Novelty Assessment Report

Paper: Unleashing Scientific Reasoning for Bio-experimental Protocol Generation via Structured Component-based Reward Mechanism
PDF URL: https://openreview.net/pdf?id=5BRMteyNOp
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

The foundation of reproducible science lies in protocols that are precise, logically ordered, and executable. The autonomous generation of these protocols through natural language queries could greatly improve the efficiency of the reproduction process. However, current leading large language models (LLMs) often generate incomplete or inconsistent protocols, limiting their utility. To address this limitation, we first introduce SciRecipe, a large-scale dataset of over 12K structured protocols spanning 27 biological subfields and encompassing both comprehension and problem-solving tasks. To further improve protocol generation, we propose the "Sketch-and-Fill" paradigm, which separates analysis, structuring, and expression to ensure each step is explicit and verifiable. Complementing this, the structured component-based reward mechanism evaluates step granularity, action order, and semantic fidelity, aligning model optimization with experimental reliability. Building on these components, we develop Thoth, trained through a staged Knowledge-to-Action process that progresses from knowledge acquisition to operational reasoning and ultimately to robust, executable protocol generation. Across multiple benchmarks, Thoth consistently surpasses both proprietary and open-source LLMs, achieving significant improvements in step alignment, logical sequencing, and semantic accuracy. Our approach paves the way for reliable scientific assistants that bridge knowledge with experimental execution.

Core Task Landscape

This paper addresses: **Autonomous Generation of Executable Biological Experimental Protocols**

A total of **29 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

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

- **AI-Driven Protocol Generation and Translation**
- **Autonomous Laboratory Systems and Platforms**
- **Experimental Design Optimization and Decision-Making**
- **Data Infrastructure and Knowledge Management**
- **Cross-Cutting Reviews and Methodological Surveys**

Complete Taxonomy Tree

- Autonomous Generation of Executable Biological Experimental Protocols Survey Taxonomy
- AI-Driven Protocol Generation and Translation
    - LLM-Based Protocol Generation from Natural Language ★ (3 papers)
    - [0] Unleashing Scientific Reasoning for Bio-experimental Protocol Generation via Structured Component-based Reward Mechanism (Anon et al., 2026) View paper
    - [19] BioPlanner: automatic evaluation of LLMs on protocol planning in biology (Odhran Oâ□□Donoghue, 2023) View paper
    - [22] Automating Biomedical Discovery with AI Agents (Qu, 2025) View paper
    - Hardware-Specific Robotic Scripting (3 papers)
    - [4] Autonomous liquid-handling robotics scripting through large language models enables accessible and safe protein engineering workflows (Gao Yuan, 2025) View paper
    - [8] LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation (Inagaki Takashi, 2023) View paper
    - [9] Large language model agents as experimental orchestrators in data-driven enzyme engineering (Yuan. Gao, 2025) View paper
    - Protocol Formalization and Standardization Languages (3 papers)
    - [13] Reproducibility in automated chemistry laboratories using computer science abstractions (Richard B. Canty, 2024) View paper
    - [18] A Language for Modeling and Optimizing Experimental Biological Protocols (Luca Cardelli, 2021) View paper
    - [25] Biocoder: A programming language for standardizing and automating biology protocols (Vaishnavi Ananthanarayanan, 2010) View paper
    - Expert-Level Protocol Translation Systems (2 papers)
    - [12] Hierarchically Encapsulated Representation for Protocol Design in Self-Driving Labs (Shi, 2025) View paper
    - [16] Expert-level protocol translation for self-driving labs (Zhangqian Bi, 2024) View paper
- Autonomous Laboratory Systems and Platforms
    - General-Purpose Autonomous Research Agents (3 papers)
    - [1] Biomni: A general-purpose biomedical ai agent (Kexin Huang, 2025) View paper
    - [3] BioLab: End-to-end autonomous life sciences research with multi-agents system integrating biological foundation models (Ruofan Jin, 2025) View paper
    - [21] BioMARS: A Multi-Agent Robotic System for Autonomous Biological Experiments (Qiu Yibo, 2025) View paper
    - Domain-Specialized Autonomous Agents (2 papers)
    - [7] SpatialAgent: An autonomous AI agent for spatial biology (Hanchen Wang, 2025) View paper

- [10] AI Agents in Drug Discovery (Srijit Seal, 2025) View paper
- Self-Driving Laboratory Infrastructures (4 papers)
- [2] Autonomous 'self-driving'laboratories: a review of technology and policy implications (A. Tobias, 2025) View paper
- [11] Development of the autonomous lab system to support biotechnology research (Keiji Fushimi, 2025) View paper
- [27] Towards robot scientists for autonomous scientific discovery (Andrew C. Sparkes, 2010) View paper
- [28] An AI-native experimental laboratory for autonomous biomolecular engineering (Wu Mingyu, 2025) View paper
- Modular and Reconfigurable Autonomous Platforms (2 papers)
- [20] Modular Autonomous Experimentation for Biological Applications (Diego Oyarzun Dinamarca, 2025) View paper
- [26] GeneForge - An AI-Native Platform for the Autonomous Execution of Biological Workflows (Manuel MenÃ©ndez, 2025) View paper
- Task-Specific Autonomous Robotic Systems (1 papers)
- [24] A machine learning-driven robotic system for autonomous nucleic acid extraction and library preparation (Jun Lu, 2025) View paper

- Experimental Design Optimization and Decision-Making
  - Reinforcement Learning for Experimental Design (2 papers)
  - [14] Deep reinforcement learning for optimal experimental design in biology (Treloar, 2022) View paper
  - [29] Self-adaptive scoutingâautonomous experimentation for systems biology (Naoki Matsumaru, 2004) View paper
  - Model-Based Experimental Optimization (1 papers)
  - [15] Autonomous Model-Based Experimental Design for Rapid Reaction Development (Sebastian Knoll, 2022) View paper
- Data Infrastructure and Knowledge Management
  - Databases and Ontologies for Autonomous Systems (1 papers)
  - [17] Genesis-DB: a database for autonomous laboratory systems (Gabriel K Reder, 2023) View paper
  - Computational Design Automation Tools (1 papers)
  - [23] EMMA-CAD: design automation for synthetic mammalian constructs (Yisha Luo, 2022) View paper
- Cross-Cutting Reviews and Methodological Surveys (2 papers)
  - [5] Data-Driven Design and Autonomous Experimentation in Soft and Biological Materials Engineering. (Andrew L. Ferguson, 2022) View paper
  - [6] Autonomous experimentation for molecular discovery applications (Canty, 2024) View paper

Narrative

Core task: autonomous generation of executable biological experimental protocols. The field encompasses efforts to translate high-level scientific goals into machine-readable instructions that robotic systems can execute. The taxonomy reveals five main branches: AI-Driven Protocol Generation and Translation focuses on leveraging large language models and structured representations to convert natural language descriptions into formal protocols, with works like BioPlanner[19] and Automating Biomedical Discovery[22] exemplifying LLM-based approaches. Autonomous Laboratory Systems and Platforms addresses the physical infrastructure and integration challenges, including liquid handling automation (Autonomous Liquid Handling[4]) and end-to-end self-driving labs (Self-Driving Labs Review[2], BioLab[3]). Experimental Design Optimization and Decision-Making explores adaptive strategies for selecting experiments, often using reinforcement learning or Bayesian methods (Deep RL Experimental Design[14], Autonomous Reaction Development[15]). Data Infrastructure and Knowledge Management tackles the challenge of organizing protocols and experimental metadata into reusable repositories (Genesis-DB[17], Biological Protocol Language[18]). Cross-Cutting Reviews and Methodological Surveys provide broader perspectives on autonomous discovery (Autonomous Molecular Discovery[6], AI Drug Discovery[10]).

Recent activity highlights a tension between end-to-end automation and modular, human-in-the-loop designs. Some systems aim for fully autonomous cycles (AI-Native Biomolecular Lab[28], Modular Autonomous Experimentation[20]), while others emphasize expert-guided translation (Expert Protocol Translation[16]) or hierarchical decomposition (Hierarchical Protocol Design[12]). Structured Component Reward[0] sits within the LLM-Based Protocol Generation cluster alongside BioPlanner[19] and Automating Biomedical Discovery[22], but distinguishes itself by introducing a reward-based framework to refine protocol generation, addressing the challenge of ensuring executability and correctness. Compared to BioPlanner[19], which focuses on planning from natural language, and Automating Biomedical Discovery[22], which surveys broader automation strategies, Structured Component Reward[0] emphasizes learning from feedback to improve protocol quality, bridging the gap between initial generation and reliable execution.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. BioPlanner: automatic evaluation of LLMs on protocol planning in biology

Authors: Odhran OâDonoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Ghareeb, et al. (8 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

The ability to automatically generate accurate protocols for scientific experiments would represent a major step towards the automation of science. Large Language Models (LLMs) have impressive capabilities on a wide range of tasks, such as question answering and the generation of coherent text and code. However, LLMs can struggle with multi-step problems and long-term planning, which are crucial for designing scientific experiments. Moreover, evaluation of the accuracy of scientific protocols is...

Relationship Analysis

Both papers belong to the LLM-Based Protocol Generation from Natural Language category, employing large language models to generate executable biological protocols from natural language inputs. They overlap in addressing the challenge of converting natural language descriptions into structured, executable experimental protocols using LLMs (GPT-3.5/GPT-4 in BioPlanner, and a custom model Thoth in the original paper). The key difference is that BioPlanner focuses on automatic evaluation through pseudocode conversion and benchmarking existing LLMs, while the original paper introduces a novel training framework (Sketch-and-Fill paradigm, SCORE mechanism, and Knowledge-to-Action learning) to develop a specialized protocol-generation model with improved reasoning capabilities.

---

2. Automating Biomedical Discovery with AI Agents

Authors: Y Qu | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ how autonomous AI systems can accelerate biological â¦ initial promise in generating biological protocols, as demonstrated by â¦ To automate biological experiment design and analysis, we â¦

Relationship Analysis

Both papers belong to the LLM-Based Protocol Generation from Natural Language category, focusing on using large language models to generate executable biological protocols from natural language instructions. They overlap in addressing the core challenge of automating protocol generation through AI-driven approaches. However, the original paper (Thoth) introduces a specific "Sketch-and-Fill" paradigm with structured component-based rewards (SCORE) for training and evaluation, while the candidate paper appears to take a broader approach to automating biomedical discovery using AI agents, potentially encompassing experiment design and analysis beyond just protocol generation.

Contributions Analysis

Overall novelty summary. The paper introduces Thoth, a system for generating executable biological protocols from natural language queries, alongside the SciRecipe dataset spanning 27 biological subfields. Within the taxonomy, it resides in the 'LLM-Based Protocol Generation from Natural Language' leaf, which contains only three papers total. This is a relatively sparse research direction compared to broader branches like 'Autonomous Laboratory Systems' (11 papers) or 'Self-Driving Laboratory Infrastructures' (4 papers). The two sibling papers—BioPlanner and a survey on automating biomedical discovery—focus on planning and broader automation strategies, suggesting limited direct competition in reward-driven protocol refinement.

The taxonomy reveals that neighboring leaves address complementary challenges: 'Hardware-Specific Robotic Scripting' targets platform-specific execution (3 papers), while 'Protocol Formalization and Standardization Languages' emphasizes machine-readable representations (3 papers). The paper's Sketch-and-Fill paradigm bridges these concerns by decomposing protocol generation into analysis, structuring, and expression phases. Unlike the 'Expert-Level Protocol Translation Systems' branch (2 papers), which assumes human-readable inputs, this work starts from natural language queries. The structured reward mechanism connects to 'Experimental Design Optimization' themes but remains distinct by focusing on protocol correctness rather than parameter inference or hypothesis testing.

Among 29 candidates examined, the Sketch-and-Fill paradigm shows one refutable candidate from 10 examined, indicating some prior work in decomposed reasoning approaches. The SciRecipe dataset (0 refutations from 10 candidates) and SCORE mechanism (0 from 9 candidates) appear more novel within this limited search scope. The statistics suggest that while the dataset and reward framework may represent fresh contributions, the staged reasoning approach has at least one overlapping precedent among the top-30 semantic matches. The scale of examination—29 papers across the entire field—means these findings reflect proximity within a focused literature sample, not exhaustive coverage.

Given the sparse population of the LLM-based protocol generation leaf and the limited search scope, the work appears to occupy a relatively open niche. The combination of a large-scale dataset, decomposed reasoning, and component-based rewards distinguishes it from sibling papers focused on planning or surveys. However, the single refutation for Sketch-and-Fill warrants attention to how the staged decomposition differs from prior hierarchical or modular approaches. The analysis covers top-30 semantic matches and does not claim completeness across all protocol generation literature.

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

Contribution 1: SciRecipe dataset for protocol generation

Description: The authors curate SciRecipe, a large-scale multi-task dataset containing over 12,000 structured experimental protocols across 27 biological subfields. The dataset covers both Protocol-Comprehension tasks (overview and specific analysis) and Problem-Solving tasks (retrieval, planning, troubleshooting, constraint, scaling, and safety), designed to serve as a foundation for training and evaluating protocol generation systems.

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. BioInformatics Agent (BIA): Unleashing the Power of Large Language Models to Reshape Bioinformatics Workflow

URL: View paper

Brief Assessment

BioInformatics Agent[45] focuses on autonomous bioinformatic analysis of single-cell RNA sequencing data through an LLM-based agent system, not on creating datasets for structured experimental protocol generation across biological subfields.

---

2. Deep learning for biology

URL: View paper

Brief Assessment

Deep Learning Biology[42] discusses applications of deep learning to biological data analysis and image processing, but does not present a dataset for structured experimental protocol generation or protocol-specific tasks.

---

3. BioProBench: Comprehensive Dataset and Benchmark in Biological Protocol Understanding and Reasoning

URL: View paper

Brief Assessment

BioProBench[39] focuses on protocol understanding and reasoning tasks (QA, ordering, error correction) rather than protocol generation training datasets. While both involve biological protocols, BioProBench's 556k instances serve benchmarking purposes across comprehension tasks, whereas SciRecipe targets training for generation systems with structured experimental protocols.

---

4. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals

URL: View paper

Brief Assessment

Sponges Sister Group[44] is a phylogenomics study focused on animal evolutionary relationships using genomic datasets, not on experimental protocol generation or structured biological datasets for training language models.

---

5. BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences

URL: View paper

Brief Assessment

BioAutoMATED[41] focuses on automated machine learning for biological sequence analysis (DNA, RNA, peptides, glycans) to predict function from sequence, not on generating structured experimental protocols or protocol-comprehension tasks as in SciRecipe.

---

6. Lock3dface: A large-scale database of low-cost kinect 3d faces

URL: View paper

Brief Assessment

Lock3dface[46] is a 3D face recognition database, not a dataset for experimental protocol generation in biology. The domains are entirely distinct.

---

7. AI Agents in Drug Discovery

URL: View paper

Brief Assessment

AI Agents in Drug Discovery[10] focuses on agentic AI architectures and their applications across drug discovery workflows, including literature synthesis, toxicity prediction, and automated protocol generation. While it mentions automated protocol generation as one application area, it does not present a large-scale dataset of structured experimental protocols comparable to SciRecipe's 12,000+ protocols across 27 biological subfields with multi-task coverage.

---

8. SciKnowEval: A Comprehensive Dataset for Evaluating Scientific Knowledge of Large Language Models

URL: View paper

Brief Assessment

SciKnowEval[40] focuses on evaluating scientific knowledge across five progressive levels (memory, comprehension, reasoning, discernment, application) using multi-choice questions and other formats, not on generating structured experimental protocols for biology.

---

9. BioPlanner: automatic evaluation of LLMs on protocol planning in biology

URL: View paper

Brief Assessment

BioPlanner[19] focuses on evaluating LLMs on protocol planning through pseudocode conversion rather than providing a large-scale multi-task dataset for protocol generation. The candidate introduces the BioProt dataset (100 protocols) for evaluation purposes, while the original paper presents SciRecipe (12,000+ protocols) as a comprehensive training and evaluation resource spanning 27 biological subfields with both comprehension and problem-solving tasks.

---

10. Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system

URL: View paper

Brief Assessment

Autonomous Carotid Ultrasonography[43] focuses on robotic ultrasound scanning with imitation learning from expert demonstrations (247k image-action pairs), not on structured experimental protocol generation across biological subfields. The datasets serve fundamentally different domains and tasks.

Contribution 2: Sketch-and-Fill reasoning paradigm

Description: The authors introduce a structured reasoning framework that decomposes protocol generation into three stages: reasoning (think), structuring key information into machine-readable JSON format (key), and expressing steps in natural language (orc). This paradigm ensures that each experimental step is explicit, verifiable, and maintains one-to-one correspondence between structured and natural language representations.

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. Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning

URL: View paper

Brief Assessment

Legal Logical Structures[49] focuses on legal question-answering with fact-rule chain prediction and semantic retrieval, not on experimental protocol generation with structured reasoning stages.

---

2. StructGPT: A General Framework for Large Language Model to Reason over Structured Data

URL: View paper

Brief Assessment

StructGPT[47] focuses on reading structured data (KGs, tables, DBs) via specialized interfaces and iterative reasoning, not on decomposing protocol generation into think-key-orc stages for experimental procedures.

---

3. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning

URL: View paper

Prior Art Analysis

Selection-Inference[48] demonstrates that a similar structured reasoning framework separating analysis, structuring, and expression was proposed earlier. The candidate paper introduces a 'selection-inference (si) framework' that decomposes reasoning into two modular stages: 1) selection, which chooses relevant information, and 2) inference, which produces new facts based on the selected information. This mirrors the original paper's three-stage decomposition (think, key, orc) where reasoning is separated from structuring and expression. Both frameworks explicitly separate the reasoning process into distinct, verifiable steps with one-to-one correspondence between structured and natural language representations.

Evidence

Evidence 1 - **Rationale**: Both papers propose structured reasoning frameworks that decompose complex tasks into modular, interpretable steps. The original's 'sketch-and-fill' and the candidate's 'selection-inference' both separate reasoning into distinct phases with explicit correspondence between components. - **Original**: we propose the "sketch-and-fill" paradigm, which formulates protocol generation as a structured reasoning process: each step is decomposed into essential components and expressed in natural language with explicit correspondence, ensuring logical coherence and experimental verifiability. - **Candidate**: we propose a selection-inference (si) framework that exploits pre-trained llms as general processing modules, and alternates between selection and inference to generate a seriesof interpretable, casualreasoningsteps leadingtothefinal answer.

Evidence 2 - **Rationale**: Both papers emphasize maintaining one-to-one correspondence between structured representations and natural language outputs, ensuring interpretability and verifiability of each reasoning step. - **Original**: in the "fill" phase, represented by , the elements of are rendered into human-readable natural language. a strict one-to-one correspondence in step count and semantics is enforced, ensuring no information is added or omitted, with the focus solely on readability. - **Candidate**: this has the benefit of producing an interpretable trace of intermediate computations, in contrast to the "black-box" computations common to end-to-end deep learning approaches. importantly, the modularity of neurosymbolic methods allows them to generalise to significantly harder problems that require...

4. Structured prompting and feedback-guided reasoning with llms for data interpretation

URL: View paper

Brief Assessment

Structured Prompting Feedback[53] focuses on structured prompting for data interpretation tasks with feedback-driven refinement, not on experimental protocol generation with explicit three-stage decomposition (think-key-orc) and one-to-one correspondence between structured and natural language representations.

5. Generating Structured Plan Representation of Procedures with LLMs

URL: View paper

Brief Assessment

Structured Plan Generation[51] focuses on transforming SOPs into DAG-based representations for workflow automation, not on a three-stage reasoning framework (think-key-orc) for protocol generation with explicit verification steps.

6. Structured path guidance for logical coherence in large language model generation

URL: View paper

Brief Assessment

Structured Path Guidance[50] focuses on guiding large language models through preset structural paths for general text generation tasks, not on decomposing biomedical protocol generation into explicit reasoning-structuring-expression stages with JSON-based intermediate representations.

7. A Retrieve-and-Edit Framework for Predicting Structured Outputs

URL: View paper

Brief Assessment

Retrieve-and-Edit Framework[56] focuses on retrieving training examples and editing them for structured output prediction (e.g., code generation), not on decomposing protocol generation into reasoning-structuring-expression stages as in the original paper's Sketch-and-Fill paradigm.

8. HiRA: A Hierarchical Reasoning Framework for Decoupled Planning and Execution in Deep Search

URL: View paper

Brief Assessment

HiRA[52] focuses on hierarchical agent coordination for web search tasks, not on structured protocol generation with explicit reasoning-structuring-expression stages as in the original paper's Sketch-and-Fill paradigm.

9. RATT: A Thought Structure for Coherent and Correct LLM Reasoning

URL: View paper

Brief Assessment

RATT[54] focuses on retrieval-augmented thought tree structures for balancing factual accuracy and logical reasoning in LLMs, not on structured protocol generation with explicit separation of analysis, structuring (JSON format), and natural language expression stages as described in the original contribution.

10. Continuum-interaction-driven intelligence: Human-aligned neural architecture via crystallized reasoning and fluid generation

URL: View paper

Brief Assessment

Continuum-Interaction Intelligence[55] focuses on dual-channel cognitive architectures integrating probabilistic generation with procedural reasoning chains, not on structured protocol generation with JSON-based intermediate representations for experimental reproducibility.

Contribution 3: SCORE mechanism for protocol evaluation and training

Description: The authors propose the Structured COmponent-based REward (SCORE) mechanism, which provides both a training reward signal and evaluation framework. SCORE jointly measures three dimensions: step granularity (controlling scale and avoiding redundancy), action ordering (ensuring logically consistent sequences), and semantic fidelity (verifying alignment of actions, objects, and parameters), moving beyond conventional text-based metrics to assess experimental executability.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. GroundedPRM: Tree-Guided and Fidelity-Aware Process Reward Modeling for Step-Level Reasoning

URL: View paper

Brief Assessment

GroundedPRM[38] focuses on process reward modeling for step-level reasoning in mathematical problem solving, not on protocol generation or experimental executability evaluation. The candidate addresses multi-step mathematical reasoning with tree-guided search and external tool verification, whereas the original contribution evaluates experimental protocols through step granularity, action ordering, and semantic fidelity.

2. Combining semantic guidance and deep reinforcement learning for generating human level paintings

URL: View paper

Brief Assessment

Semantic Guidance Painting[34] focuses on stroke-based rendering for visual art generation using reinforcement learning, not on evaluating experimental protocols or scientific workflows. The paper addresses painting agents and semantic segmentation for artistic tasks, which is fundamentally different from the SCORE mechanism's domain of bioexperimental protocol generation and evaluation.

3. RISK: A Framework for GUI Agents in E-commerce Risk Management

URL: View paper

Brief Assessment

RISK[37] focuses on GUI agents for e-commerce risk management using reinforcement fine-tuning with domain-specific reward functions (format, stepwise accuracy, process reweight, level reweight). The original paper's SCORE mechanism evaluates bioexperimental protocols through step granularity, action ordering, and semantic fidelity. These are distinct application domains with different evaluation targets.

4. Cogito Ergo Summ: Abstractive Summarization of Biomedical Papers via Semantic Parsing Graphs and Consistency Rewards

URL: View paper

Brief Assessment

Cogito Ergo Summ[33] focuses on biomedical text summarization using semantic parsing graphs and consistency rewards based on AMR graph matching, not on protocol generation with component-based evaluation of step granularity, action ordering, and semantic fidelity as described in the original contribution.

5. Optimizing large language models through highly dense reward structures and recursive thought process using monte carlo tree search

URL: View paper

Brief Assessment

Dense Reward MCTS[31] focuses on dense reward structures for general LLM optimization using MCTS, not on component-based evaluation of experimental protocols across step granularity, action ordering, and semantic fidelity dimensions.

6. Robo-Dopamine: General Process Reward Modeling for High-Precision Robotic Manipulation

URL: View paper

Brief Assessment

Robo-Dopamine[36] focuses on robotic manipulation reward modeling using multi-view visual observations and hop-based progress estimation, not on evaluating experimental protocol generation with step granularity, action ordering, and semantic fidelity metrics.

7. Curriculum guided reinforcement learning for efficient multi hop retrieval augmented generation

URL: View paper

Brief Assessment

Curriculum Guided RAG[32] focuses on multi-hop retrieval-augmented generation with a seven-factor reward vector for query rewriting, not experimental protocol generation with step granularity, action ordering, and semantic fidelity evaluation.

8. Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization

URL: View paper

Brief Assessment

Focus Discriminative Regions[35] addresses fine-grained visual categorization using deep reinforcement learning for region localization, not protocol generation or component-based reward mechanisms for experimental procedures.

9. Fine-grained reward optimization for machine translation using error severity mappings

URL: View paper

Brief Assessment

Error Severity Mappings[30] focuses on machine translation reward optimization using token-level error severity, not experimental protocol generation or component-based evaluation of step granularity, action ordering, and semantic fidelity in scientific workflows.

Appendix: Text Similarity Detection

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

References

- [0] Unleashing Scientific Reasoning for Bio-experimental Protocol Generation via Structured Component-based Reward Mechanism View paper
- [1] Biomni: A general-purpose biomedical ai agent View paper
- [2] Autonomous 'self-driving'laboratories: a review of technology and policy implications View paper
- [3] BioLab: End-to-end autonomous life sciences research with multi-agents system integrating biological foundation models View paper
- [4] Autonomous liquid-handling robotics scripting through large language models enables accessible and safe protein engineering workflows View paper
- [5] Data-Driven Design and Autonomous Experimentation in Soft and Biological Materials Engineering. View paper
- [6] Autonomous experimentation for molecular discovery applications View paper
- [7] SpatialAgent: An autonomous AI agent for spatial biology View paper
- [8] LLMs can generate robotic scripts from goal-oriented instructions in biological laboratory automation View paper
- [9] Large language model agents as experimental orchestrators in data-driven enzyme engineering View paper
- [10] AI Agents in Drug Discovery View paper
- [11] Development of the autonomous lab system to support biotechnology research View paper
- [12] Hierarchically Encapsulated Representation for Protocol Design in Self-Driving Labs View paper
- [13] Reproducibility in automated chemistry laboratories using computer science abstractions View paper
- [14] Deep reinforcement learning for optimal experimental design in biology View paper
- [15] Autonomous Model-Based Experimental Design for Rapid Reaction Development View paper
- [16] Expert-level protocol translation for self-driving labs View paper
- [17] Genesis-DB: a database for autonomous laboratory systems View paper
- [18] A Language for Modeling and Optimizing Experimental Biological Protocols View paper
- [19] BioPlanner: automatic evaluation of LLMs on protocol planning in biology View paper
- [20] Modular Autonomous Experimentation for Biological Applications View paper
- [21] BioMARS: A Multi-Agent Robotic System for Autonomous Biological Experiments View paper
- [22] Automating Biomedical Discovery with AI Agents View paper
- [23] EMMA-CAD: design automation for synthetic mammalian constructs View paper

- [24] A machine learning-driven robotic system for autonomous nucleic acid extraction and library preparation View paper
- [25] Biocoder: A programming language for standardizing and automating biology protocols View paper
- [26] GeneForge - An AI-Native Platform for the Autonomous Execution of Biological Workflows View paper
- [27] Towards robot scientists for autonomous scientific discovery View paper
- [28] An AI-native experimental laboratory for autonomous biomolecular engineering View paper
- [29] Self-adaptive scoutingâautonomous experimentation for systems biology View paper
- [30] Fine-grained reward optimization for machine translation using error severity mappings View paper
- [31] Optimizing large language models through highly dense reward structures and recursive thought process using monte carlo tree search View paper
- [32] Curriculum guided reinforcement learning for efficient multi hop retrieval augmented generation View paper
- [33] Cogito Ergo Summ: Abstractive Summarization of Biomedical Papers via Semantic Parsing Graphs and Consistency Rewards View paper
- [34] Combining semantic guidance and deep reinforcement learning for generating human level paintings View paper
- [35] Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization View paper
- [36] Robo-Dopamine: General Process Reward Modeling for High-Precision Robotic Manipulation View paper
- [37] RISK: A Framework for GUI Agents in E-commerce Risk Management View paper
- [38] GroundedPRM: Tree-Guided and Fidelity-Aware Process Reward Modeling for Step-Level Reasoning View paper
- [39] BioProBench: Comprehensive Dataset and Benchmark in Biological Protocol Understanding and Reasoning View paper
- [40] SciKnowEval: A Comprehensive Dataset for Evaluating Scientific Knowledge of Large Language Models View paper
- [41] BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences View paper
- [42] Deep learning for biology View paper
- [43] Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system View paper
- [44] A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals View paper
- [45] BioInformatics Agent (BIA): Unleashing the Power of Large Language Models to Reshape Bioinformatics Workflow View paper
- [46] Lock3dface: A large-scale database of low-cost kinect 3d faces View paper
- [47] StructGPT: A General Framework for Large Language Model to Reason over Structured Data View paper
- [48] Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning View paper
- [49] Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning View paper
- [50] Structured path guidance for logical coherence in large language model generation View paper
- [51] Generating Structured Plan Representation of Procedures with LLMs View paper
- [52] HiRA: A Hierarchical Reasoning Framework for Decoupled Planning and Execution in Deep Search View paper
- [53] Structured prompting and feedback-guided reasoning with llms for data interpretation View paper
- [54] RATT: A Thought Structure for Coherent and Correct LLM Reasoning View paper
- [55] Continuum-interaction-driven intelligence: Human-aligned neural architecture via crystallized reasoning and fluid generation View paper
- [56] A Retrieve-and-Edit Framework for Predicting Structured Outputs View paper