, , pp. 1–7

doi:

Prompt-to-Pill: Multi-Agent Drug Discovery and Clinical Simulation Pipeline

Ivana Vichentijevikj,[1] Kostadin Mishev[2] and Monika Simjanoska Misheva[2]

[1]iReason LLC, 3rd Macedonian Brigade 37, 1000, Skopje, North Macedonia, contact@ireason.mk, https://www.ireason.mk/ and [2]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, P.O. 393, 1000, Skopje, North Macedonia

Abstract

This study presents a proof-of-concept, comprehensive, modular framework for AI-driven drug discovery (DD) and clinical trial simulation, spanning from target identification to virtual patient recruitment. Synthesized from a systematic analysis of 51 LLM-based systems, the proposed *Prompt-to-Pill*[*]architecture and corresponding implementation leverages a multi-agent system (MAS) divided into DD, preclinical and clinical phases, coordinated by a central *Orchestrator*. Each phase comprises specialized large language model (LLM) for molecular generation, toxicity screening, docking, trial design, and patient matching. To demonstrate the full pipeline in practice, the well-characterized target Dipeptidyl Peptidase 4 (DPP4) was selected as a representative use case. The process begins with generative molecule creation and proceeds through ADMET evaluation, structure-based docking, and lead optimization. Clinical-phase agents then simulate trial generation, patient eligibility screening using EHRs, and predict trial outcomes. By tightly integrating generative, predictive, and retrieval-based LLM components, this architecture bridges drug discovery and preclinical phase with virtual clinical development, offering a demonstration of how LLM-based agents can operationalize the drug development workflow in silico.

Key words: Large Language Models; LLMs; Drug Discovery DD; Preclinical Phase; Clinical Phase; Multi-Agent Systems MAS; Prompt-to-Pill; ChatMED

Introduction

The ability of LLMs to learn from massive datasets and adapt to diverse inputs provides unprecedented capabilities that surpass traditional methods. Their use accelerates decision-making and reduces experimental costs across the development pipeline [Oniani et al., 2024], as evidenced by applications ranging from generating novel molecular structures [Sheikholeslami et al., 2025, Li et al., 2023] to predicting pharmacokinetics and toxicity [Cai et al., 2025, Liu et al., 2024], simulating clinical trials [Xu et al., 2025], and optimizing patient-trial matching [Datta et al., 2025, Lin et al., 2024].

The integration of LLMs into drug development pipelines has gained notable traction, especially across preclinical phases.

Gao et al. proposed a domain-guided MAS for reliable drug-target interaction (DTI) prediction, using a debate-based ensemble of LLMs. The framework partitions the DTI task into protein sequence understanding, drug structure analysis, and binding inference, handled by dedicated agents. Evaluation was conducted on the BindingDB dataset, showing improvements in both accuracy and prediction consistency compared to single-LLM baselines. The system integrates *GPT-4o*, *LLaMA-3*, and *GLM-4-Plus* [Gao et al., 2024].

Lee et al. developed *CLADD*, a retrieval-augmented MAS addressing multiple DD tasks. CLADD includes specialized teams for molecular annotation, knowledge graph querying, and prediction synthesis. Evaluations spanned property-specific captioning (BBBP, SIDER, ClinTox, BACE), target identification (DrugBank, KIBA), and toxicity classification. All agents were instantiated with *GPT-4o-mini*, showcasing the utility of general-purpose models when combined with structured RAG mechanisms [Lee et al., 2025].

Song et al. presented *PharmaSwarm*, a hypothesis-driven agent swarm for therapeutic target and compound identification. The architecture orchestrates three specialized agents (*Terrain2Drug*, *Market2Drug*, *Paper2Drug*) and a central evaluator, all integrated via a shared memory and tool-augmented validation layer. Case studies included idiopathic pulmonary fibrosis and triple-negative breast cancer, combining omics analysis, literature mining, and market signals. Agents were powered by *GPT-4*, *Gemini 2.5*, and *TxGemma* [Song et al., 2025].

Yang et al. proposed *DrugMCTS*, a novel multi-agent drug repurposing system that incorporates Monte Carlo Tree Search (MCTS) with structured agent workflows. Using *Qwen2.5-7B-Instruct* for all agents, the system conducts iterative reasoning across molecule retrieval, analysis, filtering, and protein matching. The framework was benchmarked on DrugBank and KIBA, achieving up to 55.34% recall. A case study involving Equol and CXCR3 showed successful prediction of interaction, supported by AutoDock Vina simulations with a binding score of -8.4 kcal/mol [Yang et al., 2025].

[*] https://github.com/ChatMED/Prompt-to-Pill

Inoue et al. introduced *DrugAgent*, an explainable multi-agent reasoning system for drug repurposing. Their architecture coordinates agents handling knowledge graph queries, machine learning scoring, and biomedical literature summarization. Evaluation on a kinase inhibitor dataset revealed strong interpretability and modularity. Detailed ablation studies confirmed that each agent contributes distinctly to the performance. The system employed *GPT-4o*, *o3-mini*, and *GPT-4o-mini*, and the full pipeline is available open-source [Inoue et al., 2024]. Among the surveyed systems, only DrugAgent provides a publicly accessible implementation[1].

None of the described MASs engages with clinical trial simulation, real-world evidence (RWE), or electronic health records (EHRs), thereby limiting their applicability to the preclinical stage of drug development.

This paper introduces *Prompt-to-Pill*, a unified multi-agent framework build on a systematic analysis of 51 LLM-based studies published between 2022 and 2025. The architecture integrates specialized LLM agents for molecule generation, docking, property prediction, trial construction, patient matching, and outcome forecasting through a central Orchestrator. Unlike prior frameworks confined to molecule-level reasoning, *Prompt-to-Pill* provides a proof-of-concept prototype from molecular ideation to virtual trial execution, demonstrating how modular LLM agents can operate synergistically within a closed-loop DD and development ecosystem. A complete implementation of the pipeline is available at GitHub[2].

Methods

The systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The PRISMA framework was employed to ensure transparency, methodological rigor, and reproducibility in identifying, screening, and synthesizing eligible studies. A structured multi-stage review process was followed, encompassing database search, eligibility screening, full-text assessment, and data extraction. The complete selection workflow is detailed in the accompanying PRISMA flow diagram depicted in Figure 1.

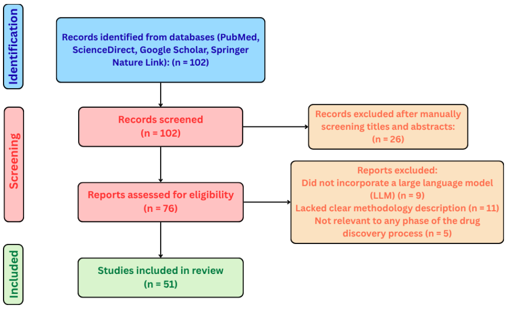

Identification
Records identified from databases (PubMed, ScienceDirect, Google Scholar, Springer Nature Link): (n = 102)
Screening
Records screened (n = 102)
Records excluded after manually screening titles and abstracts: (n = 26)
Reports assessed for eligibility (n = 76)
Reports excluded:
Did not incorporate a large language model (LLM) (n = 9)
Lacked clear methodology description (n = 11)
Not relevant to any phase of the drug discovery process (n = 5)
Included
Studies included in review (n = 51)

Fig. 1. PRISMA-based selection process.

Information Sources and Search Strategy

A structured and comprehensive literature search was conducted to identify and evaluate LLM-based approaches applied in drug design and discovery. The search was conducted between May 1 and June 15, 2025 across PubMed, ScienceDirect, Google Scholar, and Springer Nature Link. The search covered the publication period 2022 - 2025.

Search queries with predefined Boolean combinations captured studies across all DD stages. Representative search strings included: *"large language models" AND ("target identification" OR "binding site prediction"), "large language models" AND ("molecule generation" OR "de novo molecule generation"), "large language models" AND ("clinical trial design" OR "eligibility criteria extraction" OR "trial outcome prediction"), "retrieval-augmented generation" AND ("drug discovery" OR "clinical trials"), "large language models" AND ("patient recruitment" OR "clinical trial matching"* These terms were selected to align with a conceptual pipeline spanning DD, preclinical and clinical phases of pharmaceutical development.

Study Selection Process

Two reviewers independently screened the titles and abstracts of all retrieved records. Full-text reviews were then performed to assess eligibility based on the predefined inclusion and exclusion criteria.

The inclusion criteria were defined as follows: open-source studies written in English; publications or preprints published between 2022 and 2025; research incorporating LLMs for drug development tasks with clearly defined input–output structure, functional purpose, and workflow integration potential; and studies relevant to at least one stage of the DD or clinical trial process.

The exclusion criteria were: articles not written in English; studies lacking a clear methodological or architectural description; studies that are not publicly accessible; and research not directly applicable to any stage of drug development.

The PRISMA flow diagram in Figure 1 details the number of records identified, screened, excluded (with reasons), and finally taken into consideration for building the Prompt-to-Pill pipeline.

Data Extraction and Synthesis

For each included study, detailed metadata were manually extracted into structured tables, one for preclinical models and DD and another for clinical applications. Metadata fields were designed to support both technical evaluation and contextual information from each source as follows:

-**Bibliographic**: Authors, Year, Title, DOI.

-**Technical**: Base model (e.g., GPT-4, BioGPT), Task Type, RAG usage, Evaluation Metrics, Datasets.

- **Reproducibility**: GitHub/Hugging Face links, Input/Output examples.

-**Contextual**: Task Narrative, Clinical Trial Phase (I–IV), Abstract Summary.

Extracted data were then synthesized by stage as needed for the drug development pipeline. Studies were profiled and compared across multiple dimensions including application scope, base architecture, task type, and dataset diversity. The The complete metadata tables containing all reviewed studies are provided in the Section 9. This structured comparison

[1] https://anonymous.4open.science/r/DrugAgent-B2EA

[2] https://github.com/ChatMED/Prompt-to-Pill

informed the construction of the Prompt-to-Pill multi-agent framework introduced later in the paper.

Methodology

Prompt-to-Pill Architecture Foundation

The systematic review of 51 studies (2022–2025) shows a sharp growth in research, peaking in 2024 (14 preclinical/DD, 8 clinical), with 17 in 2025, and fewer in 2022 (4) and 2023 (8).

Preclinical/DD studies mainly used generative LLMs such as *LLaMA* and *GPT-2* for creative molecular tasks on open datasets (TDC, DrugBank). Models like *DrugGen* [Sheikholeslami et al., 2025] and *DrugGPT* [Li et al., 2023] generate SMILES from protein sequences, 3D structures, or text, while others introduce spatial constraints (*3DSMILES-GPT* [Wang et al., 2025b]) or RNA design (*GenerRNA* [Zhao et al., 2024]). *DrugAssist* [Ye et al., 2023] extends this process with prompt-based molecule optimization, refining compounds to improve pharmacological properties. LLMs also support ADMET prediction, synthesis feasibility, and reactivity analysis [Cai et al., 2025, Wang et al., 2025a, Chaves et al., 2024], as well as biological interaction modeling and drug repurposing [Beasley et al., 2025, Li et al., 2025, Edwards et al., 2023, Schmitt et al., 2025].

Clinical studies, by contrast, rely on discriminative or hybrid models such as *GPT-4* and *BioBERT*, often trained on structured data (e.g., ClinicalTrials.gov). About half of the 19 clinical papers propose cross-phase models addressing patient selection, outcome prediction, and document generation. LLMs assist in patient-trial matching [Datta et al., 2025, Lin et al., 2024], trial simulation [Wang et al., 2024, Reinisch et al., 2024, Xu et al., 2025], and pharmacovigilance through tools like *AskFDALabel* and *DAEDRA* [Wu et al., 2025, von Csefalvay, 2024], occasionally enhanced with RAG pipelines for context-aware text generation [Markey et al., 2025, Painter et al., 2025].

Retrieval-augmented generation (RAG) methods showed limited adoption despite their potential for complex reasoning tasks. As shown in Table 1, most studies provided input examples but fewer included output data or reproducible code, underscoring the need for transparency and standardized evaluation.

This analysis informed the design of the proposed Prompt-to-Pill architecture, implemented using the AutoGen [Wu et al., 2024] framework for scalable multi-agent AI systems. Each agent is adapted from rigorously evaluated domain models, with key performance metrics summarized in Table 2.

The datasets listed in Table 2 implicitly define the applicability domains (ADs) of the models integrated into the pipeline. For example, ChemFM's ADME and toxicity predictors are trained on specific benchmarking collections of drug-like compounds, while Panacea and MediTab operate within the disease areas and trial structures represented in CT.gov, SIGIR, and TREC. Because Prompt-to-Pill connects these components sequentially, the effective AD of the full system corresponds to the intersection of all model-specific ADs.

The Prompt-to-Pill Multi-Agent Pipeline

Constructed from the models identified and reviewed in this study, a comprehensive AI-driven pipeline for DD and clinical trial simulation is presented in Figure 2, structured into three main phases: DD Agents, Preclinical Agents and Clinical Agents, coordinated by a central *Orchestrator*, assisted by a *Planning Agent*. The workflow is task-driven, dynamically selecting the appropriate agent and its tools according to the requirements of the given task.

In our scenario, we demonstrate this process by focusing on the development of drug candidates for the DPP4 protein target (UniProt ID: P27487). For this drug development task, the pipeline begins with **Drug Discovery Agents**. Here we have 3 subgroups of agents: Hits Generation, Leads Identification and Lead Optimization.

The workflow begins with *Hits Generation*, where the *Drug-Generation Agent*, based on the DrugGen framework [Sheikholeslami et al., 2025], produces a set of candidate SMILES sequences. Then the generated SMILES are docked against the target with *Docking Agent*. The Docking Agent is responsible for evaluating the binding affinity of generated molecules against the target protein. Retrieves the target structure from the Protein Data Bank [Berman et al., 2000] or defaults to AlphaFold models [Jumper et al., 2021] when no experimental structure is available. Candidate SMILES from the *Drug-Generation Agent* are converted into 3D conformations using RDKit[3]. Binding pockets are predicted with P2Rank [Krivák and Hoksza, 2019][4], and the highest-ranked pocket defines the docking box, whose coordinates are extracted from the P2Rank output and expanded with a fixed padding margin. With receptor and ligand prepared, AutoDock Vina (v1.1.2) performs docking within the predicted pocket, generating 20 poses ranked by affinity. Using this approach, we achieved RMSD lower than 2 Å in 86.59% of cases and a mean RMSD of 1.16 Å on the Astex dataset. The docking setup and visualization, including binding sites, grid box, and ligand, are shown in Figure 3

Following the generation and docking of hits, the workflow progresses to the *Lead Identification* stage. The *Chemical Properties Agent* calculates key physicochemical descriptors (molecular weight, logP, TPSA, hydrogen bond donors and acceptors, rotatable bonds, QED, etc.) using RDkit driven tools . Molecules are filtered according to Lipinski's Rule of Five[5] [Lipinski et al., 2001] and Veber's rules[6] [Veber et al., 2002], ensuring that only drug-like compounds advance. In parallel, the *ADMET Properties Agent*, using ChemFM[Cai et al., 2025] framework, is also invoked at this stage to provide an early assessment of ADMET. Properties that this agent can predict are presented in 2. Compound that show the most favorable docking, pass physicochemical filters, and exhibit acceptable ADMET predictions is prioritized as lead.

Next is *Lead Optimizations* stage. This stage focuses on optimizing the chosen molecule to enhance its pharmacological profile while preserving strong binding affinity to the DPP4 target. The *Molecule Optimization Agent*, based on DrugAssist [Ye et al., 2023], iteratively modifies the structure to enhance bioavailability, solubility, and safety. Each optimized variant is re-evaluated by the *ADMET Properties Agent* and *Docking Agent*, and this loop continues until optimal properties are achieved.

The optimized compound with properties serve entry point into the **Preclinical Phase**, where the optimized candidate

[3] https://www.rdkit.org

[4] P2Rank success rates: 72.0% Top-n, 78.3% Top-(n+2) on COACH420; 68.6% Top-n, 74.0% Top-(n+2) on HOLO4K

[5] HBD $\leqslant 5, HBA \leqslant 10, MW \leqslant 500, logP \leqslant 5$

[6] RotB $\leqslant 10, TPSA \leqslant 140Å$

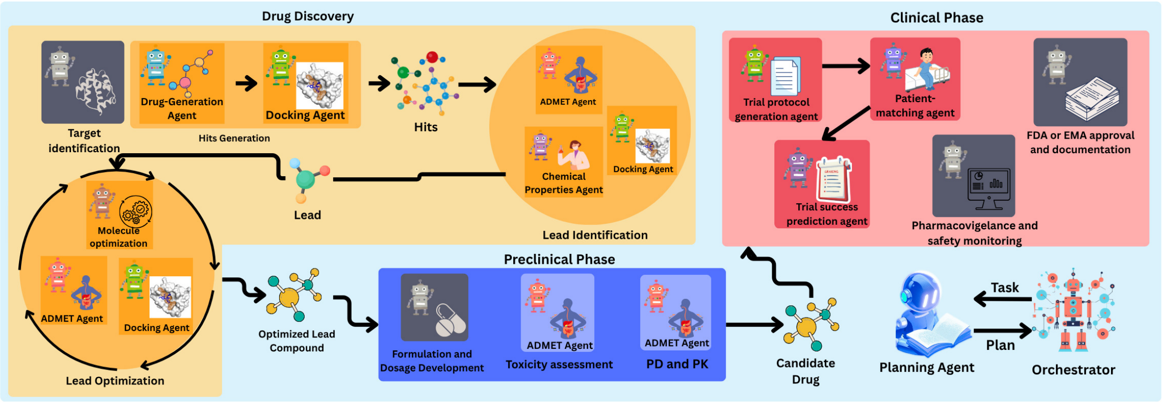


Fig. 2. Prompt-to-Pill Multi-Agent Architecture.

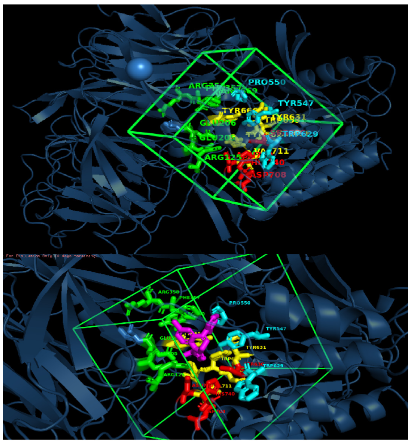


Fig. 3. Docking visualization of DPP4 (PDB ID: 2QT9) showing the predicted binding pocket (green grid box; center = 37.87, 49.09, 36.58; edge = 25.02 Å) and the docked ligand (magenta; SMILES: OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2). Pocket side chains are shown as colored sticks (colors for visual separation only) and correspond to validated binding residues ARG125, GLU205, GLU206, TYR547, TYR631, SER630, HIS740, and ASN710 [Mathur et al., 2023].

undergoes systematic pharmacokinetic and toxicity profiling using ADMET Agent's tools. Once these evaluations are completed, the workflow is shifted into the **Clinical Phase** for trial simulation.

In the Clinical Phase, the *Trial Generation Agent* constructs a trial protocol tailored to the compound and disease driven by *Panacea* model for criteria, arms and outcomes prediction. This protocol is parsed into structured data and passed to the *Patient-Matching Agent*, which also employs the *Panacea* model [Lin et al., 2024] to evaluate patient EHR descriptions and identify candidates who meet the trial's inclusion and exclusion criteria. The agent returns number of matched patients in the final report, and a set of matched patient IDs. These identifiers are saved to a file and the total number of matched patients is computed and included in the final trial report. Subsequently, the *Trial Outcome Prediction Agent* uses *MediTab* [Wang et al., 2024] to estimate the probability that the proposed trial will succeed, given its protocol structure. In line with the original MediTab formulation[Wang et al., 2024], this module operates on trial-level metadata and text and learns patterns from historical ClinicalTrials.gov and HINT benchmarks.

Finally, the matched patient data, drug properties, and trial design are provided to the *Orchestrator*, which aggregates all outputs into a structured report.

The input and output format for the Dipeptidyl peptidase 4 (DPP4) target are shown in Figure 4.

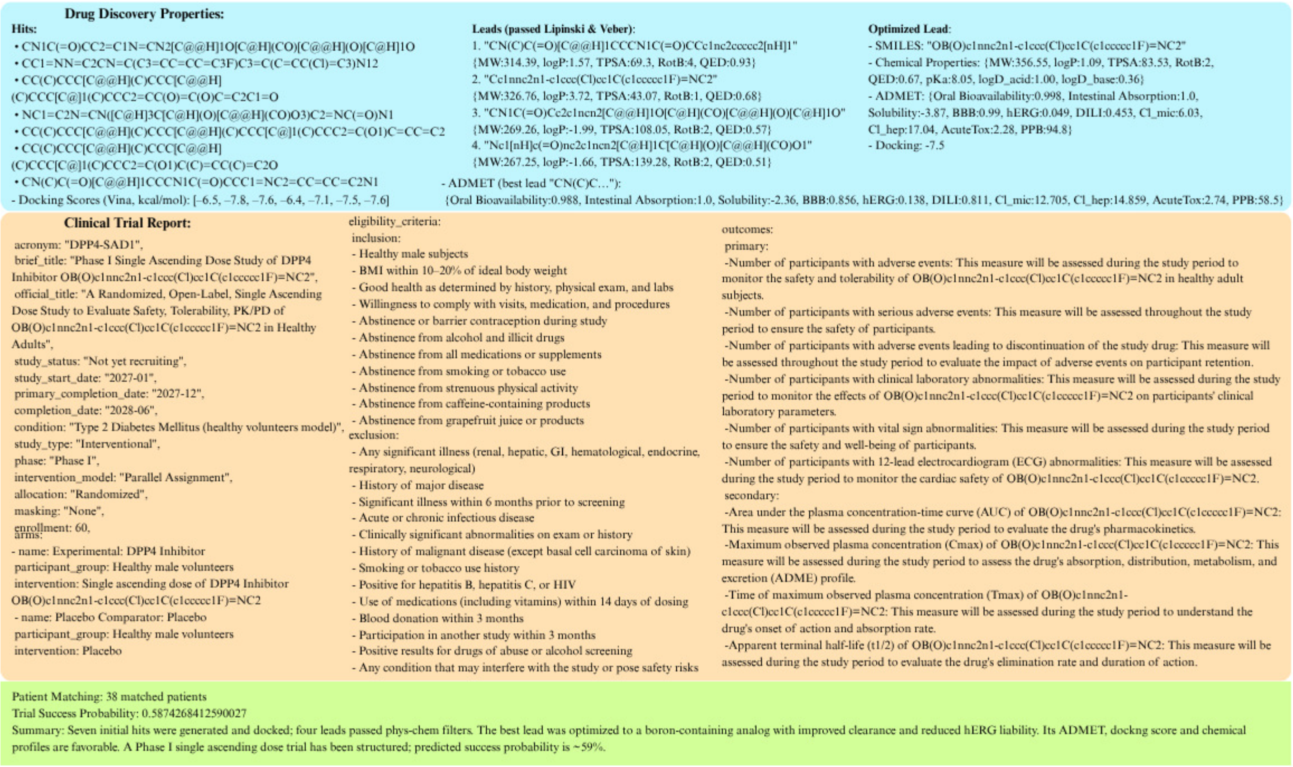
Limitations and Future Work

Prompt-to-Pill is designed as a research-oriented, hypothesis-generation framework intended for use exclusively by trained professionals such as bioinformaticians, medicinal chemists, pharmacologists, and clinicians. The system is not intended for clinical or regulatory decision-making. Instead, its outputs serve as exploratory insights that must be experimentally or clinically validated before any real-world application. The framework aims to support early ideation, academic research, and computational prototyping. While the proposed Prompt-to-Pill pipeline offers a structured, automated approach to drug discovery and clinical simulation, several limitations remain.

First, as shown in Figure 2, some agents remain conceptual placeholders (highlighted in grey), like *Target Identification Agent*, the *Formulation and Dosage Development Agent*, *FDA or EMA approval and documentation* and the *Pharmacovigilance and Safety Monitoring Agent.* Although we identified related approaches in the literature, most lack accessible implementations or compatible I/O interfaces, preventing integration. Bridging this gap remains a key direction for future work.

Second, the *Orchestrator* is currently implemented using OpenAI's o4-mini model, which has been shown to perform strongly in medical reasoning and biomedical tasks Arora et al. [2025]. However, in the trial generation phase, its role is used to producing structured protocol fields such as:study documents, brief summary, acronym, brief title, official title,study status, study start date, primary completion date, completion date, condition, study type, phase, intervention model, allocation, masking, and enrollment. While useful for structuring and simulating trial protocols, these outputs cannot substitute for expert-driven trial design.

Third, each component model operates within the applicability domain (AD) of its training data, and the pipeline therefore inherits the intersection of all ADs. Predictions



Drug Discovery Properties:

Hits:

- CN1C(=O)CC2=C1N=CN2[C@@H]1O[C@H](CO)[C@@H](O)[C@H]1O
- CC1=NN=C2CN=C(C3=CC=CC=C3F)C3=C(C=CC(Cl)=C3)N12
- CC(C)CCC[C@@H](C)CCC[C@@H](C)CCC[C@]1(C)CCC2=CC(O)=C(O)C=C2C1=O
- NC1=C2N=CN([C@H]3C[C@H](O)[C@@H](CO)O3)C2=NC(=O)N1
- CC(C)CCC[C@@H](C)CCC[C@@H](C)CCC[C@]1(C)CCC2=C(O1)C=CC=C2
- CC(C)CCC[C@@H](C)CCC[C@@H](C)CCC[C@]1(C)CCC2=C(O1)C(C)=CC(C)=C2O
- CN(C)C(=O)[C@@H]1CCCN1C(=O)CCC1=NC2=CC=CC=C2N1

- Docking Scores (Vina, kcal/mol): [–6.5, –7.8, –7.6, –6.4, –7.1, –7.5, –7.6]

Leads (passed Lipinski & Veber):

1. "CN(C)C(=O)[C@@H]1CCCN1C(=O)CCc1nc2ccccc2[nH]1" {MW:314.39, logP:1.57, TPSA:69.3, RotB:4, QED:0.93}
2. "Cc1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2" {MW:326.76, logP:3.72, TPSA:43.07, RotB:1, QED:0.68}
3. "CN1C(=O)Cc2c1ncn2[C@@H]1O[C@H](CO)[C@@H](O)[C@H]1O" {MW:269.26, logP:-1.99, TPSA:108.05, RotB:2, QED:0.57}
4. "Nc1[nH]c(=O)nc2c1ncn2[C@H]1C[C@H](O)[C@@H](CO)O1" {MW:267.25, logP:-1.66, TPSA:139.28, RotB:2, QED:0.51}

- ADMET (best lead "CN(C)C…"):
{Oral Bioavailability:0.988, Intestinal Absorption:1.0, Solubility:-2.36, BBB:0.856, hERG:0.138, DILI:0.811, Cl_mic:12.705, Cl_hep:14.859, AcuteTox:2.74, PPB:58.5}

Optimized Lead:

- SMILES: "OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2"
- Chemical Properties: {MW:356.55, logP:1.09, TPSA:83.53, RotB:2, QED:0.67, pKa:8.05, logD_acid:1.00, logD_base:0.36}
- ADMET: {Oral Bioavailability:0.998, Intestinal Absorption:1.0, Solubility:-3.87, BBB:0.99, hERG:0.049, DILI:0.453, Cl_mic:6.03, Cl_hep:17.04, AcuteTox:2.28, PPB:94.8}
- Docking: -7.5

Clinical Trial Report:

acronym: "DPP4-SAD1",
brief_title: "Phase I Single Ascending Dose Study of DPP4 Inhibitor OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2",
official_title: "A Randomized, Open-Label, Single Ascending Dose Study to Evaluate Safety, Tolerability, PK/PD of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2 in Healthy Adults",
study_status: "Not yet recruiting",
study_start_date: "2027-01",
primary_completion_date: "2027-12",
completion_date: "2028-06",
condition: "Type 2 Diabetes Mellitus (healthy volunteers model)",
study_type: "Interventional",
phase: "Phase I",
intervention_model: "Parallel Assignment",
allocation: "Randomized",
masking: "None",
enrollment: 60,
arms:
- name: Experimental: DPP4 Inhibitor
participant_group: Healthy male volunteers
intervention: Single ascending dose of DPP4 Inhibitor OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2
- name: Placebo Comparator: Placebo
participant_group: Healthy male volunteers
intervention: Placebo

eligibility_criteria:
inclusion:
- Healthy male subjects
- BMI within 10–20% of ideal body weight
- Good health as determined by history, physical exam, and labs
- Willingness to comply with visits, medication, and procedures
- Abstinence or barrier contraception during study
- Abstinence from alcohol and illicit drugs
- Abstinence from all medications or supplements
- Abstinence from smoking or tobacco use
- Abstinence from strenuous physical activity
- Abstinence from caffeine-containing products
- Abstinence from grapefruit juice or products

exclusion:
- Any significant illness (renal, hepatic, GI, hematological, endocrine, respiratory, neurological)
- History of major disease
- Significant illness within 6 months prior to screening
- Acute or chronic infectious disease
- Clinically significant abnormalities on exam or history
- History of malignant disease (except basal cell carcinoma of skin)
- Smoking or tobacco use history
- Positive for hepatitis B, hepatitis C, or HIV
- Use of medications (including vitamins) within 14 days of dosing
- Blood donation within 3 months
- Participation in another study within 3 months
- Positive results for drugs of abuse or alcohol screening
- Any condition that may interfere with the study or pose safety risks

outcomes:
primary:
-Number of participants with adverse events: This measure will be assessed during the study period to monitor the safety and tolerability of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2 in healthy adult subjects.
-Number of participants with serious adverse events: This measure will be assessed throughout the study period to ensure the safety of participants.
-Number of participants with adverse events leading to discontinuation of the study drug: This measure will be assessed throughout the study period to evaluate the impact of adverse events on participant retention.
-Number of participants with clinical laboratory abnormalities: This measure will be assessed during the study period to monitor the effects of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2 on participants' clinical laboratory parameters.
-Number of participants with vital sign abnormalities: This measure will be assessed during the study period to ensure the safety and well-being of participants.
-Number of participants with 12-lead electrocardiogram (ECG) abnormalities: This measure will be assessed during the study period to monitor the cardiac safety of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2.
secondary:
-Area under the plasma concentration-time curve (AUC) of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2: This measure will be assessed during the study period to evaluate the drug's pharmacokinetics.
-Maximum observed plasma concentration (Cmax) of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2: This measure will be assessed during the study period to assess the drug's absorption, distribution, metabolism, and excretion (ADME) profile.
-Time of maximum observed plasma concentration (Tmax) of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2: This measure will be assessed during the study period to understand the drug's onset of action and absorption rate.
-Apparent terminal half-life (t1/2) of OB(O)c1nnc2n1-c1ccc(Cl)cc1C(c1ccccc1F)=NC2: This measure will be assessed during the study period to evaluate the drug's elimination rate and duration of action.

Patient Matching: 38 matched patients
Trial Success Probability: 0.5874268412590027
Summary: Seven initial hits were generated and docked; four leads passed phys-chem filters. The best lead was optimized to a boron-containing analog with improved clearance and reduced hERG liability. Its ADMET, dockng score and chemical profiles are favorable. A Phase I single ascending dose trial has been structured; predicted success probability is ~59%.

Fig. 4. Prompt-to-Pill's I/O example for Task: "Simulate drug development for DPP4 (P27487) with patients on /path/to/patients.xml". "Trial Success Probability" correspond to MediTab's predicted likelihood of clinical trial success based on protocol text and structured trial metadata (e.g., phase, condition, enrollment, arms, outcomes), and do not represent estimates of the underlying drug's biological efficacy.

involving molecules, trial structures, or patient populations far from these distributions should be interpreted as exploratory, not definitive.

This study also presents only a single-target case (DPP-4, P27487), demonstrating feasibility but not generalizability. Future work will extend the framework to multiple targets and diseases for broader validation.

Future work will focus on completing missing agents, enlarging the AD of existing components, and performing multi-disease, multi-target validation studies.

Discussion

While LLMs have opened transformative opportunities in drug discovery and clinical research, realizing their full potential requires addressing key challenges in transparency, evaluation consistency, and reproducibility.

A key limitation is the limited reasoning ability of current models. In biomedical contexts, correctness alone is insufficient—decisions must be grounded in clear, interpretable reasoning that experts can verify. To address this, several recent models have introduced mechanisms to make reasoning more explicit. These include retrieval-augmented generation [Wang et al., 2025a, Xu et al., 2025, Feng et al., 2025], instruction-tuned multitask learning [Liu et al., 2024, Ma et al., 2024], and multi-hop rationale generation [Feng et al., 2025, Wang et al., 2023]. Such approaches represent important progress toward interpretability. However, without standardized frameworks to assess reasoning quality or consistency, trust in LLM-driven biomedical insights remains limited.

Another major challenge in LLM-based DD is the inconsistency in evaluation protocols across model types. Generative models are assessed using metrics like validity, docking scores, or QED [Sheikholeslami et al., 2025, Wang et al., 2025b, Zhao et al., 2024], while discriminative models report AUROC or F1 scores [Wang et al., 2025a, Liu et al., 2024, Ma et al., 2024], yet differ in datasets and thresholds. Knowledge-retrieval and reasoning systems often rely on qualitative outputs without standardized measures [Feng et al., 2025, Wang et al., 2023]. This fragmentation hinders comparability and progress. To address this, the field urgently needs task-specific, model-type-sensitive benchmarks.

Reproducibility and transparency also remain persistent issues. Many studies lack public access to code, models, or I/O examples, and when repositories exist, documentation is often incomplete. This fragmentation limits cumulative progress and undermines trust.

These models are the future of drug development, but there is still much work to be done. The path forward requires not only better models, but better systems around the models. This includes standardized evaluations, transparent documentation, expert-guided development, and thoughtful regulation. Only by meeting these unmet needs can we ensure that LLMs evolve from experimental tools to trusted agents in the future of biomedical discovery.

Ethical and Regulatory Considerations

As highlighted in recent discussions on responsible biomedical AI deployment [Tang et al., 2025], LLM-based systems in safety-critical domains such as clinical trial design raise central concerns around transparency, explainability, bias mitigation, and the risk of over-reliance on unvalidated outputs. The European Union's Artificial Intelligence Act (2024) explicitly designates healthcare AI as "high-risk" (Recital 58; Annex III), requiring safeguards such as traceability, human oversight, and fundamental rights impact assessments. The authors have recently published their work on AI Act compliance within the MyHealth@EU framework [Simjanoska Misheva et al., 2025], demonstrating strong ethical responsibility in advancing AI use within sensitive healthcare environments. Their tutorial addresses the dual-compliance challenge of embedding AI Act safeguards (transparency, provenance, robustness) while meeting MyHealth@EU interoperability requirements, showing how AI metadata can be integrated into HL7 CDA and FHIR messages without disrupting existing standards. The goal is not to bypass current guidelines but to ease clinicians' workload, strengthen trust in AI-assisted decisions, and ensure that compliance and safety are engineered into systems from the outset. In this context, the present pipeline is strictly positioned as a research prototype and decision-support artifact, never as an automated tool for patient eligibility or therapeutic approval. By embedding governance mechanisms early and framing the work as proof-of-concept exploration, the approach contributes to the broader dialogue on trustworthy AI in DD while acknowledging the rigorous benchmarking, reproducibility, and expert oversight still required before clinical translation.

Conclusion

To illustrate practical integration, the Prompt-to-Pill multi-agent framework was proposed, uniting specialized LLM agents to automate decision-making across preclinical and clinical stages. This architecture showcases how coordinated LLM workflows can collaborate, iterate, and self-correct within a modular design. Crucially, the successful implementation of this architecture served simultaneously as a ”limitation demonstrator," significantly highlighting the applicability domain limitations and systemic challenges that must be overcome, thereby establishing a rigorous foundation upon which reliable hypothesis generation can be built.

Looking ahead, the progress of LLM-driven drug development will depend not only on more capable models but on robust evaluation protocols, transparent sharing, and clear regulatory standards. Addressing these challenges will allow LLMs not just to accelerate, but to redefine the future of drug development.

Code Availability

The full implementation of the Prompt-to-Pill multi-agent DD and clinical simulation pipeline is available at the GitHub repository: `https://github.com/ChatMED/Prompt-to-Pill`.

Supplementary Material

Table with the retrieved papers for this review is available at the following link: **Supplementary Table 1**.

Funding

This work was supported by the European Union under the Horizon Europe programme [grant number 101159214]. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

R. K. Arora et al. HealthBench: Evaluating large language models towards improved human health. May 2025.

J.-M. T. Beasley et al. Tarragon: Therapeutic target applicability ranking and retrieval-augmented generation over networks. 2025. doi: 10.1101/2025.04.19.649662.

H. M. Berman et al. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000. doi: 10.1093/nar/28.1.235.

F. Cai et al. ChemFM as a scaling law guided foundation model pre-trained on informative chemicals. Nov. 2025.

J. M. Z. Chaves et al. Tx-LLM: A large language model for therapeutics. June 2024.

S. Datta et al. Patient2trial: From patient to participant in clinical trials using large language models. *Informatics in Medicine Unlocked*, 53:101615, 2025. ISSN 2352-9148. doi: https://doi.org/10.1016/j.imu.2025.101615.

C. Edwards et al. Synergpt: In-context learning for personalized drug synergy prediction and drug design. 2023. doi: 10.1101/2023.07.06.547759.

Y. Feng et al. A retrieval-augmented knowledge mining method with deep thinking llms for biomedical research and clinical support. *GigaScience*, 14:giaf109, 09 2025.

B. Gao et al. A multi-agent framework for reliable drug-target interaction prediction, 2024. Course proposal, Advanced Machine Learning, Tsinghua University.

Y. Inoue et al. DrugAgent: Multi-agent large language model-based reasoning for drug-target interaction prediction. Aug. 2024.

J. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.

R. Krivák and D. Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10:39, 2019.

N. Lee et al. RAG-enhanced collaborative LLM agents for drug discovery. Nov. 2025.

Y. Li et al. Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins. 2023. doi: 10.1101/2023.06.29.543848.

Z. Li et al. GraPPI: A retrieve-divide-solve GraphRAG framework for large-scale protein-protein interaction exploration. Jan. 2025.

J. Lin et al. Panacea: A foundation model for clinical trial search, summarization, design, and recruitment. 2024. doi: 10.1101/2024.06.26.24309548.

C. A. Lipinski et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1–3):3–26, 2001.

Y. Liu et al. MolecularGPT: Open large language model (LLM) for few-shot molecular property prediction. June 2024.

T. Ma et al. Y-Mol: A multiscale biomedical knowledge-guided large language model for drug development. Oct. 2024.

N. Markey et al. From rags to riches: Utilizing large language models to write documents for clinical trials. *Clinical Trials*, 22(5):626–631, Feb. 2025. ISSN 1740-7753. doi: 10.1177/17407745251320806.

V. Mathur et al. Insight into structure activity relationship of dpp-4 inhibitors for development of antidiabetic agents. *Molecules*, 28(15):5860, 2023.

D. Oniani et al. Emerging opportunities of using large language models for translation between drug molecules and indications. *Sci Rep*, 14, 2024. doi: https://doi.org/10.1038/s41598-024-61124-0.

J. L. Painter et al. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. *JAMIA Open*, 8(1): ooaf003, 02 2025.

M. Reinisch et al. Ctp-llm: Clinical trial phase transition prediction using large language models. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3667–3672, 2024. doi: 10.1109/BIBM62325.2024.10822746.

R. A. Schmitt et al. Biological database mining for llm-driven alzheimer's disease drug repurposing. 2025. doi: 10.1101/2024.12.04.626255.

M. Sheikholeslami et al. Druggen enhances drug discovery with large language models and reinforcement learning. *Sci Rep*, 15:13445, 2025. doi: https://doi.org/10.1038/s41598-025-98629-1.

M. Simjanoska Misheva et al. Ai act compliance within the myhealth@eu framework: A tutorial. *Journal of Medical Internet Research*, 2025. doi: 10.2196/81184. Advance online publication.

K. Song et al. LLM agent swarm for hypothesis-driven drug discovery. Apr. 2025.

X. Tang et al. Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16:8317, 2025.

D. F. Veber et al. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12):2615–2623, 2002.

C. von Csefalvay. DAEDRA: A language model for predicting outcomes in passive pharmacovigilance reporting. Feb. 2024.

E. Wang et al. TxGemma: Efficient and agentic LLMs for therapeutics. Apr. 2025a.

J. Wang et al. 3dsmiles-gpt: 3d molecular pocket-based generation with token-only large language model. *Chemical Science*, 16:637–648, 2025b. doi: 10.1039/D4SC06864E.

Z. Wang et al. In *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings, pages 12461–12472. Association for Computational Linguistics (ACL), 2023. doi: 10.18653/v1/2023.emnlp-main.766.

Z. Wang et al. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement. In K. Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6062–6070. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/670. Main Track.

L. Wu et al. Leveraging fda labeling documents and large language model to enhance annotation, profiling, and classification of drug adverse events with askfdalabel. *Drug Safety*, 48:655–665, 2025. doi: https://doi.org/10.1007/s40264-025-01520-1.

Q. Wu et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling*, 2024.

Z. Xu et al. *Retrieval-Reasoning Large Language Model-based Synthetic Clinical Trial Generation.* Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400722004. doi: https://doi.org/10.1145/3765612.3767193.

Z. Yang et al. DrugMCTS: a drug repurposing framework combining multi-agent, RAG and monte carlo tree search. July 2025.

G. Ye et al. DrugAssist: A large language model for molecule optimization. Dec. 2023.

Y. Zhao et al. Generrna: A generative pre-trained language model for de novo rna design. 2024. doi: 10.1101/2024.02.01.578496.

Table 1. Availability of I/O Examples, RAG Usage, and Code Links in Studies

Category	Preclinical/DD			Clinical		
	Yes	No	N/A	Yes	No	N/A
GitHub/HF link	20	12	–	9	10	–
Input examples	17	2	1	6	2	1
Output examples	6	13	1	1	7	1
RAG usage	3	29	–	6	13	–

Ivana Vichentijevikj Miss Vichentijevikj is a postgraduate student in Bioinformatics with a biomedical background. She works as a Bioinformatics Software Engineer at iReason, focusing on the development of intelligent systems for drug discovery and clinical data analysis. Her academic and research interests focus on the application of computational methods in drug development and personalized medicine. Ivana is a part of the ChatMED project, where she contributes to the integration of generative AI tools into clinical and pharmaceutical workflows. Her work reflects a strong dedication to advancing translational bioinformatics and promoting the role of AI in the future of healthcare systems.

Kostadin Mishev. Prof. Mishev is a professor of Computer Science and Engineering and an expert in the emerging interdisciplinary field that bridges artificial intelligence and DevOps, AIOps, bringing cutting-edge methodologies for operationalizing machine learning pipelines in real-world environments. As a member of the Scientific Committee of the ChatMED project, he contributes strategic and technical expertise in deploying robust, compliant, and scalable AI-driven systems within clinical workflows. Prof. Mishev is also a key industry representative. His portfolio includes the development and deployment of production-grade AI applications, notably, he has led the design of VoiceBot technologies tailored for medical contexts, production-ready for integrating into clinical decision support systems.

Monika Simjanoska Misheva. Dr. Simjanoska Misheva is a professor of Computer Science and Engineering with a research focus at the intersection of artificial intelligence, bioinformatics, and clinical decision support systems. She is currently the coordinator of the ChatMED project, a pioneering initiative that aims to integrate generative AI across all levels of a national healthcare system. Her work is driven by the ambition to ensure that the deployment of AI in healthcare is not only innovative but also compliant with the latest European regulations. Dr. Simjanoska's contributions extend across multi-agent systems, large language models, and their applications in neurology, cancer diagnosis, and personalized medicine. Her recent efforts center on creating regulatory-aware AI architectures capable of multimodal reasoning with genomic, imaging, and biosignal data.

Table 2. Evaluation metrics of the core agents used in the clinical workflow

Citation	Agent	Dataset / Task	Metric(s)	Value(s)	Type
[Sheikholeslami et al., 2025]	Molecule Generation (Druggen)	–	Validity, Novelty, Diversity	99.9%, 41.88%, 60.32%	Generation
[Cai et al., 2025]	Property Prediction (ChemFM)	Drug Oral Bioavailability	ROC-AUC	0.715 ± 0.011	Classification
		BBB	ROC-AUC	0.908 ± 0.010	Classification
		Drug Half-Life Duration	Spearman	0.551 ± 0.020	Regression
		Drug Mutagenicity	ROC-AUC	0.854 ± 0.007	Classification
		Clearance Hepatocyte	Spearman	0.495 ± 0.030	Regression
		Clearance Microsome	Spearman	0.611 ± 0.016	Regression
		DILI	ROC-AUC	0.920 ± 0.012	Classification
		hERG Channel Blockage	ROC-AUC	0.848 ± 0.009	Classification
		Drug Acute Toxicity	MAE	0.541 ± 0.015	Regression
		PPBR	MAE	7.505 ± 0.073	Regression
		P-glycoprotein Inhibition	ROC-AUC	0.931 ± 0.003	Classification
		Drug Aqueous Solubility	MAE	0.725 ± 0.011	Regression
		VDss	Spearman	0.662 ± 0.013	Regression
		CYP2C9 Inhibition	PRC-AUC	0.788 ± 0.005	Classification
		CYP3A4 Inhibition	PRC-AUC	0.878 ± 0.003	Classification
		CYP2C9 Substrate	PRC-AUC	0.414 ± 0.027	Classification
		CYP2D6 Inhibition	PRC-AUC	0.704 ± 0.003	Classification
		CYP2D6 Substrate	PRC-AUC	0.739 ± 0.024	Classification
		Human IA	ROC-AUC	0.984 ± 0.004	Classification
		CYP3A4 Substrate	ROC-AUC	0.654 ± 0.022	Classification
		Drug Permeability	MAE	0.322 ± 0.026	Regression
[Wang et al., 2024]	Trial Outcome Prediction(Meditab)	Phase I Trials	AUROC, PRAUC	0.699, 0.726	Classification
		Phase II Trials	AUROC, PRAUC	0.706, 0.733	Classification
		Phase III Trials	AUROC, PRAUC	0.734, 0.881	Classification
[Ye et al., 2023]	Molecule Optimization (DrugAssist)	-	Solubility, BBBP, All, Valid rate, Similarity	0.74, 0.80, 0.62, 0.98, 0.69	Generation
[Lin et al., 2024]	PatientMatching (Panacea)	SIGIR	BACC, F1, R, P	0.43, 0.57, 0.52, 0.66	Classification
		TREC 2021	BACC, F1, R, P	0.47, 0.58, 0.54, 0.69	Classification
	TrialDesign (Panacea)	Criteria	BLEU, ROUGE, CR	0.24, 0.44, 0.68	Generation
		Arms	BLEU, ROUGE, CR	0.28, 0.50, 0.61	Generation
		Outcome	BLEU, ROUGE, CR	0.31, 0.51, 0.55	Generation

Source: Metrics are based on results reported in the original publications.

[1] ROC-AUC/AUROC: Area under the Receiver Operating Characteristic Curve, MAE: Mean Absolute Error, PRC-AUC/PRAUC: Area Under Precision-Recall Curve, BACC: Balanced Accuracy; R: Recall, P: Precision, F1: F1 Score, BLEU: Bilingual Evaluation Understudy, ROUGE: Recall-Oriented Understudy for Gisting Evaluation, CR: Clinical Relevance