Novelty Assessment Report

Paper: MedAgentGym: A Scalable Agentic Training Environment for Code-Centric Reasoning in Biomedical Data Science
PDF URL: https://openreview.net/pdf?id=jHDZEUgS4r
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We introduce MedAgentGym, a scalable and interactive training environment designed to enhance coding-based biomedical reasoning capabilities in large language model (LLM) agents. MedAgentGym comprises 72,413 task instances across 129 categories derived from 12 authentic real-world biomedical scenarios. Tasks are encapsulated within executable sandbox environments, each featuring detailed task specifications, interactive feedback mechanisms, verifiable ground truth annotations, and scalable training trajectory generation. Extensive benchmarking of 29 LLMs reveals substantial performance disparities in biomedical data science between commercial and open-source LLMs. Leveraging efficient multi-threaded and multi-turn trajectory sampling in MedAgentGym, Med-Copilot achieves performance gains of +43.02% and +45.28% from offline and online reinforcement learning, respectively, demonstrating MedAgentGym as an effective training ground while establishing itself as a cost-effective, privacy-preserving alternative competitive with proprietary LLMs (gpt-4o). By offering a unified execution environment with a comprehensive benchmark and accessible, extensible training resources, MedAgentGym delivers an integrated platform to develop LLM-based coding assistants for advanced biomedical data science.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **coding-based biomedical reasoning in large language models**
A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Interactive Training Environments and Agent Frameworks**
- **Clinical Coding and Medical Classification**
- **Code-Driven Reasoning and Execution**
- **Clinical Reasoning and Decision Support**
- **Domain-Specific Biomedical Reasoning**
- **Training Optimization and Model Improvement**
- **Supporting Tools and Infrastructure**

Complete Taxonomy Tree

- coding-based biomedical reasoning in large language models Survey Taxonomy
- Interactive Training Environments and Agent Frameworks
  - Scalable Agentic Training Platforms ★ (2 papers)
  - [0] MedAgentGym: A Scalable Agentic Training Environment for Code-Centric Reasoning in Biomedical Data Science (Anon et al., 2026) View paper
  - [2] Medagentgym: Training llm agents for code-based medical reasoning at scale (Xu Ran, 2025) View paper
  - Multi-Agent Collaboration Architectures (5 papers)
  - [5] Agentic ai framework for end-to-end medical data inference (Soorya Ram Shimgekar, 2025) View paper
  - [12] MACD: Multi-Agent Clinical Diagnosis with Self-Learned Knowledge for LLM (Li Wenliang, 2025) View paper
  - [20] Exploring llm multi-agents for icd coding (Li Rumeng, 2024) View paper
  - [30] Biomedical reasoning in action: Multi-agent System for Auditable Biomedical Evidence Synthesis (Wysocki, 2025) View paper
  - [47] MedAide: Information Fusion and Anatomy of Medical Intents via LLM-based Agent Collaboration (Dingkang Yang, 2024) View paper
  - General-Purpose Biomedical AI Agents (3 papers)
  - [7] Biomni: A general-purpose biomedical ai agent (Kexin Huang, 2025) View paper
  - [10] Learning to be a doctor: Searching for effective medical agent architectures (Yangyang Zhuang, 2025) View paper
  - [44] LLMAgent4Bio: LLM Agents for Biological Intelligence Across Genomics, Proteomics, Spatial Biology, and Biomedicine (SA Dip, 2025) View paper
- Clinical Coding and Medical Classification
  - ICD Code Assignment (6 papers)
  - [11] GPT-Enhanced Hierarchical Deep Learning Model for Automated ICD Coding (Joshua Carberry, 2024) View paper
  - [22] Enhanced BioGPT for Automated ICD-10 Medical Coding: Harnessing Domain-Tuned Large Language Models for Smarter Medical Coding (Ramadan, 2025) View paper
  - [25] Verification is All You Need: Prompting Large Language Models for Zero-Shot Clinical Coding (Shaoxin Li, 2025) View paper
  - [26] Automatic ICD Code Generation for Lymphoma Using Large Language Models (Yu Song, 2024) View paper
  - [45] Reasoning Large Language Models for Clinical Coding (Mustafa Akram, 2025) View paper
  - [49] Leveraging Chain of Thought for Automated Medical Coding (W Song, 2025) View paper
  - Evidence-Based Medical Coding (2 papers)

- [16] Evidence extraction for automated medical coding: preliminary evaluation (Xiaorui Jiang, 2024) View paper
- [36] Evaluation and LLM-Guided Learning of ICD Coding Rationales (Li Mingyang, 2025) View paper
- Ontology-Augmented Code Mapping (2 papers)
- [4] OntologyRAG: Better and faster biomedical code mapping with retrieval-augmented generation (RAG) leveraging ontology knowledge graphs and large language â¦ (H Feng, 2025) View paper
- [32] 339P LLM-based reasoning framework for MedDRA adverse event coding: Toward scalable automation (N Dashti, 2025) View paper

• Code-Driven Reasoning and Execution
- Electronic Health Record Analysis (2 papers)
- [3] Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records (Wenqi Shi, 2024) View paper
- [15] EHRAgent: Code Empowers Large Language Models for Complex Tabular Reasoning on Electronic Health Records (Shi, 2024) View paper
- Unified Structured Knowledge Reasoning (2 papers)
- [6] Pandora: A Code-Driven Large Language Model Agent for Unified Reasoning Across Diverse Structured Knowledge (Chen Yongrui, 2025) View paper
- [28] Pandora: Leveraging Code-driven Knowledge Transfer for Unified Structured Knowledge Reasoning (Chen Yongrui, 2025) View paper
- Code as Reasoning Substrate (4 papers)
- [23] On Code-Induced Reasoning in LLMs (Waheed Abdul, 2025) View paper
- [27] Code Prompting Elicits Conditional Reasoning Abilities in Text+Code LLMs (Aditya, 2024) View paper
- [39] CodeGraph: Enhancing Graph Reasoning of LLMs with Code (Wang Zhaowei, 2024) View paper
- [50] Steering Large Language Models between Code Execution and Textual Reasoning (Chen Yong-chao, 2024) View paper
- Code-Enhanced Inductive and Conditional Reasoning (2 papers)
- [14] Code-Driven Inductive Synthesis: Enhancing Reasoning Abilities of Large Language Models with Sequences (Chen Ke-di, 2025) View paper
- [29] Code-Driven LLM Agent for One-Shot Explanatory Visual Question Answering (Z Zhou, 2025) View paper

• Clinical Reasoning and Decision Support
- Structured Clinical Reasoning Protocols (2 papers)
- [9] Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases (Yuki Sonoda, 2025) View paper
- [42] Hippocrates-o1: A Guideline-Aware, Orchestrated, Self-Refining Protocol for Specialty-Specific Clinical Reasoning (B Wang, 2025) View paper
- Medical Calculation and Quantitative Reasoning (2 papers)
- [13] From Scores to Steps: Diagnosing and Improving LLM Performance in Evidence-Based Medical Calculations (Wang, 2025) View paper
- [21] MedCalc-Eval and MedCalc-Env: Advancing Medical Calculation Capabilities of Large Language Models (Ding, 2025) View paper
- Risk Prediction and Clinical Outcome Assessment (2 papers)
- [8] Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting â¦ (P Naliyatthaliyazchayil, 2025) View paper
- [35] Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting Approach. (Parvati Naliyatthaliyazchayil, 2025) View paper
- Clinical Pathway and Procedure Reasoning (3 papers)
- [34] Multi-Agent LLM Reasoning for Clinical Procedure Sequencing from High-Granularity EHR Data (Y Zhong, 2025) View paper
- [37] ClinPath: A General-Purpose Knowledge Graph with LLM Reasoning For Understanding Clinical Interactions (S Ankireddy, 2025) View paper
- [43] From Guidelines to Code: Formalizing STOPP/START Criteria Using LLMs and RAG for Clinical Decision Support. (Samya Adrouji, 2025) View paper

• Domain-Specific Biomedical Reasoning
- Genomic and Multimodal Biological Reasoning (1 papers)
- [1] BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model (Adibvafa Fallahpour, 2025) View paper
- Biomedical Syllogistic and Logical Reasoning (1 papers)
- [48] SylloBio-NLI: Evaluating Large Language Models on Biomedical Syllogistic Reasoning (Magdalena Wysocka, 2024) View paper
- Neuro-Symbolic Medical AI (2 papers)
- [31] Enhancing Large Language Models with Neurosymbolic Reasoning for Multilingual Tasks (Agrawal Ameeta, 2025) View paper
- [40] NEURO-GUARD: Neuro-Symbolic Generalization and Unbiased Adaptive Routing for Diagnostics -- Explainable Medical AI (Midhat Urooj, 2025) View paper

• Training Optimization and Model Improvement
- Efficient Fine-Tuning and Data Selection (1 papers)
- [17] Towards Efficient Medical Reasoning with Minimal Fine-Tuning Data (Zhuang Xinlin, 2025) View paper
- Code Execution as Supervision (1 papers)
- [46] Code Execution as Grounded Supervision for LLM Reasoning (Dong-Won Jung, 2025) View paper
- Chain-of-Thought Reasoning Enhancement (2 papers)
- [33] Code to Think, Think to Code: A Survey on Code-Enhanced Reasoning and Reasoning-Driven Code Intelligence in LLMs (Dayu Yang, 2025) View paper
- [41] Chain of Thought Strategy for Smaller LLMs for Medical Reasoning. (Hurmat Ali Shah, 2025) View paper

• Supporting Tools and Infrastructure
- Data Preprocessing and Anonymization (1 papers)
- [24] A Strategy for Anonymizing Free-Text Medical Reports Using LLM-Aix (D Liehr, 2025) View paper
- Knowledge Graph and Entity Alignment (1 papers)
- [18] Unlocking the Power of Large Language Models for Entity Alignment (Xuhui Jiang, 2024) View paper
- Bioinformatics Code Generation and Educational Tools (2 papers)
- [19] BioCoder: a benchmark for bioinformatics code generation with large language models. (Xiangru Tang, 2024) View paper

- [38] Vibe Coding in nephrology education: clinician-led, AI-assisted development of open-source interactive learning tools (Francesco Pesce, 2025) View paper

Narrative

Core task: coding-based biomedical reasoning in large language models. This field explores how LLMs can leverage executable code, structured representations, and computational tools to enhance reasoning over medical data and clinical knowledge. The taxonomy reveals several complementary directions: Interactive Training Environments and Agent Frameworks focus on building scalable platforms where agents learn through interaction with medical tasks; Clinical Coding and Medical Classification address automated assignment of diagnostic codes and structured labels; Code-Driven Reasoning and Execution emphasize using executable programs to perform multi-step inference; Clinical Reasoning and Decision Support target diagnostic workflows and treatment planning; Domain-Specific Biomedical Reasoning tackles specialized problems in genomics, pathology, and other subfields; Training Optimization and Model Improvement develop techniques to refine model capabilities; and Supporting Tools and Infrastructure provide foundational resources such as benchmarks and knowledge bases. Representative works like BioReason[1] and Ehragent[3] illustrate how code generation and agent-based architectures can be combined to handle complex clinical scenarios.

A particularly active line of work centers on scalable agentic training platforms, where systems like MedAgentGym[0] and Medagentgym[2] create rich interactive environments for training agents on diverse medical reasoning tasks. These platforms contrast with more narrowly scoped clinical coding systems that focus on ICD assignment or with code-driven reasoning approaches that emphasize symbolic execution over free-form interaction. MedAgentGym[0] sits squarely within the Interactive Training Environments branch, emphasizing large-scale agent training across varied biomedical scenarios, which distinguishes it from works like Ehragent[3] that target specific EHR-based decision support or from Agentic AI Framework[5] that may prioritize general-purpose agent architectures over domain-specific medical training. The central tension across these branches involves balancing the generality of training environments with the precision required for clinical deployment, and determining whether code-based reasoning should be tightly integrated into agent learning loops or treated as a separate inference module.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Medagentgym: Training llm agents for code-based medical reasoning at scale

Authors: Xu Ran, Zhuang, Yuchen, Ran Xu, Zhong Yishan, et al. (38 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We introduce MedAgentGym, a scalable and interactive training environment designed to enhance coding-based biomedical reasoning capabilities in large language model (LLM) agents. MedAgentGym comprises 72,413 task instances across 129 categories derived from 12 authentic real-world biomedical scenarios. Tasks are encapsulated within executable sandbox environments, each featuring detailed task specifications, interactive feedback mechanisms, verifiable ground truth annotations, and scalable train...

⚠ Similarity Notice

These papers appear to be the same work or very closely related variants. Both introduce MedAgentGym as a scalable training environment for code-centric biomedical reasoning with identical core contributions: 72,413 task instances across 129 categories, executable Docker environments, and Med-Copilot achieving similar performance gains (+43.02% and +45.28% from offline and online RL). The titles, abstracts, methodology, experimental results, and technical details are nearly identical, suggesting these are likely different versions of the same paper (e.g., conference vs. workshop submission).

Contributions Analysis

Overall novelty summary. MedAgentGym introduces a scalable training environment for coding-based biomedical reasoning, comprising 72,413 task instances across 129 categories from 12 real-world scenarios. The taxonomy places this work in the 'Scalable Agentic Training Platforms' leaf, which contains only two papers including the original. This represents a relatively sparse research direction within the broader field of interactive training environments, suggesting the work addresses an emerging need for comprehensive, multi-scenario agent training platforms rather than entering a crowded space of established solutions.

The taxonomy reveals that MedAgentGym sits within the 'Interactive Training Environments and Agent Frameworks' branch, which also includes multi-agent collaboration architectures and general-purpose biomedical AI agents. Neighboring branches focus on clinical coding systems, code-driven EHR analysis, and clinical decision support protocols. The scope notes indicate MedAgentGym's emphasis on executable sandboxes and trajectory generation distinguishes it from static benchmarking approaches and from single-task frameworks that lack scalability. This positioning suggests the work bridges agent training infrastructure with practical biomedical coding applications.

Among 30 candidates examined across three contributions, none were identified as clearly refuting the work's novelty. The MedAgentGym training environment examined 10 candidates with zero refutable overlaps, as did the Med-Copilot coding agent and the unified execution environment contributions. The limited search scope means these statistics reflect top-K semantic matches rather than exhaustive coverage. The absence of refutable candidates across all contributions suggests that within the examined literature, the combination of scale, interactivity, and biomedical domain focus appears distinctive, though broader searches might reveal additional related work.

Based on the limited 30-candidate search, the work appears to occupy a relatively novel position combining large-scale interactive training with biomedical coding tasks. The sparse population of its taxonomy leaf and the lack of refutable candidates among examined papers suggest meaningful differentiation from existing approaches. However, the analysis covers semantic neighbors rather than comprehensive field coverage, and the true novelty assessment would benefit from examining additional agent training platforms and biomedical benchmarking systems beyond the top-K matches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: MedAgentGym training environment

Description: The authors present MedAgentGym as a comprehensive platform comprising 72,413 task instances across 129 categories from 12 real-world biomedical scenarios. Each task is encapsulated in executable sandbox environments with detailed specifications, interactive feedback mechanisms, verifiable ground truth annotations, and scalable training trajectory generation capabilities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning

URL: View paper

Brief Assessment

Mediq[63] focuses on interactive clinical reasoning through question-asking in patient-doctor dialogues, not on providing a comprehensive training environment for code-centric biomedical data science tasks with executable sandboxes and trajectory generation.

---

2. Interactive computer-aided diagnosis on medical image using large language models

URL: View paper

Brief Assessment

Interactive Diagnosis[61] focuses on integrating LLMs with computer-aided diagnosis systems for medical image interpretation and report generation, not on creating interactive training environments for code-centric biomedical reasoning tasks with executable sandboxes and trajectory sampling.

---

3. Medagentgym: Training llm agents for code-based medical reasoning at scale

URL: View paper

Brief Assessment

Medagentgym[2] focuses on the same training environment (MedAgentGym) as the original paper, appearing to be the same work rather than prior art that could refute novelty claims.

---

4. Improving interactive diagnostic ability of a large language model agent through clinical experience learning

URL: View paper

Brief Assessment

Clinical Experience Learning[62] focuses on interactive diagnostic scenarios requiring active information gathering from patients, not on code-centric biomedical data science tasks with executable sandbox environments.

---

5. Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions

URL: View paper

Brief Assessment

Self-Evolving Simulations[66] focuses on multi-agent clinical simulations for diagnostic conversations between doctor-patient-measurement agents, not on code-centric biomedical reasoning tasks with executable sandbox environments and trajectory generation for reinforcement learning.

---

6. â¦ patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study

URL: View paper

Brief Assessment

Virtual Patient Robotics[65] focuses on social robotics combined with LLMs for clinical reasoning training through patient simulations in medical education, not on creating executable sandbox environments for code-centric biomedical data science tasks with verifiable ground truth annotations.

---

7. Generative AI for medical education: Insights from a case study with medical students and an AI tutor for clinical reasoning

URL: View paper

Brief Assessment

AI Tutor Education[68] focuses on interactive learning tools for medical students in clinical reasoning education, not on creating executable training environments for LLM agents in biomedical data science tasks.

---

8. A Proactive Agent Collaborative Framework for Zeroâ□Shot Multimodal Medical Reasoning

URL: View paper

Brief Assessment

Proactive Agent Framework[67] focuses on multimodal medical reasoning through agent-expert collaboration for visual question answering on X-ray images, not on providing an interactive training environment for code-centric biomedical reasoning tasks with executable sandboxes and trajectory generation.

---

9. Medagents: Large language models as collaborators for zero-shot medical reasoning

URL: View paper

Brief Assessment

Medagents[60] focuses on multi-agent collaboration for zero-shot medical reasoning through role-playing discussions, not on providing an interactive training environment with executable sandboxes, trajectory generation, and scalable training infrastructure for code-centric biomedical tasks.

---

10. Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods â¦

URL: View paper

Brief Assessment

Virtual Patient Robotics[64] focuses on social robotics combined with LLMs for clinical reasoning training in medical education, not on code-centric biomedical data science tasks with executable sandbox environments.

Contribution 2: Med-Copilot coding agent

Description: The authors develop Med-Copilot, an LLM-based coding assistant trained using MedAgentGym through both offline and online reinforcement learning methods. The system demonstrates substantial performance improvements and achieves competitive results with proprietary models while maintaining privacy and cost-effectiveness.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. MBuilder: A Multi-agent System for Automated Machine Learning in Medical Imaging

URL: View paper

Brief Assessment

MBuilder[56] focuses on a multi-agent system for automated machine learning in medical imaging, not on reinforcement learning-based training of coding agents for biomedical data science tasks.

---

2. Medagentgym: Training llm agents for code-based medical reasoning at scale

URL: View paper

Brief Assessment

Medagentgym[2] describes the same Med-Copilot agent trained using MedAgentGym, appearing to be the same work rather than prior art that could refute novelty claims.

---

3. Structured preference modeling for reinforcement learning-based fine-tuning of large models

URL: View paper

Brief Assessment

Structured Preference Modeling[52] focuses on preference modeling for reinforcement learning fine-tuning of general large models across NLP tasks, not specifically on biomedical coding agents or the MedAgentGym training environment.

---

4. Enhanced vehicle routing for medical waste management via hybrid deep reinforcement learning and optimization algorithms

URL: View paper

Brief Assessment

Medical Waste Routing[54] focuses on vehicle routing optimization for medical waste management using deep reinforcement learning (Q-learning, DQN) combined with pathfinding algorithms. This is fundamentally different from Med-Copilot, which is an LLM-based coding assistant for biomedical data science tasks trained through interactive reinforcement learning in a sandbox environment.

---

5. Enhancing software development efficiency through ai-powered code generation

URL: View paper

Brief Assessment

AI-Powered Code Generation[58] focuses on general software development efficiency through AI-powered code generation techniques (rule-based systems, machine learning, neural networks, GANs, transformers) without any mention of biomedical applications, reinforcement learning training environments, or medical coding agents like Med-Copilot.

---

6. Reinforcement learning for proposing smoking cessation activities that build competencies: Combining two worldviews in a virtual coach

URL: View paper

Brief Assessment

Smoking Cessation RL[53] focuses on proposing smoking cessation activities using reinforcement learning to build competencies for quitting smoking, not on developing LLM-based coding assistants for biomedical data science tasks.

---

7. s3: You Don't Need That Much Data to Train a Search Agent via RL

URL: View paper

Brief Assessment

s3[59] focuses on training search agents for retrieval-augmented generation systems, not on developing coding assistants for biomedical data science tasks.

---

8. From llm reasoning to autonomous ai agents: A comprehensive review

URL: View paper

Brief Assessment

Autonomous AI Agents[51] is a survey paper reviewing benchmarks and frameworks across multiple domains. It does not present a specific coding agent for biomedical applications that would refute Med-Copilot's novelty.

---

9. Deep reinforcement-based conversational AI agent in healthcare system

URL: View paper

Brief Assessment

Conversational AI Healthcare[55] focuses on conversational AI agents for healthcare dialogue systems using deep learning and reinforcement learning, not on developing LLM-based coding assistants for biomedical data science tasks like Med-Copilot.

---

10. Reinforcement learning for clinical decision support in critical care: comprehensive review

URL: View paper

Brief Assessment

Clinical Decision RL[57] focuses on reinforcement learning for clinical decision support in critical care (medication optimization, drug dosing, intervention timing), not on developing LLM-based coding assistants for biomedical data science tasks.

Contribution 3: Unified execution environment with comprehensive benchmark

Description: The authors provide an integrated platform that combines executable environments, comprehensive benchmarking capabilities, and extensible training resources. This unified framework supports the development and evaluation of LLM-based coding assistants specifically designed for biomedical data science applications.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. AutoGEEval: A Multimodal and Automated Evaluation Framework for Geospatial Code Generation on GEE with Large Language Models

URL: View paper

Brief Assessment

AutoGEEval[75] focuses on geospatial code generation for Google Earth Engine, not biomedical data science applications. The domains and execution environments are fundamentally different.

---

2. Codearena: A collective evaluation platform for llm code generation

URL: View paper

Brief Assessment

Codearena[70] focuses on general LLM code generation evaluation across multiple programming languages and platforms, not specifically on biomedical data science applications with specialized medical domain requirements and EHR data handling.

---

3. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots

URL: View paper

Brief Assessment

Plot2code[69] focuses on evaluating multi-modal LLMs for generating plotting code from scientific figures, not on providing a unified training environment for biomedical data science coding assistants. The candidate addresses visualization code generation from images, while the original paper targets interactive training environments for biomedical reasoning tasks.

---

4. Evaluating language models for efficient code generation

URL: View paper

Brief Assessment

Efficient Code Generation[76] focuses on evaluating code efficiency through performance profiling and computational complexity metrics, not on providing an integrated platform for LLM-based coding assistants in biomedical data science applications.

---

5. User Centric Evaluation of Code Generation Tools

URL: View paper

Brief Assessment

User Centric Evaluation[78] focuses on evaluating LLM code generation usability through user-centric metrics (quality attributes, user experience) rather than providing a unified execution environment or training infrastructure for LLM-based coding assistants in biomedical data science.

---

6. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation

URL: View paper

Brief Assessment

ChatGPT Code Evaluation[73] focuses on evaluating code correctness through test case augmentation for general programming tasks, not on providing a unified platform for LLM-based coding assistants in biomedical data science applications.

---

7. DesignBench: A Comprehensive Benchmark for MLLM-based Front-end Code Generation

URL: View paper

Brief Assessment

DesignBench[74] focuses on front-end UI code generation across multiple frameworks (React, Vue, Angular), not biomedical data science coding assistants or executable biomedical reasoning environments.

---

8. Towards an understanding of large language models in software engineering tasks

URL: View paper

Brief Assessment

LLM Software Engineering[71] focuses on systematic literature review of LLMs in general software engineering tasks, not on providing an integrated platform with executable environments for biomedical data science coding assistants.

---

9. GeoJSEval: An Automated Evaluation Framework for Large Language Models on JavaScript-Based Geospatial Computation and Visualization Code Generation

URL: View paper

Brief Assessment

GeoJSEval[77] focuses specifically on JavaScript-based geospatial code generation with frontend libraries, not biomedical data science applications or general LLM-based coding assistants.

---

10. Opencodeinterpreter: Integrating code generation with execution and refinement

URL: View paper

Brief Assessment

Opencodeinterpreter[72] focuses on general code generation with execution and refinement capabilities, not specifically on biomedical data science benchmarking platforms or LLM-based coding assistants for medical applications.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] MedAgentGym: A Scalable Agentic Training Environment for Code-Centric Reasoning in Biomedical Data Science View paper
- [1] BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model View paper
- [2] Medagentgym: Training llm agents for code-based medical reasoning at scale View paper
- [3] Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records View paper
- [4] OntologyRAG: Better and faster biomedical code mapping with retrieval-augmented generation (RAG) leveraging ontology knowledge graphs and large language â¦ View paper
- [5] Agentic ai framework for end-to-end medical data inference View paper
- [6] Pandora: A Code-Driven Large Language Model Agent for Unified Reasoning Across Diverse Structured Knowledge View paper
- [7] Biomni: A general-purpose biomedical ai agent View paper
- [8] Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting â¦ View paper
- [9] Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases View paper
- [10] Learning to be a doctor: Searching for effective medical agent architectures View paper
- [11] GPT-Enhanced Hierarchical Deep Learning Model for Automated ICD Coding View paper
- [12] MACD: Multi-Agent Clinical Diagnosis with Self-Learned Knowledge for LLM View paper
- [13] From Scores to Steps: Diagnosing and Improving LLM Performance in Evidence-Based Medical Calculations View paper
- [14] Code-Driven Inductive Synthesis: Enhancing Reasoning Abilities of Large Language Models with Sequences View paper
- [15] EHRAgent: Code Empowers Large Language Models for Complex Tabular Reasoning on Electronic Health Records View paper

- [16] Evidence extraction for automated medical coding: preliminary evaluation View paper
- [17] Towards Efficient Medical Reasoning with Minimal Fine-Tuning Data View paper
- [18] Unlocking the Power of Large Language Models for Entity Alignment View paper
- [19] BioCoder: a benchmark for bioinformatics code generation with large language models. View paper
- [20] Exploring llm multi-agents for icd coding View paper
- [21] MedCalc-Eval and MedCalc-Env: Advancing Medical Calculation Capabilities of Large Language Models View paper
- [22] Enhanced BioGPT for Automated ICD-10 Medical Coding: Harnessing Domain-Tuned Large Language Models for Smarter Medical Coding View paper
- [23] On Code-Induced Reasoning in LLMs View paper
- [24] A Strategy for Anonymizing Free-Text Medical Reports Using LLM-Aix View paper
- [25] Verification is All You Need: Prompting Large Language Models for Zero-Shot Clinical Coding View paper
- [26] Automatic ICD Code Generation for Lymphoma Using Large Language Models View paper
- [27] Code Prompting Elicits Conditional Reasoning Abilities in Text+Code LLMs View paper
- [28] Pandora: Leveraging Code-driven Knowledge Transfer for Unified Structured Knowledge Reasoning View paper
- [29] Code-Driven LLM Agent for One-Shot Explanatory Visual Question Answering View paper
- [30] Biomedical reasoning in action: Multi-agent System for Auditable Biomedical Evidence Synthesis View paper
- [31] Enhancing Large Language Models with Neurosymbolic Reasoning for Multilingual Tasks View paper
- [32] 339P LLM-based reasoning framework for MedDRA adverse event coding: Toward scalable automation View paper
- [33] Code to Think, Think to Code: A Survey on Code-Enhanced Reasoning and Reasoning-Driven Code Intelligence in LLMs View paper
- [34] Multi-Agent LLM Reasoning for Clinical Procedure Sequencing from High-Granularity EHR Data View paper
- [35] Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting Approach. View paper
- [36] Evaluation and LLM-Guided Learning of ICD Coding Rationales View paper
- [37] ClinPath: A General-Purpose Knowledge Graph with LLM Reasoning For Understanding Clinical Interactions View paper
- [38] Vibe Coding in nephrology education: clinician-led, AI-assisted development of open-source interactive learning tools View paper
- [39] CodeGraph: Enhancing Graph Reasoning of LLMs with Code View paper
- [40] NEURO-GUARD: Neuro-Symbolic Generalization and Unbiased Adaptive Routing for Diagnostics -- Explainable Medical AI View paper
- [41] Chain of Thought Strategy for Smaller LLMs for Medical Reasoning. View paper
- [42] Hippocrates-o1: A Guideline-Aware, Orchestrated, Self-Refining Protocol for Specialty-Specific Clinical Reasoning View paper
- [43] From Guidelines to Code: Formalizing STOPP/START Criteria Using LLMs and RAG for Clinical Decision Support. View paper
- [44] LLMAgent4Bio: LLM Agents for Biological Intelligence Across Genomics, Proteomics, Spatial Biology, and Biomedicine View paper
- [45] Reasoning Large Language Models for Clinical Coding View paper
- [46] Code Execution as Grounded Supervision for LLM Reasoning View paper
- [47] MedAide: Information Fusion and Anatomy of Medical Intents via LLM-based Agent Collaboration View paper
- [48] SylloBio-NLI: Evaluating Large Language Models on Biomedical Syllogistic Reasoning View paper
- [49] Leveraging Chain of Thought for Automated Medical Coding View paper
- [50] Steering Large Language Models between Code Execution and Textual Reasoning View paper
- [51] From llm reasoning to autonomous ai agents: A comprehensive review View paper
- [52] Structured preference modeling for reinforcement learning-based fine-tuning of large models View paper
- [53] Reinforcement learning for proposing smoking cessation activities that build competencies: Combining two worldviews in a virtual coach View paper
- [54] Enhanced vehicle routing for medical waste management via hybrid deep reinforcement learning and optimization algorithms View paper
- [55] Deep reinforcement-based conversational AI agent in healthcare system View paper
- [56] MBuilder: A Multi-agent System for Automated Machine Learning in Medical Imaging View paper
- [57] Reinforcement learning for clinical decision support in critical care: comprehensive review View paper
- [58] Enhancing software development efficiency through ai-powered code generation View paper
- [59] s3: You Don't Need That Much Data to Train a Search Agent via RL View paper
- [60] Medagents: Large language models as collaborators for zero-shot medical reasoning View paper
- [61] Interactive computer-aided diagnosis on medical image using large language models View paper
- [62] Improving interactive diagnostic ability of a large language model agent through clinical experience learning View paper
- [63] Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning View paper
- [64] Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods â¦ View paper
- [65] â¦ patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study View paper
- [66] Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions View paper
- [67] A Proactive Agent Collaborative Framework for Zeroâ□Shot Multimodal Medical Reasoning View paper
- [68] Generative AI for medical education: Insights from a case study with medical students and an AI tutor for clinical reasoning View paper
- [69] Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots View paper
- [70] Codearena: A collective evaluation platform for llm code generation View paper
- [71] Towards an understanding of large language models in software engineering tasks View paper
- [72] Opencodeinterpreter: Integrating code generation with execution and refinement View paper
- [73] Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation View paper
- [74] DesignBench: A Comprehensive Benchmark for MLLM-based Front-end Code Generation View paper
- [75] AutoGEEval: A Multimodal and Automated Evaluation Framework for Geospatial Code Generation on GEE with Large Language Models View paper
- [76] Evaluating language models for efficient code generation View paper

- [77] GeoJSEval: An Automated Evaluation Framework for Large Language Models on JavaScript-Based Geospatial Computation and Visualization Code Generation View paper
- [78] User Centric Evaluation of Code Generation Tools View paper