

LLM Agents Making Agent Tools

Georg Wölflein^{1,2,3,†} Dyke Ferber^{3,4,‡} Daniel Truhn⁵
Ognjen Arandjelović² Jakob N. Kather^{1,4,6}

¹EKFZ for Digital Health, TU Dresden ²University of St Andrews

³Synagen AI ⁴NCT, Heidelberg University Hospital

⁵University Hospital Aachen ⁶University Hospital Dresden

Correspondence: georg@woelflein.de

Abstract

Tool use has turned large language models (LLMs) into powerful agents that can perform complex multi-step tasks by dynamically utilising external software components. However, these tools must be implemented in advance by human developers, hindering the applicability of LLM agents in domains demanding large numbers of highly specialised tools, like in life sciences and medicine. Motivated by the growing trend of scientific studies accompanied by public code repositories, we propose TOOLMAKER, an agentic framework that autonomously transforms papers with code into LLM-compatible tools. Given a GitHub URL and short task description, TOOLMAKER autonomously installs dependencies and generates code to perform the task, using a closed-loop self-correction mechanism for debugging. To evaluate our approach, we introduce a benchmark comprising 15 complex computational tasks spanning various domains with over 100 unit tests to assess correctness and robustness. Our method correctly implements 80% of the tasks, substantially outperforming current state-of-the-art software engineering agents. TOOLMAKER therefore is a step towards fully autonomous agent-based scientific workflows¹.

1 Introduction

Scientific discovery is the foundation for innovation and progress. Traditionally, the underlying research processes that guarantee progress have been entirely reliant on human expertise, involving the formulation of ideas and hypotheses, the collection of information and analysis of data, the planning and execution of experiments, and iterative refinement to arrive at a solution. With the recent development of autonomous agents that employ

[†]Work done while at EKFZ for Digital Health, TU Dresden and University of St Andrews. [‡]Work done while at EKFZ for Digital Health, TU Dresden and NCT Heidelberg.

¹Our code and benchmark are publicly available at <https://github.com/KatherLab/ToolMaker>.

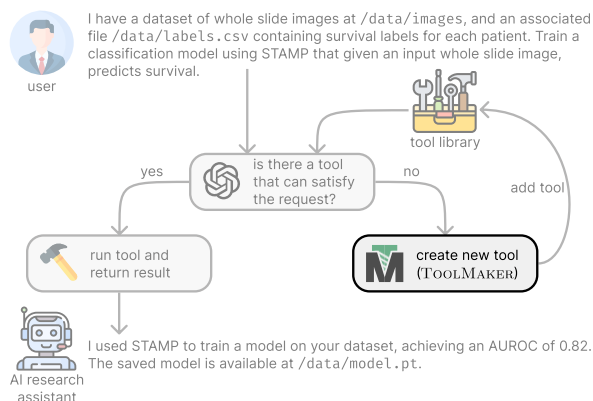


Figure 1: We envision a future where agents possess dynamic toolsets that can be expanded at runtime. *Tool creation*, studied here, is a crucial step towards this goal.

LLMs to perform tasks through multi-step reasoning and planning, and by utilising tools (external pieces of software that the model can execute), we are at the cusp of a paradigm shift where artificial intelligence (AI) can assist throughout entire research projects as a *virtual scientist* (Fig. 1), rather than being limited to addressing narrowly and *a priori* defined problems.

Although LLM agents have shown success for *specific* tasks in domains such as software engineering (Wang et al., 2024; Yang et al., 2024), healthcare (Ferber et al., 2024; Kim et al., 2024), law (Li et al., 2024), and scientific research (Swanson et al., 2024; Gao et al., 2024; Schmidgall et al., 2025), they struggle to generalise to broader classes of tasks. This limitation arises from their reliance on tools that must be explicitly designed, implemented, and integrated by human developers – often requiring extensive technical expertise – before deployment (Ferber et al., 2024; Jimenez et al., 2024). While AI assistants can support this process, current systems still depend heavily on manual intervention to ensure compatibility and functionality.

To address this, some agentic frameworks have been designed that autonomously craft their own tools (Cai et al., 2024; Yuan et al., 2024; Qian et al.,

2023). However, because these methods build each tool from scratch, they inevitably produce simple, narrowly scoped tools tailored to single-dimensional problems – an approach ill-suited to the complexity of real-world research problems.

In fact, in critical fields such as healthcare, data necessary to build tools from scratch is often inaccessible due to privacy restrictions, preventing agents from using it to build their own solutions. Moreover, the complexity of modern scientific tools has increased substantially in terms of computational requirements, data demands, and amount of code involved. Lastly, deploying tools in high-stakes applications demands rigorous validation, testing, and quality assurance – standards that current agent systems cannot realistically meet if required to develop such tools entirely from scratch.

Encouragingly, a growing emphasis on reproducibility within the scientific community has led to an increase in publicly released code accompanying research papers (Zhou et al., 2024). Consequently, a vast array of potential tools now exist as standalone solutions. However, many researchers in fields like healthcare, biology, drug development, R&D are unable to effectively use them due to the technical skills required for their deployment.

Instead of building tools entirely from scratch, we ask the following question: *Can LLM agents autonomously download, integrate, and execute complex, existing tools to empower researchers with minimal technical expertise in the future?* Towards this goal, we propose TOOLMAKER, an agentic framework that autonomously generates LLM-compatible tools from scientific papers and their associated code repositories, bypassing the need for human intermediaries to manually set up, install, and adapt them to fit the requirements of their applications. Given a task description, a scientific paper, and its associated code repository, TOOLMAKER generates an executable tool that enables LLMs to perform the task (see Fig. 2).

To evaluate TOOLMAKER, we introduce TM-BENCH, a benchmark comprising 15 diverse tasks across various medical disciplines (pathology, radiology, omics), as well as non-medical fields, e.g. LLMs and 3D vision. Unlike existing benchmarks (Jimenez et al., 2024; Zhuo et al., 2024; Jain et al., 2024) which assume pre-installed dependencies for function implementation, TOOLMAKER operates in a fully open-ended environment. Tasks in our benchmark encompass the entire workflow: downloading resources, managing and resolving dependency issues, reading through large code-

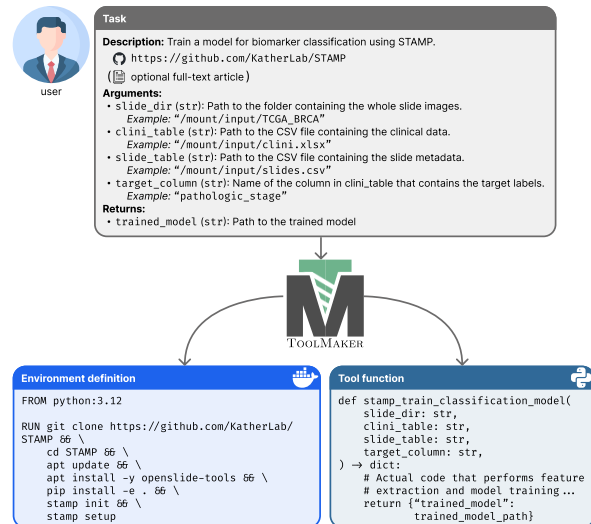


Figure 2: Given a task description, a scientific paper, a link to the associated code repository, and an example of the tool invocation, TOOLMAKER creates (i) a Docker container in which the tool can be executed, (ii) a Python function that performs the task.

bases, and implementing, testing, and debugging code. TM-BENCH includes over 100 unit tests to objectively assess the generated tools’ correctness.

2 Related work

Agents In addition to demonstrating impressive capabilities in generating human-like text, LLMs such as ChatGPT (Ouyang et al., 2022), Claude (Anthropic, 2024), Gemini (Gemini Team, 2024) and Llama (Llama Team, 2024), on their own, have shown strong potential in question answering and reasoning on problems in natural science related fields, like math (Shao et al., 2024), chemistry (Bran et al., 2024) and healthcare (Singhal et al., 2023). However, LLMs often struggle solving more complex problems directly, especially in situations that require intermediate results from multiple steps (Valmeekam et al., 2023). To address this, LLM agents have been developed which enhance an LLM’s capabilities by integrating external tools (Schick et al., 2023).

In software engineering, a number of agentic and workflow-based systems have been proposed for solving GitHub issues (Wang et al., 2024; Yang et al., 2024; Xia et al., 2024), as well as developing entire software projects (Qian et al., 2024; Nguyen et al., 2024; Hong et al., 2024). Among these, OpenHands (Wang et al., 2024) achieves state-of-the-art performance on SWE-Bench (Jimenez et al., 2024), a benchmark for solving GitHub issues.

Medical LLM agents have been developed for clinical decision-making and diagnostics,

such as building risk calculators from publications (Jin et al., 2024), oncology agents that consult guidelines and imaging tools (Ferber et al., 2024), and multi-agent systems that enable collaboration across clinicians, patients, and hospitals (Kim et al., 2024; Li et al., 2025). Beyond clinical use, agents have been proposed for bioinformatics tasks like data extraction, pipeline execution, and hypothesis testing (Ding et al., 2024; Xin et al., 2024), even automating entire scientific projects, including literature reviews, experiment design, and manuscript writing (Lu et al., 2024a; Schmidgall et al., 2025).

Nonetheless, regardless of domain, agentic systems remain constrained by the tools at their disposal. For example, when tasked to solve a pathology image classification problem, the AIDE machine learning agent (Schmidt et al., 2024) trains a standard convolutional net (*c.f.* Fig. 2 in Chan et al. (2024)). By contrast, a domain expert would instead employ pathology foundation models, as these have been designed specifically for this type of problem (Chen et al., 2024; Zimmermann et al., 2024; Wölflin et al., 2024). Thus, AIDE lacks the necessary tools to solve the task effectively.

Tool creation To address this, we consider the problem of *tool creation* – enabling LLMs to create their own tools, to dynamically expand their capabilities at runtime. Tool creation is not to be confused with *tool learning*, *i.e.* teaching LLMs to utilise appropriate, human-crafted tools more effectively which has been extensively studied in recent years (Qin et al., 2024; Schick et al., 2023). Previous work on tool creation (Cai et al., 2024; Yuan et al., 2024; Qian et al., 2023) is limited to crafting very simple tools because (i) they are crafted from scratch, and (ii) these systems cannot interact with the operating system (OS) by running bash commands, reading/writing files, *etc.* (see Table 1). Our approach addresses both of these limitations.

Method	Error handling	OS interaction	Complex tasks
CRAFT (Yuan et al., 2024)	✗	✗	✗
CREATOR (Qian et al., 2023)	✓	✗	✗
LATM (Cai et al., 2024)	✓	✗	✗
TOOLMAKER (ours)	✓	✓	✓

Table 1: Comparison of tool creation methods. *OS interaction* refers to the ability to interact with the operating system (*e.g.* read/write files, run commands, web browsing). *Complex tasks* require installing and using external dependencies (*e.g.* libraries, model weights).

Benchmarks Various benchmarks have been proposed specifically for tool creation, and software

engineering more generally. Code generation benchmarks (Zhuo et al., 2024; Jain et al., 2024) assess the ability of LLMs to generate Python functions for narrowly defined tasks (*e.g.* simple mathematical problems) using the Python standard library. Tool creation benchmarks extend this idea, enabling the LLM to decide the signature of the Python function in addition to generating the implementation itself (Yuan et al., 2024; Qian et al., 2023; Cai et al., 2024). Yet, these existing code generation and tool creation benchmarks are limited to simple Python functions – they cannot install dependencies or directly interact with the OS.

On the other hand, software engineering benchmarks assess LLM agents for solving GitHub issues (Jimenez et al., 2024), creating ML models (Tang et al., 2024; Chan et al., 2024) and performing repository-level scientific tasks (Majumder et al., 2024; Chen et al., 2025; Bogin et al., 2024; Liu et al., 2024). However, these benchmarks focus on performing *particular tasks*, as opposed to creating a reusable tool to solve a class of problems.

We combine both streams (tool creation and software engineering) by proposing a benchmark focused on real-world multi-step scientific tasks that requires agents to (i) autonomously install necessary dependencies (as opposed to implementing simple Python functions), and (ii) produce a reusable tool that can be applied with different inputs (as opposed to solving a single task instance).

3 TOOLMAKER

We design TOOLMAKER to autonomously convert stand-alone code repositories from scientific publications into LLM-compatible tools. Each tool should complete a specific, user-defined task. To do so, we require a minimal *tool definition* (see Fig. 2, top), consisting of:

- 1) a concise textual description of the task,
- 2) GitHub URL of the associated repository, and
- 3) a list of required input arguments, including an example value for each argument.

This tool definition could in principle be represented as the signature of a Python function with a docstring, like in existing code generation tasks (Zhuo et al., 2024; Jain et al., 2024). However, unlike previous work, we require the LLM to not only implement the function, but also to *set up the environment* wherein the function will be executed. The latter is necessary due to the complexity of our tasks which require *e.g.* installing external dependencies, downloading models, and

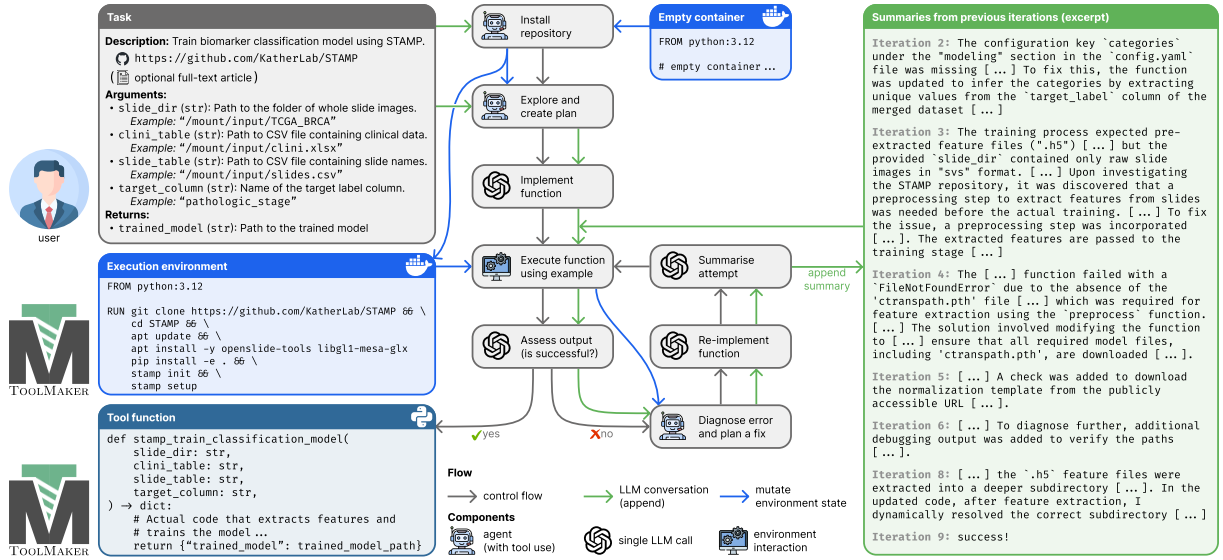


Figure 3: **TOOLMAKER workflow.** Given a task description, a scientific paper, and its associated code repository, TOOLMAKER generates an executable tool that enables a downstream LLM agent to perform the described task.

setting up configurations while considering system and hardware specifications.

We structure TOOLMAKER as an *agentic workflow* (Fig. 3) consisting of two stages: environment setup and tool implementation. During environment setup, TOOLMAKER produces a reproducible “snapshot” of the system (a Docker image) wherein the tool will run. Then, TOOLMAKER generates a Python function that implements the desired task.

3.1 Workflow components

We define the *state of the workflow* at any point in time to be a pair

$$s = (h, e) \in \mathcal{H} \times \mathcal{E}.$$

Here, $h \in \mathcal{H}$ is the *conversation history* (the ordered sequence of messages from the user, tools, and the LLM), and $e \in \mathcal{E}$ is the *environment state* (represented by a checkpointed Docker container).

TOOLMAKER is built out of fundamental *components*, each viewed as a function that acts on the workflow state as

$$\mathcal{S} \mapsto \mathcal{S} \times \mathcal{R},$$

where $\mathcal{S} = \mathcal{H} \times \mathcal{E}$ is the space of all possible workflow states, and $\mathcal{R} \supseteq \mathcal{M} \cup \mathcal{O}$ is the set of possible returns (e.g. a newly generated message in \mathcal{M} or an environment observation in \mathcal{O}). We distinguish three types of components: **LLM calls** ($\mathcal{H} \mapsto \mathcal{H} \times \mathcal{M}$), **environment interactions** ($\mathcal{E} \mapsto \mathcal{E} \times \mathcal{O}$), and **agents** ($\mathcal{H} \times \mathcal{E} \mapsto \mathcal{H} \times \mathcal{E} \times \mathcal{R}$).

3.1.1 LLM calls

An LLM can be viewed as a function

$$LLM : \mathcal{H} \rightarrow \mathcal{M},$$

which, given a conversation history, produces a single new message. As a TOOLMAKER workflow component, an LLM call $\ell : \mathcal{H} \rightarrow \mathcal{H} \times \mathcal{M}$ takes the workflow state’s conversation history h , appends $LLM(h)$, and returns the new message:

$$h \mapsto (h \oplus LLM(h), LLM(h)).$$

LLMs calls thus only update the conversation and do not modify the environment. We use OpenAI’s gpt-4o-2024-08-06 model for the LLM calls.

3.1.2 Environment interactions

An environment interaction is any action $a \in \mathcal{A}$ that can read from or write to the environment state e . We may thus model it by

$$e \mapsto (e', o),$$

where e' is the updated environment state, and $o \in \mathcal{O}$ is the observation produced by the action.

The set of environment actions are

$$\mathcal{A} = \left\{ \begin{array}{l} \text{RUN_BASH_COMMAND, LIST_DIRECTORY,} \\ \text{READ_FILE, WRITE_FILE, BROWSE,} \\ \text{GOOGLE_DRIVE_LIST_FOLDER,} \\ \text{GOOGLE_DRIVE_DOWNLOAD_FILE,} \\ \text{RUN_IMPLEMENTATION} \end{array} \right\}.$$

We distinguish between *read-only* actions and *write* actions (Huyen, 2024). While read-only actions $\mathcal{A}_r = \{\text{READ_FILE, LIST_DIRECTORY, BROWSE, GOOGLE_DRIVE_LIST_FOLDER}\}$ have $e' = e$, write actions $\mathcal{A}_w = \{\text{WRITE_FILE, GOOGLE_DRIVE_DOWNLOAD_FILE, RUN_IMPLEMENTATION}\}$ may modify e .

The **RUN_IMPLEMENTATION** action is a special action that allows TOOLMAKER to execute a candidate tool implementation.

3.1.3 Agents

An *agent* π , illustrated in Fig. 4, chains multiple LLM calls and environment interactions to accomplish a specific sub-task which is specified by a

high-level instruction, $m_\pi \in \mathcal{M}$, e.g. “install this repository and its dependencies”.

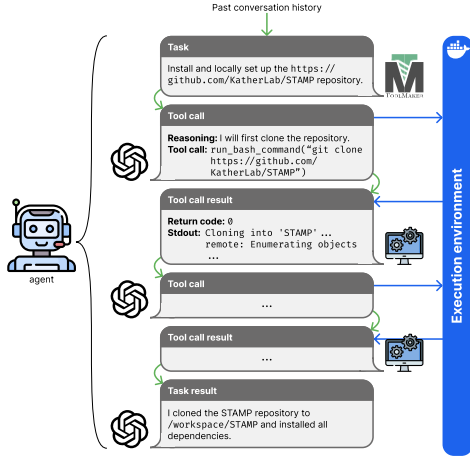


Figure 4: An agent uses a tool-augmented LLM to perform a specific sub-task, and returns the result. Messages are **appended** to the conversation history, and tool calls enable the agent to **interact** with the environment.

Formally, an agent π maps the current workflow state $s = (h, e)$ to a new state $s_T = (h_T, e_T)$ and return value $r \in \mathcal{R}$:

$$(h, e) \mapsto (h_T, e_T, r).$$

The agent follows a sequence of state transitions

$$s_0 \rightarrow s_1 \rightarrow \dots \rightarrow s_T,$$

where each state $s_t = (h_t, e_t) \in \mathcal{S}$. At step $t = 0$, the agent receives the *initial* state

$$s_0 = (h \oplus m_\pi, e).$$

At each step t , the agent employs a special *tool-augmented* LLM, denoted

$$LLM_\pi : \mathcal{H} \rightarrow \mathcal{A}_\pi \cup \mathcal{R},$$

which, given the current conversation h_t , either outputs an **action** $a_t \in \mathcal{A}_\pi$ (a tool call) or the **final result** $r \in \mathcal{R}$ of the sub-task. Here, $\mathcal{A}_\pi \subseteq \mathcal{A} \setminus \{\text{RUN_IMPLEMENTATION}\}$ excludes directly running candidate tool implementations, as this is a separate step in the TOOLMAKER workflow. We implement the choice between \mathcal{A}_π and \mathcal{R} using OpenAI’s function calling and structured output APIs respectively (OpenAI, 2025).

If the LLM proposes an action $a_t = LLM_\pi(h_t) \in \mathcal{A}$, we execute a_t on the current environment to obtain the observation and updated environment state $(e_{t+1}, o_t) = a_t(e_t)$. We then append both the tool call and its observation to the conversation, forming the new state

$$s_{t+1} = (h_t \oplus a_t \oplus o_t, e_{t+1}).$$

If instead $LLM_\pi(h_t)$ outputs a final result $r \in \mathcal{R}$, the agent terminates and returns $s_T = (h_t, e_t, r)$.

Algorithm 1 TOOLMAKER workflow.

Require: Tool definition m_{tool} , initial environment $e_\emptyset \in \mathcal{E}$

- 1: $h_\emptyset \leftarrow \{m_{\text{tool}}\}$ \triangleright initialise conversation history
- 2: $h, e, r \leftarrow \text{INSTALL_REPOSITORY}(h_\emptyset, e_\emptyset)$
- 3: $\bar{e} \leftarrow e$ \triangleright snapshot of installed environment state
- 4: $h, e, r \leftarrow \text{EXPLORE}(h_\emptyset, \bar{e})$
- 5: $h, m \leftarrow \text{PLAN}(h)$
- 6: $\bar{h} \leftarrow h$ \triangleright snapshot of conversation history
- 7: $h, m_{\text{code}} \leftarrow \text{IMPLEMENT}(h)$
- 8: $\sigma \leftarrow \emptyset$
- 9: **while** true **do**
- 10: $e \leftarrow \bar{e}$ \triangleright restore installed environment state
- 11: $h \leftarrow \bar{h} \oplus \sigma \oplus m_{\text{code}}$ \triangleright restore conversation history
- 12: $e, o \leftarrow \text{RUN_IMPLEMENTATION}(e, m_{\text{code}})$
- 13: $h, m \leftarrow \text{ASSESS_TOOL_OUTPUT}(h \oplus o)$
- 14: **if** m is successful **then**
- 15: **return** \bar{e}, m_{code}
- 16: **end if**
- 17: $h, e, r \leftarrow \text{DIAGNOSE_ERROR}(h \oplus o, e)$
- 18: $h, m_{\text{code}} \leftarrow \text{REIMPLEMENT}(h)$
- 19: $h, m_{\text{summary}} \leftarrow \text{SUMMARISE}(h)$
- 20: $\sigma \leftarrow \sigma \oplus m_{\text{summary}}$
- 21: **end while**

3.2 TOOLMAKER workflow

In this section, we describe our workflow in detail, which at a high level is illustrated in Fig. 3, and in pseudocode in Algorithm 1, using the three types of components (LLM calls, environment interactions, and agents) introduced above.


TOOLMAKER’s initial conversation history h_\emptyset is a system prompt that contains the tool definition m_{tool} . We provide the full prompts in Appendix D.


Environment setup To obtain the state of the execution environment necessary for the tool to execute, we employ the `INSTALL_REPOSITORY` agent (line 2) that is instructed to install and set up the repository. This agent clones and explores the repository, reads documentation, and downloads any dependencies it deems necessary such as models, datasets, and libraries. Each of these steps involve planning and learning from previous observations such as error logs arising during execution.

The agent begins with a clean environment state e_\emptyset (a python:3.12 Docker image). Importantly, we record all write actions (\mathcal{A}_w) that the agent performs. Since each of these actions may be expressed as a bash command, we simply concatenate their bash representations to obtain the environment definition in the form of a bash script or Dockerfile.


Initial implementation We first instruct an agent (`EXPLORE`) to explore the repository and gather all information necessary to implement the tool. Note that we do not carry over the conversation history from the previous stage, in order to not pollute the context with a large number of messages


(by calling  EXPLORE on h_0 , not h on line 4).



Next we perform an LLM call ( PLAN) to create a step-by-step plan for the implementation. We keep all messages (including actions and observations) in the conversation history, so this information can be used to create the plan.

Then, we instruct the LLM ( IMPLEMENT) to write the Python code for the tool based on the plan, producing our first *candidate implementation*.

Closed-loop self-improvement Now, we enter the closed-loop self-improvement phase. First, we reset the execution environment to the *environment definition* \bar{e} because the agent may have performed write actions in the past. We also restore the conversation history to immediately after generating the implementation plan, but include summaries of past appempts (described later).

After running the candidate Python function in the execution environment using the example invocation provided in the tool definition (line 12), we instruct the LLM to assess whether the execution was successful ( ASSESS_TOOL_OUTPUT). Specifically, we ask the LLM to check whether the result returned by the tool is in line with the task description (*i.e.* if the result is plausible), and whether the standard output and standard error streams contain any indications of errors. If the LLM deemed tool execution successful, we have arrived at our final tool implementation, and exit the loop. Otherwise, we continue the self-improvement loop.

Next, we instruct the  DIAGNOSE_ERROR agent to gather information about the error in order to diagnose its root cause and formulate a plan to fix it. Importantly, we do not reset the execution environment – the agent is able to check intermediate files and outputs created during tool execution.

Then, the LLM re-implements the tool based on the current implementation, error diagnosis, and plan to fix the error ( REIMPLEMENT). Finally, we ask the LLM to summarise the current step ( SUMMARISE), and append this summary to the conversation history for the next iteration.

3.3 Execution environment

An important implementation detail is the *execution environment*, which is the environment in which (i) actions (\mathcal{A}) are performed throughout the TOOLMAKER workflow, and (ii) wherein the final tool created by TOOLMAKER will be executed.

The execution environment itself is *stateful*. Specifically, write actions $\mathcal{A}_w = \{\text{📁}, \text{📄}, \text{🗑️}, \text{🐍}\}$ may mutate environment state. However, we re-

quire the ability to roll back to previous states, *e.g.* on line 10 of Algorithm 1, the execution environment is restored to the “freshly installed” state \bar{e} . Furthermore, the execution environment should be sandboxed from the host system (for security reasons), and it should be reproducible (so the generated tool can be executed on any machine).

We satisfy these requirements by implementing the execution environment as a Docker container that TOOLMAKER controls via an HTTP server running inside the container, which can run the pre-defined actions \mathcal{A} . State restoration is achieved via Docker’s checkpointing functionality.

4 Benchmark

To evaluate our approach, we collect a dataset of 15 diverse tasks spanning multiple scientific disciplines, which we refer to as TM-BENCH. The tasks were curated in close collaboration with researchers in medicine and life sciences to reflect realistic problems in these fields, with a focus on the medical domain (pathology, radiology, omics), while also including some tasks from other areas such as 3D vision, imaging, tabular data analysis, and natural language processing to ensure broader coverage of real-world scientific challenges.

Before including a task in TM-BENCH, we manually implemented the intended tool using the associated GitHub repository to ensure the task is well-defined and solvable. This vetting process gave us confidence that each task is meaningful, correctly specified, and feasible. The resulting benchmark covers a range of difficulty levels, from simple tasks that can be achieved by calling an existing method, to more complex, multi-step tasks that require orchestrating multiple function calls, transforming data, and utilising GPUs.

Task definitions As shown in Fig. 2 (top), each task definition consists of: (i) a one-sentence task description, (ii) the URL to the code repository, (iii) a list of input arguments, alongside an example invocation (see below), and (iv) a description of the expected output. An overview of the tasks and associated papers can be found in Table 2, and we provide a full list of all task definitions with their example invocations in Appendix C.

Invocations A task *invocation* specifies a concrete value for each input argument, as well as *external files and directories* that should be made accessible from within the execution environment during the invocation. Indeed, most tasks in TM-BENCH require external files, *e.g.*

Task		TOOLMAKER (ours)					OpenHands (Wang et al., 2024)				
		Invoc.	Tests	Cost	Actions	Tokens	Invoc.	Tests	Cost	Actions	Tokens
Pathology	conch_extract_features (Lu et al., 2024b)	3/3	9/9	\$0.35	15 (1 _○)	171,226	3/3	9/9	\$0.08	5	51,701
	musk_extract_features (Xiang et al., 2025)	3/3	6/6	\$1.19	29 (6 _○)	696,386	✗	✗	\$0.15	7	97,386
	pathfinder_verify_biomarker (Liang et al., 2023)	0/2	4/6	\$0.61	27 (1 _○)	356,825	0/2	4/6	\$0.08	6	49,414
	stamp_extract_features (El Nahhas et al., 2024)	3/3	12/12	\$1.12	20 (4 _○)	631,138	0/3	3/12	\$0.07	6	42,793
	stamp_train_classification_model (El Nahhas et al., 2024)	3/3	9/9	\$2.27	33 (9 _○)	1,249,521	0/3	0/9	\$0.15	8	87,915
	uni_extract_features (Chen et al., 2024)	3/3	9/9	\$0.61	16 (4 _○)	326,806	✗	✗	\$0.25	10	177,119
Radiology	medsam_inference (Ma et al., 2024)	3/3	6/6	\$0.96	18 (6 _○)	508,954	✗	✗	\$0.07	5	41,096
	nnunet_train_model (Isensee et al., 2020)	0/2	0/4	\$2.90	35 (9 _○)	1,792,291	0/2	0/4	\$0.12	8	79,231
Omics	cytopus_db (Kunes et al., 2023)	3/3	12/12	\$0.41	10 (3 _○)	185,912	✗	✗	\$0.36	8	236,217
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	2/3	13/15	\$0.66	20 (1 _○)	336,754	✗	✗	\$0.11	6	69,493
Other	flowmap_overfit_scene (Smith et al., 2024)	2/2	6/6	\$0.70	18 (5 _○)	358,552	✗	✗	\$0.36	15	250,787
	medsss_generate (Jiang et al., 2025)	3/3	6/6	\$0.53	25 (3 _○)	282,771	3/3	6/6	\$0.15	10	104,505
	modernbert_predict_masked (Warner et al., 2024)	3/3	9/9	\$0.66	20 (4 _○)	356,228	✗	✗	\$0.13	10	82,930
	retfound_feature_vector (Zhou et al., 2023)	3/3	6/6	\$0.97	31 (5 _○)	561,936	0/3	0/6	\$0.08	4	46,521
	tabpfn_predict (Hollmann et al., 2025)	3/3	9/9	\$0.23	10 (1 _○)	95,257	3/3	9/9	\$0.07	4	36,320

Table 2: Performance of the tools created by TOOLMAKER and the OpenHands baseline (Wang et al., 2024) on the benchmark tasks. ✗ indicates that the environment installation failed. We use ○ to indicate the number of self-correcting iterations. Green cells indicate that the tool implementation is correct (all unit tests pass), yellow indicates that at least one unit test failed, and red indicates that all unit tests failed.

stamp_train_classification_model takes an input dataset of whole slide images (WSIs) and a clinical data table, on which to train a classification model using the STAMP (El Nahhas et al., 2024) pipeline. Analysing and utilising datasets is a fundamental aspect of many real-world scientific tasks, which is why TM-BENCH explicitly supports this functionality, unlike many existing code generation benchmarks (Zhuo et al., 2024; Jain et al., 2024).

Each task definition includes a single example invocation, which may be used in the tool creation process. Crucially, this specification does not include the expected return value, as the goal is to autonomously implement and execute the task without prior knowledge of the correct output.

Assessing correctness TM-BENCH specifies 2-3 additional test invocations per task, which are different to the example invocation and held-out from the tool creation process. We purposefully chose different input argument values for the test invocations (different datasets, images, paths, etc.) to assess whether the implementations would generalise to other inputs, and to ensure that implementations did not ‘hard-code’ the example invocation.

For each test invocation, TM-BENCH includes unit tests to assess whether the tool produces the expected output by checking various properties of the return value and output files. We opted for unit tests over simple equality checks (e.g. strict or near-exact matches to reference outputs, as used in previous benchmarks (Bogin et al., 2024)) because unit tests can accommodate more complex criteria, such as verifying the shape of generated feature vectors or checking that a segmentation model produces plausibly sized masks. Specifically, we employ unit tests to verify correctness through assertions on: *structure* (dimensions and types of return val-

ues), *values* (range, accuracy, and statistical properties of return values), *files* (existence, format, and content of files produced by the tool, if applicable), and *execution* (errors/crashes).

To ensure an unbiased assessment of tool implementations, the unit tests and test invocations are used strictly for evaluation and are not available during tool creation. TM-BENCH comprises 15 tasks, with a total of 42 test invocations (average 2.8 per task) and 124 unit tests (average 8.3 per task). We consider a tool implementation correct only if it passes all unit tests of its test invocations.

5 Results

TM-BENCH can evaluate any “tool maker” that produces an environment definition 🐳 and a tool implementation 🐍. However, to the best of our knowledge, no existing approaches are specifically designed to address the “paper repository → LLM tool” problem. In order to nonetheless facilitate comparison with prior work, we adapt the OpenHands (Wang et al., 2024) to this setting. OpenHands is a software engineering agent that achieves SOTA performance on SWE-bench (Jimenez et al., 2024). We instruct OpenHands to generate the same artifacts as TOOLMAKER: an environment definition 🐳 (expressed as a bash script to be run in a fresh python:3.12 Docker image to create the environment state required for the tool to execute) and a tool implementation 🐍 (a Python function). To ensure a fair comparison, we reuse large parts of the TOOLMAKER prompts in the prompts we supply to the OpenHands, and add additional instructions to encourage OpenHands to test the artifacts it creates. We use gpt-4o for the OpenHands baseline, but also ablate the choice of LLM in Section 5.1. The full prompts for TOOLMAKER and

OpenHands are listed in Appendices D and E.

Performance In Table 2, we report the performance of TOOLMAKER and OpenHands on all tasks in TM-BENCH, reporting correctness, cost, number of tokens, number of actions performed in the tool creation process (both stages), and the number of self-correcting iterations. We consider a test invocation successful (“Tests” column marked **green**) if all of the unit tests that are associated with it pass. Similarly, a tool implementation is correct (“Invoc.” column marked **green**) if *all* of its test invocations are successful, *i.e.* all of the unit tests associated with its test invocations pass.

TOOLMAKER significantly outperforms OpenHands, achieving an accuracy of 80% (correctly implementing 12/15 tasks) while OpenHands was only able to correctly implement 20% (3/15 tasks).

For the `esm_fold_predict` (Verkuil et al., 2022) task, TOOLMAKER generates a partially correct implementation (**yellow**) that passes two out of three test invocations. The goal of this task is to predict the contact map of a protein from its sequence. Upon inspection, we determined that the failed test invocation was different from the other invocations: it contained a mask token in the input sequence which was not present in the task definition’s example invocation. However, when including such a mask token in the example invocation and re-running TOOLMAKER, the tool implementation passed all test invocations. This highlights that the example invocation in the task definition needs to be representative of the task.

By contrast, OpenHands fails to generate correct tool implementations for most tasks, primarily due to errors at the environment setup stage. Nearly half of its environment definitions were invalid, causing installation scripts to crash before execution. Even among the tasks where OpenHands successfully generated an environment definition, only three implementations passed all unit tests.

This poor performance can largely be attributed to issues during environment setup. Specifically, OpenHands often created installation scripts without testing them, omitted essential setup commands previously executed manually, or overlooked dependencies necessary for tool execution. In contrast, TOOLMAKER inherently avoids such pitfalls by automatically capturing every installation command and resetting the execution environment between iterations, ensuring reproducible and robust tool implementations.

Multi-step tools A remarkable feature of TOOLMAKER is that it is able to create tools that require multiple steps to complete. For example, the `stamp_train_classification_model` task provides a dataset of pathology WSIs and a table of clinical data, and requires the tool implementation to use the STAMP pipeline (El Nahhas et al., 2024) to train a classification model that predicts a specific biomarker from the WSI images. This task requires multiple steps to complete: after downloading and installing the STAMP repository and its dependencies, the tool implementation needs to use STAMP to (1) perform feature extraction on the WSI images, and (2) train a classification model using the extracted features and the clinical data. The self-correcting loop allows TOOLMAKER to realise that it needs to perform feature extraction before it can train a classification model, and to subsequently implement the tool function to perform both steps, illustrated in Fig. 3 (right). For this particular task, TOOLMAKER performs 9 self-correcting iterations, executing 33 actions in total, before arriving at the final implementation.

Cost TOOLMAKER performs an average of 21.8 actions during tool creation, costing on average \$0.94 per tool, while OpenHands performs 7.5 actions on average (\$0.15 per tool). The three tools that OpenHands correctly implemented were among the cheapest for TOOLMAKER, requiring the fewest actions and self-correcting iterations. This shows OpenHands can implement very “easy” tools, but fails to generalise to more complex tasks.

5.1 Ablations

Paper summaries Since each task is based on one or more research papers, we perform an ablation study to determine whether we can inject useful information from the papers into the tool creation process. Instead of directly including the full paper text in the prompts which would require too many tokens, we first provide the full text to gpt-4o and instruct it to summarise it with respect to the task at hand. Then, we provide these task-specific and paper-specific summaries in the prompts for TOOLMAKER and OpenHands.

The results in Table 3 indicate that including paper summaries does not increase the performance of either approach. However, it does decrease the average number of actions and, for TOOLMAKER, the average number of self-correcting iterations required to create the tools. For example, while TOOLMAKER required 9 iterations (33 actions) to

Method	Tools	Invoc.	Tests	Cost	Actions
TOOLMAKER* (ours)	12/15	37/42	116/124	\$0.94	21.8
(with paper summary)	11/15	34/42	113/124	\$0.71	18.1
(using o3-mini)	9/15	28/42	107/124	\$0.55	14.1
OpenHands* (Wang et al., 2024)	3/15	9/42	31/124	\$0.15	7.5
(with paper summary)	3/15	9/42	31/124	\$0.12	6.6
(using o3-mini)	1/15	2/42	15/124	\$0.04	1.9
(using Claude 3.5 Sonnet)	2/15	6/42	19/124	\$0.13	5.2

Table 3: Ablation results. Rows marked with asterisk correspond to the results in Table 2. We report the number of correct tools, invocations, and tests, as well as the per-tool average cost and number of actions performed.

create the `stamp_train_classification_model` tool, this decreased to only 5 iterations (15 actions) when using the paper summary (see Appendix B.1).

Choice of LLM We also evaluate TOOLMAKER and OpenHands using OpenAI’s o3-mini model instead of gpt-4o, and find that while this reduces cost, it also degrades performance in both cases. Finally, since OpenHands achieved SOTA performance on SWE-bench (Jimenez et al., 2024) using Claude 3.5 Sonnet (Anthropic, 2024), we re-run the OpenHands baseline using this model, but find that it performs worse than using gpt-4o (see Table 3).

6 Conclusion

In this work, we showed that autonomous tool creation can go beyond simple Python functions and produce tools for real-world scientific tasks. We introduced TOOLMAKER, a framework that autonomously transforms scientific code repositories into LLM-compatible tools, potentially drastically reducing the technical overhead in future for developing agents with specialised tool-sets. In evaluations across multiple scientific domains, TOOLMAKER surpassed the state-of-the-art software engineering agent, OpenHands, achieving 80% accuracy. Additionally, we release TM-BENCH as a comprehensive benchmark to spur further advancements in agentic tool creation.

We acknowledge that automated tool creation in life sciences carries significant risks that require careful consideration. The ability to autonomously implement complex biochemical tools could potentially be misused for creating harmful agents or bioweapons. Additionally, fully automated research systems might inadvertently generate dangerous compounds or protocols without proper oversight. These risks underscore the importance of developing robust safety measures and ethical guidelines alongside technical capabilities. Nonetheless, by removing technical barriers to tool creation, TOOLMAKER brings us closer to a future where the pace of scientific discovery is limited by

computational capacity, not human resources.

Acknowledgments

We thank Junhao Liang, Michaela Unger, and David Charatan for contributing tasks to TM-BENCH. We also appreciate Jan Clusmann, Tim Lenz, and Lina Hadji-Kyriacou for their feedback on the manuscript, and thank Nathaly Dongo and Annelies Blätterlein for logo design.

Funding GW is supported by SCADS.AI, Lothian NHS, and in part by funding from the European Union’s Horizon 2020 research and innovation programme (KATY, 101017453). JNK is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A; DECIPHER-M, 01KD2420A; NextBIG, 01ZU2402A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048), the European Union’s Horizon Europe research and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (grant number NIHR203331). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Limitations

While TOOLMAKER addresses the challenge of tool creation, we acknowledge that fully autonomous scientific discovery remains constrained by physical experimentation. TOOLMAKER does not address this aspect, but we believe that with an increasing proportion of life science research being conducted *in silico*, it provides a building block for autonomous scientific workflows in future.

Our framework assumes that the referenced code repositories are reasonably well-structured, up-to-

date, and documented. In practice, however, open-source repositories may be poorly documented or incomplete, making them challenging to install autonomously. In fact, there is no guarantee that any given repository will be installable and usable as a tool. For TM-BENCH, we manually curated the tasks such that we were able to successfully install and use the repository ourselves. This way, we ensured the tasks were *possible* in the first place.

While TM-BENCH contains over 100 unit tests, passing these tests does not guarantee correctness in all real-world scenarios. Scientific workflows often involve edge cases or unexpected patterns that are not captured by a small set of tests. Moreover, high-stakes applications such as clinical research would naturally demand additional layers of rigorous validation and oversight by domain experts.

Finally, while TM-BENCH pins the exact commits of the referenced repositories, external factors such as repository deletion, force-pushing changes, or renaming branches, could affect reproducibility.

References

- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. [Accessed 20-01-2025].
- Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. SUPER: Evaluating agents on setting up and executing tasks from research repositories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12622–12645. Association for Computational Linguistics.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.
- Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2024. Large language models as tool makers. In *The Twelfth International Conference on Learning Representations*.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2024. *Mle-bench: Evaluating machine learning agents on machine learning engineering*. Preprint, arXiv:2410.07095.
- Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025. *Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery*. Preprint, arXiv:2410.05080.
- Ning Ding, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong, Yuxin Zuo, Zhangren Chen, Ermo Hua, Xingtai Lv, Youbang Sun, Yang Li, Dong Li, Fuchu He, and Bowen Zhou. 2024. *Automating exploratory proteomics research via language models*. Preprint, arXiv:2411.03743.
- Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligero, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. 2024. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nature Protocols*.
- Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie-Elisabeth Leßman, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. 2024. *Autonomous artificial intelligence agents for clinical decision making in oncology*. Preprint, arXiv:2404.04667.
- Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.
- Gemini Team. 2024. *Gemini: A family of highly capable multimodal models*. Preprint, arXiv:2312.11805.
- Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. 2022. *A high-level programming language for generative protein design*. Preprint, bioRxiv:2022.12.21.521526.
- Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang

- Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.
- Chip Huyen. 2024. *AI engineering*. O'Reilly Media, Sebastopol, CA.
- Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. 2020. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Live-codebench: Holistic and contamination free evaluation of large language models for code. *Preprint*, arXiv:2403.07974.
- Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds³: Towards medical small language models with self-evolved slow thinking. *Preprint*, arXiv:2501.12051.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.
- Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and Zhiyong Lu. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Preprint*, arXiv:2402.13225.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Russell Z. Kunes, Thomas Walle, Max Land, Tal Nawy, and Dana Pe'er. 2023. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*, 42(7):1084–1095.
- Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2024. Legalagentbench: Evaluating llm agents in legal domain. *Preprint*, arXiv:2412.17259.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.
- Junhao Liang, Weisheng Zhang, Jianghui Yang, Meilong Wu, Qionghai Dai, Hongfang Yin, Ying Xiao, and Lingjie Kong. 2023. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nature Machine Intelligence*, 5(4):408–420.
- Tianyang Liu, Canwen Xu, and Julian McAuley. 2024. Repobench: Benchmarking repository-level code auto-completion systems. In *The Twelfth International Conference on Learning Representations*.
- Llama Team. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024a. The ai scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.
- Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. 2024b. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.
- Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1).
- Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *Preprint*, arXiv:2407.01725.
- Minh Huynh Nguyen, Thang Phan Chau, Phong X. Nguyen, and Nghi D. Q. Bui. 2024. Agilecoder: Dynamic collaborative agents for software development based on agile methodology.
- OpenAI. 2025. Openai developer platform. <https://platform.openai.com/docs>. [Accessed 15-02-2025].
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15174–15186. Association for Computational Linguistics.

- Cheng Qian, Chi Han, Yi Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023. CREATOR: Tool creation for disentangling abstract and concrete reasoning of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. 2024. Tool learning with foundation models. *ACM Comput. Surv.*, 57(4).
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. [Agent laboratory: Using llm agents as research assistants](#). *Preprint*, arXiv:2501.04227.
- Dominik Schmidt, Zhengyao Jiang, and Yuxiang Wu. 2024. Introducing Weco AIDE — weco.ai. <https://www.weco.ai/blog/technical-report>. [Accessed 20-01-2025].
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. 2024. [Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent](#). *Preprint*, arXiv:2404.15259.
- Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2024. [The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation](#).
- Xiangru Tang, Yuliang Liu, Zefan Cai, Yanjun Shao, Junjie Lu, Yichi Zhang, Zexuan Deng, Helan Hu, Kaikai An, Ruijun Huang, Shuzheng Si, Sheng Chen, Haozhe Zhao, Liang Chen, Yan Wang, Tianyu Liu, Zhiwei Jiang, Baobao Chang, Yin Fang, Yujia Qin, Wangchunshu Zhou, Yilun Zhao, Arman Cohen, and Mark Gerstein. 2024. [Ml-bench: Evaluating large language models and agents for machine learning tasks on repository-level code](#). *Preprint*, arXiv:2311.09835.
- Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. 2023. [On the planning abilities of large language models \(a critical investigation with a proposed benchmark\)](#). *Preprint*, arXiv:2302.06706.
- Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. 2022. [Language models generalize beyond natural proteins](#). *Preprint*, bioRxiv:2022.12.21.521521.
- Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. 2024. [Openhands: An open platform for ai software developers as generalist agents](#). *Preprint*, arXiv:2407.16741.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Georg Wölflein, Dyke Ferber, Asier Rabasco Meneghetti, Omar S. M. El Nahhas, Daniel Truhn, Zunamys I. Carrero, David J. Harrison, Ognjen Arandjelović, and Jakob Nikolas Kather. 2024. A good feature extractor is all you need for weakly supervised pathology slide classification. In *European Conference on Computer Vision (ECCV)*. Springer. BioImage Computing Workshop.
- Chun Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2024. [Agentless: Demystifying llm-based software engineering agents](#).
- Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. 2025. A vision-language foundation model for precision oncology. *Nature*.

- Qi Xin, Quyu Kong, Hongyi Ji, Yue Shen, Yuqi Liu, Yan Sun, Zhilin Zhang, Zhaorong Li, Xunlong Xia, Bing Deng, and Yinqi Bai. 2024. [Bioinformatics agent \(bia\): Unleashing the power of large language models to reshape bioinformatics workflow](#).
- John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. 2024. SWE-agent: Agent-computer interfaces enable automated software engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi Fung, Hao Peng, and Heng Ji. 2024. CRAFT: Customizing LLMs by creating and retrieving from specialized toolsets. In *The Twelfth International Conference on Learning Representations*.
- Siqi Zhou, Lukas Brunke, Allen Tao, Adam W. Hall, Federico Pizarro Bejarano, Jacopo Panerati, and Angela P. Schoellig. 2024. What is the impact of releasing code with publications?: Statistics from the machine learning, robotics, and control communities. *IEEE Control Systems*, 44(4):38–46.
- Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Naomi Allen, John E. J. Gallacher, Thomas Littlejohns, Tariq Aslam, Paul Bishop, Graeme Black, Panagiotis Sergouniotis, Denize Atan, Andrew D. Dick, Cathy Williams, Sarah Barman, Jenny H. Barrett, Sarah Mackie, Tasanee Braithwaite, Roxana O. Carare, Sarah Ennis, Jane Gibson, Andrew J. Lotery, Jay Self, Usha Chakravarthy, Ruth E. Hogg, Euan Paterson, Jayne Woodside, Tunde Peto, Gareth McKay, Bernadette McGuinness, Paul J. Foster, Konstantinos Balaskas, Anthony P. Khawaja, Nikolas Pontikos, Jugnoo S. Rahi, Gerassimos Lascaratos, Praveen J. Patel, Michelle Chan, Sharon Y. L. Chua, Alexander Day, Parul Desai, Cathy Egan, Marcus Frutiger, David F. Garway-Heath, Alison Hardcastle, Sir Peng T. Khaw, Tony Moore, Sobha Sivaprasad, Nicholas Strouthidis, Dhanes Thomas, Adnan Tufail, Ananth C. Viswanathan, Bal Dhillon, Tom Macgillivray, Cathie Sudlow, Veronique Vitart, Alexander Doney, Emanuele Trucco, Jeremy A. Guggenheim, James E. Morgan, Chris J. Hammond, Katie Williams, Pirro Hysi, Simon P. Harding, Yalin Zheng, Robert Luben, Phil Luthert, Zihan Sun, Martin McKibbin, Eoin O’Sullivan, Richard Oram, Mike Weedon, Chris G. Owen, Alicja R. Rudnicka, Naveed Sattar, David Steel, Irene Stratton, Robyn Tapp, Max M. Yates, Axel Petzold, Savita Madhusudhan, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163.
- Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widayarsi, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kadour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. 2024. [Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions](#). *Preprint*, arXiv:2406.15877.
- Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. 2024. [Virchow2: Scaling self-supervised mixed magnification models in pathology](#). *Preprint*, arXiv:2408.00738.



A TOOLMAKER

A.1 Detailed workflow description

We provide a detailed description of every step in the TOOLMAKER workflow to supplement Algorithm 1 and our discussion thereof in Section 3.2.


A.1.1 Setting up the environment


The environment definition is a state of the world (*e.g.* the operating system) that is required for the tool created by TOOLMAKER to execute. We can represent this state as a sequence of actions (*e.g.* bash commands or instructions in a Dockerfile, as shown in Fig. 2, left) that mutate a known initial state (*e.g.* a freshly installed operating system) to the state required for the tool to execute.


To obtain the state of the execution environment necessary for the tool to execute, we employ an agent  that is instructed to install and set up the repository (we provide the full prompt in Appendix D). This agent will clone and explore the repository, read documentation, and download any dependencies it deems necessary such as models, datasets, and libraries. Each of these steps involve planning and learning from previous mistakes such as error logs arising during execution. The agent begins with a clean state (a python:3.12 Docker image). Importantly, we record all actions  that the agent performs. Since each of the write actions can be expressed as a bash command, we can simply concatenate the bash representations of all recorded write actions to obtain the environment definition in the form of a bash script or Dockerfile.

A.1.2 Initial tool implementation

Equipped with the environment definition, which allows TOOLMAKER to reset the state of the execution environment to the state in which the tool should be executed, it can now implement the tool itself. Note that we do not carry over the conversation history from the previous stage, in order to not pollute the context window with a large number of messages that are irrelevant for this stage.


 **Gather information** We first instruct an agent to explore the installed repository and gather all information necessary to implement the tool. We include the *tool definition* (see Fig. 3, top left) as a Python function signature with a docstring in the initial prompt, so that it can use the information it has already gathered to create the plan.


 **Create a plan** Then, we perform an LLM call to create a step-by-step plan for the tool implementation. Here, we keep all of the agent’s messages (including actions and observations) in the conversation history, so that it can use the information it has already gathered to create the plan.

 **Implement the tool function** Next, we instruct the LLM to implement the tool based on the plan. Again, we keep the entire conversation history in the context window of the LLM call, so that it can refer to previous messages. We now have our first *candidate implementation* of the tool function.

We use OpenAI’s o1-mini-2024-09-12 model for the planning step as well as the implementation step, to take advantage of its reasoning and code generation capabilities.

A.1.3 Closed-loop self-improvement

 **Run the tool** Before executing the candidate implementation, we *reset* the execution environment to the *environment definition* because the agent may have performed write actions in the past (either in the process of exploring the repository, or in a previous iteration of the loop). Then, we run the candidate Python function in the execution environment, using the example invocation provided in the tool definition.

 **Assess tool execution** We instruct the LLM to assess whether the execution was successful, based on the result returned by the function, as well as the standard output and standard error streams produced during function execution. Specifically, we ask the LLM to check whether the result returned by the tool is in line with the task description (*i.e.* if the result is plausible), and whether the standard output and standard error streams contain any indications of errors. If the LLM determines that the tool execution was successful, we have arrived at our final tool implementation, and exit the loop. Otherwise, we continue the self-improvement loop.

Task		TOOLMAKER (ours)					OpenHands (Wang et al., 2024)				
		Invoc.	Tests	Cost	Actions	Tokens	Invoc.	Tests	Cost	Actions	Tokens
Pathology	conch_extract_features (Lu et al., 2024b)	3/3	9/9	\$0.57	15 (2 _o)	274,256	✗	✗	\$0.10	6	65,957
	musk_extract_features (Xiang et al., 2025)	0/3	3/6	\$0.68	19 (3 _o)	355,561	✗	✗	\$0.12	6	77,416
	pathfinder_verify_biomarker (Liang et al., 2023)	0/2	4/6	\$0.75	25 (1 _o)	473,741	0/2	4/6	\$0.11	7	69,545
	stamp_extract_features (El Nahhas et al., 2024)	3/3	12/12	\$1.13	20 (4 _o)	649,284	0/3	3/12	\$0.08	7	52,596
	stamp_train_classification_model (El Nahhas et al., 2024)	3/3	9/9	\$0.76	15 (5 _o)	393,150	0/3	0/9	\$0.25	11	143,934
	uni_extract_features (Chen et al., 2024)	3/3	9/9	\$0.53	14 (3 _o)	268,481	3/3	9/9	\$0.14	5	87,344
Radiology	medsam_inference (Ma et al., 2024)	3/3	6/6	\$0.40	11 (2 _o)	181,604	0/3	0/6	\$0.09	4	50,053
	nnunet_train_model (Isensee et al., 2020)	0/2	0/4	\$0.32	13	213,901	0/2	0/4	\$0.11	4	65,458
Omics	cytopus_db (Kunes et al., 2023)	3/3	12/12	\$0.89	15 (8 _o)	501,078	0/3	0/12	\$0.13	7	87,369
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	2/3	13/15	\$0.96	22 (1 _o)	563,759	✗	✗	\$0.10	5	63,723
Other	flowmap_overfit_scene (Smith et al., 2024)	2/2	6/6	\$2.14	42 (12 _o)	1,204,247	✗	✗	\$0.07	4	46,316
	medsss_generate (Jiang et al., 2025)	3/3	6/6	\$0.76	28 (3 _o)	423,235	3/3	6/6	\$0.15	12	101,581
	modernbert_predict_masked (Warner et al., 2024)	3/3	9/9	\$0.26	11 (1 _o)	106,456	✗	✗	\$0.14	9	84,959
	retfound_feature_vector (Zhou et al., 2023)	3/3	6/6	\$0.29	11 (2 _o)	126,270	0/3	0/6	\$0.15	7	96,780
	tabpfn_predict (Hollmann et al., 2025)	3/3	9/9	\$0.26	10 (1 _o)	104,749	3/3	9/9	\$0.08	5	49,357

Table 4: Results (with paper summary in context).

Task		TOOLMAKER (ours)					OpenHands (Wang et al., 2024)				
		Invoc.	Tests	Cost	Actions	Tokens	Invoc.	Tests	Cost	Actions	Tokens
Pathology	conch_extract_features (Lu et al., 2024b)	0/3	6/9	\$0.22	15 (2 _o)	232,441	✗	✗	\$0.04	0	28,880
	musk_extract_features (Xiang et al., 2025)	0/3	3/6	\$0.24	18 (2 _o)	247,840	✗	✗	\$0.03	2	22,820
	pathfinder_verify_biomarker (Liang et al., 2023)	0/2	4/6	\$0.10	11 (1 _o)	85,312	✗	✗	\$0.04	2	25,797
	stamp_extract_features (El Nahhas et al., 2024)	3/3	12/12	\$0.18	14 (1 _o)	187,972	✗	✗	\$0.04	2	23,343
	stamp_train_classification_model (El Nahhas et al., 2024)	3/3	9/9	\$0.38	17 (4 _o)	403,138	0/3	0/9	\$0.04	2	32,516
	uni_extract_features (Chen et al., 2024)	3/3	9/9	\$0.58	12 (8 _o)	563,488	0/3	0/9	\$0.03	2	22,905
Radiology	medsam_inference (Ma et al., 2024)	0/3	3/6	\$0.87	15 (14 _o)	868,977	0/3	0/6	\$0.04	2	23,410
	nnunet_train_model (Isensee et al., 2020)	0/2	0/4	\$2.74	25 (30 _o)	3,165,597	0/2	0/4	\$0.06	2	36,563
Omics	cytopus_db (Kunes et al., 2023)	3/3	12/12	\$0.22	9 (4 _o)	214,546	0/3	0/12	\$0.04	2	23,522
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	2/3	13/15	\$0.24	11 (1 _o)	270,976	0/3	9/15	\$0.03	2	22,344
Other	flowmap_overfit_scene (Smith et al., 2024)	2/2	6/6	\$0.29	18 (3 _o)	295,054	2/2	6/6	\$0.04	2	24,332
	medsss_generate (Jiang et al., 2025)	3/3	6/6	\$0.60	15 (8 _o)	653,697	✗	✗	\$0.03	2	21,574
	modernbert_predict_masked (Warner et al., 2024)	3/3	9/9	\$0.54	14 (4 _o)	589,902	✗	✗	\$0.03	2	22,617
	retfound_feature_vector (Zhou et al., 2023)	3/3	6/6	\$0.51	10 (8 _o)	490,555	✗	✗	\$0.03	2	23,345
	tabpfn_predict (Hollmann et al., 2025)	3/3	9/9	\$0.54	8 (8 _o)	583,062	0/3	0/9	\$0.03	2	23,468

Table 5: Results (using o3-mini).

🔍 Diagnose error We instruct an agent to gather information about the error in order to diagnose its root cause, and to formulate a plan to fix the error. Importantly, we do not reset the execution environment – the agent is able to check intermediate files and outputs created during tool execution.

🔧 Re-implement the tool function We perform an LLM call to re-implement the tool based on the current implementation, the error diagnosis, and the plan to fix the error.

📝 Summarise the attempt Given the conversation history of the current attempt, we instruct the LLM to summarise the attempt (*i.e.* the diagnosed error and steps taken to fix the error).

This concludes the current attempt. We reset the state of the execution environment to the environment definition. We also reset the conversation history to the state before the current attempt started (*i.e.* immediately after the initial implementation of the tool function). However, we append the summaries of all past attempts including the current one to the conversation history, and also include the current version of the code implementation. Then, we continue with the next iteration of the loop, *i.e.* go back to the start of Section A.1.3.

B Extended results

B.1 Per-task ablation results

In Tables 4 to 6, we provide detailed extended results for the ablations in a format similar to Table 2 in the main paper.

B.2 Raw unit test results

We provide the raw unit test results for all tasks in Tables 7 and 8 for the main experiments and Tables 9 to 13 for the ablations.

B.3 Transitions between tool calls

In Fig. 5, we show the transitions between tool calls by TOOLMAKER.

Task		OpenHands (Wang et al., 2024)				
		Invoc.	Tests	Cost	Actions	Tokens
Pathology	conch_extract_features (Lu et al., 2024b)	3/3	9/9	\$0.12	4	59,911
	musk_extract_features (Xiang et al., 2025)	0/3	0/6	\$0.09	3	45,985
	pathfinder_verify_biomarker (Liang et al., 2023)	0/2	4/6	\$0.09	4	49,661
	stamp_extract_features (El Nahhas et al., 2024)	✗	✗	\$0.05	2	24,799
	stamp_train_classification_model (El Nahhas et al., 2024)	0/3	0/9	\$0.13	6	66,376
	uni_extract_features (Chen et al., 2024)	0/3	0/9	\$0.11	8	82,516
Radiology	medsam_inference (Ma et al., 2024)	0/3	0/6	\$0.10	4	50,830
	nnunet_train_model (Isensee et al., 2020)	0/2	0/4	\$0.07	2	28,216
Omics	cytopus_db (Kunes et al., 2023)	0/3	0/12	\$0.08	4	49,682
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	0/3	0/15	\$0.13	5	68,098
Other	flowmap_overfit_scene (Smith et al., 2024)	✗	✗	\$0.08	4	41,152
	medsss_generate (Jiang et al., 2025)	3/3	6/6	\$0.07	3	37,926
	modernbert_predict_masked (Warner et al., 2024)	0/3	0/9	\$0.61	20	542,207
	retfound_feature_vector (Zhou et al., 2023)	✗	✗	\$0.09	4	50,891
	tabpfn_predict (Hollmann et al., 2025)	0/3	0/9	\$0.10	5	56,150

Table 6: Results (using Claude 3.5 Sonnet).

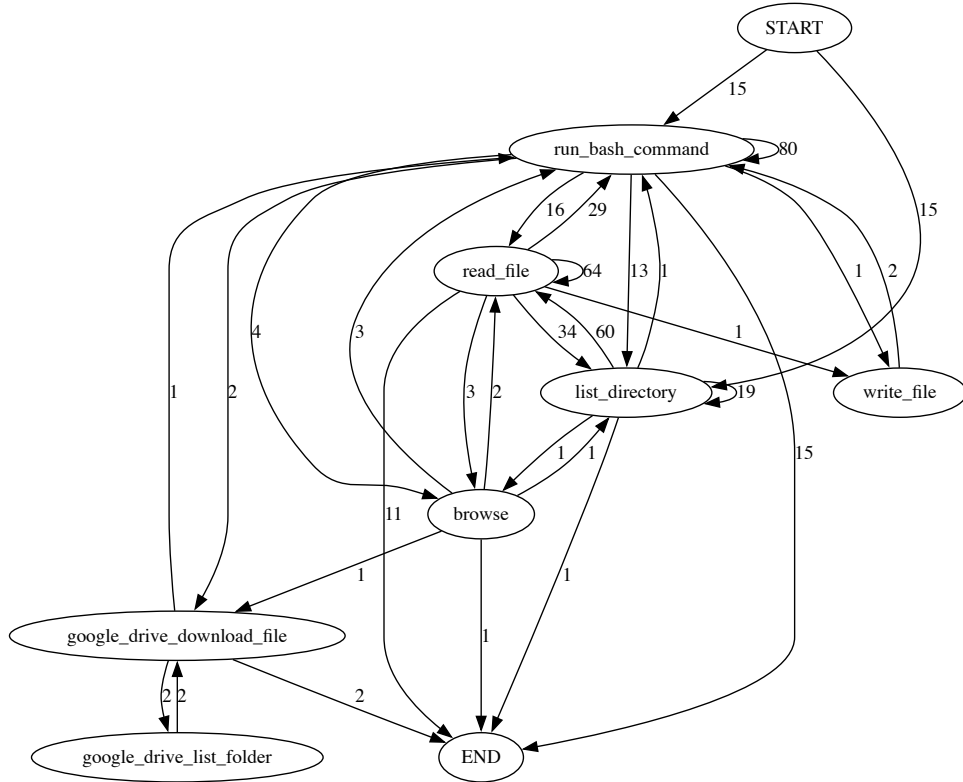


Figure 5: Transitions between tool calls by TOOLMAKER.

Category	Task	Call	Test	Passed
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	pathfinder_verify_biomarker (Liang et al., 2023)	cancer100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
Radiology	medsam_inference (Ma et al., 2024)	cancer100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	nnet_train_model (Isensee et al., 2020)	prostate	test_status	✓
			test_trained_model_exists	✓
			test_status	✓
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	protein2_with_mask	test_contact_map_values	✓
			test_sequence_representation_values	✓
			test_status	✓
Other	Flowmap_overfit_scene (Smith et al., 2024)	l1ff_fern	test_type_contact_map	✓
			test_type_sequence_representation	✓
			test_status	✓
		l1ff_orchids	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	medss_generate (Jiang et al., 2025)	motor_vehicle_accident	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	modernbert_predict_masked (Warner et al., 2024)	future_of_ai	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	retfound_feature_vector (Zhou et al., 2023)	cucumber_different_filename	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	tabpfn_predict (Hollmann et al., 2025)	diabetes	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	medss_generate (Jiang et al., 2025)	motor_vehicle_accident	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	modernbert_predict_masked (Warner et al., 2024)	future_of_ai	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	retfound_feature_vector (Zhou et al., 2023)	cucumber_different_filename	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓
	tabpfn_predict (Hollmann et al., 2025)	diabetes	test_contact_map_values	✓
			test_status	✓
			test_type_contact_map	✓

Table 7: Raw results (TOOLMAKER, without paper summary in context).

Category	Task	Call	Test	Passed
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_png	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	pathfinder_verify_biomarker (Liang et al., 2023)	cancer_fraction_score	test_pvalue_cancer_str	✓
			test_status	✓
			test_types	✓
		cancer_tum_fraction_score	test_pvalue_cancer_tum	✓
			test_status	✓
			test_types	✓
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
		cancer_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
	stamp_train_classification_model (El Nahhas et al., 2024)	cancer_braf	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
		cancer_kras	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
Radiology	medsam_inference (Ma et al., 2024)	cucumber	test_output_file	✓
			test_status	✓
			test_output_file	✓
		png	test_output_file	✓
			test_status	✓
			test_status	✓
	nnunet_train_model (Isensee et al., 2020)	prostate	test_status	✓
			test_trained_model_exists	✓
			test_status	✓
		spleen	test_trained_model_exists	✓
			test_status	✓
			test_status	✓
	uni_extract_features (Chen et al., 2024)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
Omics	cytopus_db (Kunes et al., 2023)	8_and_CD4_T	test_output_file_exists	✓
			test_status	✓
			test_types	✓
		Treg_and_plasma_and_8_naive	test_output_file_exists	✓
			test_status	✓
			test_types	✓
	esm_Fold_predict (Verkuil et al., 2022; Hie et al., 2022)	protein2_with_mask	test_contact_map_values	✓
			test_sequence_representation_values	✓
			test_status	✓
		protein2	test_type_contact_map	✓
			test_type_sequence_representation	✓
			test_contact_map_values	✓
		protein3	test_sequence_representation_values	✓
			test_status	✓
			test_type_contact_map	✓
			test_type_sequence_representation	✓
			test_contact_map_values	✓
			test_status	✓
Other	Flowmap_overfit_scene (Smith et al., 2024)	l1ff_fern	test_correct_number_of_frames	✓
			test_status	✓
			test_types_and_shapes	✓
		l1ff_orchids	test_correct_number_of_frames	✓
			test_status	✓
			test_types_and_shapes	✓
	medss_generate (Jiang et al., 2025)	motor_vehicle_accident	test_response_is_str	✓
			test_status	✓
			test_status	✓
		nslc	test_response_is_str	✓
			test_status	✓
			test_status	✓
	modernbert_predict_masked (Warner et al., 2024)	future_of_ai	test_prediction_contains_original_sentence	✓
			test_prediction	✓
			test_status	✓
		meaning_of_life	test_prediction_contains_original_sentence	✓
			test_prediction	✓
			test_status	✓
	retfound_feature_vector (Zhou et al., 2023)	cucumber_different_filename	test_shape_and_type	✓
			test_status	✓
			test_shape_and_type	✓
		jpg	test_status	✓
			test_shape_and_type	✓
			test_status	✓
tabpfn_predict (Hollmann et al., 2025)	diabetes		test_number_of_probs	✓
			test_status	✓
			test_types	✓
	heart_disease		test_number_of_probs	✓
			test_status	✓
			test_types	✓
	parkinsons		test_number_of_probs	✓
			test_status	✓
			test_types	✓
			test_status	✓
			test_status	✓
			test_types	✓

Table 8: Raw results (OpenHands (Wang et al., 2024), without paper summary in context).

Category	Task	Call	Test	Passed
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_png	test_feature_values	✓
			test_shape_and_type	✓
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_shape_and_type	✗
			test_status	✓
			test_shape_and_type	✗
		tcga_brca_patch_jpg	test_shape_and_type	✓
			test_status	✗
		tcga_brca_patch_png	test_shape_and_type	✗
			test_status	✓
	pathfinder_verify_biomarker (Liang et al., 2023)	crc_str_fraction_score	test_pvalue_crc_str	✗
			test_status	✓
			test_types	✓
		crc_tum_fraction_score	test_pvalue_crc_tum	✗
			test_status	✓
			test_types	✓
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
		crc_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
		crc	test_status	✓
			test_types	✓
	stamp_train_classification_model (El Nahhas et al., 2024)	crc_braf	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
		crc_kras	test_num_params	✓
			test_status	✓
		crc_msi	test_trained_model_exists	✓
			test_status	✓
Radiology	medsam_inference (Ma et al., 2024)	cucumber	test_output_file	✓
			test_status	✓
			test_output_file	✓
		png	test_status	✓
			test_status	✓
		nnunet_train_model (Isensee et al., 2020)	test_status	✗
			test_trained_model_exists	✗
	uni_extract_features (Chen et al., 2024)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
		tcga_brca_patch_png	test_feature_values	✓
			test_shape_and_type	✓
	Omic	cytopus_db (Kunes et al., 2023)	test_status	✓
			test_status	✓
			test_status	✓
			test_status	✓
			test_status	✓
			test_status	✓
		esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	test_contact_map_values	✗
			test_sequence_representation_values	✗
			test_status	✓
			test_type_contact_map	✓
			test_type_sequence_representation	✓
			test_type_sequence_representation	✓
Other	Flowmap_overfit_scene (Smith et al., 2024)	l1ff_fern	test_correct_number_of_frames	✓
			test_status	✓
			test_types_and_shapes	✓
		l1ff_orchids	test_correct_number_of_frames	✓
			test_status	✓
		medsss_generate (Jiang et al., 2025)	test_status	✓
			test_status	✓
	modernbert_predict_masked (Warner et al., 2024)	future_of_ai	test_prediction_contains_original_sentence	✓
			test_prediction	✓
			test_status	✓
		meaning_of_life	test_prediction_contains_original_sentence	✓
			test_prediction	✓
		walking	test_prediction_contains_original_sentence	✓
			test_prediction	✓
	retfound_feature_vector (Zhou et al., 2023)	cucumber_different_filename	test_shape_and_type	✓
			test_status	✓
			test_shape_and_type	✓
		jpg	test_status	✓
			test_shape_and_type	✓
		png	test_status	✓
			test_status	✓
	tabpfn_predict (Hollmann et al., 2025)	diabetes	test_number_of_probs	✓
			test_status	✓
			test_types	✓
		heart_disease	test_number_of_probs	✓
			test_status	✓
		parkinsons	test_status	✓
			test_types	✓

Table 9: Raw results (TOOLMAKER, with paper summary in context).

Category	Task	Call	Test	Passed
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_jpg	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_png	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_jpg	test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_png	test_shape_and_type	✗
			test_status	✗
	pathfinder_verify_biomarker (Liang et al., 2023)	crc_str_fraction_score	test_pvalue_crc_str	✗
			test_status	✓
			test_types	✓
		crc_tum_fraction_score	test_pvalue_crc_tum	✓
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_status	✓
			test_num_processed_slides	✗
			test_output_files_exist	✗
			test_output_files_have_correct_shape_and_type	✓
		crc_single	test_status	✗
			test_num_processed_slides	✗
			test_output_files_exist	✗
			test_output_files_have_correct_shape_and_type	✓
		crc	test_status	✗
			test_num_processed_slides	✗
			test_output_files_exist	✗
	stamp_train_classification_model (El Nahhas et al., 2024)	crc_braf	test_output_files_have_correct_shape_and_type	✗
			test_status	✗
			test_trained_model_exists	✗
		crc_kras	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
		crc_msi	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
	uni_extract_features (Chen et al., 2024)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_png	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
Radiology	medsam_inference (Ma et al., 2024)	cucumber	test_output_file	✗
			test_status	✗
		other_output_file	test_output_file	✗
			test_status	✗
	nnunet_train_model (Isensee et al., 2020)	png	test_output_file	✗
			test_status	✗
		prostate	test_status	✗
			test_trained_model_exists	✗
		spleen	test_status	✗
			test_trained_model_exists	✗
Omics	cytopus_db (Kunes et al., 2023)	8_and_CD4_T	test_output_file_contains_all_keys	✗
			test_output_file_exists	✗
			test_status	✗
		Treg_and_plasma_and_B_naive	test_output_file_contains_all_keys	✗
			test_output_file_exists	✗
			test_status	✗
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	leukocytes	test_status	✗
			test_output_file_contains_all_keys	✗
			test_output_file_exists	✗
		protein2	test_status	✗
			test_type	✗
			test_type_contact_map	✗
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	protein2_with_mask	test_contact_map_values	✗
			test_sequence_representation_values	✗
			test_status	✗
		protein2	test_type_contact_map	✗
			test_type_sequence_representation	✗
			test_contact_map_values	✗
		protein3	test_status	✗
			test_type_contact_map	✗
			test_type_sequence_representation	✗
		protein3	test_contact_map_values	✗
			test_status	✗
			test_type_contact_map	✗
Other	Flowmap_overfit_scene (Smith et al., 2024)	l1ff_fern	test_type_sequence_representation	✗
			test_correct_number_of_frames	✗
		l1ff_orchids	test_status	✗
			test_types_and_shapes	✗
	medsss_generate (Jiang et al., 2025)	motor_vehicle_accident	test_status	✗
			test_response_is_str	✓
		nslc	test_status	✓
			test_response_is_str	✓
	modernbert_predict_masked (Warner et al., 2024)	pediatric_rash	test_status	✓
			test_status	✓
			test_status	✓
		future_of_ai	test_prediction_contains_original_sentence	✗
			test_prediction	✗
			test_status	✗
	retfound_feature_vector (Zhou et al., 2023)	meaning_of_life	test_prediction_contains_original_sentence	✗
			test_prediction	✗
			test_status	✗
		walking	test_prediction_contains_original_sentence	✗
			test_prediction	✗
			test_status	✗
	retfound_feature_vector (Zhou et al., 2023)	cucumber_different_filename	test_shape_and_type	✗
			test_status	✗
		jpg	test_shape_and_type	✗
			test_status	✗
	tabpfn_predict (Hollmann et al., 2025)	png	test_shape_and_type	✗
			test_status	✗
		diabetes	test_number_of_probs	✓
			test_status	✓
		heart_disease	test_types	✓
			test_status	✓
	tabpfn_predict (Hollmann et al., 2025)	parkinsons	test_number_of_probs	✓
			test_status	✓
		parkinsons	test_status	✓
			test_types	✓

Table 10: Raw results (OpenHands (Wang et al., 2024), with paper summary in context).

Category	Task	Call	Test	Passed
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✗
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✗
			test_shape_and_type	✓
			test_status	✓
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_feature_values	✗
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_shape_and_type	✗
			test_status	✓
			test_shape_and_type	✓
	pathfinder_verify_biomarker (Liang et al., 2023)	crc_str_fraction_score	test_pvalue_crc_str	✗
			test_status	✓
			test_types	✓
		crc_tum_fraction_score	test_pvalue_crc_tum	✗
			test_status	✓
			test_types	✓
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
		crc_single	test_num_processed_slides	✓
			test_output_files_exist	✓
			test_output_files_have_correct_shape_and_type	✓
	stamp_train_classification_model (El Nahhas et al., 2024)	crc_braf	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
		crc_kras	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
Radiology	medsam_inference (Ma et al., 2024)	cucumber	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
		other_output_file	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
	nnunet_train_model (Isensee et al., 2020)	prostate	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
		spleen	test_num_params	✓
			test_status	✓
			test_trained_model_exists	✓
	uni_extract_features (Chen et al., 2024)	kather100k_muc	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		tcga_brca_patch_jpg	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
Omics	cytopus_db (Kunes et al., 2023)	8_and_CD4_T	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
		Treg_and_plasma_and_B_naive	test_feature_values	✓
			test_shape_and_type	✓
			test_status	✓
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	protein2_with_mask	test_output_file	✗
			test_status	✓
			test_output_file_exists	✓
		protein2	test_output_file_exists	✓
			test_status	✓
			test_types	✓
		protein3	test_output_file_contains_all_keys	✓
			test_output_file_exists	✓
			test_status	✓
		protein2	test_output_file_contains_all_keys	✓
			test_output_file_exists	✓
			test_status	✓
Other	Flowmap_overfit_scene (Smith et al., 2024)	l1ff_fern	test_contact_map_values	✗
			test_sequence_representation_values	✗
			test_status	✓
		l1ff_orchids	test_type_contact_map	✓
			test_type_sequence_representation	✓
			test_contact_map_values	✓
	medsss_generate (Jiang et al., 2025)	motor_vehicle_accident	test_sequence_representation_values	✓
			test_status	✓
			test_type_contact_map	✓
		nslcl	test_type_sequence_representation	✓
			test_response_is_str	✓
			test_status	✓
	modernbert_predict_masked (Warner et al., 2024)	pediatric_rash	test_response_is_str	✓
			test_status	✓
			test_status	✓
		future_of_ai	test_prediction_contains_original_sentence	✓
			test_prediction	✓
			test_status	✓
	retfound_feature_vector (Zhou et al., 2023)	meaning_of_life	test_prediction_contains_original_sentence	✓
			test_prediction	✓
			test_status	✓
		walking	test_prediction_contains_original_sentence	✓
			test_prediction	✓
			test_status	✓
Tabular	retfound_feature_vector (Zhou et al., 2023)	cucumber_different_filename	test_shape_and_type	✓
			test_status	✓
			test_shape_and_type	✓
		jpg	test_status	✓
			test_shape_and_type	✓
			test_status	✓
	tabpfn_predict (Hollmann et al., 2025)	png	test_shape_and_type	✓
			test_status	✓
			test_status	✓
		diabetes	test_number_of_probs	✓
			test_status	✓
			test_types	✓
		heart_disease	test_number_of_probs	✓
			test_status	✓
			test_types	✓
		parkinsons	test_number_of_probs	✓
			test_status	✓
			test_types	✓

Table 11: Raw results (TOOLMAKER, using o3-mini instead of gpt-4o).

Category	Task	Call	Test	Passed
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_jpg	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_jpg	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
	pathfinder_verify_biomarker (Liang et al., 2023)	crc_str_fraction_score	test_pvalue_crc_str	✗
			test_status	✗
			test_types	✗
		crc_tum_fraction_score	test_pvalue_crc_tum	✗
			test_status	✗
			test_types	✗
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_num_processed_slides	✗
			test_output_files_exist	✗
			test_output_files_have_correct_shape_and_type	✗
		crc_single	test_num_processed_slides	✗
			test_output_files_exist	✗
			test_output_files_have_correct_shape_and_type	✗
	stamp_train_classification_model (El Nahhas et al., 2024)	crc_braf	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
		crc_kras	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
Radiology	medsam_inference (Ma et al., 2024)	cucumber	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
		other_output_file	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
	nnunet_train_model (Isensee et al., 2020)	prostate	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
		spleen	test_num_params	✗
			test_status	✗
			test_trained_model_exists	✗
	uni_extract_features (Chen et al., 2024)	kather100k_muc	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
		tcga_brca_patch_jpg	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
Omics	cytopus_db (Kunes et al., 2023)	tcga_brca_patch_png	test_feature_values	✗
			test_shape_and_type	✗
			test_status	✗
		8_and_CD4_T	test_output_file	✗
			test_status	✗
			test_types	✗
	esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	Treg_and_plasma_and_B_naive	test_output_file_contains_all_keys	✗
			test_output_file_exists	✗
			test_status	✗
		leukocytes	test_output_file_contains_all_keys	✗
			test_output_file_exists	✗
			test_status	✗
		protein2_with_mask	test_contact_map_values	✗
			test_sequence_representation_values	✗
			test_status	✗
		protein2	test_type_contact_map	✓
			test_type_sequence_representation	✓
			test_contact_map_values	✗
Other	Flowmap_overfit_scene (Smith et al., 2024)	protein3	test_sequence_representation_values	✗
			test_status	✓
			test_type_contact_map	✓
		protein2	test_contact_map_values	✗
			test_status	✓
			test_type_contact_map	✓
	medsss_generate (Jiang et al., 2025)	protein3	test_type_sequence_representation	✓
			test_status	✓
			test_type_sequence_representation	✓
		motor_vehicle_accident	test_correct_number_of_frames	✓
			test_status	✓
			test_types_and_shapes	✓
	modernbert_predict_masked (Warner et al., 2024)	nslc	test_correct_number_of_frames	✓
			test_status	✓
			test_types_and_shapes	✓
		pediatric_rash	test_response_is_str	✗
			test_status	✗
			test_status	✗
	retfound_feature_vector (Zhou et al., 2023)	future_of_ai	test_prediction_contains_original_sentence	✗
			test_prediction	✗
			test_status	✗
		meaning_of_life	test_prediction_contains_original_sentence	✗
			test_prediction	✗
			test_status	✗
Tabular	tabpfn_predict (Hollmann et al., 2025)	walking	test_prediction_contains_original_sentence	✗
			test_prediction	✗
			test_status	✗
		cucumber_different_filename	test_shape_and_type	✗
			test_status	✗
			test_shape_and_type	✗
		jpg	test_shape_and_type	✗
			test_status	✗
			test_shape_and_type	✗
		png	test_shape_and_type	✗
			test_status	✗
			test_status	✗
		diabetes	test_number_of_probs	✗
			test_status	✗
			test_types	✗
		heart_disease	test_number_of_probs	✗
			test_status	✗
			test_types	✗
		parkinsons	test_number_of_probs	✗
			test_status	✗
			test_types	✗
			test_status	✗
			test_status	✗
			test_types	✗

Table 12: Raw results (OpenHands (Wang et al., 2024), using o3-mini instead of gpt-4o).

Category	Task	Call	Test	Passed		
Pathology	conch_extract_features (Lu et al., 2024b)	kather100k_muc	test_feature_values	✓		
			test_shape_and_type	✓		
			test_status	✓		
		tcga_brca_patch_jpg	test_feature_values	✓		
			test_shape_and_type	✓		
			test_status	✓		
		tcga_brca_patch_png	test_feature_values	✓		
			test_shape_and_type	✓		
			test_status	✓		
	musk_extract_features (Xiang et al., 2025)	kather100k_muc	test_shape_and_type	✗		
			test_status	✗		
		tcga_brca_patch_jpg	test_shape_and_type	✗		
			test_status	✗		
		tcga_brca_patch_png	test_shape_and_type	✗		
			test_status	✗		
	pathfinder_verify_biomarker (Liang et al., 2023)	crc_str_fraction_score	test_pvalue_crc_str	✗		
			test_status	✓		
			test_types	✓		
		crc_tum_fraction_score	test_pvalue_crc_tum	✗		
			test_status	✓		
			test_types	✓		
	stamp_extract_features (El Nahhas et al., 2024)	brca_single	test_num_processed_slides	✗		
			test_output_files_exist	✗		
			test_output_files_have_correct_shape_and_type	✗		
			test_status	✗		
		crc_single	test_num_processed_slides	✗		
			test_output_files_exist	✗		
			test_output_files_have_correct_shape_and_type	✗		
			test_status	✗		
		crc	test_num_processed_slides	✗		
			test_output_files_exist	✗		
			test_output_files_have_correct_shape_and_type	✗		
			test_status	✗		
	stamp_train_classification_model (El Nahhas et al., 2024)	crc_braf	test_num_params	✗		
			test_status	✗		
			test_trained_model_exists	✗		
		crc_kras	test_num_params	✗		
			test_status	✗		
			test_trained_model_exists	✗		
		crc_msi	test_num_params	✗		
			test_status	✗		
			test_trained_model_exists	✗		
	uni_extract_features (Chen et al., 2024)	kather100k_muc	test_feature_values	✗		
			test_shape_and_type	✗		
			test_status	✗		
		tcga_brca_patch_jpg	test_feature_values	✗		
			test_shape_and_type	✗		
test_status			✗			
tcga_brca_patch_png		test_feature_values	✗			
		test_shape_and_type	✗			
		test_status	✗			
Radiology	medsam_inference (Ma et al., 2024)	cucumber	test_output_file	✗		
			test_status	✗		
		other_output_file	test_output_file	✗		
			test_status	✗		
		png	test_output_file	✗		
	test_status		✗			
	nnunet_train_model (Isensee et al., 2020)		prostate	test_status	✗	
		test_trained_model_exists		✗		
		spleen	test_status	✗		
			test_trained_model_exists	✗		
		Omics	cytopus_db (Kunes et al., 2023)	8_and_CD4_T	test_output_file_contains_all_keys	✗
					test_output_file_exists	✗
				Treg_and_plasma_and_B_naive	test_output_file_contains_all_keys	✗
					test_output_file_exists	✗
			leukocytes	test_status	✗	
test_types				✗		
esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	protein2_with_mask			test_contact_map_values	✗	
				test_sequence_representation_values	✗	
		test_status	✗			
		test_type_contact_map	✗			
	protein2	test_type_sequence_representation	✗			
		test_contact_map_values	✗			
		test_sequence_representation_values	✗			
		test_status	✗			
esm_fold_predict (Verkuil et al., 2022; Hie et al., 2022)	protein3	test_type_contact_map	✗			
		test_type_sequence_representation	✗			
		test_contact_map_values	✗			
		test_sequence_representation_values	✗			
	protein3	test_status	✗			
		test_type_contact_map	✗			
		test_type_sequence_representation	✗			
		test_status	✗			
Other	flowmap_overfit_scene (Smith et al., 2024)	l1ff_fern	test_correct_number_of_frames	✗		
			test_status	✗		
		l1ff_orchids	test_types_and_shapes	✗		
			test_status	✗		
	medsss_generate (Jiang et al., 2025)	motor_vehicle_accident	test_types_and_shapes	✗		
			test_response_is_str	✓		
		nslc	test_status	✓		
			test_response_is_str	✓		
	modernbert_predict_masked (Warner et al., 2024)	pediatric_rash	test_status	✓		
			test_response_is_str	✓		
		future_of_ai	test_prediction_contains_original_sentence	✗		
			test_prediction	✗		
	retfound_feature_vector (Zhou et al., 2023)	meaning_of_life	test_prediction_contains_original_sentence	✗		
			test_prediction	✗		
		walking	test_status	✗		
			test_prediction_contains_original_sentence	✗		
	tabpfn_predict (Hollmann et al., 2025)	cucumber_different_filename	test_prediction	✗		
			test_status	✗		
		jpg	test_shape_and_type	✗		
			test_status	✗		
	tabpfn_predict (Hollmann et al., 2025)	png	test_shape_and_type	✗		
			test_status	✗		
		diabetes	test_shape_and_type	✗		
			test_status	✗		
	tabpfn_predict (Hollmann et al., 2025)	heart_disease	test_number_of_probs	✗		
			test_status	✗		
		parkinsons	test_types	✗		
			test_number_of_probs	✗		
parkinsons	test_status	✗				
	test_types	✗				

Table 13: Raw results (OpenHands (Wang et al., 2024), using Claude 3.5 Sonnet instead of gpt-4o).

C TM-BENCH

Below are the complete task definitions for all tasks in our benchmark, TM-BENCH.

C.1 Pathology

conch_extract_features

Perform feature extraction on an input image using CONCH.


Arguments:

- `input_image (str)`: Path to the input image
Example: `'/mount/input/TUM/TUM-TCGA-ACRLPPQE.tif'`

Returns:

- `features (list)`: The feature vector extracted from the input image, as a list of floats

 <https://github.com/mahmoodlab/CONCH>

 Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. 2024b. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874

musk_extract_features

Perform feature extraction on an input image using the vision part of MUSK.


Arguments:

- `input_image (str)`: Path to the input image
Example: `'/mount/input/TUM/TUM-TCGA-ACRLPPQE.tif'`

Returns:

- `features (list)`: The feature vector extracted from the input image, as a list of floats

 <https://github.com/lilab-stanford/MUSK>

 Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang, and Ruijiang Li. 2025. A vision-language foundation model for precision oncology. *Nature*

pathfinder_verify_biomarker

Given WSI probability maps, a hypothesis of a potential biomarker, and clinical data, determine (1) whether the potential biomarker is significant for patient prognosis, and (2) whether the potential biomarker is independent among already known biomarkers.


Arguments:

- **heatmaps (str):** Path to the folder containing the numpy array (`.numpy`) files, which contains the heatmaps of the trained model (each heatmap is HxWxC where C is the number of classes)
Example: `/mount/input/TCGA_CRC`
- **hypothesis (str):** A python file, which contains a function `def hypothesis_score(prob_map_path: str) -> float` which expresses a mathematical model of a hypothesis of a potential biomarker. For a particular patient, the function returns a risk score.
Example: `/mount/input/mus_fraction_score.py`
- **clini_table (str):** Path to the CSV file containing the clinical data
Example: `/mount/input/TCGA_CRC_info.csv`
- **files_table (str):** Path to the CSV file containing the mapping between patient IDs (in the PATIENT column) and heatmap filenames (in the FILENAME column)
Example: `/mount/input/TCGA_CRC_files.csv`
- **survival_time_column (str):** The name of the column in the clinical data that contains the survival time
Example: `OS.time`
- **event_column (str):** The name of the column in the clinical data that contains the event (e.g. death, recurrence, etc.)
Example: `vital_status`
- **known_biomarkers (list):** A list of known biomarkers. These are column names in the clinical data.
Example: `['MSI']`

Returns:

- **p_value (float):** The p-value of the significance of the potential biomarker
- **hazard_ratio (float):** The hazard ratio for the biomarker

 <https://github.com/LiangJunhao-THU/PathFinderCRC>

 Junhao Liang, Weisheng Zhang, Jianghui Yang, Meilong Wu, Qionghai Dai, Hongfang Yin, Ying Xiao, and Lingjie Kong. 2023. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nature Machine Intelligence*, 5(4):408–420

stamp_extract_features

Perform feature extraction using CTransPath with STAMP on a set of whole slide images, and save the resulting features to a new folder.


Arguments:

- `output_dir (str)`: Path to the output folder where the features will be saved
Example: `'/mount/output/TCGA-BRCA-features'`
- `slide_dir (str)`: Path to the input folder containing the whole slide images
Example: `'/mount/input/TCGA-BRCA-SLIDES'`

Returns:

- `num_processed_slides (int)`: The number of slides that were processed

 <https://github.com/KatherLab/STAMP>

 Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligerio, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. 2024. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nature Protocols*

stamp_train_classification_model

Train a model for biomarker classification. You will be supplied with the path to the folder containing the whole slide images, alongside a path to a CSV file containing the training labels.


Arguments:

- `slide_dir (str)`: Path to the folder containing the whole slide images
Example: `'/mount/input/TCGA-BRCA-SLIDES'`
- `clini_table (str)`: Path to the CSV file containing the clinical data
Example: `'/mount/input/TCGA-BRCA-DX_CLINI.xlsx'`
- `slide_table (str)`: Path to the CSV file containing the slide metadata
Example: `'/mount/input/TCGA-BRCA-DX_SLIDE.csv'`
- `target_column (str)`: The name of the column in the clinical data that contains the target labels
Example: `'TP53_driver'`
- `trained_model_path (str)`: Path to the `*.pkl` file where the trained model should be saved by this function
Example: `'/mount/output/STAMP-BRCA-TP53-model.pkl'`

Returns:

- `num_params (int)`: The number of parameters in the trained model

 <https://github.com/KatherLab/STAMP>

 Omar S. M. El Nahhas, Marko van Treeck, Georg Wölflein, Michaela Unger, Marta Ligerio, Tim Lenz, Sophia J. Wagner, Katherine J. Hewitt, Firas Khader, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. 2024. From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. *Nature Protocols*

uni_extract_features

Perform feature extraction on an input image using UNI.


Arguments:

- `input_image (str)`: Path to the input image
Example: `'/mount/input/TUM/TUM-TCGA-ACRLPPQE.tif'`

Returns:

- `features (list)`: The feature vector extracted from the input image, as a list of floats

 <https://github.com/mahmoodlab/UNI>

 Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862

C.2 Radiology

medsam_inference


Use the trained MedSAM model to segment the given abdomen CT scan.

Arguments:

- `image_file (str)`: Path to the abdomen CT scan image.
Example: `'/mount/input/my_image.jpg'`
- `bounding_box (list)`: Bounding box to segment (list of 4 integers).
Example: `[25, 100, 155, 155]`
- `segmentation_file (str)`: Path to where the segmentation image should be saved.
Example: `'/mount/output/segmented_image.png'`

Returns: *empty dict*

 <https://github.com/bowang-lab/MedSAM>

 Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1)

nnunet_train_model


Train a nnUNet model from scratch on abdomen CT scans. You will be provided with the path to the dataset, the nnUNet configuration to use, and the fold number to train the model on.

Arguments:

- `dataset_path` (str): The path to the dataset to train the model on (contains dataset.json, imagesTr, imagesTs, labelsTr)
Example: `'/mount/input/Task02_Heart'`
- `unet_configuration` (str): The configuration of the UNet to use for training. One of `'2d'`, `'3d_fullres'`, `'3d_lowres'`, `'3d_cascade_fullres'`
Example: `'3d_fullres'`
- `fold` (int): The fold number to train the model on. One of 0, 1, 2, 3, 4.
Example: `0`
- `output_folder` (str): Path to the folder where the trained model should be saved
Example: `'/mount/output/trained_model'`

Returns: *empty dict*

 <https://github.com/MIC-DKFZ/nnUNet>

 Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. 2020. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211

C.3 Omics

cytopus_db

Initialize the Cytopus KnowledgeBase and generate a JSON file containing a nested dictionary with gene set annotations organized by cell type, suitable for input into the Spectra library.


Arguments:

- `celltype_of_interest` (list): List of cell types for which to retrieve gene sets
Example: `['B_memory', 'B_naive', 'CD4_T', 'CD8_T', 'DC', 'ILC3', 'MDC', 'NK', 'Treg', 'gdT', 'mast', 'pDC', 'plasma']`
- `global_celltypes` (list): List of global cell types to include in the JSON file.
Example: `['all-cells', 'leukocyte']`
- `output_file` (str): Path to the file where the output JSON file should be stored.
Example: `'/mount/output/Spectra_dict.json'`

Returns:

- `keys` (list): The list of keys in the produced JSON file.

 <https://github.com/wallet-maker/cytopus>

 Russell Z. Kunes, Thomas Walle, Max Land, Tal Nawy, and Dana Pe'er. 2023. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*, 42(7):1084–1095

esm_fold_predict

Generate the representation of a protein sequence and the contact map using Facebook Research's pretrained esm2_t33_650M_UR50D model.


Arguments:

- `sequence (str)`: Protein sequence to for which to generate representation and contact map.
Example: 'MKTVRQERLKSIVRILERSKEPVSGAQLAEELSVSRQVIVQDIAYLRSLGYNIVATPRGYVLAGG'

Returns:

- `sequence_representation (list)`: Token representations for the protein sequence as a list of floats, i.e. a 1D array of shape L where L is the number of tokens.
- `contact_map (list)`: Contact map for the protein sequence as a list of list of floats, i.e. a 2D array of shape LxL where L is the number of tokens.

 <https://github.com/facebookresearch/esm>

 Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. 2022. [Language models generalize beyond natural proteins](#). *Preprint*, bioRxiv:2022.12.21.521521

 Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. 2022. [A high-level programming language for generative protein design](#). *Preprint*, bioRxiv:2022.12.21.521526

C.4 Other

retfound_feature_vector

Extract the feature vector for the given retinal image using the RETFound pretrained vit_large_patch16 model.


Arguments:

- `image_file (str)`: Path to the retinal image.
Example: `'/mount/input/retinal_image.jpg'`

Returns:

- `feature_vector (list)`: The feature vector for the given retinal image, as a list of floats.

 https://github.com/rmaphoh/RETFound_MAE

 Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Naomi Allen, John E. J. Gallacher, Thomas Littlejohns, Tariq Aslam, Paul Bishop, Graeme Black, Panagiotis Sergouniotis, Denize Atan, Andrew D. Dick, Cathy Williams, Sarah Barman, Jenny H. Barrett, Sarah Mackie, Tasanee Braithwaite, Roxana O. Carare, Sarah Ennis, Jane Gibson, Andrew J. Lotery, Jay Self, Usha Chakravarthy, Ruth E. Hogg, Euan Paterson, Jayne Woodside, Tunde Peto, Gareth McKay, Bernadette McGuinness, Paul J. Foster, Konstantinos Balaskas, Anthony P. Khawaja, Nikolas Pontikos, Jugnoo S. Rahi, Gerassimos Lascaratos, Praveen J. Patel, Michelle Chan, Sharon Y. L. Chua, Alexander Day, Parul Desai, Cathy Egan, Marcus Fruttiger, David F. Garway-Heath, Alison Hardcastle, Sir Peng T. Khaw, Tony Moore, Sobha Sivaprasad, Nicholas Strouthidis, Dhanes Thomas, Adnan Tufail, Ananth C. Viswanathan, Bal Dhillon, Tom Macgillivray, Cathie Sudlow, Veronique Vitart, Alexander Doney, Emanuele Trucco, Jeremy A. Guggenheim, James E. Morgan, Chris J. Hammond, Katie Williams, Pirro Hysi, Simon P. Harding, Yalin Zheng, Robert Luben, Phil Luthert, Zihan Sun, Martin McKibbin, Eoin O'Sullivan, Richard Oram, Mike Weedon, Chris G. Owen, Alicja R. Rudnicka, Naveed Sattar, David Steel, Irene Stratton, Robyn Tapp, Max M. Yates, Axel Petzold, Savita Madhusudhan, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163

medsss_generate

Given a user message, generate a response using the MedSSS_Policy model.


Arguments:

- `user_message (str)`: The user message.
Example: `'How to stop a cough?'`

Returns:

- `response (str)`: The response generated by the model.

 <https://github.com/pixas/MedSSS>

 Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds³: Towards medical small language models with self-evolved slow thinking. *Preprint*, arXiv:2501.12051

modernbert_predict_masked

Given a masked sentence string, predict the original sentence using the pretrained ModernBERT-base model on CPU.


Arguments:

- `input_string (str)`: The masked sentence string. The masked part is represented by "[MASK]".
Example: 'Paris is the [MASK] of France.'

Returns:

- `prediction (str)`: The predicted original sentence

 <https://github.com/AnswerDotAI/ModernBERT>

 Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.](#) *Preprint*, arXiv:2412.13663

flowmap_overfit_scene

Overfit FlowMap on an input scene to determine camera extrinsics for each frame in the scene.


Arguments:

- `input_scene (str)`: Path to the directory containing the images of the input scene (just the image files, nothing else)
Example: '/mount/input/llff_flower'

Returns:

- `n (int)`: The number of images (frames) in the scene
- `camera_extrinsics (list)`: The camera extrinsics matrix for each of the n frames in the scene, must have a shape of nx4x4 (as a nested python list of floats)

 <https://github.com/dcharatan/flowmap>

 Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. 2024. [Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent.](#) *Preprint*, arXiv:2404.15259

tabpfn_predict

Train a predictor using TabPFN on a tabular dataset. Evaluate the predictor on the test set.


Arguments:

- `train_csv` (str): Path to the CSV file containing the training data
Example: `'/mount/input/breast_cancer_train.csv'`
- `test_csv` (str): Path to the CSV file containing the test data
Example: `'/mount/input/breast_cancer_test.csv'`
- `feature_columns` (list): The names of the columns to use as features
Example: `['mean radius', 'mean texture', 'mean perimeter', 'mean area', 'mean smoothness', 'mean compactness', 'mean concavity', 'mean concave points', 'mean symmetry', 'mean fractal dimension', 'radius error', 'texture error', 'perimeter error', 'area error', 'smoothness error', 'compactness error', 'concavity error', 'concave points error', 'symmetry error', 'fractal dimension error', 'worst radius', 'worst texture', 'worst perimeter', 'worst area', 'worst smoothness', 'worst compactness', 'worst concavity', 'worst concave points', 'worst symmetry', 'worst fractal dimension']`
- `target_column` (str): The name of the column to predict
Example: `'target'`

Returns:

- `roc_auc` (float): The ROC AUC score of the predictor on the test set
- `accuracy` (float): The accuracy of the predictor on the test set
- `probs` (list): The probabilities of the predictor on the test set, as a list of floats (one per sample in the test set)

 <https://github.com/PriorLabs/TabPFN>

 Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326

D TOOLMAKER prompts

Install Repository System Instructions

You're a diligent software engineer AI. You can't see, draw, or interact with a browser, but you can read and write files, and you can run commands, and you can think. The user will specify a task for you to complete. You likely need to run several actions in order to complete the task. You will only be able to execute a single action at a time.

Use the tools (actions) that are at your disposal. Each time you invoke a tool, provide a one-sentence summary of why you are invoking it and what you expect to accomplish by invoking it.

Your working directory is `{LOCAL_WORKSPACE_DIR!s}`.

You will continue the process of invoking tools until you have completed the task.

Install Repository User Instructions

Clone and locally set up the `definition.repo.name` repository from GitHub. Follow these steps:

1. Git clone the repository `definition.repo.info()` into the directory `'install_path'`.
2. Check the README (find it if it is not in the root directory) and closely follow the recommended instructions to set up the entire repository correctly for the user.
3. Follow the instructions in the README to correctly set up the repository for the user. Perform any necessary installations, configurations, downloads or setups as described. If the repository is in Python, prefer using `'pip'` as opposed to `conda`, `virtualenv`, or similar. Install the repository and its dependencies globally. Do not use Docker or similar container tools (even if the README suggests it); instead, install the repository and its dependencies globally.
4. Make sure that you complete every step, so that a user could directly use this repository without the need to do further setups, installations or downloads. This includes downloading any necessary pretrained models. However, do NOT download any datasets. If you encounter any issues, try to solve them.

`environment_variables_prompt(definition.repo)`

You should set up the repository in such a way that it can be used to implement the following task later on: `<intended_task> <description> definition.description </description> <arguments>`
"

```
n".join(f"name          (arg.type):          arg.description"          for          name,
arg          in          definition.arguments.items())          </arguments>          <returns>
definition.description_of_returns()          </returns>          </intended_task>          IMPORTANT:
```

Your task right now is to only set up the repository, NOT implement this task.

When you are done, provide a brief summary of what you did and what you accomplished, as well as the absolute path to the cloned and installed repository.

Explore Repository User Instructions

Background The repository 'definition.repo.name' is fully set up and installed at 'get_local_install_path(definition.repo)!s'. We need to wrap a specific functionality from this repository into a standalone python function, that can be called independently. This function will be called 'definition.name', and it is described as follows: <description> definition.description </description>

The function will have the following arguments: <arguments> "

n".join(f"<argument>arg!r</argument>" for arg in definition.arguments) </arguments>

installed_repository_summary

High-level approach In order to implement this function, you will follow these steps: 1. Explore the repository to gather all relevant information needed to write the plan. 2. Write a plan for the body/implementation of the function. This plan should be in the form of very high-level pseudo-code, that describes how the function will work. 3. Write the function, based on the plan.

Task Right now, you are at step 1: Explore the repository to gather all relevant information needed to write the plan. This step is very important, and you must be thorough because you will rely on this information later when implementing the function. To gather all relevant information needed for the implementation, explore the repository, but only look at relevant files. Use the tools at your disposal to read files, list directories, search, etc. HINT 1: If the repository contains a README file, that is often a good starting point. Note that there may be zero or more README files. Always check for README files, and prefer to follow the instructions therein. HINT 2: If the repository provides a command line interface, prefer to invoke that via subprocess, rather than calling the underlying python functions. Only as a last resort, wrap python functions. HINT 3: Do NOT attempt to read image files, audio files, etc. Do not unnecessarily read files that are not relevant for the task. **However, make sure to read ALL files (e.g. documentation, code, configuration files, etc.) that are necessary in order to implement the function. It should be possible to implement the function based only on the plan and the files you read!** **You should read relevant code files in order to understand how the functionality you are wrapping is implemented. If you are planning to wrap specific functions, be sure to read the relevant code in order to understand what the input and output arguments/formats are. This is especially relevant if the function you are wrapping produces output files that you will need to read.** Do NOT write the function yet. Your task is specifically to explore the repository to gather information.

Once you have gathered ALL relevant information, respond with a one-paragraph summary of what you found.

Remember, the function should do the following: <description> definition.description </description>

As such, the signature of the function will be: "python definition!s "

Explore Repository User Instructions

Using the information you gathered previously, your task is now to write an outline (plan) for the body/implementation of the function. This plan should be in the form of very high-level pseudo-code, that describes how the function will work. It should be a numbered list of steps, each of which describes what you will do in that step. Respond with just this list of steps, nothing else.

Remember, the function should do the following: 'definition.description'

As such, the signature of the function will be: "python definition!s "

Summarize Problem User Instructions

Provide a one-paragraph summary of the most recent problem that occurred, your diagnosis of it, and how you attempted to fix it with this code change. Be specific. Include any file paths and other details that are relevant to the problem/solution. Your summary should contain all information needed to implement the fix, and include the key insights/observations made for diagnosing the problem. Begin your response with "The problem was..."

Summarize Problem User Instructions

Now that you have identified the problem as well as a plan to fix the function, you need to write the updated implementation of the function. Remember, the function is called 'definition.name', and it is described as follows: 'definition.description'. The function will have the following arguments: <arguments> "

n".join(f"<argument>arg!r</argument>" for arg in definition.arguments) </arguments>

As such, the signature of the function will be: "python definition!s "

Your task is now to write the Python function. To do so, use the information you gathered above to fix the function.

coding_instructions(definition)

As a reminder, the current draft of the function is: "python code_draft "

Remember, your diagnosis is: <diagnosis> diagnosis.diagnosis </diagnosis>

And your plan to fix the issue is: <plan> diagnosis.plan </plan>

Respond with the updated function code only, without any other text.

Coding Instructions

You **must** output a valid, standalone python function that is callable without any modification by a user. The requirements for the code are: 1. Import the required modules/libraries. 2. You are only allowed to write a single python function. It must start with 'def ...' and end with 'return ...'. 3. You are not allowed to output free texts, test code for the function or anything outside of the function definition. 4. The function needs to be a standalone function that can be called independently. 5. Make sure all required imports are included in the function. 6. The function must perform the task you are given. As a reminder, the task is: 'definition.description'. 7. Make sure the function accepts all required parameters as inputs. 8. The function must have type hints and a docstring. 9. The function must be named exactly 'definition.name'. 10. The function must be a valid python function, that can be executed by a python interpreter.

environment_variables_prompt(definition.repo)

Additional instructions: * Write the function in such a way that it can easily be debugged later. This means that you should include a lot of print statements for logging purposes. Especially for long-running tasks, it is important to print the progress periodically. * When catching exceptions in the code (with 'try' and 'except'), make sure to output the entire stack trace to stderr, so that it can be used to diagnose any issues, e.g. using 'traceback.format_exc()'. * When running commands and scripts (e.g. using subprocesses), make sure to stream the stdout and stderr to the parent process, so that it can be used to diagnose any issues. Use the utility function 'run_and_stream_command' provided by the 'subprocess_utils' module. It accepts the same arguments as 'subprocess.Popen', and returns a tuple '(return_code, output)' (return_code is an integer, output is a string containing stdout and stderr combined). The 'run_and_stream_command' automatically handles the streaming of stdout and stderr to the parent process. Be sure to appropriately set the 'cwd' argument. Example usage: “python from subprocess_utils import run_and_stream_command # you must import this return_code, output = run_and_stream_command("echo hello && echo world", shell=True, env="MY_VAR": "my_value", cwd="/workspace/my_project") # shell=True is the default “ * Make sure that you do not run interactive commands. If some python function that you are calling itself runs interactive commands, try and find a way to avoid calling that function. If, as a last resort, you cannot avoid calling that function, mock/patch the external interactive function to ensure that it does not run interactive commands. * Always prefer to import existing functions into the function you are writing, or run existing scripts/modules (e.g. via the subprocess functionality described above), instead of writing your own implementations. Only if this does not work, or there is no existing function that can be imported, write your own implementation.

Implement Function User Instructions

Now that you have identified the plan for the implementation, you need to write the actual implementation of the function. This needs to be a standalone python function, that can be called independently. This function will be called 'definition.name', and it is described as follows: 'definition.description'. The function will have the following arguments: "

n".join((" - " + repr(arg)) for arg in definition.arguments)

As such, the signature of the function will be: “python definition!s “

Your task is now to write the Python function. To do so, follow the plan you identified earlier for the implementation: <plan> plan </plan>

coding_instructions(definition)

Remember, you should use the repository 'definition.repo.name' (installed at 'get_local_install_path(definition.repo)!s') to complete the task. Finally, ensure your function is ready-to-use without any modifications by a user. In many cases, wrapping an existing function, script or module in a subprocess is enough. Respond with the code of the function only, without any other text.

Diagnose User Instructions

Your initial code implementation did not work. This was attempt number `len(problem_summaries)` to fix the problem.

Here is a summary of the previous problems, and your attempts to fix them. Keep this in mind as we proceed, and avoid repeating the same mistakes. `<summaries>` '

`n'.join(f'<summary number=i>summary</summary>' for i, summary in enumerate(problem_summaries)) </summaries>`

The current version of your code (after `len(problem_summaries)` attempts) is below. **IMPORTANT:** this is the most up-to-date version of your code, so focus on it when diagnosing the problem. “python code “

Upon executing this updated function, I received another error. As a diligent software engineer AI, your task is now to diagnose the issue and fix the function. You can't see, draw, or interact with a browser, but you can read and write files, and you can run commands, and you can think. You will be provided with the stdout and stderr from the function execution. First, use your tools (e.g. running commands, listing directories, reading files, etc.) to gather information about the issue, in order to diagnose it. Specifically, try to find out the root cause of the issue. Often, this requires reading relevant code files in the repository to understand how the problem occurred, and if any assumptions you made in your implementation of the function are incorrect. Then, formulate a plan to fix the issue, and finally respond with that plan.

NOTE: The plan you write should be the immediate plan to modify the function to fix the issue. After you provide the plan, you will then be asked to provide the code to implement the plan, and I will execute that code. I will then give you the output of the code execution, and you will be asked to provide a new plan to fix the new issue. Therefore, if after exploring the codebase you still don't know what's wrong, your plan should be to modify the function to provide more logging to help you diagnose the problem next time it is executed.

IMPORTANT: While you are able to interact with the environment (writing files, running commands, etc.), any changes you make will be lost when the function is executed again, as the environment will be reset. Therefore, use this opportunity only to gather information about the issue, and not to fix it. **HINT:** After gathering information, you may decide to use a slightly different approach to fix the issue – if this is the case, include this in your plan! **HINT:** Always prefer importing code from the repository, rather than implementing it yourself. The information you gather may contain code that you can import to fix the issue.

Output (stdout and stderr) of the function execution: `<output> output.stdout </output>`

Initial assessment why the function call was not successful: `<assessment> assessment.reasoning </assessment>`

As mentioned above, your immediate task is to diagnose the issue, and formulate a plan to fix it.

Function Execution Assessment User Instructions

I executed the function you wrote. Based on the output and returned result, assess whether the function call was successful or not. Specifically, you should assess whether the function performed the task it was supposed to perform. Also make sure that the returned result is plausible and matches the stdout/stderr output logs, if applicable. As a reminder, the task is the following:

<task_description> definition.description </task_description>

Description of expected result: <expected_result_description>
definition.description_of_returns() </expected_result_description>

Returned result: <result> truncate_str(repr(output.result), max_length=10000) </result>

Output (stdout and stderr) of the function execution: <output> truncate_str(output.stdout, max_length=10000) </output>

****IMPORTANT:** You must also ensure that the returned result itself is correct. This includes ensuring that the result dict contains the correct keys and values, and that the values have the correct types and shapes! If any of these are incorrect, the function call is NOT successful! If this is the case, include this in your reasoning.**** """,**

E OpenHands baseline prompt

Below is the prompt used for the OpenHands baseline (Wang et al., 2024).

Your task is to create a tool from the repository `definition.repo.name` which implements the function `definition.name` to perform the following task: `definition.description`. While you may perform any necessary installations, configurations, downloads or setups, your deliverables are the following two files: 1. A bash script, named `/workspace/install.sh` that will install all necessary dependencies for the tool to run. 2. A Python file, named `/workspace/code.py` that will contain the code for the tool.

Part 1: Install the repository Clone and locally set up the `definition.repo.name` repository from GitHub. Follow these steps: 1. Git clone the repository `definition.repo.info()`. 2. Check the README (find it if it is not in the root directory) and closely follow the recommended instructions to set up the entire repository correctly for the user. 3. Follow the instructions in the README to correctly set up the repository for the user. Perform any necessary installations, configurations, downloads or setups as described. If the repository is in Python, prefer using `'pip'` as opposed to `conda`, `virtualenv`, or similar. Install the repository and its dependencies globally. 4. Make sure that you complete every step, so that a user could directly use this repository without the need to do further setups, installations or downloads. This includes downloading any necessary models. However, do NOT download any datasets. If you encounter any issues, try to solve them.

`environment_variables_prompt(definition.repo)`

Part 2: Implement the tool function You need to implement a standalone python function, that can be called independently. This function will be called `definition.name`, and it is described as follows: `definition.description`. The function will have the following arguments: "

`n".join((f"- arg_name (arg.type): arg.description") for arg_name, arg in definition.arguments.items())`

As such, the signature of the function will be: `python definition!s` "You ****must**** output a valid, standalone python function that is callable without any modification by a user. The requirements for the code are: 1. Import the required modules/libraries. 2. You are only allowed to write a single python function. It must start with `'def ...'` and end with `'return ...'`. 3. You are not allowed to output free texts, test code for the function or anything outside of the function definition. 4. The function needs to be a standalone function that can be called independently. 5. Make sure all required imports are included in the function. 6. The function must perform the task you are given. As a reminder, the task is: `definition.description`. 7. Make sure the function accepts all required parameters as inputs. 8. The function must have type hints and a docstring. 9. The function must be named exactly `definition.name`. 10. The function must be a valid python function, that can be executed by a python interpreter.

`environment_variables_prompt(definition.repo)`

Remember, you should use the repository `definition.repo.name` to complete the task. Finally, ensure your function is ready-to-use without any modifications by a user. In many cases, wrapping an existing function, script or module in a subprocess is enough. Note: It may be useful to run the function with the following example invocation to test it: `python3 from code import definition.name definition.name(", ".join(f"k=v!r" for k, v in definition.example.arguments.items()))` "

IMPORTANT: - The only two files that you need to produce are `/workspace/install.sh` and `/workspace/code.py` (though you may create other files as well, or install additional dependencies in the process). - You may use any tools at your disposal to complete the task. - From within a fresh environment (i.e. a fresh Docker image of `python:3.12`) that contains the `/workspace` directory which is empty except for your `'install.sh'` and `'code.py'` files, it should be possible to run the `'install.sh'` script, and then run the `'code.py'` file, without any additional prior installations or dependencies. - The `'code.py'` file should NOT contain any imports at the top of the file. The first line of the file should be the function signature (of the `definition.name` function). In the body of the function, you may import any necessary modules.