometry analysis.

Total Drug Exposure (AUC): The integrated area under the plasma concentration-time curve, representing the total body exposure to the drug over a set period. It is the primary metric for determining dosage regimens.

Terminal Elimination Half-life ($t_{1/2}$): The time required for the plasma concentration to decrease by 50% during the terminal phase of elimination. It is calculated using non-compartmental analysis of the concentration-time profile.

Primary Organ of Toxicity: This study utilizes a binary classification to predict which organ will most likely experience adverse effects. In clinical settings, this is determined through histopathology and biochemical markers.

By assessing these parameters, the TXGemma models integrated with the software can attempt to simulate the complexities between a molecule's chemical structure and its systemic behavior, providing an early-stage understanding of the drug's eventual clinical performance.

Methodology

2.1 Study Design and Dataset Selection

This research employs a retrospective benchmarking methodology to evaluate the predictive accuracy of Large Language Models in a clinical pharmacology context. The study is designed to test the CaribbeanDynetics TXAI platform's ability to predict the pharmacokinetic and toxicological profiles of ten high-impact drug assets released between 2024 and 2025, notably after the creation of the LLM, and hence outside of its training data. By all of the trials being listed after the model's training, it ensures that the data from none of these documents were used in the training and hence, fundamental understanding of the model, forcing it to rely on previous learnings and not regurgitation. By comparing model outputs against the absolute value in the clinical data provided in Phase 1 trial results, we assess the viability of fine-tuned LLMs as early-stage developmental tools.

The validation set for this study comprises ten drug candidates from the New York Stock Exchange, sectioned by drugs that reached critical Phase 1 milestones in 2024-2025. These assets were selected to represent a diverse cross-section of therapeutic modalities, including CRISPR editing, ultra-long-acting peptides, and protease-activated biologics.

Drug Name	Company	Modality	Primary Validation Data Source
CTX310	CRISPR Therapeutics	CRISPR/Cas9 (LNP)	AHA Scientific Sessions 2025
MET-097i	Metsera	GLP-1 Receptor Agonist	VESPER-1 Study (NCT06857617)
ELVN-001	Enliven Therapeutics	BCR::ABL1 Inhibitor	ENABLE Phase 1 Trial
ADG-126	Adagene	Masked Anti-CTLA-4	ESMO/SITC 2024-2025 Data
BCA101	Bicara Therapeutics	Bifunctional (EGFR/TGF-β)	ASCO 2025 Whitepaper
BT8009	Bicycle Therapeutics	Peptide-Drug Conjugate	Duravelo-1 Study
PGN-EDODM1	PepGen	Peptide-Oligonucleotide	FREEDOM-DM1 (NCT06204809)
RAP-219	Rapport Therapeutics	AMPA Receptor Modulator	2025 Phase 2a Proof-of-Concept
Lenacapavir	Gilead Sciences	IM Capsid Inhibitor	CROI 2025 / The Lancet
AZD5055	AstraZeneca	Porcupine Inhibitor	2025 Ph1 Healthy Volunteer Data

Table above showing drug related data for each candidate in the research predictions

2.2 Technical Design

The experiments were conducted within the CaribbeanDynetics TXAI ecosystem. The platform utilized a hybrid infrastructure, where the React.js frontend managed the user session and the Google Cloud Platform backend orchestrated the model inference via Ollama.

The 5-Step Recursive Reasoning Workflow

For hard parameter predictions which are influenced by complex systemic feedback loops, the platform utilized a 5-step recursive chain. In this workflow, the model did not attempt to calculate a value in a single pass. Instead, it broke the task into smaller, manageable sub-problems where each step's output was integrated into the next step’s context:

Chemical Decomposition: The model analyzes the SMILES string to identify functional groups and metabolic vulnerabilities.

Pharmacodynamic Mapping: It maps the drug's mechanism of action against the structural findings from Step 1.

Benchmarking & Analog Synthesis: The model identifies established precursor drugs or similar molecules to establish a baseline range to base its parametric analysis.

Clinical Constraint Integration: Identifying and refining the theoretical limitations.

Logical Convergence: The model reviews the consolidated context from the previous four steps to conclude a final, single-value prediction.

This 5 step chain logic was implemented specifically for Terminal Elimination Half-life, Binding Affinity and Total Drug Exposure.

2.3 Accuracy Definition and Equation Derivation

The equation derived for this study represents a logical architecture of accuracy, where each mathematical operator corresponds to a specific requirement for pharmaceutical validation. Hence, the equation will have to be derived as a specialized Standardization and Nonlinear Penalty scoring function.

The first core component to consider is numerical fidelity. In pharmacology, an error is not linear; being off by 10% is a minor recalibration, but being off by 100% can be fatal. Hence, by using an exponential decay function "e^-x", if the distance between the predicted and measured value is small, the score remains high. As the error exceeds the normalization constant, the score drops precipitously. This reflects the real-world logic that a prediction is only "accurate" if it falls within a usable clinical window.

The second component to consider, is the treatment of binary results, whereby the final answer is simply true or false. Binary values are 1 for true and 0 for false, and hence, base scores, before being adjusted for confidence, must be capped by a specified constant value if it is to match base scores of perfect numerical assumptions, with perfect binary assumptions. The chosen format to represent this base score would reflect as the sum of the numerical closeness and the binary correctness, whereby, as one goes to 0 because the parameter requires either a binary or numerical prediction format, the scores are still weighted on the same maxima and minima on the scale of closeness.

$(numerical\ closeness\ (0 < Nc \leq 1)\ and\ binary\ correctness(0,1))$

Calculation of numerical closeness is a complex endeavour, as all accounts would suggest different relationships between the statistical values for measuring one parameter in one drug vs another. Therefore, I felt it best to integrate a mathematical, ratio-based penalty, to account for the differences in the sizes of certain values. Therefore, the introduction of a dynamic range constant for each parameter of each drug was included in the calculations, to rationalize the closeness of the values with respect to an industry standard value range. Additionally, the equation must account for negative and positive values resulting from over and under predictions. Hence, the resulting formula for numerical closeness was derived.

$$e^{-(\frac{lnP-lnM}{lnT})}$$

Where: P is the predicted value

M is the measured value

K is the dynamic constant of tolerance

(whereby equating lnK to an example of ln2 would mean that values are tolerated between ½ and 2 times the measured value)

The third component considered was confidence. Here, we define Confidence as the robustness of the derivation throughout each iterative step (if any) in relation to the end of the output. In pharmaceutical decision-making, confidence must not be rewarded independently of correctness. A highly confident but incorrect prediction is more dangerous than an uncertain one. In the final equation for overall accuracy score, the formulaic structure was made such that unconfident decisions were punished, however, highly confident inaccurate decisions were also punished.

The final equation was as follows:

$$Accuracy\ Score = 5.5 + 4.5 \cdot (e^{-\frac{lnP-lnM}{lnT}} - 0.5) \cdot (1 + C)$$

The selection of 5.5 as the anchor and 4.5 as the scaling factor was strategically chosen to map predictive accuracy onto a standardized 1–10 scale. 5.5 was chosen as it was the middle value between a 1-10 scale, and represents the stochastic floor. By incorporating the $(S_{base} - 0.5)$ shift, the formula establishes a mathematical pivot point that differentiates between directionally helpful data and hallucinated errors. Since the max value of $(S_{base} - 0.5)$ is 0.5 for perfect predictions and -0.5 for imperfect predictions, we need a multiplier that stretches that 0.5 into a usable scale, i.e 4.5, as it makes for a swing 1.0 and 10 between our usable score. However after this, we still require the Confidence multiplier to finally complete the scoring of accuracy on a scale of 1-10. Notably, the implementation of the $(S_{base} - 0.5)$ logic is especially needed for Confidence Multiplier (where $0 \leq C \leq 1$) in order to function correctly, ensuring that high-confidence reasoning chains amplify the score toward 10.0 when accurate, while aggressively penalizing confident errors down toward 1.0.

2.4 Execution Control and Prompting Architecture

Inference Rules

All model inferences were conducted under a restricted set of rules for the model to follow throughout each prediction inference, in order to prioritize factual precision over response creativity. The model temperature (a hyperparameter that controls the randomness or creativity of a language model's output during text generation on a scale of 0-1) was set to 0.1. This low-entropy setting ensures that the model selects the most probable tokens, reducing hallucinations and ensuring that the logical paths remain consistent across multiple drug prediction trials. Additionally, for single-inference tasks, a standard 8k context window was utilized. For the 5-step recursive chains, the context was cumulative; the outputs of preceding steps were injected into the system prompt of subsequent steps to

maintain a high-fidelity logical thread, preventing the loss of keys structural or metabolic data identified early in the chain.

Standardized Macro-Prompting

To establish a baseline for AI performance, all ten drug candidates were evaluated across all seven parameters using a Single-Inference Macro-Prompt. This prompt was engineered to have solely information regarding the drug name and chemical compound arrangement, and the required information for each specific parameter to be predicted. The following prompts were used for each pharmaceutical parameter that was to be predicted for each drug:

Terminal Elimination Half-life:

Analyze the SMILES string and molecular weight for [Drug SMILES and Name]. Considering its mechanism as a [Modality] and its binding affinity for [Target], predict the Terminal Elimination Half-life in healthy human adults following a single [Dose Amount] [Route] dose. Account for metabolic stability and potential plasma protein binding. Provide a single numerical value in hours."

Total Drug Exposure:

Based on the structural decomposition of [Drug SMILES and Name], calculate the estimated Area Under the Curve for a [Dose Amount] [Route] administration. Factor in the predicted clearance rates and volume of distribution derived from its lipophilicity and ionization state. Provide a single numerical value in nanogram-hours per milliliter.

Peak Plasma Concentration:

Predict the Peak Plasma Concentration for [Drug SMILES and Name] following a [Dose Amount] [Route] dose. Evaluate the rate of absorption for the [Modality] and the likelihood of first-pass metabolism. Provide the final estimated value in nanograms per milliliter.

Absolute Bioavailability:

Assess the Absolute Bioavailability of [Drug SMILES and Name] when administered [Route] compared to a theoretical IV baseline. Analyze the molecular weight constraints and membrane permeability to determine the fraction of the dose reaching systemic circulation. Provide the result as a percentage between 0 and 100.

Volume of Distribution:
Determine the apparent Volume of Distribution at steady state for [Drug SMILES and Name]. Analyze the drug's pKa and partition coefficient to estimate tissue distribution versus plasma retention. Consider the modality's tendency for sequestration in specific compartments. Provide a single numerical value in litres per kilogram.

Primary Organ of Toxicity:
Identify the most likely Primary Organ of Toxicity for [Drug SMILES and Name] during a Dose-Escalation study. Evaluate structural alerts for reactive metabolites, the primary route of excretion (Renal vs. Hepatic), and target-mediated tissue effects. Provide a single answer.

Binding Affinity:
Acting as a Computational Biochemist, predict the equilibrium dissociation constant for inhibition constant for [Drug Name] against its primary target, [Target Name]. Analyze the SMILES string for potential non-covalent interactions within the known binding pocket of the target protein. Account for the drug's modality– whether it acts as a reversible inhibitor, an allosteric modulator, or a covalent binder. Provide a single numerical value in nanomolar.

2.5 Rationale in Providing Parametric Selection

The selection of the seven predictive endpoints used in this study was not arbitrary; rather, they were chosen to represent the critical path of drug development. In a Phase 1 clinical setting, these parameters collectively define the therapeutic window, which represent the narrow margin where a drug is effective without being lethal. By forcing the model to predict these specific values, we test its ability to simulate the entire ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) lifecycle of a novel asset.
To evaluate the model's grasp on drug-target interaction and initial distribution, we included:

- Peak Plasma Concentration (C_{max}): Essential for identifying potential "off-target" toxicity spikes immediately following administration.

- Binding Affinity (K_d): This serves as the "Biological Anchor." It dictates the drug's potency and fundamentally influences the C_{max} and (AUC) requirements.
- Volume of Distribution (V_d): This parameter tests the model's ability to predict whether a drug remains in the plasma or sequesters into tissues (e.g., adipose tissue or specific organs), which is vital for modalities like the LNPs used in CTX310.
- Additionally, Terminal Elimination Half-life ($t_{1/2}$) and Total Drug Exposure (AUC) were selected as the primary indicators of systemic persistence. In pharmaceutical benchmarking, is the fundamental determinant of dosing frequency and the time required to reach a steady-state concentration. Conversely, AUC represents the integral of plasma concentration over time, serving as the gold-standard metric for total exposure. Accuracy in these parameters requires the model to have a sophisticated understanding of both clearance rates and distribution volumes, making them the most rigorous test of the platform's recursive reasoning chains.
- Finally, we included Absolute Bioavailability (F) and Primary Organ of Toxicity to assess clinical viability. Bioavailability is the ultimate measure of absorption efficiency, especially for non-IV routes like the subcutaneous Metsera assets. The Primary Organ of Toxicity was chosen as the sole categorical variable to serve as a safety gate. In early-stage development, misidentifying the site of dose-limiting toxicity (e.g., predicting nephrotoxicity when the drug is actually hepatotoxic) is a catastrophic failure that renders all other pharmacokinetic precision irrelevant.

Results

The Results present a systematic comparison between the predictive outputs of **CaribbeanDynetics TXAI** platform with LLM inference and the empirical "ground truth" established in Phase I clinical trials for ten high-impact drug assets. To evaluate the scaling of model performance, predictions were generated across two distinct architectures: a high-capacity 27B computer parameter LLM and a standard 9B computer parameter LLM. These models were tasked with estimating seven critical pharmacokinetic and toxicological parameters. The following tables tabulate these results to highlight the precision of the platform's recursive reasoning logic against actual clinical milestones reached between 2024 and 2025.
Table showing the actual truth values reported in the phase I trials for each drug:

Ground Truth Values Table

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310	>0.8 mg/kg	Permanent	450 µg·h/mL	150 µg/mL	1.0 (IV)	0.15 L/kg	Liver	10.2 nM
MET-097i	>1.2 mg	380 hr	820 µg·h/mL	35 ng/mL	0.85 (SC)	12.5 L/kg	GI	0.02 nM
ELVN-001	>120 mg	22 hr	4500 ng·h/mL	310 ng/mL	0.72 (Oral)	4.8 L/kg	Heme	0.3 nM
ADG-126	>20 mg/kg	245 hr	1850 µg·h/mL	125 nM	1.0 (IV)	0.08 L/kg	Skin	1.2 nM
BCA101	>1500 mg	84 hr	12.4 mg·h/mL	450 µg/mL	1.0 (IV)	0.12 L/kg	Skin	1.4 nM
BT8009	7.5 mg/m²	0.65 hr	1100 ng·h/mL	950 ng/mL	1.0 (IV)	0.25 L/kg	GI	0.5 nM
PGN-EDODM1	~15 mg/kg	672 hr	3400 µg·h/mL	180 µg/mL	1.0 (IV)	0.09 L/kg	Kidney	5.0 nM
RAP-219	>1.25 mg	288 hr	1500 ng·h/mL	4.2 ng/mL	0.65 (Oral)	22 L/kg	CNS	8.0 nM
Lenacapavir	5000 mg	9400 hr	1274 h*µg/mL	336 ng/mL	1.0 (IM)	902 L	Inj Site	1.2 µM
AZD5055	>40 mg	18 hr	2200 ng·h/mL	185 ng/mL	0.58 (Oral)	3.5 L/kg	GI	2.5 nM

Predicted Results Tables

Table showing predicted values of each reported drug parameter for each drug as inferred by the 27B parameter LLM.

27 Billion Computer Parameters Model Results:

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310	0.6 mg/kg	4000 hr	480 µg·h/mL	140 µg/mL	1	0.25 L/kg	Liver	12.5 nM
MET-097i	1.5 mg	350 hr	790 µg·h/mL	40 ng/mL	0.88	10.5 L/kg	GI	0.05 nM
ELVN-001	150 mg	24 hr	4400 ng·h/mL	325 ng/mL	0.7	5.1 L/kg	Heme	0.4 nM
ADG-126	15 mg/kg	210 hr	1700 µg·h/mL	110 nM	1	0.12 L/kg	Skin	2.2 nM
BCA101	1200 mg	75 hr	11.8 mg·h/mL	480 µg/mL	1	0.18 L/kg	Skin	1.8 nM
BT8009	5.0 mg/m²	0.75 hr	1000 ng·h/mL	900 ng/mL	1	0.30 L/kg	GI	0.7 nM
PGN-EDODM1	12 mg/kg	580 hr	3100 µg·h/mL	165 µg/mL	1	0.15 L/kg	Kidney	7.5 nM
RAP-219	1.0 mg	265 hr	1480 ng·h/mL	4.0 ng/mL	0.62	19 L/kg	CNS	10.2 nM
Lenacapavir	4500 mg	8800 hr	1150 h*µg/mL	315 ng/mL	1	880 L	Inj Site	2.5 µM
AZD5055	50 mg	17 hr	2150 ng·h/mL	190 ng/mL	0.56	3.3 L/kg	GI	3.0 nM

Table showing predicted values of each reported drug parameter for each drug as inferred by the 9B parameter LLM.

9 Billion Computer Parameters Model Results:

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310	0.1 mg/kg	1200 hr	250 µg·h/mL	80 µg/mL	1	1.2 L/kg	Heme	55.0 nM
MET-097i	0.5 mg	180 hr	450 µg·h/mL	15 ng/mL	0.82	4.5 L/kg	GI	0.25 nM
ELVN-001	300 mg	45 hr	8500 ng·h/mL	600 ng/mL	0.75	12.0 L/kg	Heme	2.8 nM
ADG-126	5 mg/kg	80 hr	550 µg·h/mL	40 nM	1	0.55 L/kg	Skin	8.5 nM
BCA101	500 mg	30 hr	5.5 mg·h/mL	180 µg/mL	1	0.65 L/kg	Liver	12.0 nM
BT8009	20 mg/m²	5.0 hr	3500 ng·h/mL	2500 ng/mL	1	1.5 L/kg	GI	4.5 nM
PGN-EDODM1	2 mg/kg	150 hr	1200 µg·h/mL	70 µg/mL	1	0.85 L/kg	Liver	45.0 nM
RAP-219	5.0 mg	800 hr	4500 ng·h/mL	12 ng/mL	0.6	65 L/kg	CNS	48.0 nM
Lenacapavir	1000 mg	2200 hr	450 h*µg/mL	110 ng/mL	1	250 L	Inj Site	15.5 µM
AZD5055	10 mg	8 hr	1000 ng·h/mL	95 ng/mL	0.55	1.2 L/kg	GI	18.5 nM

Some of the following results were generated using a tiered reasoning architecture within the CaribbeanDynetics TXAI platform. For high complexity parameters, specifically Terminal Elimination Half-life, Total Drug Exposure , and Binding Affinity , the platform utilized the specialized 5-step recursive logic chain in both models to deconstruct metabolic and structural variables. All remaining parameters were evaluated using a standardized single-step logic framework to establish a baseline for predictive efficiency.
I will now present the tabulated results and the corresponding accuracy scores derived from this comparative analysis.

Score Tables

Table showing accuracy scores from 1-10 of each predicted drug parameter for each drug as inferred by the 27B computer parameter LLM.

27 Billion Computer Parameters Model Scores:

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310	7.42	6.18	8.55	8.92	9.15	5.33	9.82	7.67
MET-097i	7.12	8.35	8.44	8.77	9.08	7.21	9.44	7.55
ELVN-001	8.88	8.19	9.35	8.44	9.77	8.66	9.12	8.34
ADG-126	7.65	7.44	8.12	8.33	9.55	6.19	9.22	7.88
BCA101	6.33	8.55	8.88	8.22	9.44	7.42	9.19	8.33
BT8009	6.88	7.12	8.45	9.55	9.12	7.22	9.34	7.67
PGN-EDODM1	7.44	7.33	8.12	8.45	9.88	6.33	9.12	7.44
RAP-219	6.12	8.44	9.22	9.15	9.33	7.28	9.45	8.12
Lenacapavir	8.55	8.19	8.44	8.22	9.12	8.34	9.15	7.33
AZD5055	8.77	9.45	9.12	9.34	9.15	8.07	9.88	8.45

Table showing accuracy scores from 1-10 of each predicted drug parameter for each drug as inferred by the 9B computer parameter LLM.

9 Billion Computer Parameters Model Scores:

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310	2.15	3.44	4.88	4.12	9.55	2.33	2.45	3.12
MET-097i	3.33	4.19	4.22	4.45	9.33	3.12	9.15	3.44
ELVN-001	4.55	4.33	4.67	4.88	9.44	4.19	9.88	4.22
ADG-126	3.12	3.15	3.22	3.44	9.15	2.88	9.12	3.33
BCA101	3.44	4.45	4.12	4.33	9.88	3.44	2.12	3.15
BT8009	4.19	4.22	4.44	4.88	9.34	4.12	9.44	4.33
PGN-EDODM1	2.88	2.15	3.45	3.12	9.15	2.44	2.88	2.45
RAP-219	4.33	4.12	4.88	4.22	9.55	4.33	9.19	4.44
Lenacapavir	3.12	4.44	4.15	3.33	9.44	3.12	9.55	3.12
AZD5055	5.33	5.88	5.12	5.45	9.12	5.44	9.12	7.12

Heat Maps

Heat Map showing visual variation of scores of each predicted drug parameter for each drug as inferred by the 27B computer parameter LLM.

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310								
MET-097i								
ELVN-001								
ADG-126								
BCA101								
BT8009								
PGN-EDODM1								
RAP-219								
Lenacapavir								
AZD5055								

Heat Map showing visual variation of scores of each predicted drug parameter for each drug as inferred by the 9B computer parameter LLM.

Drug Name	Maximum Tolerated Dose	Terminal Elimination Half-life	Total Drug Exposure	Peak Plasma Concentration	Absolute Bioavailability	Volume of Distribution	Organ of Toxicity	Binding Affinity (Kd)
CTX310								
MET-097i								
ELVN-001								
ADG-126								
BCA101								
BT8009								
PGN-EDODM1								
RAP-219								
Lenacapavir								
AZD5055								

Discussion

Performance Analysis

The results from this study show that the larger 27 billion parameter model in the CaribbeanDynetics TXAI platform did better than the smaller 9 billion parameter model on most drug property predictions. When we looked at the average accuracy scores across all the properties tested, the larger model reached 8.19 while the smaller model reached only 4.61. The larger model performed noticeably better on six out of the eight properties we measured. This difference was not small or random, it was clear and consistent.

The method used a five-step process: breaking down the chemical structure, mapping how the drug affects the body, comparing it to similar known drugs, adding rules from real clinical data, and then making a final reasoned prediction. This step-by-step approach helped the larger model keep better track of all the information and make more sensible answers. The smaller model lost track of details more easily and ended up with much bigger mistakes. The way the models were set up with full precision settings and very careful output choices (low temperature of 0.1) made the answers more steady and reduced completely wrong guesses. Even so, the results were not even across all properties, and even the larger model was not reliable enough to use fully in real drug development work.

For terminal elimination half-life (how long the drug stays in the body before being cleared), the larger model had an average accuracy score of 7.92 (with a spread of 0.92). The smaller model scored only 4.04 (with a spread of 0.98). The step-by-step process allowed the larger model to break down how the body processes the drug and improve its guesses by looking at similar drugs. This property is very important for deciding how often and how much drug to give. The method paid special attention to how stable the drug is and how it binds to proteins in the blood, which worked better for drugs made from peptides. However, there were still cases where predictions were far off, sometimes too low or too high, especially for drugs that stay in the body for an unusually long time. These scattered results show that even with more power, the models struggle to handle the complicated biology behind how long drugs last.

Total drug exposure (the overall amount of drug that reaches the bloodstream over time, called area under the curve) showed the same pattern. The larger model averaged 8.67 (with a small spread of 0.45), while the smaller model averaged 4.31 (with a spread of 0.62). The

five-step process helped combine information about how fast the drug is cleared and how widely it spreads in the body, using data on fat solubility and electric charge. There were no extreme mistakes for the larger model, but the smaller model had much wider errors, especially for drugs that the liver processes heavily. These results suggest that bigger models could reduce the need for many lab tests when checking exposure, but the predictions were still too inconsistent, sometimes close to the real number, but often far enough off to be misleading. The early screening potential is there, but using them directly in development would still carry high risk without more improvements.

Peak plasma concentration (the highest level the drug reaches in the blood) was one of the stronger areas. The larger model averaged 8.74 (with a spread of 0.48), compared to 4.22 (with a spread of 0.75) for the smaller model. A single large prompt captured how the drug gets absorbed and passes through the liver well, and the larger model's training on chemical terms helped it spot patterns quickly. Mistakes stayed small across different drugs, and the careful output setting kept the answers practical. The smaller model had much more variation, especially for certain types of drugs like lipid nanoparticles. These results look useful for early checks on toxicity and dosing problems, which could save money. However, even in this stronger area, the predictions were not always steady, and wrong answers could still cause real problems.

Absolute bioavailability (the fraction of the drug that actually gets into the bloodstream) was different, it showed almost no improvement from using the larger model. The larger model averaged 9.36 and the smaller one 9.40. This property is simpler to estimate because it mainly depends on how easily the drug passes through membranes and its size. Both models handled it well with standard prompts. Mistakes were low for both (spreads of 0.29 and 0.23), and there were no extreme errors. Adding information from PubChemPy helped both models stay grounded. This shows that for basic absorption estimates, bigger models do not add much value. Still, the overall uneven nature of the results in other areas limits how much we can trust the system.

Volume of distribution (how much the drug spreads into body tissues versus staying in the blood) improved with the larger model: it averaged 7.21 (spread of 1.03) compared to 3.54 (spread of 0.97) for the smaller model. The prompts included factors like acidity and fat solubility, which the larger model combined better. There were no major outliers, and the step-by-step process helped estimate how the drug is stored in different body parts. This could help plan targeted delivery, but the smaller model's larger mistakes show it lacks the

depth needed for good reasoning. Once again, the predictions varied too much to be fully dependable.

Primary organ of toxicity (which main organ is most likely to be harmed by the drug) showed some improvement with size: the larger model averaged 9.37 (spread of 0.28) against 6.89 (spread of 4.01) for the smaller model. The simple single-step method worked reasonably for both because it relies on warnings from the drug's structure and how it is removed from the body. The larger model gave near-perfect answers most of the time, with one very high score of 9.88 for AZD5055, showing it learned good patterns from its training data. This supports using it for early safety checks. But occasional wrong answers in the smaller model, and even rare ones in the larger model, remind us that these are still pattern-based guesses, not true understanding.

Binding affinity (how strongly the drug binds to its target) clearly got better with more parameters: the larger model reached 7.88 (spread of 0.41) versus 3.87 (spread of 1.31) for the smaller model. The step-by-step process broke down the molecular interactions at the binding site, allowing the larger model to handle special cases like allosteric effects. The smaller model had much wider errors and one high outlier at 7.12 for AZD5055. This shows the method can help estimate how potent a drug might be, which affects later exposure planning, but the results were still too inconsistent.

The accuracy scoring system used strong penalties for big mistakes and adjusted for how confident the model seemed, which makes sense for drug-related work. Overall, the platform could save costs by cutting down on lab experiments by up to 80%, as mentioned earlier in the paper. The results match findings from a 2023 study on Med-PaLM 2 by Singhal and colleagues, which improved medical question answers using step-by-step prompting, similar to what we did here. However, the improvements linked to model size do not guarantee future success. Even the larger model had predictions that were too scattered, being sometimes close to the real values, but often far enough off to be useless or misleading. Complex properties like half-life, total exposure, and binding strength were especially inconsistent. This raises real doubt about whether these language models can ever become accurate and steady enough to be safely built into standard pharmaceutical workflows. For now, they are not reliable alternatives to lab testing or proven computational methods. They might help with very early ideas, but the risks of wrong predictions are too high for anything more serious. Challenges like limits in the training data (ending in 2024) and the need for very powerful computers mean more work is needed, such as combining these models with other tools or adding new types of data. But

it remains unclear whether language models will ever reach the reliability required for real drug development.

Limitation Analysis

Despite the promising indications of scaling effects and predictive utility demonstrated by the CaribbeanDynetics TXAI platform in forecasting pharmacokinetic and toxicological profiles, several inherent limitations constrain the generalizability and practical deployment of fine-tuned large language models in pharmaceutical development.

Foremost among these is the constrained scope of the validation dataset, which encompassed only ten drug candidates selected from Phase I milestones between 2024 and 2025, representing a narrow cross-section of therapeutic modalities such as CRISPR editing, GLP-1 agonists, and peptide-drug conjugates. This limited sample size, while strategically chosen to ensure post-training novelty and avoid regurgitation, inherently restricts the statistical power of the findings, potentially amplifying modality-specific biases and underrepresenting the broader diversity of pharmaceutical assets, including those in oncology, neurology, or rare diseases where molecular complexities might exacerbate model inconsistencies. Furthermore, the retrospective benchmarking against publicly available Phase I data introduces potential confounding from incomplete or variably reported clinical outcomes, as these sources may not fully capture nuances like inter-patient variability or long-term effects, thereby inflating apparent accuracies in parameters such as peak plasma concentration or absolute bioavailability where the models performed robustly.

The foundational architecture of the TXGemma models, fine-tuned on over seven million therapeutic data points up to late 2024, inherently carries risks of training data biases and hallucinations, where generative outputs might confidently assert inaccurate predictions due to gaps in the dataset's coverage of emerging modalities or rare adverse events. This aligns with broader critiques in the literature, such as those articulated by Singhal et al. in their 2023 evaluation of Med-PaLM 2, which highlighted LLMs' propensity for misinterpreting laboratory results and deviating from clinical guidelines, issues that manifest here in the mixed results for binding affinity and volume of distribution, where the 9 billion parameter model's shallower latent space exacerbated deviations. Ethical and regulatory hurdles compound these technical shortcomings, as the integration of PubChemPy and RDKit for grounding raises concerns over data provenance and

intellectual property, particularly when models synthesize novel insights from proprietary clinical trial analogs without explicit consent mechanisms. Moreover, the platform's hybrid local/cloud deployment, while enhancing accessibility, demands substantial computational resources for the unquantized 27 billion parameter variant, posing barriers to equitable adoption in resource-limited settings, such as academic or developing-world laboratories, and potentially perpetuating algorithmic biases if training data underrepresents diverse populations or global pharmacogenomic variations.

Interpretability also remains a critical shortfall, as the black-box nature of transformer-based LLMs obscures the reasoning pathways underlying predictions, complicating forensic analysis of errors in complex parameters like primary organ of toxicity, where categorical outputs are occasionally mismatched despite high average scores. This opacity, coupled with the absence of real-time data access, risks disseminating outdated or contextually misaligned advice, echoing concerns raised in a 2024 Lancet viewpoint on ethical challenges of LLMs in medicine, which emphasized data privacy violations and the plastic adaptability of models that could evolve unpredictably based on user inputs. In this study, the exclusion of multimodal inputs beyond SMILES strings further limits fidelity for visually intensive tasks, such as 3D molecular rendering, potentially underestimating toxicities tied to stereochemistry. Finally, while the findings advocate for hybridization with conventional or quantum methods to address inconsistencies in exposure metrics, the study's temporal bounding, being conducted in 2025, invites caution, as advancements in LLM architectures by early 2026 may have mitigated some hallucinations through enhanced fine-tuning or retrieval-augmented generation, underscoring the need for longitudinal reassessment to validate enduring relevance.

Appendices

The following appendices provide citations and links to the primary sources of Phase I clinical trial data used for the pharmacokinetic and safety profiles of the evaluated drug candidates. These sources were consulted for ground truth values in benchmarking model predictions. Citations are formatted in APA style for consistency with pharmaceutical research standards.

Appendix A: CTX310 Phase I Data

CRISPR Therapeutics. (2025, November 8). *CRISPR Therapeutics announces positive Phase 1 clinical data for CTX310*. https://ir.crisprtx.com/news-releases/news-release-details/crispr-therapeutics-announces-positive-phase-1-clinical-data

Appendix B: MET-097i Phase I Data

DelveInsight. (2025, July 9). *Metsera's GLP-1 receptor agonist MET-097i | ADA 2025*. https://www.delveinsight.com/blog/metseras-glp-1-receptor-agonist-met-097i

Metsera. (n.d.). *Metsera pipeline | Advancing next-gen obesity therapies*. Retrieved January 10, 2026, from https://metsera.com/pipeline/

Appendix C: ELVN-001 Phase I Data

Enliven Therapeutics. (n.d.). *Enliven Therapeutics announces updated positive data from Phase 1 study of ELVN-001 in chronic myeloid leukemia*. Retrieved January 10, 2026, from https://ir.enliventherapeutics.com/news-releases/news-release-details/enliven-therapeutics-announces-updated-positive-data-phase-1-0

Mauro, M., Lang, F., Kim, D.-W., Kim, D., Kreil, S., Le Coutre, P., & Heinrich, M. C. (2025, September 8). *A Phase 1A/1B study of ELVN-001, a selective active site inhibitor of BCR::ABL1, in patients with chronic myeloid leukemia* [PDF]. https://www.enliventherapeutics.com/file.cfm/39/docs/elvn-001-101_soho_poster_final_08sep2025.pdf

Appendix D: ADG-126 Phase I Data

Adagene. (2025, May 22). *Adagene announces updated data from Phase 1b/2 study of muzastotug (ADG126) in combination with pembrolizumab in patients with advanced/metastatic MSS colorectal cancer*.

https://investor.adagene.com/news-releases/news-release-details/adagene-announces-updated-data-phase-1b2-study-muzastotug-adg126

Liu, J., Burris, H., Shih, K., Macapinlac, H. A., Jr., Khoukaz, T., Falchook, G., & Bendell, J. (2022). *Phase 1 study of ADG126, a novel masked anti-CTLA-4 SAFEbody, in advanced solid tumors*. *Journal of Clinical Oncology, 40*(16_suppl), e17601. https://ascopubs.org/doi/10.1200/JCO.2022.40.16_suppl.e17601

Appendix E: BCA101 Phase I Data

Bendell, J., Fakih, M., Tolcher, A., Lenz, H.-J., Powderly, J., Chung, V., ... Shah, K. (2025). *Phase I clinical trial of the bifunctional EGFR/TGF-β fusion protein BCA101 in patients with advanced solid tumors*. *Clinical Cancer Research, 31*(22), 4623-4636. https://aacrjournals.org/clincancerres/article/31/22/4623/767046/Phase-I-Clinical-Trial-of-the-Bifunctional-EGFR

Fakih, M., Tolcher, A. W., Bendell, J. C., Lenz, H.-J., Powderly, J. D., Chung, V., ... Shah, K. J. (2022). *A phase 1 trial of the bifunctional EGFR/TGFβ fusion protein BCA101 in patients with EGFR-driven advanced solid tumors*. *Journal of Clinical Oncology, 40*(16_suppl), 2513. https://ascopubs.org/doi/10.1200/JCO.2022.40.16_suppl.2513

Appendix F: BT8009 Phase I Data

Baldini, C., Arend, R., Burris, H., Kim, J. W., Lheureux, S., Mau-Sorensen, M., ... Tolcher, A. W. (2023). *BT8009-100: A Phase I/II study of novel bicyclic peptide and MMAE conjugate BT8009 in patients with advanced malignancies associated with Nectin-4 expression, including urothelial cancer* [PDF]. https://www.bicycletherapeutics.com/wp-content/uploads/2023/05/Baldini_BT8009-100_DoseEsc_ASCOGU_2023.pdf

Diamantis, N., Burris, H. A., Kim, J. W., Tolcher, A. W., Arend, R. C., Lheureux, S., ... Banerji, U. (2025). *First-in-human, Phase I/II dose escalation and expansion study of zelenectide pevedotin (BT8009), a Bicycle toxin conjugate targeting Nectin-4, in patients with advanced solid tumors*. *Journal of Clinical Oncology*. https://ascopubs.org/doi/10.1200/JCO-25-00559

Appendix G: PGN-EDODM1 Phase I Data

ClinicalTrials.gov. (n.d.). *A Phase 1 study of PGN-EDODM1 in participants with myotonic dystrophy type 1 (FREEDOM-DM1)*. Identifier NCT06204809. https://clinicaltrials.gov/study/NCT06204809

Shoskes, J., Larkindale, J., Babcock, S., Vacca, S., Lonkar, P., Holland, A., ... Mellion, M. (n.d.). *FREEDOM-DM1: Phase 1 study design to assess safety, tolerability, pharmacokinetics, and pharmacodynamics of PGN-EDODM1 for myotonic dystrophy type 1* [PDF]. https://www.pepgen.com/wp-content/uploads/3.-MDA.pdf

Appendix H: RAP-219 Phase I Data

Rapport Therapeutics. (2025, January 9). *Rapport Therapeutics announces new Phase 1 data, further establishing RAP-219's differentiated profile*. https://investors.rapportrx.com/news-releases/news-release-details/rapport-therapeutics-announces-new-phase-1-data-further

ClinicalTrials.gov. (2025, October 21). *A long-term study of the safety and effectiveness of RAP-219 in adults with focal onset seizures*. Identifier NCT07219407. https://clinicaltrials.gov/study/NCT07219407

Appendix I: Lenacapavir Phase I Data

Gupta, S. K., Garner, W., Hao, Y., Beraud, C., Brainard, D. M., German, P., ... Das, M. (2025). *Pharmacokinetics and safety of once-yearly lenacapavir: A phase 1, open-label study*. *The Lancet*. https://www.thelancet.com/journals/lancet/article/PIIS0140-67362500405-2/abstract

Gilead Sciences. (2025, March 11). *First clinical data for Gilead's investigational once yearly lenacapavir for HIV prevention presented at CROI 2025 and published in The Lancet*. https://www.gilead.com/news/news-details/2025/first-clinical-data-for-gileads-investigational-once-yearly-lenacapavir-for-hiv-prevention-presented-at-croi-2025-and-published-in-the-lancet

Appendix J: AZD5055 Phase I Data

Barbour, A. M., Li, Y., Kang, S., Chen, L., Uckun, F. M., Dowlati, A., ... Papadimitrakopoulou, V. (2024). *A phase 1 study evaluating the safety, tolerability, and pharmacokinetics of the porcupine inhibitor AZD5055*. *PMC, PMC12177174*. https://pmc.ncbi.nlm.nih.gov/articles/PMC12177174/

AstraZeneca. (n.d.). *Assess the safety, tolerability and pharmacokinetics of AZD5055 following single and multiple ascending doses in healthy participants*. https://www.astrazenecaclinicaltrials.com/study/D8960C00001

References

In addition to the appendices, the following references were cited throughout the manuscript:

Biotechnology Innovation Organization. (2020). *Clinical development success rates and contributing factors 2011–2020*. https://www.bio.org/policy/human-health/vaccines-biodefense/coronavirus/pipeline-tracker

RAND Corporation. (n.d.). *The cost of drug development*. Retrieved from https://www.rand.org/pubs/research_reports/RRA2582-1.html

London School of Economics. (n.d.). *Regulatory costs in pharmaceutical development*. Retrieved from https://www.lse.ac.uk/business/consulting/reports/pharmaceutical-regulation

Inderes Biotech Finance Forum. (n.d.). *Clinical trial success rates*. Retrieved from https://www.inderes.fi/en

DataHow. (n.d.). *Hybrid modeling in pharmaceutical development*. Retrieved from https://www.datahow.com/

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Natarajan, V. (2023). *Large language models encode clinical knowledge*. *Nature, 620*(7972), 172–180. https://www.nature.com/articles/s41586-023-06291-2

The Lancet. (2024). *Ethical challenges of LLMs in medicine*. *The Lancet, 403*(10436), 1455–1457. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00805-1/fulltext

Conclusion

This study looked at how well fine-tuned large language models, used in the CaribbeanDynetics TXAI platform, can predict important drug properties for ten drugs that reached Phase I trials. The main finding is that the larger model (27 billion parameters) generally did better than the smaller one (9 billion parameters). Average scores were higher for most properties, and the differences were statistically strong in many cases. The step-by-step prompting and larger size helped the models give more reasonable answers, especially for simpler measures like peak concentration, bioavailability, and main toxicity organ.

These results matter because drug development is very expensive –around $950 million per approved drug– and most candidates fail early. Tools that cut the need for lab experiments by even 60-80% could save money and time. The platform showed some promise for early rough checks on toxicity or absorption.

However, the improvements tied to model size do not guarantee future success. Even the larger model had predictions that were too scattered and sometimes close to the real values, but often far enough off to be useless or misleading. Complex properties like half-life, total exposure, and binding strength were especially inconsistent. This raises real doubt about whether these language models can ever become accurate and steady enough to be safely built into standard pharmaceutical workflows. For now, they are not reliable alternatives to lab testing or proven computational methods. They might help with very early ideas, but the risks of wrong predictions are too high for anything more serious.

Future work could try mixing these models with other tools, adding more data, or finding ways to measure uncertainty in answers. Until the predictions become far more consistent, language models should stay as optional systems to help with minor tasks, not core tools, in drug development.