DMPKBench: A Multi-Modal Benchmark for Evaluating LLMs and Agents in Drug Discovery DMPK Tasks

Jie Li[1†*], Baiming Chen[2†], Zhiyang Zou[1], Rumin Zhang[1], Sheng Ding[1], Jinjiang Guo[1*]

Abstract

With the rapid progress of large language models (LLMs) and multi-agent systems, there is an increasing demand for fair and comprehensive evaluation of their capacity to address complex tasks in specialized scientific domains. Drug metabolism and pharmacokinetics (DMPK) constitutes a critical stage in drug discovery, requiring interdisciplinary reasoning and integration of diverse knowledge. Thus, we constructed DMPKBench, a comprehansive benchmark designed to evaluate LLMs and multi-agent performance in DMPK-related tasks. Grounded in real-world drug development pipeline, DMPKBench covers five core competencies essential to domain experts: experimental design and troubleshooting, interpretation of experimental results, ADMET multi-parameter optimization, pharmacokinetic (PK) modeling and simulation, and preclinical-to-clinical PK translation to the human body. The DMPKBench offers over 120k question-answer pairs, with four dimensions quality-controlled by specialists and one validated through experimental evidence. We comprehensively evaluated five models across major DMPKBench modules, revealing accuracy ranges from 11% to 89%. A Significant performance gap is observed, with models excelling in knowledge-driven benchmarks but struggling in multi-modal tasks like drug structure understanding, real-world DMPK data table and PK curve analysis, and multi-step quantitative reasoning. Overall, DMPKBench offers a high-quality, domain-specific foundation for advancing LLMs and multi-agent systems in drug discovery and is publicly available at: `https://github.com/GHDDI-AILab/DMPKBench`.

1 Introduction

The rapid advancements in LLMs and multi-agent systems have revolutionized the capabilities of artificial intelligence (AI), enabling unprecedented progress across diverse scientific fields. These developments hold significant potential for improving data analysis, automating complex reasoning, and accelerating scientific discovery. However, there remains a critical need for specialized, quantitative benchmarks to evaluate LLMs within vertical, scientific domains, particularly those that demand high levels of domain expertise and multi-modal reasoning. A recent LLM benchmark indicates that while significant efforts have been made in the medical (1), the field of drug discovery remains vastly under-explored. Specifically, the creation of LLM benchmarks for the DMPK domain is essential for assessing the real-world applicability of LLMs and utility in this complex area. Such benchmarks are vital not only for supervised fine-tuning (SFT) and reinforcement fine-tuning (RFT) of LLMs but also for optimizing multi-agent systems that can be used to simulate, predict, and recommend in the context of pharmaceutical research. These benchmarks will provide the foundational datasets

[*] Corresponding author: jie.li@ghddi.org & jinjiang.guo@ghddi.org. [1] The Global Health Drug Discovery Institute, Beijing, China, 100192. [2] School of Medicine, The Chinese University of Hong Kong, Shenzhen, China, 518172.

[†] These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

necessary for the development, validation, and continuous improvement of models and systems (2) (3). Therefore, the establishment of robust benchmarks in the drug discovery domain is an urgent priority for enabling AI systems to operate effectively and reliably.

In the context of drug discovery, DMPK represents a critical bottleneck, as it encompasses a complex range of tasks that require high-dimensional reasoning and the integration of diverse data types. DMPK studies involve not only traditional PK modeling but also the interpretation of experimental data, often drawn from a variety of sources, including in vitro and in vivo testing, animal studies, and clinical trials. The complexity of this field arises from the necessity to model the absorption, distribution, metabolism, and excretion (ADME) of compounds in humans, requiring a deep understanding of biological systems and computational modeling techniques. Moreover, DMPK tasks are inherently multi-modal, as they often involve the application of multi-step formulas, the interpretation of experimental tables, the analysis of PK curves, and the integration of graphical data, all of which demand a high level of expertise and reasoning across multiple domains (3). Additionally, these tasks frequently require the extrapolation of animal data to human predictions, involving sophisticated statistical and computational models that account for both temporal variations and inter-individual differences in PK. The challenge of DMPK is magnified by the high-dimensional nature of the data and the intricate relationships between various biological systems and drug interactions, making it a prime candidate for the application of advanced AI methodologies.

Furthermore, the need for accurate predictions in DMPK studies extends beyond basic PK to encompass broader aspects of drug development, including toxicology, efficacy and safety profiling (4). AI and machine learning methods, including LLMs and multi-agent, offer a promising approach to integrating and analyzing these complex datasets, providing a means to automate time-consuming tasks, predict potential outcomes, and identify key trends in large volumes of experimental data. However, to ensure that these models are truly effective and reliable in this domain, comprehensive benchmarks are needed. These benchmarks would facilitate the comparison of different LLM architectures and multi-agent systems, guiding the optimization of models for more accurate predictions and better decision-making in drug development. The DMPK domain, with its combination of multi-step reasoning, multi-modal data analysis, and the need for highly specialized domain knowledge, serves as an ideal testing ground for the application of LLMs in drug discovery.

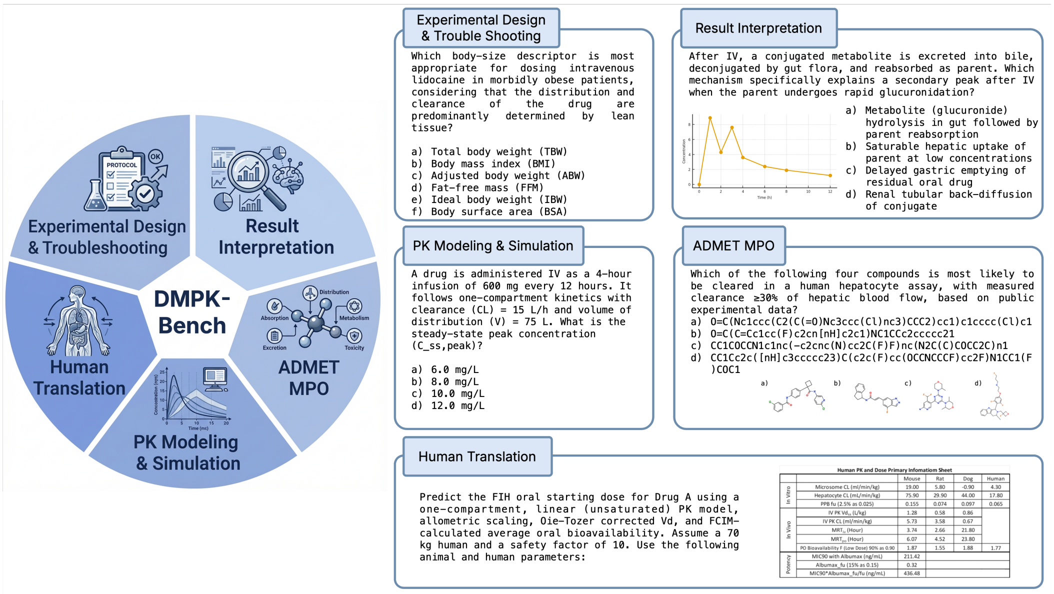
In this study, we proposed the DMPKBench (Figure 1). The creation of dedicated benchmark will support the development of more efficient and accurate AI models and help to establish a clear framework for evaluating the performance of these systems in solving complex real-world drug discovery problems. By providing a consistent set of evaluation criteria and standardized datasets, such a benchmark would enhance the reproducibility of research, facilitate cross-disciplinary collaboration, and contribute to the wider acceptance of AI-based solutions in pharmaceutical research. With the potential to revolutionize the way drugs are developed and tested, the establishment of DMPKBench would mark a significant step forward in leveraging AI for scientific discovery (5) (6).

2 Related work

In recent years, several benchmarking efforts have emerged to evaluate the capabilities of LLMs within biomedical and healthcare domains, each contributing uniquely to the advancement of AI in scientific discovery.

2.1 PubmedQA

PubMedQA is a benchmark dataset introduced in 2019 for biomedical question answering that requires reasoning over research abstracts (7). It contains 273.5K instances across three subsets: expert-annotated PQA-L (1k), artificially generated PQA-A (211.3K), and unlabeled PQA-U (61.2K). Each instance presents a yes/no/maybe question derived from paper titles, context from PubMed abstracts, and labels based on conclusion paragraphs. PubMedQA addresses critical limitations in biomedical QA by necessitating scientific reasoning. Its expert-validated subset (PQA-L) serves as a high-quality evaluation benchmark, while the synthetic PQA-A enables scalable training. The dataset has significantly advanced biomedical language models and is widely adopted in medical AI benchmarks. However, several limitations persist: its information extraction relies primarily on rule-based matching, which offers limited flexibility.



Experimental Design & Trouble Shooting

Which body-size descriptor is most appropriate for dosing intravenous lidocaine in morbidly obese patients, considering that the distribution and clearance of the drug are predominantly determined by lean tissue?

a) Total body weight (TBW)
b) Body mass index (BMI)
c) Adjusted body weight (ABW)
d) Fat-free mass (FFM)
e) Ideal body weight (IBW)
f) Body surface area (BSA)

Result Interpretation

After IV, a conjugated metabolite is excreted into bile, deconjugated by gut flora, and reabsorbed as parent. Which mechanism specifically explains a secondary peak after IV when the parent undergoes rapid glucuronidation?

a) Metabolite (glucuronide) hydrolysis in gut followed by parent reabsorption
b) Saturable hepatic uptake of parent at low concentrations
c) Delayed gastric emptying of residual oral drug
d) Renal tubular back-diffusion of conjugate

PK Modeling & Simulation

A drug is administered IV as a 4-hour infusion of 600 mg every 12 hours. It follows one-compartment kinetics with clearance (CL) = 15 L/h and volume of distribution (V) = 75 L. What is the steady-state peak concentration (C_ss,peak)?

a) 6.0 mg/L
b) 8.0 mg/L
c) 10.0 mg/L
d) 12.0 mg/L

ADMET MPO

Which of the following four compounds is most likely to be cleared in a human hepatocyte assay, with measured clearance ≥30% of hepatic blood flow, based on public experimental data?

a) O=C(Nc1ccc(C2(C(=O)Nc3ccc(Cl)nc3)CCC2)cc1)c1cccc(Cl)c1
b) O=C(C=Cc1cc(F)c2cn[nH]c2c1)NC1CCc2ccccc21
c) CC1COCCN1c1nc(-c2cnc(N)cc2C(F)F)nc(N2C(C)COCC2C)n1
d) CC1Cc2c([nH]c3ccccc23)C(c2c(F)cc(OCCNCCCF)cc2F)N1CC1(F)COC1

Human Translation

Predict the FIH oral starting dose for Drug A using a one-compartment, linear (unsaturated) PK model, allometric scaling, Oie-Tozer corrected Vd, and FCIM-calculated average oral bioavailability. Assume a 70 kg human and a safety factor of 10. Use the following animal and human parameters:

Human PK and Dose Primary Infomatiom Sheet

		Mouse	Rat	Dog	Human
In Vitro	Microsome CL (ml/min/kg)	19.00	5.80	-0.90	4.30
	Hepatocyte CL (mL/min/kg)	75.90	29.90	44.00	17.80
	PPB fu (2.5% as 0.025)	0.155	0.074	0.097	0.065
In Vivo	IV PK Vd_{ss} (L/kg)	1.28	0.58	0.86	
	IV PK CL (ml/min/kg)	5.73	3.58	0.67	
	MRT_{iv} (Hour)	3.74	2.66	21.80	
	MRT_{po} (Hour)	6.07	4.52	23.80	
	PO Bioavailability F (Low Dose) 90% as 0.90	1.87	1.55	1.88	1.77
Potency	MIC90 with Albumax (ng/mL)	211.42			
	Albumax_fu (15% as 0.15)	0.32			
	MIC90*Albumax_fu/fu (ng/mL)	436.48			

Figure 1: Overview of DMPKBench

2.2 LAB-Bench

To assess LLM capacity in practical biological tasks beyond textbook-style science questions, Laurent et al. (2024) introduced the Language Agent Biology Benchmark (LAB-Bench) (8). This benchmark is designed to evaluate LLM performance on tasks essential for scientific research, comprising 2,457 question-answer pairs. LAB-Bench assesses capabilities including literature recall and reasoning, figure interpretation, database access and navigation, and comprehension and manipulation of DNA and protein sequences.

2.3 BioProBench

Introduced in 2025, BioProBench is a benchmark dataset designed for biological protocol understanding and reasoning (9). It encompasses five core tasks: protocol question answering, step ordering, error correction, protocol generation, and protocol reasoning. Built upon 26,933 original protocols, BioProBench provides 556,171 high-quality structured instances. This benchmark reveals fundamental challenges in current LLMs, including: procedural knowledge comprehension, deep domain adaptation, reliability of structured reasoning, and handling of complex precision and safety constraints.

2.4 HealthBench

To evaluate the performance and safety of large language models (LLMs), OpenAI recently developed HealthBench, a benchmark comprising 5,000 multi-turn conversations between a model and either an individual user or a healthcare professional. LLM responses were assessed using conversation-specific rubrics established by 262 physicians, encompassing 48,562 rubric criteria. The benchmark result of LLM in HealthBench indicates recent models have improved rapidly in human health across performance, cost, and reliability. HealthBench features two variants:

- HealthBench Consensus comprises 3,671 examples featuring exclusively consensus-based criteria. HealthBench Consensus can be thought of as a version with a higher level of physician validation.
- HealthBench Hard, a curated subset of 1,000 particularly challenging examples from HealthBench remains notably difficult for current state-of-the-art models, with no evaluated model achieving scores exceeding 32%.

Despite these notable contributions, a comprehensive benchmark for drug discovery remains largely absent, especially for critical domains like DMPK. While there are efforts like ChemBERTa (10), which addresses chemical structure prediction, and MolBench (11), focused on molecular property prediction, they do not specifically cater to the complex, multi-modal, and multi-step reasoning inherent in DMPK. Moreover, BioGPT (12), designed for biomedicine, is another relevant effort but is not tailored to drug discovery or DMPK-related tasks. These existing benchmarks primarily focus on isolated aspects of drug development, yet they fail to provide a holistic, multi-faceted evaluation for a comprehensive field like DMPK, which requires multi-dimensional reasoning, complex data integration, and expert-level domain knowledge. This gap highlights the critical need for DMPKBench, a specialized benchmark that can address the unique challenges and complexities within the DMPK domain, which remains unexplored in the context of LLM benchmarking.

3 DMPKBench Construction

DMPKBench is designed based on extensive testing experience within the pharmaceutical industry, focusing on common scientific challenges in pharmaceutical research. By translating real-world drug pipeline tasks into generalized drug discovery problems, it avoids the use of proprietary industrial data while ensuring that the benchmark remains closely aligned with the needs of the industry, providing valuable insights for model development and performance evaluation in the context of actual pharmaceutical pipelines.

DMPKBench encompasses five critical competencies that are central to the expertise required in DMPK: 1) experimental design and troubleshooting, 2) interpretation of experimental results, 3) ADMET multi-parameter optimization, 4) pharmacokinetic (PK) modeling and simulation, 5) and preclinical-to-clinical PK translation to the human body. Benchmark data were derived from textbooks, peer-reviewed journal articles, historical research datasets, expert-level examinations, and internal drug pipeline review discussions. All data were standardized into an LLM-compatible JSONL format, with multi-modal inputs (tables, figures, PK curves) stored in Markdown for enhanced accessibility. Rigorous quality control was ensured through multi-round expert validation, formula verification, and inter-rater agreement checks.

3.1 DMPK Experimental Design and Troubleshooting

The dimension of DMPK experimental design and troubleshooting primarily involves fundamental competencies in the DMPK domain knowledge. In the context of actual drug discovery pipelines, DMPK is tasked with identifying issues within the DMPK profile and designing experiments for in-depth exploration of these concerns. In this domain, we utilize two primary sources: ABT (American Board of Toxicology, https://www.abtox.org/) exams and CoT-driven question-answer pairs derived from PubMed abstracts, along with high-quality full-text articles. The types of benchmark included multiple-choice questions (MCQs), true/false assessments, fill-in-the-blank and short-answer questions, ensuring a comprehensive evaluation of both theoretical and practical DMPK knowledge.

ABT Certification Examination The ABT (American Board of Toxicology) certification examination serves as an expert-level certification, primarily assessing competencies in toxicology. The benchmark dataset is sourced from preparatory question banks, specifically from textbooks, with the content initially in PDF format being converted into markdown format and subsequently transformed into an LLM-compatible JSON format for ease of use. Beyond the standard question-answer pairs, the dataset includes additional metadata such as question IDs, resource tags, and detailed explanations and conceptual clarifications. In total, the dataset comprises 3190 MCQs and 751 fill-in-the-blank questions. This collection includes closed-ended and open-ended QAs, covering a broad range of query types. It includes both knowledge-driven questions that test factual understanding and reasoning questions that require advanced problem-solving and critical thinking.

CoT-enhanced QAs from DMPK Related Journal This portion of the benchmark includes both a CoT-enhanced question bank derived from the last five years of scientific literature. In our earlier research, DrugBench (13), we established a fully automated pipeline for clustering PubMed abstracts, extracting key scientific concepts, and generating chain-of-thought (CoT) reasoning questions and answers. We now apply this workflow to create a CoT-enhanced question answer pairs in the DMPK domain, resulting in 701 multiple-choice QAs and 401 short-answer QAs.

3.2 Interpretation of Experimental Result

In real-world drug pipelines, the interpretation of DMPK data and PK curves represents a core task for domain experts, particularly when addressing multi-modal inputs such as tables and graphical profiles with abnormal signal values. These tasks require routine data interpretation, multifactorial analysis and causal reasoning to uncover underlying mechanisms. To emulate this complexity, we first validated our methodology on internal pipeline data before extending it to case studies from published journals, e.g. Drug Metabolism and Disposition (DMD) and the Journal of Medicinal Chemistry (JMC). Special emphasis was placed on atypical cases where PK curves deviated from standard expectations, as these provide valuable challenges for model evaluation. From these cases, question–answer pairs were systematically constructed and standardized into JSON format. To preserve confidentiality, compounds were anonymized into generic placeholders (e.g., Drug A). Multi-modal content, including tables and PK curves, was stored in markdown format, with QA pairs designed to remain interpretable independently of the embedded figures and tables.

3.3 ADMET MPO

One of the major tasks for DMPK experts involves multi-parameter optimization (MPO) of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties to achieve the desired safety window. Using an internally standardized knowledge-graph dataset (PharmNet) and our developed ADMET prediction AI platform (Eight Formutions of AI KongMing, `https://datascience.ghddi.org`), we extracted ADMET-specific data comprising both binary classification and regression tasks, which were reformulated into LLM-compatible question-answer pairs in JSON format to assess model performance in ADMET evaluation. To ensure reliability, all distractors were supported by experimentally validated values rather than artificially constructed entries. The classification dataset generated 127,600 MCQs. Molecular structures were represented using SMILES notation. Given the limitations of some LLMs in processing SMILES and SMARTS, conversion to IUPAC names was also performed to facilitate a fairer evaluation.

3.4 PK Modeling and Simulation

Another critical task for DMPK experts in real-world drug pipelines is multi-step mathematical derivation for PK modeling and simulation. Given recent advances in LLMs' mathematical reasoning capabilities, such calculations provide an effective testbed for evaluating formula derivation and numerical computation. In this component, datasets were constructed from published drug case studies, which were then generalized to anonymized examples (e.g., Drug A). Each question was designed to provide the necessary parameter values while requiring the model to identify and derive the relevant formulas, rather than supplying them directly. To ensure sufficient complexity, only problems involving at least three computational steps were retained. Explanations included the explicit formulas and calculation processes, while answers were manually verified for correctness. Instances in which discrepancies arose between the provided derivation and the final answer were manually corrected to guarantee accuracy.

3.5 Preclinical-to-clinical PK Translation

The final component of DMPKBench focuses on preclinical-to-clinical PK translation to the human body, including tasks like in vitro-in vivo extrapolation (IVIVE), first-in-human (FIH) dose prediction, and human drug–drug interaction (DDI) inference. The tasks require integration of preclinical and clincial data, interspecies scaling, and physiological reasoning. All datasets were curated from the public domain and restricted to cases with complete experimental values required for quantitative extrapolation. Drug identities were anonymized (e.g., Drug A).

For FIH dose prediction, each case included molecular weight, potency, plasma protein binding (PPB), microsomal and hepatocyte clearance across humans and at least three preclinical species, together with animal in vivo PK data. Given the high dimensionality and integrity requirements of the dataset, a total of 10 case studies were collected. Human data were used as ground truth, with error tolerances in 2-fold, distractor options were designed to deviate by larger than 3-fold. All entries were converted into standardized JSON format with markdown support for multi-modal tables and schematics, allowing consistent benchmarking of LLMs in this domain.

Human translation requires the application of advanced PK principles, including one-compartment linear PK models, allometric scaling and Oie-Tozer model. These settings go beyond simple numeric prediction, encompassing formula derivation, stepwise reasoning, and tool/software application. Given this multi-dimensional complexity, the human translation benchmark is particularly suited for multi-agent frameworks that can integrate mathematical reasoning with domain knowledge.

4 Experiment

4.1 DMPKBench Overview

Grounded in practical drug discovery requirements, DMPKBench encompasses five core competencies: experimental design and troubleshooting, interpretation of experimental results, ADMET multi-parameter optimization, PK modeling and simulation, and preclinical-to-clinical PK translation to the human body. Following systematic construction and quality-control optimization. The benchmark provides a balanced and comprehensive coverage of tasks, with data statistics summarized in Table 1.

Table 1: Dataset Composition

Domain	Resource	MCQ	Open-ended QA	Quality
Experiment Design	ABT Test	3,190	751	Expert QC
	DrugBench[1]	701	401	LLM QC
Result Interpretation	Public case	100	–	Expert QC
PK Modeling	Public case	200	–	Expert QC
ADMET MPO	SPR data	127,600	–	Experimental Evidence
Human Translation	Public case	–	10	Expert QC

[1] Derived from PubMed (13)

The benchmark integrates multiple sources of difficulty, including expert-level certification exams and questions derived from the last five years of scientific literature, thereby capturing both foundational and frontier knowledge. Real-world case studies from the public domain are incorporated to benchmark the interpretation of experimental results, such as PK curves and DMPK parameter tables. Multi-step quantitative problems challenge the mathematical reasoning and formula derivation skills required for PK modeling and simulation. In addition, complex translational tasks, e.g. FIH dose prediction, challenge integration of diverse experimental inputs and advanced modeling approaches.

The benchmark spans both close-ended QAs (e.g., multiple-choice and true/false questions) and open-ended QAs (e.g., fill-in-the-blank and short-answer questions), ensuring that model evaluation is not restricted to narrow knowledge retrieval but extends to reasoning, interpretation, and synthesis. Taken together, DMPKBench represents a high-quality, multi-modal, domain-specific benchmark for systematically evaluating the performance of LLMs and multi-agent systems in drug discovery and development.

4.2 Benchmark Performance

We conducted a comparative performance analysis of several state-of-the-art LLMs, including OpenAI o3 and DeepSeek R1 (Table 2). We found that LLMs consistently excel in knowledge-based tasks (such as the ABT Test and PubMed), yet their performance varies significantly across other tasks. In particular, they demonstrate particularly poor performance in result interpretation, with accuracy approximately 23%.

Benchmark of Close-ended Questions In DMPKBench, close-ended questions include MCQs and true/false items. Evaluation was restricted to MCQs. All MCQ subsets were fully evaluated, except for ADMET multi-parameter optimization, where stratified random sampling was applied with 10 representative questions per property. The performance of five models across five dimensions is shown

Table 2: Performance of Different Models

Model	ABT Test (Acc %)	PubMed (Acc %)	Result Interpretation (Acc %)	ADMET MPO (Acc %)	PK Modeling (Acc %)
OpenAI o3	**89.47**	**85.73**	**31.65**	42.09	53.25
OpenAI GPT-4.1	82.15	85.59	27.80	37.50	38.96
Claude 4 Sonnet	86.03	84.31	11.39	**43.51**	40.26
Deepseek R1	87.69	80.46	17.72	38.13	40.26
Deepseek V3	84.59	85.16	29.11	37.50	**54.55**
Average	85.99	84.25	23.53	39.75	47.46

Note: The number of evaluated questions is smaller than the full dataset for some tasks because certain LLM outputs contained system errors that prevented scoring.

in 2. Overall, all models achieve strong performance on knowledge-driven benchmarks (ABT Test and PubMed), with accuracy consistently above 80%, indicating robust domain knowledge retrieval. In contrast, performance drops substantially when tasks extended beyond purely knowledge-driven or text-centric reasoning. Result interpretation remains the most challenging module, with accuracy below 32% for all models. ADMET multi-parameter optimization shows moderate performance, where Claude 4 Sonnet achieves the highest accuracy (43.51%). Deepseek-V3 displays the best performance (54.55%) in PK modeling, suggesting relative strengths in multi-step formula calculation and reasoning. These results highlight a clear gap between factual knowledge proficiency and complex multi-modal DMPK reasoning, underscoring the need for benchmarks that reflect real-world drug discovery challenges.

Benchmark of Open-ended Questions Open-ended benchmark evaluation in DMPKBench includes fill-in-the-blank and short-answer questions. For fill-in-the-blank items, model-generated answers are assessed by a secondary LLM to determine semantic equivalence with the reference answers, accounting for synonyms and paraphrased expressions. For example, responses such as “Metallothionein (MT), CdMT” are correctly judged as equivalent to “metallothionein, Cd-metallothionein”. For the 751 fill-in-the-blank items, OpenAI o3-mini model was used both to generate the answers and to assess semantic equivalence between the generated answers and the ground-truth solutions. Under this setting, 418 items were judged correct, corresponding to an accuracy of 55.66%. Since LLM-based adjudication does not fully ensure correctness and no standardized evaluation protocol exists for open-ended questions, performance metrics are not reported here.

Short-answer questions pose greater challenges due to their inherent subjectivity and sensitivity to linguistic style. We tried two approaches. Expert judgment provides high-quality and reliable assessment but is time-consuming, not scalable and subjective. Automated evaluation is an alternative approach. Reference answers are decomposed into key points, and models are required to generate both narrative responses and explicit key points. A secondary LLM is applied to evaluate semantic alignment and coverage to approximate answer accuracy.

Benchmark of Agent Performance DMPKBench also provides a framework for evaluating agent performance in DMPK tasks. It includes knowledge-driven questions from ABT tests and PubMed-extracted CoTs, as well as multi-modal tasks such as PK graph and table interpretation, formula calculation and derivation. The benchmark enables the assessment of whether domain-optimized agents extend the general LLMs’ reasoning capabilities and improve performance across multiple levels. It also incorporates complex multi-step reasoning challenges. These include ADMET multi-property optimization, which involves up to tens of simultaneous property predictions, and preclinical-to-clinical translation, which requires sequential inference.

Conventional LLMs often struggle with intricate tasks such as human dose prediction. Such tasks typically demand more than ten reasoning steps and prolonged computational effort. Multi-agent frameworks equipped with modular tools and collaborative task decomposition can enhance overall performance. These frameworks are capable of predicting human dose from preclinical data within a two-fold error margin the accepted gold standard for accuracy in human dose prediction.

4.3 Unresolved challenge subset

DMPKBench further defines an unresolved challenge subset, comprising questions that were answered incorrectly by all evaluated models (OpenAI o3, GPT-4.1, Claude 4 Sonnet, DeepSeek-R1, and DeepSeek-V3). These results are based on a single run and may vary across runs. The prevalence of unresolved questions differs markedly by task type: knowledge-driven tasks (e.g., ABT test and DrugBench from PubMed CoT) account for only 4.8% of all questions, whereas multi-modal QA constitutes 28.26%. This indicates that multi-modal tasks remain a major limitation for current LLMs, particularly in PK curve interpretation, multi-step PK modeling and ADMET multi-parameter optimization. A breakdown analysis of these question-answer pairs indicates that the models can correctly parse numerical values from figures and tables but fail on complex reasoning, which likely drives the observed errors. The unresolved challenge subset is also publicly released on GitHub to support further study of LLM performance on multi-modal DMPK benchmark.

5 Conclusion

To address the lack of domain-specific benchmarks for rapidly evolving LLMs and multi-agent systems in drug discovery, we developed DMPKBench, a multidisciplinary reasoning benchmark grounded in the real-world drug pipeline requirements. DMPKBench encompasses five expert-level competencies essential to DMPK specialists: experimental design and troubleshooting, interpretation of experimental results, ADMET multi-parameter optimization, PK modeling and simulation, and preclinical-to-clinical PK translation to the human body. It is inherently multimodal, integrating tasks such as SMILES interpretation, PK curve analysis, parameter table comprehension, and multi-step mathematical derivations. With over 120,000 question–answer pairs, DMPKBench combines scale with rigor: four dimensions are QC-verified by domain experts, and one is supported by experimental evidence. Interestingly, we observed that the LLMs demonstrated strong performance on knowledge-driven tasks, while their performance declined significantly on the multi-modal tasks. This may be attributed to the models' currently limited thinking and reasoning capabilities, or to the relatively scarce availability of multi-modal training data. In summary, DMPKBench provides a robust, comprehensive, and high-quality foundation for evaluating and advancing AI in pharmaceutical research. To ensure reproducibility and reduce prompt-related variability, relevant datasets as well as validation prompts used in this study are available and regularly updated at `https://github.com/GHDDI-AILab/DMPKBench`.

DMPKBench addresses a critical gap by providing an industry-relevant benchmark for evaluating large language models and AI agents. However, several challenges remain. Firstly, a disconnect persists between academic benchmark design and real-world industrial applications, meaning that conventional performance metrics may not fully capture practical utility. Secondly, evaluating open-ended responses remains challenging due to subjectivity in evaluation. Moreover, complex multi-agent tasks, such as ADMET MPO and first-in-human dose prediction, involve multistep reasoning and iterative validation, further complicating consistent and automated evaluation. Finally, discrepancies can arise between benchmark design and downstream applications in supervised or reinforcement fine-tuning and agent development. To address these limitations, DMPKBench will be continuously refined to keep pace with advances in AI and the evolving demands of real-world drug discovery.

References

[1] Zhang, M., Shen, Y., Li, Z., Sha, H., Hu, B., Wang, Y., ... & Huang, X. (2025). LLMEval-Med: A Real-world Clinical Benchmark for Medical LLMs with Physician Validation. *arXiv preprint arXiv:2506.04078*

[2] Cai, H., Cai, X., Chang, J., Li, S., Yao, L., Wang, C., ... & Ke, G. (2024). Sciassess: Benchmarking llm proficiency in scientific literature analysis. arXiv preprint arXiv:2403.01976.

[3] Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., ... & Singhal, K. (2025). Healthbench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775.

[4] Baillie, T. A. (2008). Metabolism and toxicity of drugs. Two decades of progress in industrial drug metabolism. *Chemical research in toxicology*, *21*(1), 129-137.

[5] Blanco-Gonzalez, A., Cabezon, A., Seco-Gonzalez, A., Conde-Torres, D., Antelo-Riveiro, P., Pineiro, A., & Garcia-Fandino, R. (2023). The role of AI in drug discovery: challenges, opportunities, and strategies. Pharmaceuticals, 16(6), 891.

[6] Chou, W. C., & Lin, Z. (2023). Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. Toxicological Sciences, 191(1), 1-14.

[7] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146.

[8] Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M., Hammerling, M. J., Narayanan, S., ... & Rodriques, S. G. (2024). Lab-bench: Measuring capabilities of language models for biology research. arXiv preprint arXiv:2407.10362.

[9] Liu, Y., Lv, L., Zhang, X., Yuan, L., & Tian, Y. (2025). BioProBench: Comprehensive Dataset and Benchmark in Biological Protocol Understanding and Reasoning. arXiv preprint arXiv:2505.07889.

[10] Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

[11] Jiang, X., Tan, L., Cen, J., & Zou, Q. (2023). Molbench: A benchmark of ai models for molecular property prediction. In International Symposium on Benchmarking, Measuring and Optimization (pp. 53-70). Singapore: Springer Nature Singapore.

[12] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics, 23(6), bbac409.

[13] Li, J., Chen, B., Liu, R., Zou, Z., & Guo, J.(2025). DrugBench: A Data-Mining Pipeline for Generating CoT-driven LLM Benchmark in Drug Discovery [Manuscript in preparation].