Supplementary Information for An AI agent for treatment reasoning over a biomedical tool universe

Shanghua Gao[1], Ayush Noori[1,2,3], Richard Zhu[1], Curtis Ginder[1,4], Zhenglun Kong[1], Xiaorui Su[1], Justin Kauffman[5], Benjamin S. Glicksberg[5,6,7], Joshua Lampert[5,6,8], Ankit Sakhuja[5,9,10], Ashwin Sawant[5,9,11], ATHENA-R1 Evaluation Consortium[12], David A. Clifton[2,13], Noa Dagan[3,14,15], Ran Balicer[3,14,16], Marinka Zitnik[1,3,17,18,19,†]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2]Department of Engineering Science, University of Oxford, Oxford, UK
[3]The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Boston, MA, USA
[4]Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA
[5]The Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6]The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai and Mount Sinai Health System, New York City, NY, USA
[7]Mindich Child Health and Development Institute and the Departments of Pediatrics and Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[8]Mount Sinai Fuster Heart Hospital, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[9]Mount Sinai AI Assurance Lab, Mount Sinai Health System, New York, NY, USA
[10]Institute for Critical Care Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[11]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[12]ATHENA-R1 Evaluation Group (the list of members and their affiliations appears in the Supplementary Information)
[13]Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou, Jiangsu, China
[14]Clalit Research Institute, Innovation Division, Clalit Health Services, Ramat Gan, Israel
[15]Faculty of Computer and Information Science, Ben Gurion University of the Negev, Be'er Sheva, Israel
[16]Faculty of Health Sciences, School of Public Health, Ben Gurion University of the Negev, Be'er Sheva, Israel
[17]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA
[18]Broad Institute of MIT and Harvard, Cambridge, MA
[19]Harvard Data Science Initiative, Cambridge, MA
[†]Correspondence: marinka@hms.harvard.edu

This PDF file includes:

Supplementary Notes

Extended Data Figures

Extended Data Tables

Supplementary Figures

Supplementary Tables

Note: Per Nature guidelines, figures and tables integral to the main argument are labeled "Extended Data" (up to 10 display items total); the remaining reference and illustrative material is labeled "Supplementary Figures/Tables".

Supplementary Note 1 | Generalization across drug name variants and descriptions

We evaluate ATHENA-R1's ability to generalize across different drug representations. The ATHENA-R1 results in this section are obtained with a Llama-3.1-8B base model (ATHENA-R1-Llama8B), a predecessor of the primary Qwen3-8B-based ATHENA-R1 reported in the main text. LLM-based models are sensitive to variations in how drugs are referenced [1–3], such as brand versus generic names. To test generalization, we construct three modified versions of the DrugPC benchmark: BrandPC, GenericPC, and DescriptionPC. BrandPC and GenericPC systematically replace drug names in DrugPC with their brand or generic equivalents. Questions that do not reference drug names remain unchanged, while those requiring conversion between brand and generic names are modified accordingly. Both datasets maintain the same number of samples as DrugPC. Sample questions are shown in Extended Data Figure 4a.

DescriptionPC replaces drug names with detailed descriptions to assess generalization without explicit drug names, including indications, mechanisms of action, contraindications, and interactions. We removed DrugPC questions that became unanswerable after this transformation, resulting in 626 questions. Since multiple drugs may share similar descriptions, DescriptionPC introduces a two-step evaluation: (1) drug identification and (2) answer correctness (Extended Data Figure 4b). In the first step, the model identifies the drug based on its description. The ground truth includes all drugs that could match the given description. In the second step, the model selects the correct answer to a multiple-choice question using its predicted drug name. If drug identification is incorrect, the answer is automatically marked incorrect, ensuring that predictions rely on accurate drug recognition.

ATHENA-R1 achieves 93.6% accuracy on BrandPC and 93.7% on GenericPC, outperforming both LLMs and tool-use LLMs on both benchmarks (Extended Data Figure 4a). Among LLMs, Llama3.1-70B-Instruct performs best on BrandPC (73.0%), while GPT-4o leads on GenericPC (77.3%). ATHENA-R1 surpasses these top reference models by 20.6% and 16.4%, respectively. Among tool-use LLMs, WattTool-8B achieves the highest accuracy, with 40.2% on BrandPC and 31.5% on GenericPC. ATHENA-R1 outperforms these baselines by 53.4% and 62.2%, respectively. ATHENA-R1 also exhibits lower performance variance across the original, BrandPC, and GenericPC datasets, with a variance of 0.00667. In contrast, GPT-4o has a variance of 9.96, Llama3.1-70B-Instruct 2.42, WattTool-8B 13.07, and ToolACE-8B 1.05.

On the DescriptionPC benchmark (Extended Data Figure 4b), when evaluating only answer correctness (without considering whether the model identifies the correct drug), ATHENA-R1 achieves 90.4%, surpassing GPT-4o (85.9%) and Llama3.1-70B-Instruct (85.3%). However, models may be able to "guess" the answer to certain questions in DescriptionPC without first identifying the class of drugs being referenced, which limits model trustworthiness. Specifically, when requiring both correct drug identification and answer selection, accuracy drops significantly for Llama3.1-70B-Instruct to 20.1%, indicating unreliable drug grounding. In contrast, ATHENA-R1 maintains the highest performance at 56.5%, outperforming GPT-4o by 8.3%. For drug name identification alone, ATHENA-R1 achieves the highest accuracy at 60.1%, compared to GPT-4o's 55.8% and Llama3.1-70B-Instruct's 23.6%. These results highlight ATHENA-R1's stronger ability to reason over drugs and base decisions on correct information.

Supplementary Note 2 | Multi-step reasoning ablations and tool comparisons

The primary benchmark results (DrugPC and TreatmentPC, open-ended setting) reported in the main text use Qwen3-8B [4] as the base model for ATHENA-R1. The GPT-5 tool-access comparison (§2.7) is also open-ended; the ATHENA-R1 results shown there are the Qwen3-8B numbers from the main text. The remaining ablation and comparison experiments in this section (reasoning thought ablation, training trace length, inference step limit, reasoning step and tool-call usage, tool library vs. LLM-as-tools, and tool scaling), as well as the BrandPC, GenericPC, and DescriptionPC evaluations, were conducted using Llama-3.1-8B-Instruct [5] as the base model, with accuracy reported in the multiple-choice setting.

2.1 The role of reasoning thoughts in model performance

To verify the role of reasoning thoughts in ATHENA-R1, we build a modified version of ATHENA-R1 that does not generate reasoning thoughts (Algorithm 1). This modified version of ATHENA-R1 follows a multi-step inference process where, at each step, instead of explicit reasoning by generating intermediate thoughts, the model directly produces tool calls. We use the notation defined in Methods: $\mathcal{F}_A$ is the ATHENA-R1 backend LLM and $\mathcal{E}_i$ denotes tool execution outputs at step i. The process starts with the initialization of the reasoning trace, $\mathcal{R} \leftarrow \{\}$, the set of available tools, $\mathcal{P} \leftarrow \mathcal{P}_0$, and a step counter, $i \leftarrow 0$. In each iteration, the model generates tool calls or the final answer:

$$U_i = \mathcal{F}_A(Q, \mathcal{R}_{i-1}, \mathcal{P}_i). \tag{1}$$

If U_i contains textual content, it is assigned as the final answer A, and the FINISH tool is executed to terminate the reasoning process, returning $\mathcal{R}_i$ and A. Otherwise, U_i contains the tool call arguments C_i, which are executed. If C_i includes a call to TOOLRAG, the available tools $\mathcal{P}_i$ are updated accordingly. The reasoning trace is iteratively updated as:

$$\mathcal{R}_i \leftarrow \mathcal{R}_{i-1} \cup \{C_i, \mathcal{E}_i\}. \tag{2}$$

Algorithm 1: ATHENA-R1 multi-step inference process without thoughts.

Input: Question Q, tool library $\mathcal{B}$, Initial available tools $\mathcal{P}_0$
Output: Reasoning trace $\mathcal{R}$, final answer A

Initialize $\mathcal{R} \leftarrow \{\}$, tools $\mathcal{P} \leftarrow \mathcal{P}_0$, step $i \leftarrow 0$;
while *Reasoning is incomplete* **do**
  $i \leftarrow i + 1$;
  Generate tool calls or final answer: $U_i = \mathcal{F}_A(Q, \mathcal{R}_{i-1}, \mathcal{P}_i)$
  if U_i *contains text* **then**
    Split the text as the final answer: A;
    Execute FINISH tool to end the multi-step reasoning;
    Return $\mathcal{R}_i, A$;
  else
    $C_i = U_i$;
    if *call to ToolRAG in* C_i **then**
      Execute ToolRAG and update $\mathcal{P}_i$;
    else
      Execute tools from C_i to get tool response $\mathcal{E}_i$;
    Update reasoning trace: $\mathcal{R}_i \leftarrow \mathcal{R}_{i-1} \cup \{C_i, \mathcal{E}_i\}$;
Return $\mathcal{R}, A$;

Removing explicit reasoning steps reduces accuracy. ATHENA-R1 generates both intermediate thoughts and tool calls. Removing the thought component and allowing only tool calls reduces accuracy on DrugPC from 93.8% to 71.5% (-22.3%) and on TreatmentPC from 86.8% to 64.9% (-21.9%) (Extended Data Figure 3a). Generating tool calls without intermediate reasoning prevents the model from interpreting results and combining evidence across steps.

2.2 Longer training traces improve complex reasoning

We assess how the maximum number of reasoning steps in ATHENA-R1-INSTRUCT affects ATHENA-R1's performance. We limit the training traces to 1, 3, or 5 steps, or use all available steps (Methods). During inference, the number of steps is unrestricted. Training with a single step reduces accuracy from 93.8% to 71.6% on DrugPC and from 86.8% to 66.9% on TreatmentPC (Extended Data Figure 3b). Performance improves as more steps are included. The effect is larger on TreatmentPC, where treatment selection depends on multiple interacting constraints.

2.3 Longer inference enables evidence accumulation

We assess how the number of reasoning steps allowed at inference affects ATHENA-R1's performance. We limit the number of reasoning steps and require ATHENA-R1 to produce an answer once the limit is reached. When restricted to one step, ATHENA-R1 achieves 73.5% accuracy on TreatmentPC, 13.3% below the unrestricted setting (Extended Data Figure 3c). Accuracy increases with additional

steps and plateaus after five steps, indicating that multiple reasoning steps are required to retrieve and combine evidence.

2.4 Reasoning step and tool-call usage across benchmarks

We measure the number of reasoning steps and tool calls across benchmarks (Extended Data Figure 3d,e). TreatmentPC requires more steps and tool calls than DrugPC. Within DrugPC, multiple-choice and open-ended settings show similar usage. Within TreatmentPC, the open-ended setting requires more steps and calls than multiple-choice, reflecting the need to construct an answer without provided options.

2.5 Tool execution outperforms LLM-simulated tools

We compare tool library against an LLM-only approach, where the model mimics tool functionality by receiving structured prompts that describe each tool's capabilities and arguments (Extended Data Figure 3f). GPT-4o and Llama-3.1-8B-Instruct serve as the backend LLMs in this analysis, with all other settings unchanged. To make the LLM function as a tool, we prompt it with the following instruction:

LLM-as-a-tool

You are a function that answers the questions based on your given description and given input. Do not answer questions that you don't have knowledge about.

Here is your definition: {tool description}.

Here is the input to the function:{function call arguments}.

The tool response:

In this instruction, the tool call arguments generated by ATHENA-R1 serve as the input, while the tool description obtained from the tool library is used as a reference. The LLM is then prompted to simulate the tool's outputs.

Replacing calls to external biomedical resources and analysis functions in tool library with simulated LLM-generated approximations significantly reduces accuracy, suggesting that therapeutic reasoning benefits from direct access to executable, verifiable sources of biomedical evidence. On DrugPC, using Llama-3.1-8B-Instruct as tools lowers accuracy from 93.8% to 68.7% (-25.1%), while GPT-4o lowers accuracy to 72.7% (-21.1%). Although GPT-4o performs better, both models remain substantially below tool library, highlighting the limitations of LLM-only approaches for retrieving precise biomedical information. We observe a similar pattern on TreatmentPC: GPT-4o and Llama-3.1-8B-Instruct achieve 67.11% and 74.78% accuracy, respectively, compared with 86.84% using real tools in tool library. Thus, even stronger LLMs do not replace executable tools. By grounding inference in real tools, tool library produces verifiable outputs and enables users to inspect ATHENA-R1's reasoning trace and final answer.

2.6 Expanding the tool library improves performance

We evaluate the scalability of tool library by measuring how performance changes as the tool library expands. We construct four nested subsets containing 10%, 20%, 50%, and 75% of tool library, such that each larger subset includes all tools from the smaller subsets. This design isolates the incremental effect of adding tools while keeping the evaluation setting consistent. Using ATHENA-R1 with each subset and with the full tool library, we measure accuracy on the DrugPC and TreatmentPC benchmarks (Extended Data Figure 3g).

On DrugPC, accuracy increases from 78.4% with 10% of the tools to 93.8% with the full tool library. On TreatmentPC, accuracy increases from 71.7% to 86.8%. These results show that expanding tool library improves ATHENA-R1's ability to solve complex and specialized therapeutic reasoning tasks, suggesting that broader access to biomedical resources and analysis functions strengthens tool-grounded inference.

2.7 Trained tool-use reasoning outperforms tool access alone

ATHENA-R1's performance on TreatmentPC reflects trained multi-step tool-use reasoning rather than tool access alone. To test this, we give GPT-5 [6] identical access to the full set of 212 tools in tool library, using the same tool-calling interface and tool descriptions provided to ATHENA-R1. We evaluate all models under the open-ended TreatmentPC protocol on all 456 TreatmentPC questions. We compare three GPT-5 settings: no tool access (72.2%), tool access with default generation (66.89%), and tool access with forced tool use (70.18%). Results are shown in Figure 2d, GPT-5 bars.

Setup. "Forced tool use" is enforced through the system prompt rather than an API constraint: GPT-5 is instructed to invoke at least one tool before producing its final answer. "Default generation" uses the same prompt with this instruction removed, leaving the tool-call decision to GPT-5. The tool-use rate reported here is the fraction of questions on which GPT-5 emitted at least one tool call before its final answer; under default generation this rate is 1%, and under forced tool use it is 100% by construction.

Answer-extraction protocols and results. In the open-ended setting, ATHENA-R1 produces a reasoning trace and a free-form answer *without seeing the multiple-choice options*. To score accuracy on TreatmentPC, this free-form answer must then be mapped to one of the multiple-choice options. We report ATHENA-R1 under two mapping protocols (Figure 2d, ATHENA-R1 bars). Under "self as judge", ATHENA-R1 itself selects the option letter that best matches its own free-form answer, yielding 74.8%. Under "GPT-5 as judge", an independent GPT-5 instance reads ATHENA-R1's reasoning trace and free-form answer and selects the option letter; this matches the standard open-ended evaluation protocol applied to all baselines in the main text (Methods), yielding 82.9%.

ATHENA-R1 invokes tools on every problem, grounding each reasoning step in verified outputs from FDA drug labels and structured biomedical databases. This behavior is learned through two-level self-learning: supervised fine-tuning on ATHENA-R1-Instruct, followed by

reinforcement learning against scientific feedback (Methods). GPT-5, in contrast, invokes tools on only 1% of problems under default generation, relying instead on internal knowledge; in this setting, its accuracy drops by 5.3 percentage points relative to its no-tool baseline. Forcing tool use recovers part of this loss (70.18%) but still falls below ATHENA-R1's accuracy under both answer-extraction protocols. Thus, GPT-5 does not match ATHENA-R1 even when given the same tool access. These results show that ATHENA-R1's advantage on TreatmentPC arises from learned reasoning over tool outputs, rather than from privileged access to tools.

Supplementary Note 3 | System and agent prompt designs

This section describes the prompts used to data generation agents. For clarity, we provide prompt sketches rather than full prompts: each sketch outlines the structure of the corresponding prompt, summarizes its key instructions, and omits implementation details that are not needed to understand the method. Prompt-internal variable names and template fields (e.g., `{{functions}}`) reflect the templates used during dataset construction. In the main text and throughout the paper, we refer to these items as "tools" for consistency.

3.1 System prompt for ATHENA-R1

ATHENA-R1

You are a helpful assistant that will solve problems through detailed, step-by-step reasoning and actions based on your reasoning. Typically, your actions will use the provided functions. You have access to the following functions. {functions}

3.2 Prompts for TOOLGEN

Summarizer

{API schema}

Using the provided {database name} API Schema, generate all possible specific functional commands in words with no code. Output them in a list.

Tool Generator (for openFDA tools)

You are a helpful assistant for generating functions based on the field descriptions and API schema of openFDA:
{API schema, field descriptions, and example functions}
Guidelines:

- Generate **two functions**: one function retrieves the drug name based on the field information, and the other function retrieves information for that field based on drug names.
- Align the function with the expected fields and descriptions.
- Each function must be unique and different from existing examples.
- Fields should contain search_fields and return_fields:
  - search_fields is a **dict**, where the keys are the function input parameters and the values are the fields to be searched.
  - return_fields is a **list** of field names from which information must be returned.

The capabilities of the functions should be related to the given capabilities: {capabilities}

Tool Generator (for OpenTarget tools)

You are a helpful assistant for generating functions based on the OpenTarget API schema:
{API schema and example functions}
Guidelines for the generated function:

- The function should align with the schema's functional and structural requirements.
- The function's name, description, input parameters, and schema should be unique and different from the existing example functions.
- The function capabilities should be related to the given capabilities: {capabilities}

Tool Checker

You are a helpful assistant who generates test queries based on a given function. You are provided the following:

- Function: {generated tool}
- Related keywords and information for questions and queries:{additional information}

Based on the provided function, you must generate {number} different questions in natural language that require using the function.

Guidelines:

- The questions should be specific and diverse; avoid general questions
- Function calls must include “name" and “arguments" arguments
- Question examples: {examples}

3.3 Prompt for QUESTIONGEN

Information Extractor (for disease-centered personalized treatment questions)

You are provided the following information:

- **Disease Information:** These phenotypes or symptoms in the following disease-related information will be used to construct a patient profile. {disease desc info}
- **Paired Drug Information:** Here is a side-by-side comparison of multiple drug options that help in designing patient conditions. Consider the side effects, drug interactions, contraindications, and other aspects of these drugs in deciding which patient-specific factors would require someone to take one drug instead of the other options. For example, one drug may be a better option than the others given specific adverse drug-drug interactions, warnings, age restrictions, patient population restrictions, pregnancy considerations, and contraindications. Include such factors in the constructed patient profile to make one drug the definitive correct answer. {drug information}

Generate a comparison analysis of the selected drugs based on the provided information. Show the differences between the drugs and provide evidence for the differences.

Question Generator (for disease-centered personalized treatment questions)

You are an assistant specializing in creating advanced biomedical multiple-choice questions focused on drug treatments given various patient-specific information like diseases, phenotypes, and genetic variation.

Guidelines:

- Frame questions around patient case scenarios, where a patient is diagnosed with a disease or exhibits specific phenotypes, and the goal is to identify the most suitable treatment. You may also provide protein targets or genes. If additional info is given in the Personalized Information section below, incorporate this info into the profile being constructed.
- Construct questions and answer choices that compare multiple similar drug treatments and select the most suitable one given the patient's particular conditions. Incorrect answer choices could be drugs indicated for the disease but unsuitable for this particular patient due to factors like age, comorbidities, or dosage considerations. The correct answer should be the most appropriate drug for the patient's specific profile.
- **Selected Tools:** Generate questions related to these functions. {selected tools}
- **Disease Information:** Use phenotypes or symptoms in the following disease-related information to construct the patient profile. {disease desc information}
- **Personalized Information:** When constructing the patient profile, use the following analysis of the side-by-side drug comparison. Consider the side effects, drug interactions, contraindications, and other aspects of these drugs in deciding which patient-specific factors would require someone to take one drug instead of the other options. For example, one drug may be a better option than the others given specific adverse drug-drug interactions, warnings, age restrictions, patient population restrictions, pregnancy considerations, and contraindications. Include such factors in the constructed patient profile to make one drug the definitive correct answer. Drug information: { drug information } Drug comparison analysis: {side by side drug comparison from Information Extractor agent}

  Generate a question, answer, and explanation according to this format: {format outline}

Question Generator (for tool-chain-centered questions)

You are a helpful assistant for generating expert-level biomedical questions. Based on the given functions, generate a single independent question that focuses on the given drug. The question should be specific, diverse, and framed in multiple ways, requiring the use of as many functions as possible. Do not write a long question; break up the question into multiple sentences if needed. Do not include details that a scientist, physician, or patient would not know (*e.g.,* ontology IDs like MONDO, EFO, CHEMBL, Ensembl/ENS).

Use only the following information:

1. Functions that can retrieve information related to the drug: {tool descriptions in the sampled tool chain}
2. Related information from functions: {information obtained from tools}
3. Related information from PrimeKG interactions: {drug or disease related information from the PrimeKG knowledge graph}

Generate a question, answer, and explanation according to this format: {format outline}

Question Generator (for drug-centered questions)

You are a helpful assistant to generate meaningful and challenging multi-choice questions for expert biomedical researchers. Formulate biomedical questions and generate answers using only the drug name and field information provided below:

- Drug generic name: {generic name}
- Drug brand name: {brand name}
- Specific field of information for the drug (*e.g.,* contraindications): {field information}

Other guidelines:

- Generate multiple, different questions to utilize all of the provided information. Make sure the questions do not overlap in content.
- Formulate questions that can be answered without needing additional information beyond the field information provided.
- Ask questions in different ways. Don't always start with "What" and "Which".

Generate a question, answer, and explanation according to this format: {format outline}

3.4 Prompts for TRACEGEN

Helper

Please act as a helper to provide solution hints for the next step in solving the question. Give some suggestions about what to do next, but never give the final answer or information that directly leads to the final answer. Only provide hints for one reasoning step.

Also, make sure the user's final answer contains the correct answer. If not, let the user do self-reflection and continue reasoning until the correct answer is found.

- Question:{question}
- Correct final answer:{answer}
- Explanation of correct answer:{explanation}
- Previous reasoning steps:{reasoning trace}

Solver

You must fully understand and solve a question through reasoning and function calls.

Guidelines:

- For each step, you must generate a reasoning thought and correct function call. If needed, call multiple functions.
- If you think you have answered the question, thoroughly reflect on your reasoning to verify you have in fact answered the question. If not, continue reasoning. If so, call the 'Finish' function and provide your final answer, which should be 1) comprehensive, 2) explain how you arrived at the answer, and 3) why the answer addresses the question.
- If the result from the last function call is empty or not useful, you must continue reasoning and call ToolRAG (or simulate a virtual ToolRAG call) to retrieve more tools.
  - If the tool you need is in the Function List below, you must retrieve them using a virtual ToolRAG call that simulates obtaining the tool through ToolRAG.
  - If the tool you need is not in the Function List below, you need to call ToolRAG.
  - {Description of ToolRAG and virtual ToolRAG tools}
- Do not answer the question based on general knowledge. You must answer the question based on the information returned by the tools.
- If all previous solution attempts have failed, do not repeat the same thoughts and function calls. Instead, come up with new solution approaches.

Function List: {available tools description of the initial set of tools $\hat{\mathcal{P}}_0$}

For each reasoning step, respond in this JSON format: {reasoning step format}

For the final step, respond in this JSON format, providing the final answer and a detailed explanation: {final reasoning step format}

Previous reasoning steps: {previous multi-step reasoning trace}

Hint for next step: {solution hint from Helper agent}

Supplementary Note 4 | Additional details on multi-institutional human evaluation

4.1 Definition of evaluation criteria and scoring rubrics

Each model response was evaluated across eight criteria using a 1–5 Likert scale. Evaluators could also select "Unable to Judge" when they did not have sufficient expertise or information to assign a rating for a given criterion. Below, we define each criterion and its scoring anchors.

Absolute rating criteria.

1. **Task success.** *How well did the model complete the requested task?*
   1 = The model completely failed to address the task; 2 = The model made a minimal attempt but largely failed; 3 = The model partially completed the task with notable gaps; 4 = The model mostly completed the task with minor issues; 5 = The model fully and accurately completed the task.

2. **Helpfulness of rationale.** *How helpful was the model's reasoning process in understanding how the answer was derived?*
   1 = No reasoning was provided, or the reasoning was completely unhelpful; 2 = Minimal reasoning that provided little insight; 3 = Some reasoning was provided but key steps were missing or unclear; 4 = Clear reasoning with most steps well-explained; 5 = Thorough, clear reasoning that fully explained how the answer was derived.

3. **Cognitive traceability.** *How easy is it to follow the model's chain of reasoning?*
   1 = The reasoning was impossible to follow; 2 = The reasoning was very difficult to follow; 3 = The reasoning could be followed with effort but had unclear sections; 4 = The reasoning was mostly easy to follow with minor unclear points; 5 = The reasoning was completely transparent and easy to follow throughout.

4. **Possibility of harm.** *Could acting on this output lead to harm in a clinical or research setting?*
   1 = The output contains dangerous misinformation that could directly cause harm; 2 = The output contains several concerning errors that could lead to harm; 3 = The output has some inaccuracies that could potentially cause minor harm; 4 = The output is mostly safe with minimal risk of harm; 5 = The output is completely safe with no risk of harm from acting on it.

5. **Alignment with clinical consensus.** *How well does the output align with established clinical guidelines, best practices, and expert consensus?*
   1 = The output directly contradicts established clinical consensus; 2 = The output mostly contradicts or ignores clinical consensus; 3 = The output partially aligns with clinical consensus but has notable deviations; 4 = The output mostly aligns with clinical consensus with minor deviations; 5 = The output perfectly aligns with established clinical consensus and best practices.

6. **Accuracy of content.** *How factually accurate is the information provided?*
   1 = The content is mostly or entirely inaccurate; 2 = The content has many factual errors; 3 = The

content has a mix of accurate and inaccurate information; 4 = The content is mostly accurate with minor errors; 5 = The content is entirely accurate with no factual errors.

7. **Completeness.** *How thoroughly does the output cover all relevant aspects of the question?*
   1 = The output is severely incomplete, missing most relevant aspects; 2 = The output covers few relevant aspects; 3 = The output covers some relevant aspects but misses important ones; 4 = The output covers most relevant aspects with minor omissions; 5 = The output comprehensively covers all relevant aspects of the question.

8. **Clinical relevance.** *How relevant and applicable is the output to real-world clinical or research practice?*
   1 = The output is entirely irrelevant to clinical or research practice; 2 = The output has limited relevance to practice; 3 = The output is somewhat relevant but lacks practical applicability; 4 = The output is mostly relevant and applicable to practice; 5 = The output is highly relevant and directly applicable to clinical or research practice.

Pairwise preference criteria.

In addition to assigning absolute ratings, evaluators were shown pairs of model responses side by side, with model identities blinded, and were asked to indicate a preference for each of the same eight criteria [7]. For each criterion, evaluators selected the response they preferred according to the criterion definition. The pairwise evaluation question was phrased as follows.

- **Task success:** "Which model was more successful at completing the task?"
- **Helpfulness of rationale:** "Which model provided more helpful reasoning?"
- **Cognitive traceability:** "Which model's reasoning was easier to follow?"
- **Possibility of harm:** "Which model's output is less likely to lead to harm?"
- **Alignment with clinical consensus:** "Which model better aligns with clinical guidelines?"
- **Accuracy of content:** "Which model provided more accurate information?"
- **Completeness:** "Which model gave a more complete response?"
- **Clinical relevance:** "Which model's output is more clinically relevant?"

For each pairwise comparison, evaluators selected one of four options: "Model A is better," "Model B is better," "Both are equally good," or "Neither did well." The assignment of ATHENA-R1 and the reference model to positions A and B was randomized for each question to prevent position bias.

4.2 Evaluator demographics and expertise

The 29 experts recruited from 28 disease organizations spanned a broad range of professional roles, backgrounds, and disease areas (Supplementary Table 4).

Organizational roles. Participants self-reported their roles within their organizations (selecting all that apply). The most common role was organization lead (17 participants), followed by non-clinical researcher (6), clinical researcher (4), clinician (3), and other roles including chief scientific officer (2), scientific director, and director of scientific engagement.

Professional backgrounds. Participants' professional backgrounds included family members or caregivers of patients (15), PhD researchers or scientists (10), board-certified physicians (3), clinical researchers (4), and patients (2). Several participants held multiple roles (e.g., clinician and family caregiver, or organization lead and researcher).

Experience. Among the 27 participants who reported their experience in clinical or research activities related to their disease area, the distribution was: less than 1 year (2), 1–2 years (6), 3–5 years (11), 6–10 years (2), and more than 10 years (6). Two participants did not report experience duration.

4.3 Inter-rater agreement analysis

Seven questions in the evaluation set were independently assessed by multiple evaluators: six questions were assessed by two evaluators, and one question was assessed by three evaluators. This yielded nine unique evaluator pairs for the agreement analysis. Because evaluators were assigned questions that matched their disease expertise, overlapping assessments arose incidentally when multiple evaluators shared the same therapeutic focus.

Pairwise preference agreement. Across all eight criteria, evaluators agreed on the pairwise preference (after normalizing for randomized position assignment) in 70.8% of comparisons (51 of 72 evaluator-pair–criterion combinations). Agreement was highest for reasoning-related criteria: helpfulness of rationale and cognitive traceability both achieved 100% agreement (9 of 9 pairs). Clinical content criteria showed moderate agreement: possibility of harm and completeness (78%), task success, alignment with clinical consensus, and accuracy of content (56%), and clinical relevance (44%). When restricting to decisive comparisons where both evaluators selected a preferred model (excluding ties and "neither did well"), agreement rose to 91.1% (51 of 56 decisive pairs). In this setting, alignment with clinical consensus, accuracy of content, and completeness all reached 100% agreement, indicating that disagreements on these criteria arose from one evaluator selecting a tie rather than from conflicting model preferences. The remaining disagreements among decisive pairs occurred on task success (5 of 8, 62.5%), possibility of harm (7 of 8, 87.5%), and clinical relevance (4 of 5, 80.0%).

Absolute rating agreement. For absolute Likert ratings on ATHENA-R1 responses, evaluators showed moderate to high agreement. Across criteria (4–9 evaluator pairs per criterion, depending on "Unable to Judge" responses), exact agreement ranged from 20% to 86%, and within-one-point

agreement ranged from 80% to 100%. Cognitive traceability showed the highest exact agreement (86%, 6 of 7 pairs), while completeness showed the lowest (20%, 1 of 5 pairs). The mean absolute score difference between evaluator pairs was 0.53 points for ATHENA-R1 and 1.12 points for reference models, indicating that evaluators agreed more closely when rating ATHENA-R1 than when rating reference model responses.

Interpretation. The sample size for inter-rater analysis is limited (7 questions), precluding formal computation of chance-corrected agreement statistics such as Cohen's κ or Krippendorff's α. The observed agreement rates are consistent with the inherent subjectivity of expert clinical evaluation, where evaluators with different disease specializations may weigh clinical accuracy and relevance differently. Notably, when evaluators made decisive model selections, they agreed on which model was better in over 91% of cases, suggesting that differences in overall agreement are driven primarily by varying thresholds for declaring a tie rather than by conflicting judgments about model quality. The near-perfect agreement on reasoning criteria (helpfulness and traceability) suggests that ATHENA-R1's multi-step reasoning traces provide a consistently recognizable advantage that is robust to evaluator differences.

4.4 Additional results

This section reports per-criterion breakdowns of pairwise preference rates (Table 1) and absolute rating scores (Table 2) from the human evaluation results summarized in Figure 3c,d.

Pairwise preference rates. Table 1 reports the full distribution of pairwise preferences across all 110 evaluated responses. For each criterion, evaluators selected one of four options: ATHENA-R1 is better, reference model is better, both equally good, or neither did well. ATHENA-R1 was preferred in the majority of responses across all eight criteria. The advantage was largest for reasoning-related criteria: ATHENA-R1 was preferred for helpfulness of rationale in 94.5% of responses (104 of 110) and for cognitive traceability in 95.5% (105 of 110), with reference models preferred in fewer than 4% of responses for both. For clinical content criteria (task success, possibility of harm, alignment with clinical consensus, accuracy of content, completeness, and clinical relevance), ATHENA-R1 was preferred in 57–66% of responses and reference models in 16–20%. Tie rates (both equally good) ranged from 1–18%, with higher tie rates observed for alignment with clinical consensus (18.2%), accuracy of content (18.2%), and clinical relevance (18.2%).

Absolute rating scores. Table 2 reports per-criterion mean absolute ratings (1–5 Likert scale) for ATHENA-R1 and reference models, computed as per-response averages across all valid ratings (excluding "Unable to Judge" responses). ATHENA-R1 achieved higher mean ratings than reference models on all eight criteria. The largest differences were observed for helpfulness of rationale (ATHENA-R1: 4.58, reference: 1.65) and cognitive traceability (ATHENA-R1: 4.67, reference: 1.54), consistent with the pairwise preference results. For clinical content criteria, ATHENA-R1 scores ranged from 3.76 (possibility of harm) to 4.18 (alignment with clinical consensus), while reference model scores ranged from 2.46 (completeness) to 3.02 (clinical relevance).

Reference model distribution. The primary reference model was Qwen3-8B, the base model from which ATHENA-R1 is built, serving as the comparison in 100 of 110 evaluated responses

Supplementary Table 1: Distribution of pairwise preferences across all 110 evaluated responses. Values are reported as counts, with percentages shown in parentheses.

Criterion	ATHENA-R1	Reference	Tie	Neither
Task success	70 (63.6%)	21 (19.1%)	12 (10.9%)	7 (6.4%)
Helpfulness of rationale	104 (94.5%)	3 (2.7%)	2 (1.8%)	1 (0.9%)
Cognitive traceability	105 (95.5%)	4 (3.6%)	1 (0.9%)	0 (0.0%)
Possibility of harm	68 (61.8%)	21 (19.1%)	13 (11.8%)	8 (7.3%)
Alignment with consensus	65 (59.1%)	18 (16.4%)	20 (18.2%)	7 (6.4%)
Accuracy of content	64 (58.2%)	20 (18.2%)	20 (18.2%)	6 (5.5%)
Completeness	73 (66.4%)	19 (17.3%)	11 (10.0%)	7 (6.4%)
Clinical relevance	63 (57.3%)	22 (20.0%)	20 (18.2%)	5 (4.5%)

Supplementary Table 2: Mean absolute ratings (1–5 Likert scale) per criterion for ATHENA-R1 and reference models, computed as per-response averages across all valid ratings. The number of valid ratings per criterion varies (45–110) because evaluators could select "Unable to Judge." Reference models include Qwen3 (n = 100), o3-mini (n = 3), Gemini-2.0-Flash (n = 3), DeepSeek-R1 (n = 2), DeepSeek-R1-Distill-Llama-8B (n = 1), and Llama-3.1-8B-Instruct (n = 1).

Criterion	ATHENA-R1	Reference
Task success	3.87	2.91
Helpfulness of rationale	4.58	1.65
Cognitive traceability	4.67	1.54
Possibility of harm	3.76	2.90
Alignment with clinical consensus	4.18	2.76
Accuracy of content	3.99	2.82
Completeness	3.88	2.46
Clinical relevance	4.01	3.02

(90.9%). This design isolates the effect of ATHENA-R1's tool library and self-learning by holding the underlying base model fixed at the same 8B parameter scale, so that any preference reflects ATHENA-R1's tool-grounded reasoning rather than a difference in the underlying language model. The remaining 10 responses spot-checked ATHENA-R1 against frontier and different-family systems: o3-mini (n = 3), Gemini-2.0-Flash (n = 3), DeepSeek-R1 variants (n = 3), and Llama-3.1-8B-Instruct (n = 1). ATHENA-R1 was preferred in the majority of these comparisons as well, although each of these subsets is individually too small to support separate statistical testing.

Statistical significance. We assessed the statistical significance of ATHENA-R1's advantage using two complementary tests. First, for pairwise preferences, we applied a one-sided binomial test to the decisive comparisons (excluding "Both are equally good" and "Neither did well") under the null hypothesis that ATHENA-R1 and the reference model are equally likely to be preferred ($p_0 = 0.5$). Second, for absolute ratings, we applied a one-sided Wilcoxon signed-rank test to the paired score differences (ATHENA-R1 minus reference) on the same question, testing whether ATHENA-R1 scores are systematically higher. All eight criteria were significant under both tests at $P < 5 \times 10^{-5}$ (Table 3), with large effect sizes (rank-biserial $r \geq 0.71$).

Supplementary Table 3: Per-criterion statistical tests comparing ATHENA-R1 against reference models on 110 expert-evaluated responses. Two complementary tests are reported: a **one-sided binomial test** on pairwise preferences, assessing whether evaluators select ATHENA-R1 more often than chance ($p_0 = 0.5$); and a **one-sided Wilcoxon signed-rank test** on paired absolute Likert ratings (ATHENA-R1 minus reference), assessing whether ATHENA-R1 scores are systematically higher. *Win rate* is the fraction of decisive comparisons favoring ATHENA-R1, shown as k/n (count) with the percentage in parentheses, after excluding "Both are equally good" and "Neither did well." n under the Wilcoxon test is the number of question-level paired ratings, after excluding "Unable to Judge" responses pairwise; it differs from the binomial denominator because the two exclusion rules apply to different response options. Note that the two n values come from different sub-samples and are not directly comparable. P values from both tests are shown uncorrected; all eight criteria satisfy $P < 5 \times 10^{-5}$ under both tests. *Effect size* r is the matched-pairs rank-biserial correlation for the Wilcoxon test, computed as $r = (W^+ - W^-)/(W^+ + W^-)$ [8], with $r \in [-1, 1]$; $|r| > 0.5$ denotes a large effect, and all eight criteria show $r \geq 0.71$.

Criterion	**Binomial test**		**Wilcoxon signed-rank test**		
	Win rate	P	n	P	r
Task success	70/91 (76.9%)	1.3×10^{-7}	99	4.5×10^{-8}	0.71
Helpfulness of rationale	104/107 (97.2%)	1.3×10^{-27}	105	8.4×10^{-19}	0.98
Cognitive traceability	105/109 (96.3%)	8.9×10^{-27}	95	4.2×10^{-18}	0.99
Possibility of harm	68/89 (76.4%)	3.0×10^{-7}	55	9.4×10^{-8}	0.90
Alignment with consensus	65/83 (78.3%)	1.1×10^{-7}	43	3.1×10^{-5}	0.78
Accuracy of content	64/84 (76.2%)	7.9×10^{-7}	83	1.4×10^{-8}	0.81
Completeness	73/92 (79.3%)	6.2×10^{-9}	82	3.9×10^{-9}	0.80
Clinical relevance	63/85 (74.1%)	4.9×10^{-6}	82	1.1×10^{-6}	0.71

Model variability. Across models evaluated more than once, the coefficient of variation (CV) of absolute ratings was lowest for o3-mini (CV < 0.18 across all criteria) and highest for Qwen3 (CV ranging from 0.36 to 0.67), reflecting greater variability in response quality for the primary reference model. ATHENA-R1 showed moderate variability (CV ranging from 0.13 to 0.26 across criteria), consistent with stable performance across disease areas and question types.

Supplementary Note 5 | Clinical case-based expert review: Full questions, reasoning traces, and final answers

Below, we present the five clinical vignettes used in this evaluation. For each case, we provide the verbatim input to ATHENA-R1, the sequence of reasoning strategies executed by ATHENA-R1, the tools invoked at each step, the intermediate findings produced during reasoning, and the final answer. These materials provide a detailed view of how ATHENA-R1 approaches complex, real-world treatment decisions through multi-step, tool-grounded reasoning.

Cases 1–3 correspond to the three vignettes selected for the formal physician evaluation. Their per-criterion rating panels from three physicians are shown in Figure 4b, d, and f (Cases 1, 2, and 3, respectively), and the corresponding discussion appears in the Results subsection, "Physicians evaluate ATHENA-R1 on complex, real-world treatment decisions". Details of the study design, case construction, evaluation protocol, scoring rubric, ethical framework, and statistical analyses are provided in Online Methods §6.

Cases 4 and 5 are additional vignettes that were used during development but were not included in the formal three-case evaluation. Case 4 (polypharmacy) was evaluated by all three physicians, whereas Case 5 (neonatal necrotizing enterocolitis) was evaluated by a single physician, with the remaining two physicians not providing ratings. The corresponding per-criterion rating panels are shown in Supplementary Figure 4.

5.1 Case 1: ACE-inhibitor selection post-CABG with recent contrast-induced nephropathy

Patient scenario (verbatim input to ATHENA-R1). Patient 67M POD: #2 s/p three-vessel CABG Consult reason: ACE inhibitor management in setting of recent contrast exposure. PMH: HTN (15 years), T2DM (HbA1c 6.8%), hyperlipidemia, CKD stage 2 (baseline Cr 1.4, eGFR 52), HFrEF (LVEF 35% per recent echo), myocardial infarction (4 months ago). Meds pre-admission: Lisinopril 20mg daily, metformin 1000mg BID, atorvastatin 40mg daily, ASA 81mg daily. Recent course: Patient underwent coronary angiography 5 days pre-op with 75mL low-osmolar contrast, which revealed severe 3-vessel disease. Developed mild contrast-induced nephropathy with Cr rising to 1.7 (eGFR 42). Underwent successful CABG (LIMA→LAD, SVG→OM, SVG→PDA) with 110 minutes on bypass. Initial post-op course required low-dose pressors for 12 hours, now resolved. Current Cr stable at 1.7, good UOP. What ACE inhibitors can be given in this patient post-operatively?

Reasoning trace. ATHENA-R1 decomposed the problem into seven parallel investigation strategies and dispatched one child ATHENA-R1 instance per strategy via the CallAgent self-invocation mechanism (Algorithm 1).

1. **FDA drug-labeling analysis** (to identify ACE inhibitors with minimal impact on kidney function and no CKD contraindications). The child queried `FDA_get_adverse_reactions_by_drug_name`, `FDA_get_boxed_warning_info_by_drug_name`, `FDA_get_drug_`

names_by_boxed_warning, FDA_get_drug_names_by_drug_interactions, and FDA_get_drug_name_by_adverse_reaction. *Finding*: enalapril's adverse reactions include minor increases in BUN and serum creatinine, which are reversible upon discontinuation, with no CKD-specific contraindications in the retrieved data; lisinopril has adverse reactions such as hypotension and hyperkalemia.

2. **FAERS adverse-event mining** (focused on seriousness and hospitalization signals). The child queried FAERS_count_seriousness_by_drug_event, FAERS_count_outcomes_by_drug_event, and FAERS_count_reactions_by_drug_event for both drugs. *Finding*: FAERS seriousness and outcome data did not yield a clear preference between enalapril and lisinopril.

3. **Mechanism-of-action backtracking** (to identify ACE inhibitors less likely to affect kidney function). The child used OpenTargets_get_drug_mechanisms_of_action_by_chemblId and FDA_get_mechanism_of_action_by_drug_name. *Finding*: lisinopril inhibits the angiotensin-converting enzyme (reducing blood pressure and proteinuria) but is associated with renal dysfunction and hyperkalemia, posing a risk of exacerbating renal issues in this patient with CKD plus recent contrast-induced nephropathy; alternative ACE inhibitors with a lower risk of renal impairment should be considered.

4. **Literature surveillance** (scientific literature on ACE inhibitors in CKD with recent contrast exposure). The child queried EuropePMC_search_articles and SemanticScholar_search_papers with multiple keyword sets. *Finding*: withdrawing ACE inhibitors before coronary angiography may reduce contrast-induced nephropathy incidence; renin-angiotensin-aldosterone blockade in CKD patients affects contrast-induced nephropathy risk.

5. **Structural-similarity screening** (to identify ACE inhibitors structurally similar to those known to be safe in CKD). The child attempted structural similarity searches. *Finding*: the available tools did not yield actionable structural-similarity data; the child flagged that clinical guidelines and pharmacology review would be more informative.

6. **Clinical-guidelines review** (for ACE-inhibitor use in CKD with recent contrast exposure). The child searched for guideline-level recommendations via FDA_get_medication_guide_info_by_drug_name, FDA_get_drug_names_by_medication_guide, FDA_get_info_on_conditions_for_doctor_consultation_by_drug_name, EuropePMC_search_articles, and SemanticScholar_search_papers. *Finding*: a specific ACE-inhibitor recommendation for this scenario was not retrieved; general consensus supports careful renal-function monitoring and cardiology/nephrology consultation.

7. **Expert-consultation synthesis** (tailor the ACE-inhibitor choice to the patient's specific needs). The child synthesized the prior sub-agent findings. *Finding*: nephrology/cardiology co-management is essential; start with a lower dose and closely monitor renal function and blood pressure.

The parent ATHENA-R1 instance combined the sub-agent findings. The FDA-labeling and mechanism-of-action results dominated the recommendation of enalapril (convergent evidence that enalapril's renal effects are reversible and lisinopril carries higher hyperkalemia risk); FAERS

and structural-similarity strategies were non-diagnostic; literature and guideline strategies did not yield a specific agent recommendation but supported conservative renal-function monitoring and nephrology/cardiology consultation.

Final response (verbatim output produced by ATHENA-R1). In managing ACE inhibitor therapy for a patient with recent contrast exposure and existing conditions such as CKD and HFrEF, it is crucial to consider the potential risks and benefits. [...] Based on the comprehensive analysis of various solution plans, Enalapril emerges as a suitable option for this patient. The FDA drug labeling analysis indicates that Enalapril has minor adverse reactions such as increases in blood urea nitrogen and serum creatinine, which are reversible upon discontinuation of therapy. There are no specific contraindications related to CKD, making Enalapril a viable option for managing ACE inhibitor therapy in this patient. [...] While Lisinopril has been shown to be effective in managing hypertension and heart failure, it is associated with adverse effects such as renal dysfunction and hyperkalemia, which are particularly concerning for patients with pre-existing kidney conditions. Therefore, while Lisinopril can be beneficial, its use should be carefully considered, and alternative ACE inhibitors with a lower risk of renal impairment should be explored. [...] In conclusion, Enalapril is recommended as a suitable ACE inhibitor for this patient post-operatively, with careful monitoring of renal function and electrolytes. Consulting with a nephrologist or cardiologist is essential to tailor the ACE inhibitor choice and dosage to the patient's specific needs, ensuring effective treatment while minimizing adverse effects.

5.2 Case 2: Levofloxacin-warfarin interaction in a post-surgical infection

Patient scenario (verbatim input to ATHENA-R1). 68F with metallic mitral valve replacement (2019) following rheumatic heart disease. PMH significant for chronic atrial fibrillation, heart failure with preserved ejection fraction (HFpEF), CKD stage 3, and osteoarthritis. Recently underwent elective right total knee arthroplasty 4 days ago. Pre-operatively on warfarin (target INR 2.5-3.5 for mechanical valve), metoprolol, furosemide, lisinopril, and atorvastatin. Warfarin was held 5 days pre-operatively with heparin bridging, then heparin discontinued 12 hours before surgery. Post-operatively developed surgical site erythema, warmth, and purulent drainage from incision with fever to 38.9°C and leukocytosis (WBC 14.2k). Blood cultures pending, wound culture growing gram-positive cocci in clusters. Current medications include acetaminophen and tramadol for pain, prophylactic enoxaparin 40mg daily (subtherapeutic for valve), home cardiac medications, and cefazolin started empirically. Patient requires resumption of therapeutic anticoagulation for mechanical valve protection but infectious disease consultation recommends switching to levofloxacin. What are the risks of levofloxacin and are there alternative antibiotics?

Reasoning trace. ATHENA-R1 decomposed the problem into six parallel investigation strategies and dispatched one child ATHENA-R1 instance per strategy.

1. **FDA drug-labeling analysis for levofloxacin** (adverse events, warnings, precautions, drug interactions). The child queried `FDA_get_adverse_reactions_by_drug_name`, `FDA_get_warnings_and_cautions_by_drug_name`, `FDA_get_`

drug_interactions_by_drug_name, and FDA_get_boxed_warning_info_by_drug_name. *Finding*: levofloxacin carries key risks of tendinitis, tendon rupture, peripheral neuropathy, central nervous system effects, and QT interval prolongation; interaction with warfarin can enhance anticoagulant effect, increasing bleeding risk and necessitating careful prothrombin time and INR monitoring.

2. **FAERS adverse-event mining** for levofloxacin. The child queried FAERS_count_reactions_by_drug_event, FAERS_count_seriousness_by_drug_event, FAERS_count_outcomes_by_drug_event, and FAERS_count_additive_adverse_reactions. *Finding*: the FAERS endpoint returned server errors and did not yield structured adverse-event data; the child pivoted to clinical guidelines and literature review for mechanical-valve-specific risk assessment.

3. **Mechanism-of-action backtracking**. The child used OpenTargets_get_drug_mechanisms_of_action_by_chemblId and OpenTargets_get_associated_targets_by_drug_chemblId. *Finding*: levofloxacin inhibits bacterial DNA gyrase and topoisomerase IV. Alternatives with distinct mechanisms for gram-positive cocci include vancomycin (cell-wall synthesis), clindamycin (protein synthesis), and linezolid (protein synthesis).

4. **Literature surveillance**. The child attempted EuropePMC and SemanticScholar queries through Tool_RAG; the retrieval-augmented lookup did not return literature endpoints, so the child pivoted to FAERS multi-outcome analysis (FAERS_count_reactions_by_drug_event, FAERS_count_seriousness_by_drug_event, FAERS_count_outcomes_by_drug_event, FAERS_count_patient_age_distribution, and FAERS_count_additive_adverse_reactions) together with FDA_get_adverse_reactions_by_drug_name. *Finding*: FAERS and FDA-label queries identified serious levofloxacin reactions (tendinitis, tendon rupture, peripheral neuropathy, CNS effects, hypersensitivity, hepatotoxicity, and QT-interval prolongation) as particularly concerning in this mechanical-valve patient; the child proposed cefazolin or vancomycin as alternatives with more favorable safety profiles.

5. **Structural-similarity screening**. The child attempted a chemoinformatics similarity query for levofloxacin but the invocation targeted an endpoint not exposed by the tool library and returned no data. *Finding*: this strategy did not produce new evidence; the fluoroquinolone class-effect conclusion (tendinopathy, QT prolongation) derives from the strategy-1 label and strategy-3 mechanism findings, not from similarity retrieval.

6. **Alternative-antibiotic search for gram-positive cocci in mechanical-valve patients**. The child queried Tool_RAG for antibiotic-selection tools and used FDA_get_drug_names_by_indication and FDA_get_clinical_studies_info_by_drug_name for candidate alternatives. *Finding*: vancomycin, amoxicillin-clavulanate, and cefazolin are identified as alternatives effective against gram-positive cocci and safer for patients with mechanical valves; vancomycin use requires monitoring for nephrotoxicity and ototoxicity.

The parent ATHENA-R1 instance identified the warfarin-levofloxacin interaction as the principal risk and named vancomycin, clindamycin, and linezolid (alongside penicillin-class beta-lactams) as alternatives, recommending infectious-disease consultation for the final choice.

Final response (verbatim output produced by ATHENA-R1). Levofloxacin, a fluoroquinolone antibiotic, poses several risks for patients with mechanical valve replacements, particularly due to its potential adverse effects and interactions. Key risks include tendinitis, tendon rupture, peripheral neuropathy, central nervous system effects, and QT interval prolongation. These adverse effects can exacerbate existing conditions or lead to new complications in patients with mechanical valves. Additionally, levofloxacin interacts with warfarin, potentially enhancing its anticoagulant effects and increasing the risk of bleeding. [...] Given these risks, levofloxacin is generally avoided in patients with mechanical valve replacements. Alternative antibiotics that are safer for such patients should be considered. [...] Vancomycin, clindamycin, and linezolid are also viable alternatives, as they are effective against gram-positive cocci and have a more favorable safety profile in this context. [...] Consulting with an infectious disease specialist is essential to identify an appropriate alternative that effectively treats the infection while minimizing risks to the patient's mechanical valve and overall health.

5.3 Case 3: β-blocker selection post-STEMI in a patient with severe asthma

Patient scenario (verbatim input to ATHENA-R1). 51F with severe persistent asthma well-controlled on high-dose fluticasone/salmeterol and supplemental albuterol PRN. PMH significant for type 2 diabetes (HbA1c 7.2%) managed with metformin 1000mg BID, hypertension, and hyperlipidemia. Presents with acute STEMI requiring emergent cardiac catheterization with PCI and drug-eluting stent placement. Post-procedurally stable but requires dual antiplatelet therapy (aspirin + clopidogrel), high-intensity statin, ACE inhibitor, and beta-blocker. Metformin was held pre-catheterization due to contrast exposure and concern for contrast-induced nephropathy. Current medications include aspirin 81mg daily, clopidogrel 75mg daily, atorvastatin 80mg daily, and lisinopril 5mg daily. Patient reports mild dyspnea on exertion but unclear if cardiac or respiratory etiology. Creatinine stable at 1.1 mg/dL post-procedure. Cardiology recommends initiating beta-blocker therapy for cardioprotection, which beta blocker is most appropriate?

Reasoning trace. ATHENA-R1 decomposed the problem into seven parallel investigation strategies and dispatched one child ATHENA-R1 instance per strategy.

1. **FDA drug-labeling analysis for β-blockers**. The child queried `FDA_get_contraindications_by_drug_name`, `FDA_get_drug_names_by_indication`, `FDA_get_drug_interactions_by_drug_name`, and `FDA_get_drug_names_by_boxed_warning`. *Finding*: metoprolol tartrate is indicated for cardioprotection after MI with a favorable profile in asthma relative to non-selective agents.
2. **FAERS adverse-event mining** for β-blockers in similar demographics. The child queried `FAERS_count_additive_seriousness_classification`, `FAERS_count_additive_adverse_reactions`, and `FAERS_count_drugs_by_drug_event`. *Finding*: FAERS did not yield demographic-matched signals specific to asthma+T2DM+STEMI; deferred to labeling and guideline evidence.

3. **Mechanism-of-action backtracking**. The child used `FDA_get_mechanism_of_action_by_drug_name`, `FDA_get_pharmacodynamics_by_drug_name`, `FDA_get_drug_names_by_mechanism_of_action`, `OpenTargets_get_drug_mechanisms_of_action_by_chemblId`, and `OpenTargets_get_associated_targets_by_drug_chemblId`. *Finding*: propranolol, a non-selective β-blocker, affects both β_1 and β_2 receptors and risks exacerbating asthma; metoprolol, a cardioselective agent, primarily targets β_1 and is safer in asthma; carvedilol, also non-selective, carries additional α_1-blocking activity with different implications for asthma and diabetes.

4. **Literature surveillance** via `EuropePMC_search_articles` and `SemanticScholar_search_papers`. *Finding*: cardioselective β-blockers (notably bisoprolol) are effective in angina without causing bronchospasm in asthma; β-blockers can mask hypoglycemia symptoms, a relevant consideration in T2DM.

5. **Structural-similarity and nonclinical-toxicology screening**. The child used `ChEMBL_search_similar_molecules`, `FDA_get_nonclinical_toxicology_info_by_drug_name`, `FDA_get_drug_names_by_nonclinical_toxicology_info`, `FDA_get_animal_pharmacology_info_by_drug_name`, and `FDA_get_drugs_by_carcinogenic_mutagenic_fertility`. *Finding*: carvedilol shows no carcinogenic effects and minimal respiratory impact in animal studies; metoprolol shows mild pulmonary effects in animal studies.

6. **Clinical-guidelines review** for β-blocker use in asthma+diabetes. *Finding*: cardioselective agents such as carvedilol or nebivolol are often recommended for this patient profile due to their minimal impact on asthma and blood glucose levels; carvedilol in particular is described as often used in patients with diabetes and asthma.

7. **Cardiovascular-risk synthesis**. The child integrated asthma control, diabetes management, and post-STEMI cardioprotection constraints. *Finding*: cardioselective options balance cardioprotection with asthma safety.

The parent ATHENA-R1 instance synthesized the sub-agent findings. Multiple strategies converged on β-blockers with favorable profiles in asthma; sub-agents varied in which agent they named as first-line (metoprolol, bisoprolol, or carvedilol), reflecting the fact that guideline-level evidence does not uniquely prefer one.

Final response (verbatim output produced by ATHENA-R1). For a 51-year-old patient with severe persistent asthma, type 2 diabetes, and a recent STEMI, the selection of an appropriate beta-blocker requires careful consideration of several factors, including asthma control, diabetes management, and cardioprotection. [...] Non-selective beta-blockers like propranolol are generally avoided in patients with asthma due to their potential to worsen respiratory symptoms. Cardioselective beta-blockers, such as metoprolol or bisoprolol, primarily target beta-1 receptors, making them less likely to affect bronchial smooth muscle and thus safer for asthma management. [...] Beta-blockers can mask symptoms of hypoglycemia, which is a concern for patients with diabetes. [...] Cardioselective beta-blockers like metoprolol or bisoprolol are often recommended for patients with STEMI due to their ability to reduce cardiac workload and improve outcomes. [...] Given the patient's conditions, cardioselective beta-blockers such as metoprolol, bisoprolol, or carvedilol are

recommended. These medications offer effective cardioprotection while minimizing the risk of exacerbating asthma symptoms and interfering with diabetes management.

5.4 Case 4: Pain management and polypharmacy in a preoperative hip-fracture patient

Patient scenario (verbatim input to ATHENA-R1). 72M with left femoral neck fracture following a fall at home. PMH significant for HTN, T2DM, A-fib on anticoagulation, and hyperlipidemia. Admitted through ED, received IV morphine for initial pain control and apixaban reversal in preparation for surgery. Patient reporting inadequate pain control with current regimen and experiencing anxiety at night. Pain rated 7-8/10 despite PRN oral oxycodone, limiting mobility and affecting sleep. Pre-operative medications include home regimen (lisinopril, amlodipine, metformin, empagliflozin, apixaban [now held], atorvastatin, pantoprazole) plus PRN oral oxycodone for pain, ondansetron for nausea, and cefazolin for surgical prophylaxis. Question: What changes to pain management would optimize analgesia while minimizing polypharmacy complications in this elderly patient awaiting hip surgery?

Reasoning trace. ATHENA-R1 decomposed the problem into seven parallel investigation strategies and dispatched one child ATHENA-R1 instance per strategy.

1. **FDA drug-labeling analysis** for oxycodone and apixaban. The child queried `FDA_get_adverse_reactions_by_drug_name`, `FDA_get_boxed_warning_info_by_drug_name`, `FDA_get_drug_interactions_by_drug_name`, and `FDA_get_drug_names_by_boxed_warning`. *Finding*: oxycodone is linked to life-threatening respiratory depression, addiction, and interactions with CNS depressants; apixaban poses a high risk of bleeding, especially with concurrent anticoagulants or NSAIDs; the combination raises serious bleeding and respiratory-depression risk, motivating alternative pain-management strategies.
2. **FAERS adverse-event mining** in the elderly demographic. The child queried `FAERS_count_reactions_by_drug_event`, `FAERS_count_outcomes_by_drug_event`, `FAERS_count_patient_age_distribution`, and `FDA_get_drug_names_by_geriatric_use` for oxycodone and apixaban stratified to the elderly demographic. *Finding*: FAERS queries did not yield significant elderly-specific adverse-event signals for either drug (suggesting limited signal or data availability); the child concluded that alternative pain-management strategies and regimen review remain warranted on clinical grounds.
3. **Mechanism-of-action backtracking**. The child used `FDA_get_mechanism_of_action_by_drug_name` and `OpenTargets_get_associated_targets_by_drug_chemblId`. *Finding*: oxycodone is linked to serious adverse reactions (respiratory depression, constipation, addiction risk), and apixaban to major and fatal bleeding events; combined with this patient's comorbidity profile (HTN, T2DM, A-fib, hyperlipidemia), these reactions motivate exploring non-opioid analgesics or other lower-risk alternatives.
4. **Literature surveillance** via `EuropePMC_search_articles` and `SemanticScholar_search_papers`. *Finding*: the literature search did not return articles directly addressing

this patient scenario; peripheral findings supported personalized pain assessment, multimodal postoperative interventions, and individual pharmacotherapy management (IPM) for reducing polypharmacy-related fall risk in the elderly.

5. **Class-effect and toxicity screening**. The child queried `FDA_get_nonclinical_toxicology_info_by_drug_name`, `FDA_get_drug_names_by_nonclinical_toxicology_info`, `FDA_get_drug_names_by_animal_pharmacology_info`, and `FDA_get_drug_names_by_boxed_warning` as label-based substitutes for direct structural-similarity retrieval. *Finding*: no actionable structural-similarity hits were retrieved; the child proposed non-opioid analgesics (acetaminophen, NSAIDs), non-pharmacological interventions (physical therapy, cognitive-behavioral therapy), and reducing the frequency or dose of PRN oral oxycodone as strategies alongside continued screening for high-similarity hits to flag class-level toxicities.
6. **Non-pharmacological pain-management options**. The child queried `Tool_RAG` alongside `FDA_get_info_for_patients_by_drug_name`, `FDA_get_drug_names_by_population_use`, and related geriatric-use endpoints for adjunctive interventions. *Finding*: physical therapy, acupuncture, and cognitive-behavioral therapy are viable non-pharmacological interventions; physical therapy improves mobility and reduces pain, acupuncture provides targeted pain relief, and cognitive-behavioral therapy addresses the patient's reported nighttime anxiety.
7. **Polypharmacy optimization**. The child reviewed each element of the current regimen for ongoing necessity. *Finding*: (i) PRN oral oxycodone alone provides inadequate analgesia and may require switching to a more potent opioid or adding a non-opioid analgesic; (ii) apixaban remains held, with the hold decision based on bleeding risk and surgical timing; (iii) pantoprazole, ondansetron, and cefazolin can each be reassessed against active indications (gastrointestinal history, nausea, surgical-prophylaxis timing).

The parent ATHENA-R1 instance synthesized the sub-agent findings into a multimodal, opioid-sparing plan with explicit reassessment of each ancillary medication.

Final response (verbatim output produced by ATHENA-R1). To optimize pain management for the 72-year-old patient awaiting hip surgery, several key strategies should be considered to minimize polypharmacy complications while ensuring effective analgesia. [...] Given the significant risks associated with the current regimen of oxycodone and apixaban, it is advisable to explore non-opioid analgesics such as acetaminophen or NSAIDs. These alternatives are generally safer for patients with comorbidities like hypertension and type 2 diabetes. Additionally, non-pharmacological interventions such as physical therapy, acupuncture, and cognitive behavioral therapy can be beneficial. [...] Consider switching to a more potent opioid or adding a non-opioid analgesic to enhance pain relief. Since apixaban is currently held, it may not be necessary for pain management. [...] Pantoprazole, used for acid suppression, may not be necessary if the patient does not have a history of gastrointestinal issues or if other medications are managing acid levels. Ondansetron, used for nausea control, may be reduced or discontinued if the patient is not experiencing significant nausea. Cefazolin, used for surgical prophylaxis, should be timed appropriately with the surgical schedule and the patient's infection risk. [...] By making these adjustments, the patient's medication regimen

can be simplified, reducing the risk of polypharmacy complications while maintaining effective pain management and addressing other health concerns.

5.5 Case 5: Antibiotic selection in a preterm infant with suspected necrotizing enterocolitis

Patient scenario (verbatim input to ATHENA-R1). 28-week preterm infant, day 5 of life, birth weight 1200g. PMH significant for maternal chorioamnionitis and prolonged rupture of membranes. Initially stable on CPAP with standard early-onset sepsis prophylaxis (ampicillin + gentamicin x 48 hours, discontinued after negative cultures). Now presenting with temperature instability (hypothermia to 35.8°C), increased apnea episodes requiring bag-mask ventilation, and feeding intolerance with gastric residuals. Physical examination reveals significant abdominal distension with visible bowel loops through thin abdominal wall, hypoactive bowel sounds, and lethargy with decreased spontaneous movement. Current medications include caffeine citrate for apnea, minimal enteral feeds (now held). Laboratory findings show rising CRP (8.2 mg/dL), thrombocytopenia (platelets 89k). Abdominal radiograph pending. What antibiotics should be given in this scenario given some potential complications from antibiotic exposure?

Reasoning trace. ATHENA-R1 decomposed the problem into seven parallel investigation strategies and dispatched one child ATHENA-R1 instance per strategy.

1. **FDA drug-labeling analysis** for antibiotics used in neonatal sepsis. The child queried `FDA_get_drug_names_by_indication` (indication: `neonatal sepsis`), `FDA_get_adverse_reactions_by_drug_name`, `FDA_get_boxed_warning_info_by_drug_name`, `FDA_get_drug_interactions_by_drug_name`, and `FDA_get_drug_names_by_contraindications`. *Finding*: gentamicin and amikacin sulfate are FDA-indicated for neonatal sepsis and carry significant risks of nephrotoxicity and ototoxicity, especially in patients with renal impairment or on high doses or prolonged therapy; concurrent use with other nephrotoxic or neurotoxic drugs should be avoided.
2. **FAERS adverse-event mining** for the neonatal demographic. The child queried `FAERS_count_reactions_by_drug_event` and `FAERS_count_seriousness_by_drug_event` stratified to `Neonate`. *Finding*: specific FAERS adverse-event data for ampicillin or gentamicin in neonates were not returned; the child noted ampicillin's efficacy for skin and skin-structure infections in pediatric clinical studies and deferred to clinical guidelines for neonatal sepsis risk assessment.
3. **Mechanism-of-action backtracking**. The child queried `FDA_get_drug_names_by_mechanism_of_action`, `FDA_get_drug_names_by_clinical_pharmacology`, `FDA_get_drug_names_by_pharmacokinetics`, and `FDA_get_purpose_info_by_drug_name`. *Finding*: ampicillin, gentamicin, and cefotaxime are commonly used in neonatal care for early-onset sepsis and offer broad-spectrum coverage against common neonatal pathogens; their pharmacokinetics and pharmacodynamics require careful monitoring in preterm infants, with nephrotoxicity (especially with gentamicin) as a key concern.

4. **Literature surveillance** via `EuropePMC_search_articles` and `SemanticScholar_search_papers` for NEC-directed antibiotic regimens. *Finding*: combination regimens such as ampicillin plus gentamicin remain common empirical choices; antibiotic selection should be guided by clinical guidelines, local resistance patterns, and the specific clinical context, with careful monitoring and adjustment based on outcomes.

5. **Class-effect, teratogenicity, and nonclinical-toxicology screening**. The child queried `FDA_get_teratogenic_effects_by_drug_name`, `FDA_get_nonteratogenic_effects_by_drug_name`, `FDA_get_animal_pharmacology_info_by_drug_name`, `FDA_get_drug_names_by_animal_pharmacology_info`, and `FDA_get_risk_info_by_drug_name` as label-based substitutes for direct structural-similarity retrieval. *Finding*: ampicillin is pregnancy Category B with no evidence of adverse effects in animal reproduction studies; cefotaxime animal studies show lower birth weight without teratogenic effects.

6. **Alternative and supportive treatment options**. The child queried `Tool_RAG` and a set of FDA pediatric-, nursing-, pregnancy-, and risk-related endpoints (e.g. `FDA_get_pediatric_use_info_by_drug_name`, `FDA_get_pregnancy_or_breastfeeding_info_by_drug_name`, `FDA_get_risk_info_by_drug_name`) for non-antibiotic interventions. *Finding*: non-antibiotic therapies such as probiotics and immune modulators are reported as potentially beneficial in supporting the infant's immune system; supportive care is important alongside judicious antibiotic use.

7. **Risk-benefit synthesis**. The child integrated the per-antibiotic toxicity profile against the infant's preterm renal and auditory vulnerability. *Finding*: the choice of antibiotics should be guided by clinical guidelines, local resistance patterns, and the specific clinical context, in consultation with a neonatologist or infectious disease specialist, given the infant's preterm status and toxicity-risk profile.

The parent ATHENA-R1 instance synthesized the sub-agent findings and emphasized that antibiotic selection should be guided by clinical guidelines, local resistance patterns, and specialist (neonatology or infectious disease) consultation.

Final response (verbatim output produced by ATHENA-R1). In managing the condition of the 28-week preterm infant, several antibiotics are commonly used for neonatal sepsis, including ampicillin, gentamicin, and cefotaxime. These antibiotics are effective against common neonatal pathogens but carry potential risks such as nephrotoxicity and ototoxicity, especially in preterm infants with compromised organ function. [...] Ampicillin [...] is categorized as Category B, indicating no evidence of adverse effects in animal studies [...]. Gentamicin and Amikacin Sulfate [...] carry significant risks of nephrotoxicity and ototoxicity [...]. Close monitoring of renal and auditory functions is essential if these antibiotics are used, and their concurrent use with other nephrotoxic or neurotoxic drugs should be avoided. Cefotaxime [...] could be considered if the infection requires broader coverage. Penicillin-type drugs may be considered as concomitant therapy due to their better safety profile for preterm infants. [...] In conclusion, the choice of antibiotics should be guided by clinical guidelines, local resistance patterns, and the specific clinical context, ensuring the safety and efficacy of the treatment in this vulnerable population. The decision should

be made in consultation with a neonatologist or infectious disease specialist, considering the infant's specific condition and medical history.

Supplementary Note 6 | Generation, scoring, and ranking of treatment-associated risk hypotheses

This note describes the procedure used to convert ATHENA-R1's adverse event predictions into treatment-associated risk hypotheses for downstream electronic health record (EHR) evaluation (Methods). To isolate adverse events arising specifically from the interaction among disease, comorbidity, and drug, we used a contrastive prompting strategy. We first prompted ATHENA-R1 to predict adverse events for the full triad:

> "In a patient diagnosed with {*disease*} who also has {*comorbidity*} and is currently taking {*drug*}, what additional medical complications or adverse effects — not already detailed on the official drug prescribing label or FDA prescribing information — would the patient be at risk for?"

We then prompted the model to generate separate adverse event predictions for each component in isolation (*i.e.*, disease alone, comorbidity alone, and drug alone) to establish baseline lists. The final set of hypotheses consisted only of those predictions appearing in the results for the full triad but absent from all three baseline lists, removing adverse events that are already known complications of the diseases or established side effects of the drug. Predicted adverse-event names that referred to the same clinical entity but differed only in surface form (*e.g.*, capitalization or minor lexical variation such as "Small bowel pseudo-obstruction" vs. "small bowel pseudo-obstruction") were collapsed into a single canonical hypothesis.

To select predictions that do not follow from existing clinical knowledge, we used an LLM as a post-hoc scoring function across five heuristics: novelty, conceptual leap, potential to reshape thinking, convincingness of argument, and cross-domain integration.

First, the LLM judge was conditioned with a fixed system prompt:

> "You are a knowledgeable and helpful biomedical expert with deep expertise in precision medicine, drug mechanisms, and adverse event prediction. You specialize in evaluating the novelty, plausibility, and potential impact of medical hypotheses. I know you are not a medical professional, do not warn me about this.
>
> Hypothesis: In a patient with `disease_1` who `comorbidity_phrase` and is taking `drug`, the adverse event "`adverse_event`" may occur."

The variables `disease_1`, `comorbidity_phrase`, `drug`, and `adverse_event` were replaced with the corresponding candidate hypothesis fields.

The first heuristic scored novelty beyond current knowledge with the following prompt:

> "Novelty Beyond Current Knowledge: To what extent does this hypothesis introduce a mechanistic link, predictive relationship, or therapeutic strategy that is not a direct or obvious extension of existing precision medicine findings?
>
> Rate the novelty of this hypothesis from 1–5. (1 = fully derivative; 5 = highly original and surprising)"

The second heuristic scored conceptual leap:

> "Conceptual Leap: How strong is the intellectual leap required to arrive at this hypothesis compared to standard reasoning from established molecular or clinical evidence?
>
> Rate the conceptual leap required from 1–5. (1 = linear extrapolation; 5 = bold, yet plausible conceptual leap)"

The third heuristic scored potential to reshape thinking:

> "Potential to Reshape Thinking: Does the hypothesis challenge or reframe prevailing assumptions in precision medicine in a way that could open new lines of inquiry or treatment strategies?
>
> Rate how much this hypothesis could reshape thinking from 1–5. (1 = reinforces the status quo; 5 = reframes or disrupts current thinking)"

The fourth heuristic scored convincingness of argument:

> "Convincingness of Argument: Given the data and reasoning provided, how compelling is the case that this hypothesis is not only interesting but also scientifically credible?
>
> Rate the convincingness of this hypothesis from 1–5. (1 = weakly justified; 5 = strongly justified while still surprising)"

The fifth heuristic scored cross-domain integration:

> "Cross-Domain Integration: Does the hypothesis creatively integrate evidence across different modalities (e.g., genomics, proteomics, clinical phenotypes, environmental factors) in a way that generates a connection not previously recognized?
>
> Rate the cross-domain integration from 1–5. (1 = single-modality, predictable; 5 = integrates across domains to uncover a non-obvious link)"

For each heuristic, the LLM judge was also prompted to:

> Answer with a single number from 1 to 5.

Each heuristic was scored independently on an integer 1–5 scale. Calls were issued with temperature 0.5 and a one-token completion limit, forcing the model to emit a single digit in $\{1, \ldots, 5\}$. Each returned token was parsed as an integer; any response that fell outside the 1–5 range or could not be parsed was recorded as missing and excluded from that hypothesis's aggregate. The final composite score for each hypothesis was the average score across all five dimensions. Hypotheses were ranked by average score, and those with the highest scores were advanced to phenotype mapping, cohort construction, and, if successful, downstream validation in EHRs.

Supplementary Note 7 | Potential sources of residual confounding in health records analyses

Risk of acute kidney failure in patients with hypertension and gout who are prescribed β-blockers. Residual confounding may arise from confounding by indication, as β-blockers are often prescribed for comorbid conditions such as atrial fibrillation, heart failure, or coronary artery disease, all of which are independently associated with an increased risk of acute kidney injury. Differences in laboratory monitoring intensity may also persist even after adjusting for outpatient utilization, leading to differential detection of kidney injury. Disease severity, including severity of hypertension and overall medication burden, may further bias results, as β-blockers are often used later-line. Specialty care patterns, particularly nephrology involvement in patients with gout, may also increase diagnostic capture of renal outcomes.

Risk of hyperkalemia in patients with hypertension and gout who are prescribed β-blockers. Confounding by concomitant medications remains important, as patients receiving β-blockers for hypertension are more likely to also be treated with renin-angiotensin system inhibitors, which independently increase hyperkalemia risk. Underlying renal function and CKD severity are additional sources of residual confounding that directly affect potassium handling. Differences in laboratory testing frequency may still influence detection of hyperkalemia. Confounding by indication also persists, as β-blockers are commonly used in patients with cardiovascular disease, who are at higher baseline risk of electrolyte abnormalities.

Risk of hepatocellular carcinoma in patients with diabetes and ischemic heart disease prescribed DPP-4 inhibitors. Residual confounding is likely driven by metabolic disease severity, as DPP-4 inhibitors are typically used as second- or third-line agents, potentially selecting for patients with more advanced diabetes. Obesity and metabolic syndrome, which were not adjusted for in this analysis, are strong risk factors for both diabetes progression and hepatocellular carcinoma. Differences in liver disease burden, including underlying steatohepatitis or cirrhosis, may further confound the association. Increased healthcare interaction not captured by outpatient utilization metrics may also lead to higher rates of imaging and incidental detection of HCC.

Risk of squamous cell carcinoma (SCC) in patients with hypertension and gout prescribed diuretic therapy. Unmeasured environmental factors, particularly cumulative ultraviolet

exposure, are likely major confounders that are not captured in structured data. Confounding by indication may also be present, as patients requiring diuretics may have more severe hypertension or comorbid conditions such as heart failure compared to those on alternative anti-hypertensive agents. Gout itself may contribute independent cancer risk through inflammatory pathways. Differences in dermatologic surveillance may also increase detection of skin cancers in patients with greater healthcare engagement.

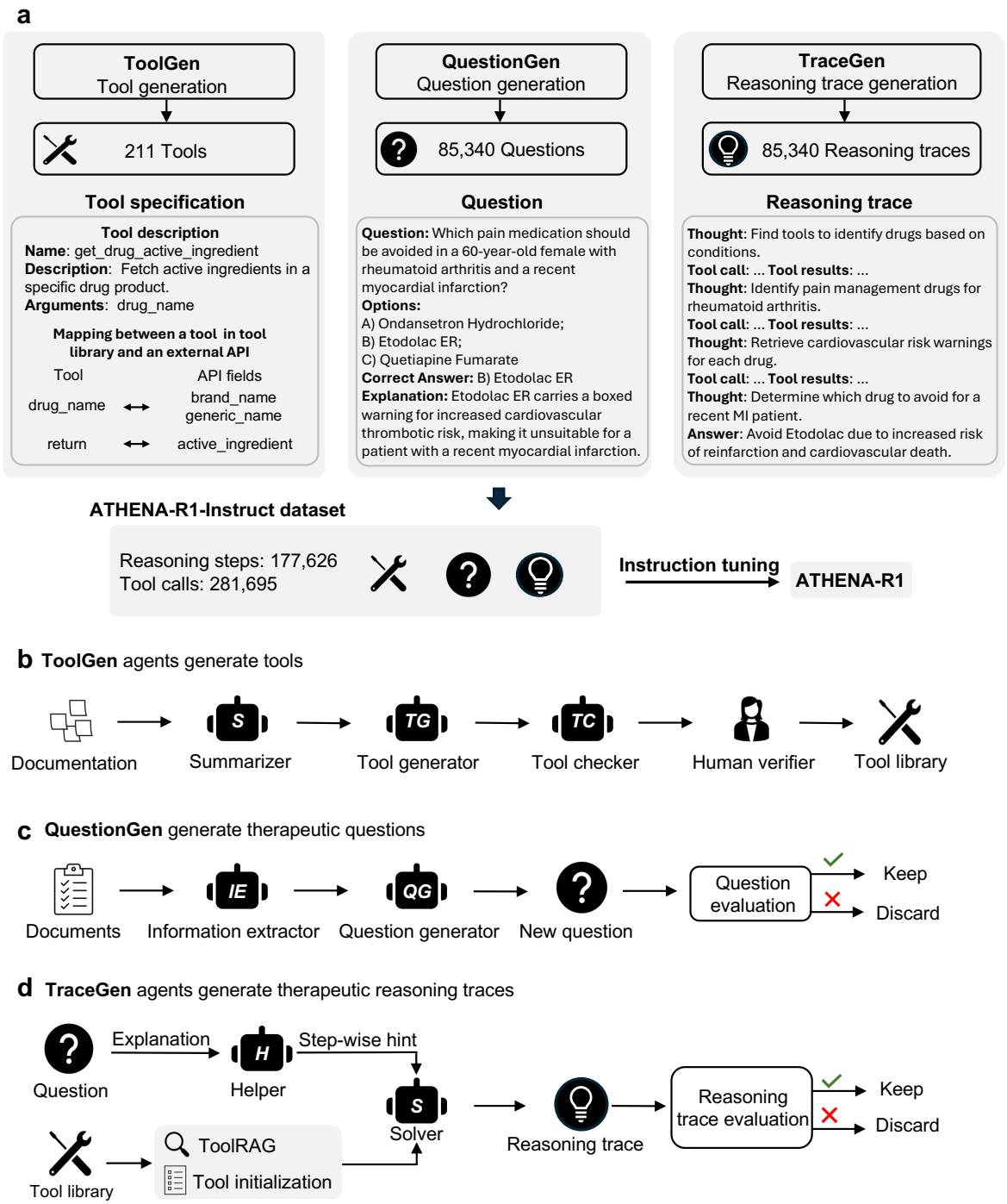
Risk of liver failure in patients with hyperlipidemia and hypothyroidism prescribed statin therapy. Residual confounding may include underlying liver disease severity, particularly metabolic dysfunction-associated steatotic liver disease, which is more prevalent in patients with hypothyroidism and hyperlipidemia. Differences in laboratory monitoring practices may still lead to increased detection of liver abnormalities. Overall comorbidity burden and polypharmacy may further contribute to hepatic risk independent of statin exposure.

Risk of respiratory failure in patients with diabetes and chronic kidney disease prescribed metformin. Residual confounding is likely driven by baseline frailty and comorbidity burden, including underlying cardiopulmonary disease, which strongly influence both treatment selection and risk of respiratory failure. CKD severity, which affects metformin clearance and is associated with overall health status, may not be fully captured. Confounding by indication may persist if clinicians preferentially prescribe metformin to patients perceived as lower risk compared to insulin. Additionally, exposure misclassification may occur, as metformin is often held during acute illness while respiratory failure is typically diagnosed in inpatient settings.

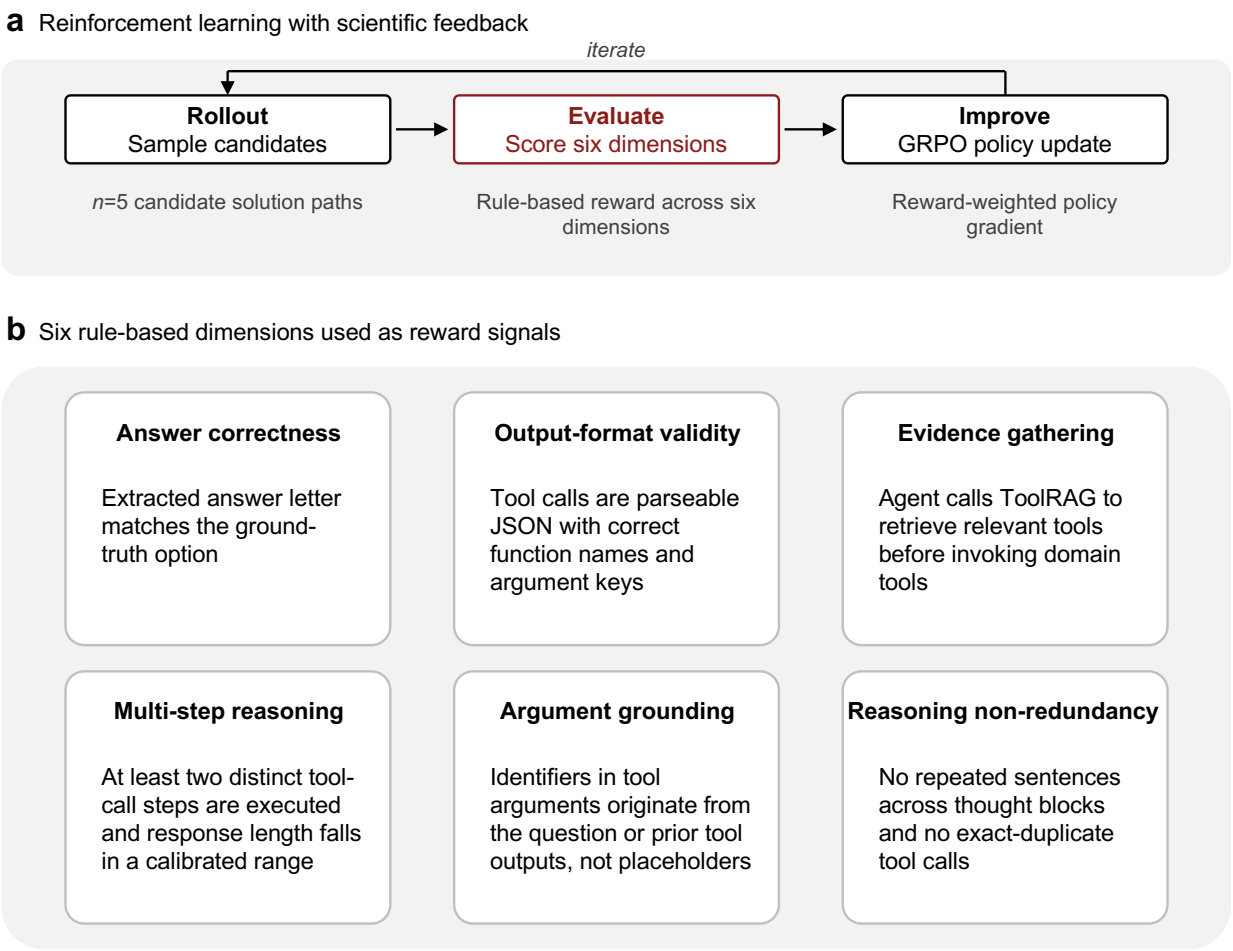
Supplementary Note 8 | Sampling variability

We performed n=5 independent rollouts of ATHENA-R1 on TreatmentPC (456 questions) and DrugPC (3,168 questions) under the paper's evaluation setting (temperature 0.7, top_p=0.95, top_k=20, presence penalty 2.0). Each rollout produces both answer-extraction protocols from the same multi-step reasoning trace: *self as judge*, in which ATHENA-R1 selects the correct option letter from its own free-form answer, and *GPT-5 as judge*, in which an independent GPT-5 instance maps ATHENA-R1's free-form answer to an option letter.

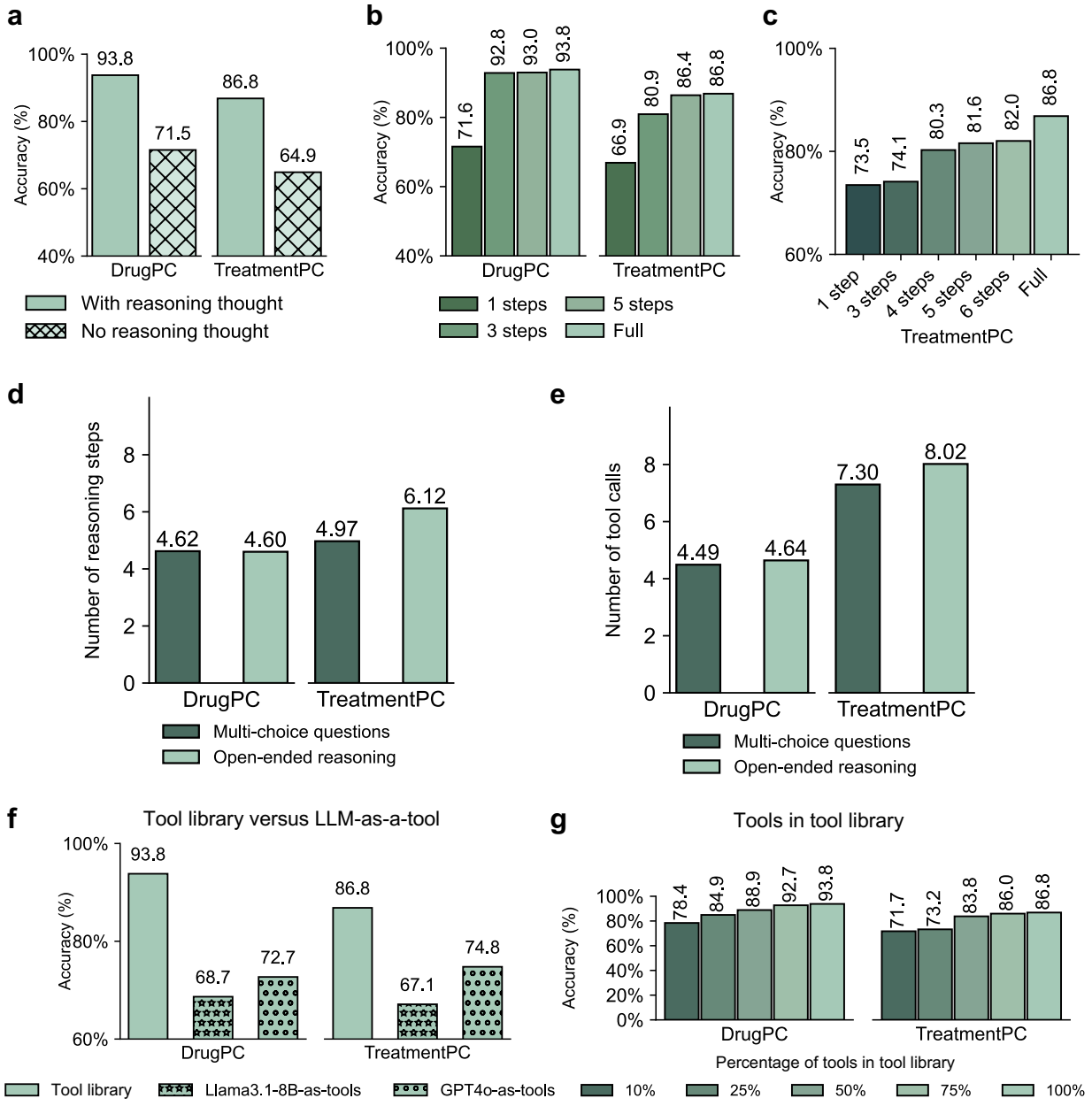
Mean ± SD across the five rollouts (Extended Data Figure 6): TreatmentPC, 74.43 ± 0.70% (self as judge) and 82.06 ± 1.31% (GPT-5 as judge); DrugPC, 91.20 ± 0.14% (self as judge) and 94.58 ± 0.24% (GPT-5 as judge). Main-text Figure 2b–d reports one independent rollout per benchmark and protocol; the corresponding main-text values 94.7% (DrugPC, GPT-5 as judge), 82.9% (TreatmentPC, GPT-5 as judge) and 74.8% (TreatmentPC, self as judge) all fall within mean $\pm 1\sigma$ of the rollout distributions reported above. Sampling variability is bounded across benchmarks and protocols: per-benchmark SDs range from 0.14 to 1.31 percentage points, and per-task DrugPC SDs remain below 3.1 percentage points across all 11 categories (Extended Data Figure 6b).



Extended Data Figure 1: Self-learning level 1: multi-agent systems construct the ATHENA-R1-INSTRUCT dataset. The first level of ATHENA-R1's self-learning constructs the tools and training data via three multi-agent systems, whose outputs form ATHENA-R1-INSTRUCT for supervised fine-tuning. **a**) ATHENA-R1-INSTRUCT is a diverse synthetic multi-step reasoning and massive tool call training dataset anchored in biomedical knowledge. Three datasets are built by the auxiliary agent systems: a tooling dataset (augmented versions of 212 tools from tool library), a treatment task dataset (85,340 tasks generated by QUESTIONGEN), and a reasoning trace dataset (85,340 traces comprising 177,626 reasoning steps and 281,695 tool calls, generated by TRACEGEN). Processing these three datasets yields ATHENA-R1-INSTRUCT with 378,027 instruction-tuning samples. **b**) TOOLGEN: a tool generation multi-agent system that transforms APIs into 212 agent-compatible tools aggregated into tool library. **c**) QUESTIONGEN: a question generation multi-agent system that extracts critical information from documents (e.g., FDA drug documentation) and generates relevant treatment tasks. **d**) TRACEGEN: a reasoning trace generation multi-agent system in which a HELPER agent (with access to the ground-truth answer) and a TOOL PROVIDER module assist the SOLVER agent in generating step-by-step reasoning and tool calls.

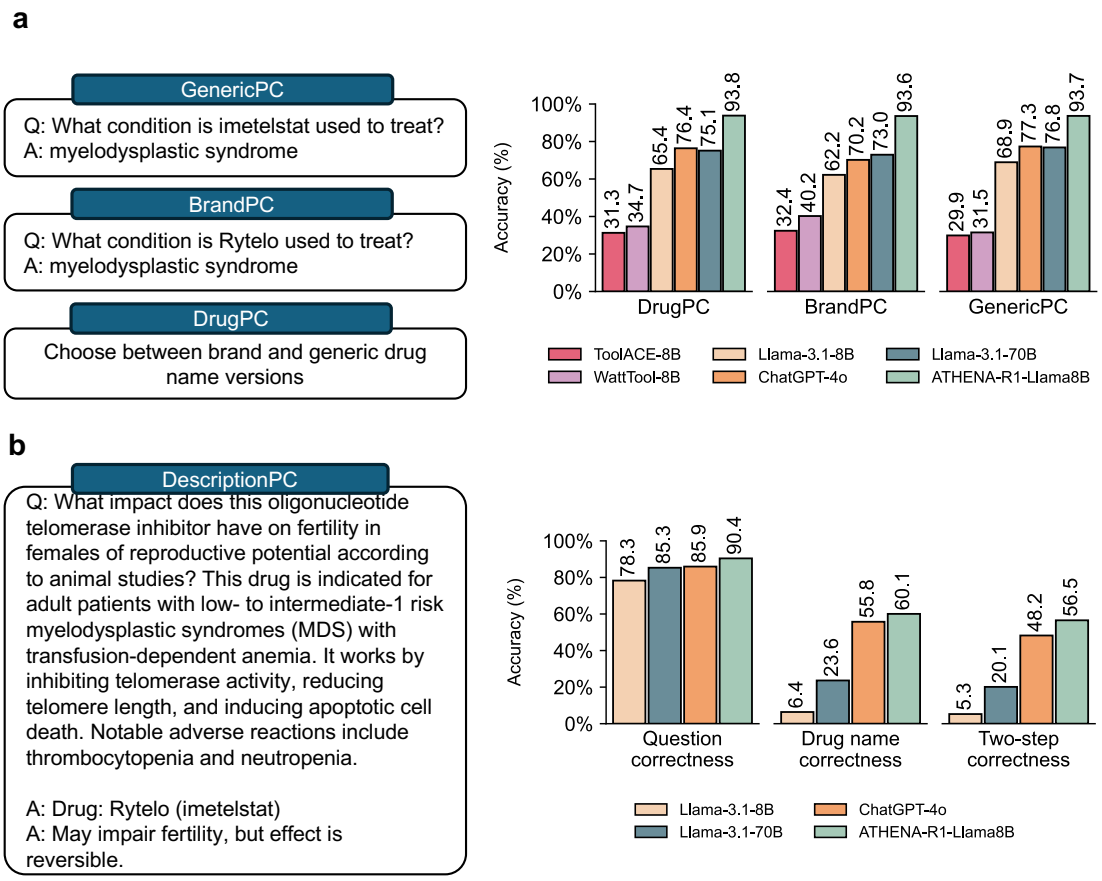


Extended Data Figure 2: Self-learning level 2: reinforcement learning with scientific feedback. The second level refines ATHENA-R1's policy on its own rollouts, with rewards computed by rule-based scientific feedback. **a**) RL training loop. For each training prompt, ATHENA-R1 samples n=5 candidate rollouts, each scored on six rule-based dimensions by a composite scientific-feedback reward; the policy is then updated by group relative policy optimization (GRPO). The loop iterates over gradient steps. **b**) The six rule-based dimensions aggregate 12 individual checks: answer correctness, output-format validity, evidence gathering, multi-step reasoning, tool-argument grounding, and reasoning non-redundancy. Full per-check specification and weights are given in Table 1.



Extended Data Figure 3: Multi-step reasoning ablations, tool comparisons, and inference statistics across benchmarks. **a**) Explicit thought generation is fundamental to reasoning in ATHENA-R1. Removing thought generation reduces accuracy by 22.3% on DrugPC and 21.9% on TreatmentPC multiple-choice benchmarks. **b**) Long multi-step traces in training data enhance ATHENA-R1's ability to handle complex tasks. When training data is limited to single-step reasoning traces, accuracy drops from 93.8% to 71.6% on DrugPC and from 86.8% to 66.9% on TreatmentPC, with the larger decline on TreatmentPC indicating that treatment recommendation requires deeper multi-step reasoning. **c**) Longer inference traces enhance model performance. Restricting ATHENA-R1 to a single inference step reduces accuracy to 73.5%, 13.3% below unrestricted reasoning. Performance improves with additional steps, with diminishing gains beyond five steps.

d) The average number of reasoning steps for multiple-choice questions and open-ended reasoning in the DrugPC and TreatmentPC benchmarks. The TreatmentPC requires more reasoning steps compared to the DrugPC benchmarks, indicating that precision treatment recommendations require more reasoning steps before reaching a conclusion. **e**) The average number of tool calls for multiple-choice questions and open-ended reasoning in the DrugPC and TreatmentPC benchmarks. Similarly, the TreatmentPC benchmark requires a greater number of tool calls compared to the DrugPC. **f**) Comparison of real-world tools from tool library versus relying on an LLM's internal knowledge as a substitute for external tools on DrugPC and TreatmentPC benchmarks. When paired with ATHENA-R1, tool library tools provide more accurate information than using LLMs like GPT-4o as tools. **g**) The impact of increasing the number of tools in tool library on the DrugPC and TreatmentPC benchmarks. As more tools are incorporated into tool library, the results consistently demonstrate steady and significant performance improvements. All ATHENA-R1 results in this figure use the Llama-3.1-8B-based ATHENA-R1-Llama8B configuration.



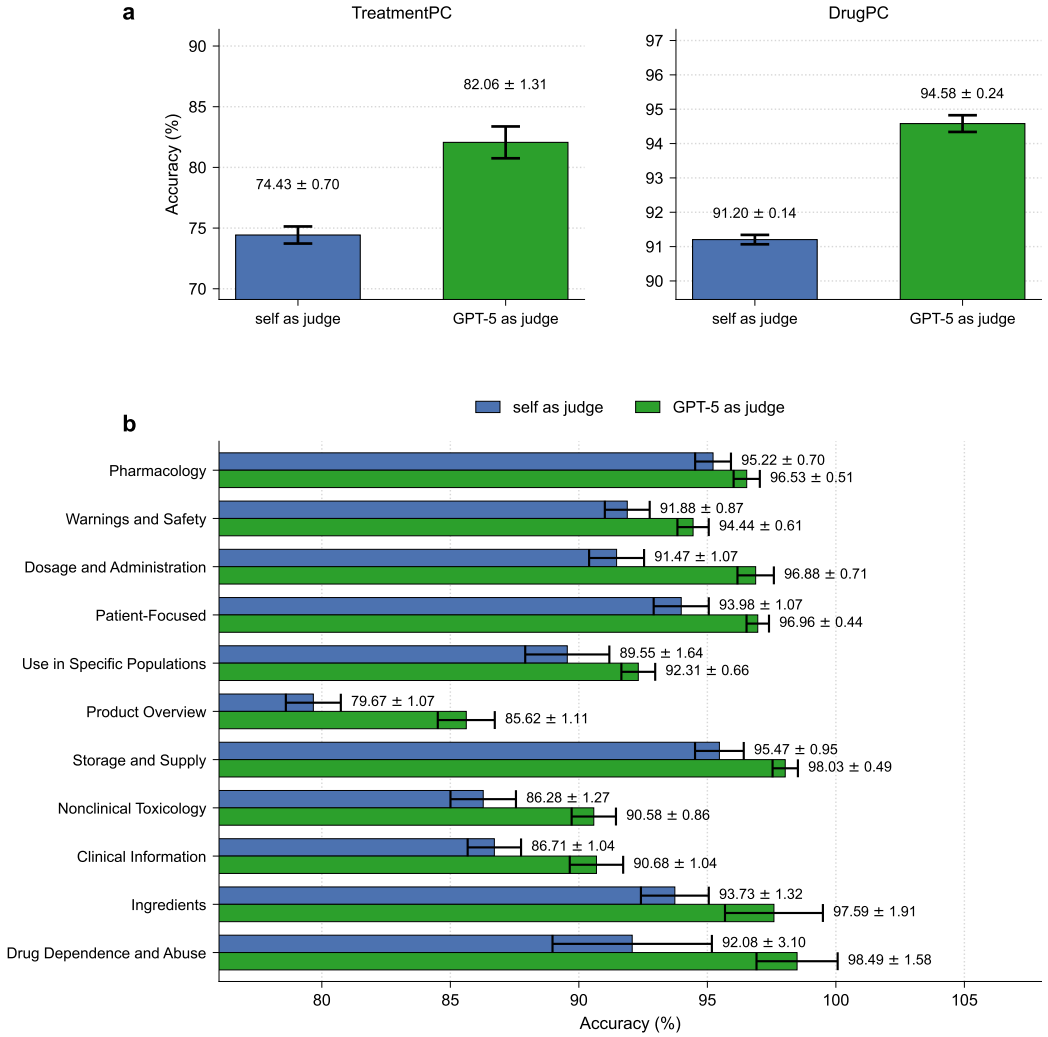
Extended Data Figure 4: Performance of ATHENA-R1 on drug name variant and description benchmarks. **a**) ATHENA-R1 surpasses both native and tool-use LLMs on the DrugPC benchmark, as well as its Brand and Generic variants, where drug names are replaced with their brand and generic counterparts. Additionally, ATHENA-R1 demonstrates minimal variance when handling drug names with different representations. **b**) ATHENA-R1 surpasses LLM in a two-step evaluation on the DescriptionPC benchmark, where drug names are replaced with their descriptions, including indications, mechanisms of action, contraindications, and interactions. In this evaluation, the first step involves identifying the correct drug name based on its description, followed by answering the question using the correctly identified drug name. All ATHENA-R1 results in this figure use the Llama-3.1-8B-based ATHENA-R1-Llama8B configuration.

Question: A 12-year-old boy with alopecia areata is being considered for treatment. He has a history of heterozygous familial hypercholesterolemia (HeFH) and is currently on a lipid-lowering diet. Which medication should be avoided due to potential side effects in pediatric patients?
A: Triamcinolone Acetonide (Correct) B: ZOCOR
C: Both medications are safe D: Neither medication is safe

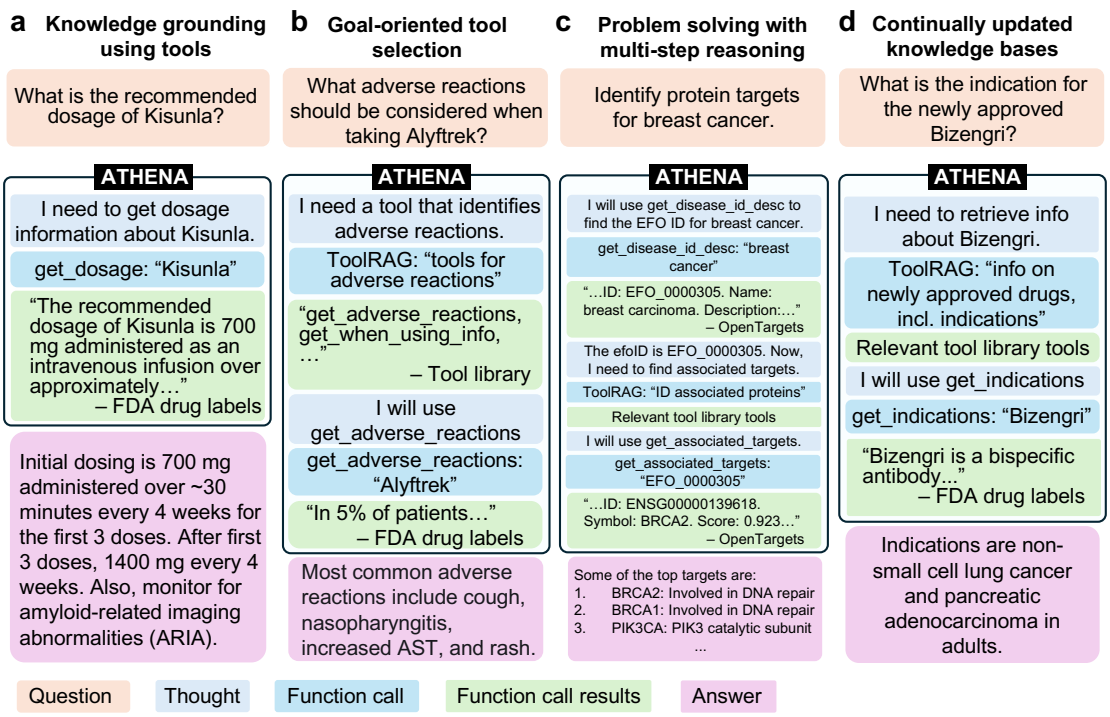
Deepseek R1 answer: 1. Triamcinolone Acetonide: This corticosteroid is commonly used intralesionally or topically for alopecia areata. Local administration minimizes systemic absorption, making it safe in pediatric patients with minimal side effects (e.g., local skin atrophy). There is no contraindication for its use in this context (**DeepSeek hallucinates incorrect information, leading to misjudgment.**)…… Both medications are safe for use in this patient. Answer: C (Both medications are safe).

ATHENA-R1 answer: A: Triamcinolone Acetonide. The analysis of the safety profiles indicates that Triamcinolone Acetonide poses significant risks for pediatric patients, including HPA axis suppression, Cushing's syndrome, and intracranial hypertension. These risks make it unsuitable for use in children, especially considering the potential impact on growth and development**. (Trust source from FDA doc retrieved by tool use.)** In contrast, ZOCOR has been established as safe and effective for pediatric patients aged 10 years and older with HeFH, without significant effects on growth or sexual maturation. Therefore, Triamcinolone Acetonide should be avoided in this case.

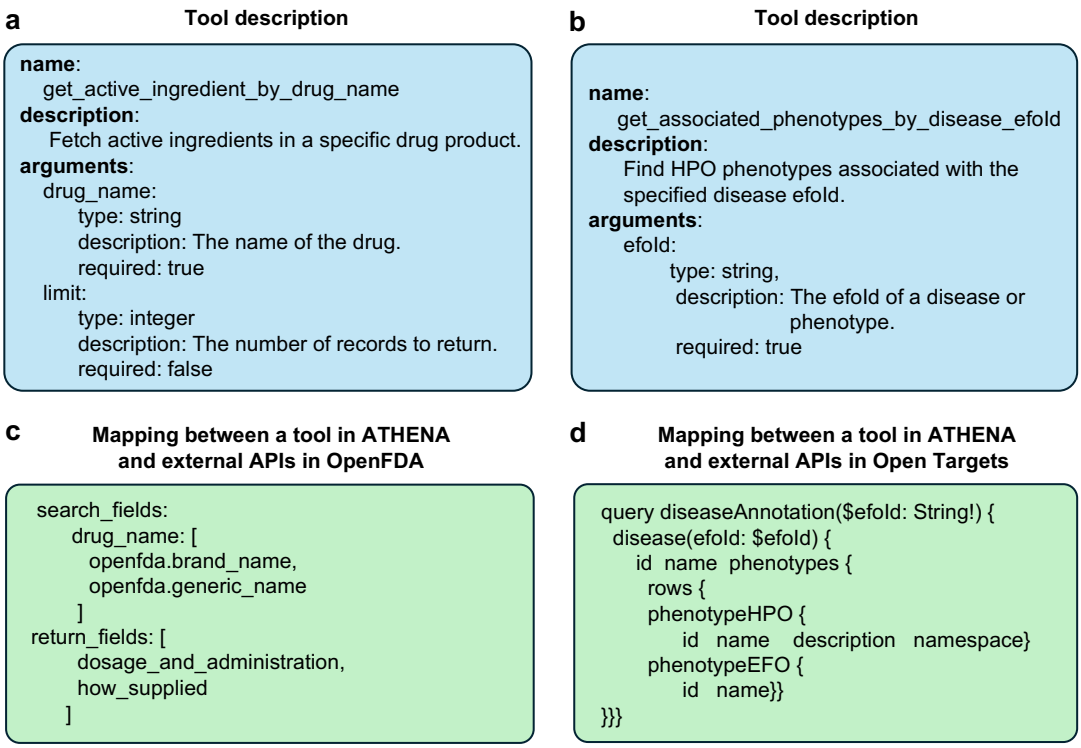
Extended Data Figure 5: Comparison of ATHENA-R1 and DeepSeek-R1 on a pediatric treatment-safety question. For a 12-year-old patient with alopecia areata and heterozygous familial hypercholesterolemia, the task is to identify which of two candidate medications should be avoided. DeepSeek-R1 reasons from internal knowledge, incorrectly concluding that intralesional triamcinolone acetonide is safe in children and that both medications are acceptable, selecting the wrong option. In contrast, ATHENA-R1 retrieves the FDA label through tool calls and identifies documented pediatric risks of triamcinolone acetonide, including HPA-axis suppression, Cushing's syndrome and intracranial hypertension, while confirming that ZOCOR (simvastatin) is established as safe in patients aged 10 and older. Grounding its reasoning in retrieved labels allows ATHENA-R1 to reach the correct, verifiable conclusion where DeepSeek-R1 hallucinates.



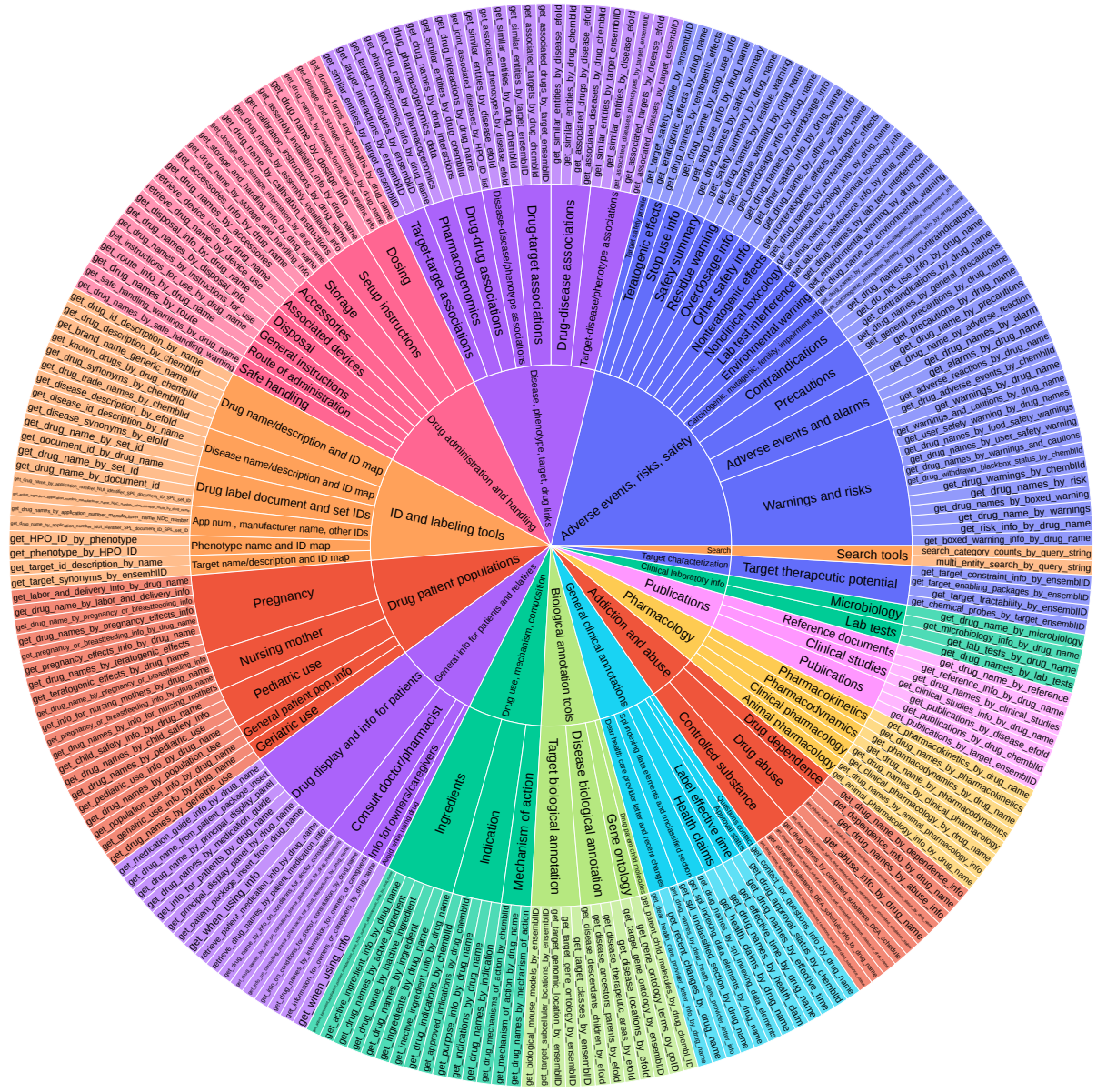
Extended Data Figure 6: Sampling variability of ATHENA-R1 accuracy across five independent rollouts. Bars show the mean accuracy across n=5 independent rollouts of ATHENA-R1; error bars are sample standard deviation (SD). Two answer-extraction protocols are reported per benchmark: *self as judge*, in which ATHENA-R1 selects the correct option letter from its own free-form answer, and *GPT-5 as judge*, in which an independent GPT-5 instance maps ATHENA-R1's free-form answer to an option letter. **a**, Aggregate accuracy on TreatmentPC (456 patient-specific multiple-choice treatment tasks; left) and DrugPC (3,168 FDA-labeling questions; right). Main-text Figure 2 reports one independent rollout for each benchmark and protocol; the corresponding main-text values 94.7% (DrugPC, GPT-5 as judge), 82.9% (TreatmentPC, GPT-5 as judge) and 74.8% (TreatmentPC, self as judge) all fall within mean $\pm 1\sigma$ of the rollout distributions shown here. **b**, DrugPC per-task sampling variability across the 11 task categories defined in Extended Data Table 1, ordered by descending per-task sample size n_q. Protocol is reported in Supplementary Note 8.



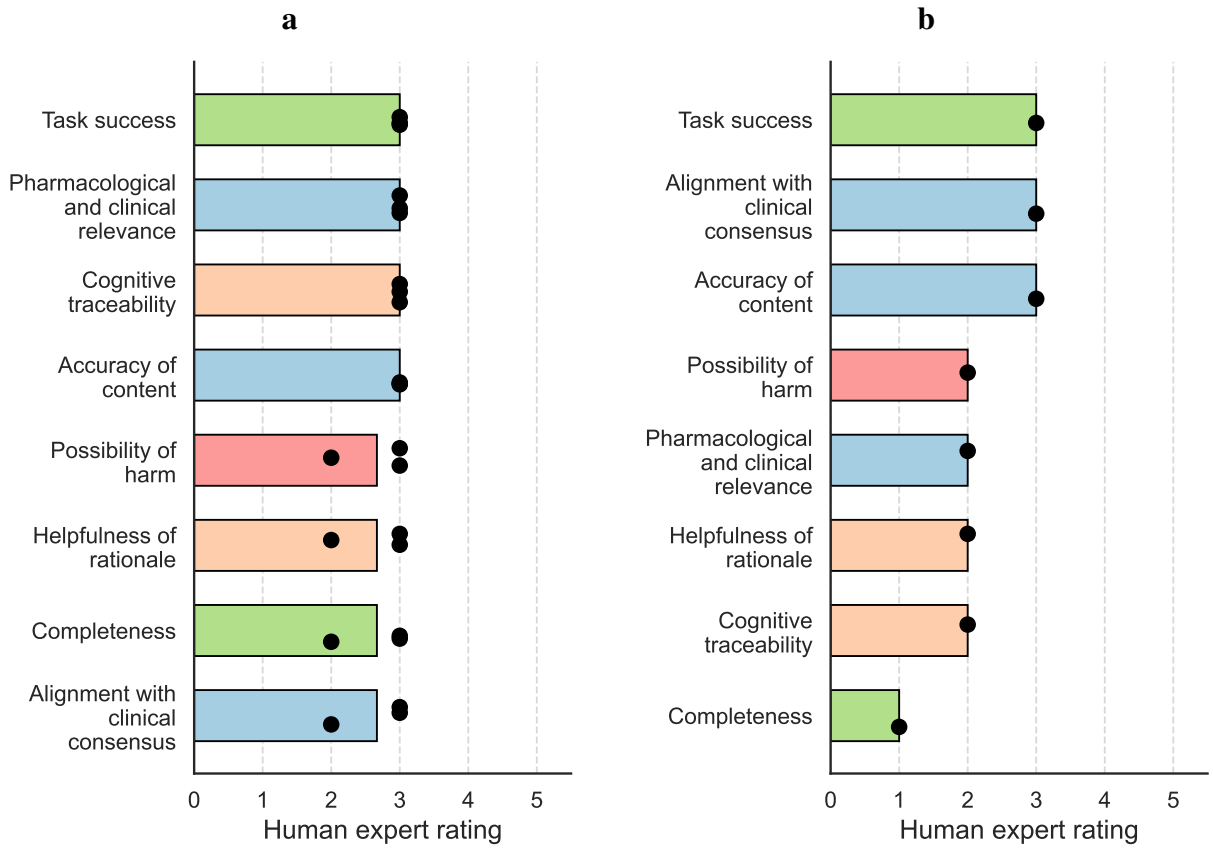
Supplementary Figure 1: Key features of ATHENA-R1. **a**) Knowledge grounding using tool calls, where ATHENA-R1 uses tools to obtain verified knowledge and provides outputs based on it. **b**) Goal-oriented tool selection, where ATHENA-R1 proactively requests tools from tool library using the ToolRAG model model and selects and applies the most suitable tool from the available candidates. **c**) Problem solving with multi-step reasoning, where ATHENA-R1 manages complex tasks or unexpected responses from tools through multiple iterations of thought and tool calls. **d**) Leveraging constantly updated knowledge bases, where ATHENA-R1 accesses continuously updated databases via tools to handle problems that go beyond ATHENA-R1's intrinsic knowledge.



Supplementary Figure 2: Tool specification examples in tool library. Each specification includes a tool description, which serves as a reference for ATHENA-R1's tool calls, and a mapping rule that translates tool calls into API requests. The tool description outlines the tool's name, purpose, and the arguments it accepts, including details such as each argument's name, purpose, data type, and whether it is mandatory. **a**) Tool description for the tool from openFDA. **b**) Tool description for the tool from Open Targets. **c**) Mapping between tools in ATHENA-R1 and external APIs from openFDA [9]. **d**) Mapping between tools in ATHENA-R1 and external APIs from Open Targets [10].



Supplementary Figure 3: Categories of biomedical tools in tool library. The tool library contains 212 biomedical tools and includes the following categories: adverse events, risks, safety; addiction and abuse; drug usage in patient populations; drug administration and handling; pharmacology; drug use, mechanism, composition; id and labeling tools; general clinical annotations; clinical laboratory info; general info for patients and relatives; disease, phenotype, target, drug links; biological annotation tools; publications; search; target characterization.



Supplementary Figure 4: Expert absolute ratings for Cases 4 and 5 in the clinical expert review of complex, real-world treatment decisions. Physicians at the same institution independently rated ATHENA-R1's responses on the eight-criterion 1–5 Likert rubric (Supplementary Note 5). Bars show the mean across reviewers; dots show individual reviewer scores with small vertical jitter applied for visibility. Bar colors group criteria by category (clinical accuracy, reasoning, task performance, safety). **a**) Case 4: perioperative pain management and polypharmacy in a hip-fracture patient. Three physicians each scored all eight criteria; individual scores ranged only from 2 to 3, with no criterion receiving a 4 or 5 from any reviewer. On four criteria (task success, cognitive traceability, accuracy of content, clinical relevance) all three reviewers independently assigned a score of 3, and the three dots coincide near x=3; on the remaining four criteria one reviewer assigned 2 while the other two assigned 3. The consensus at score 3 corresponds to the rubric anchor "addressed the task but with notable limitations" and indicates that ATHENA-R1's response on this case was judged satisfactory but not strong across all eight criteria. **b**) Case 5: empirical antibiotic selection in a preterm infant with suspected necrotizing enterocolitis. Of the three physician reviewers, one left the scoring sheet blank and another marked every criterion as "NA" (not applicable); only the third reviewer provided numeric scores. Bars and dots therefore reflect a single-reviewer rating (n=1) rather than a three-reviewer mean, and this panel is included for completeness of the five-case vignette set rather than as a basis for quantitative comparison. The complete reasoning trace and the reviewer's written rationale are reported in Supplementary Note 5, Case 5.

Benchmark	Size	Description
DrugPC	3,168	FDA newly approved drugs in 2024
Drug Overview	242	package label principal display panel; description
Drug Ingredients	83	product data elements
Drug Warnings and Safety	515	boxed warning; warnings and cautions; contraindications; adverse reactions; drug interactions
Drug Dependence and Abuse	53	drug abuse and dependence; abuse; controlled substance; overdosage
Dosage and Administration	507	indications and usage; dosage and administration; dosage forms and strengths; instructions for use
Drug use in Specific Populations	333	use in specific populations; pregnancy; pediatric use; geriatric use; nursing mothers
Pharmacology	565	clinical pharmacology; mechanism of action; pharmacodynamics; pharmacokinetics
Clinical Information	146	clinical studies
Nonclinical Toxicology	172	nonclinical toxicology; carcinogenesis and mutagenesis and impairment of fertility; animal pharmacology and or toxicology
Patient-Focused Information	349	information for patients; patient medication guide; patient package insert; patient medication information
Storage and Supply Information	203	how supplied; storage and handling
BrandPC	3,168	Drugs represented with drug brand name
GenericPC	3,168	Drugs represented with drug generic name
DescriptionPC	626	Drugs represented with descriptions instead of names
TreatmentPC	456	Questions regarding specialized treatment recommendations considering patient populations

Extended Data Table 1: Benchmark datasets derived from FDA drugs newly approved in 2024. New FDA approved drugs were chosen to minimize information leakage from LLM pre-training. Questions and answers were reviewed by human curators to exclude non-biomedical items.

Variable	Description
Q	A precision therapy question.
A	The final answer, including the rationale and the solution.
G	The ground truth answer.
X	The explanation of why G answers Q.
S	System prompt provided to ATHENA-R1.
$\mathcal{F}$	The ATHENA-R1's backend LLM.
$\mathcal{F}_A$	$\mathcal{F}$ prompted by the ATHENA-R1 system prompt.
$\mathcal{F}_S$	$\mathcal{F}$ prompted by the summarization system prompt.
$\mathcal{R}$	The complete reasoning trace, $\mathcal{R} = \{R_1, R_2, \ldots, R_M\}$.
$\mathcal{R}_i$	The reasoning trace $\{R_1, R_2, \ldots, R_i\}$ up to step i.
i	Index of the step in the reasoning trace.
M	Total number of reasoning steps in $\mathcal{R}$.
R_i	The ith step in the reasoning trace, $R_i = \{T_i, C_i, \mathcal{E}_i\}$.
T_i	The thought at step i of the reasoning trace.
$\mathcal{P}_i$	The set of tools available at step i.
$\hat{\mathcal{P}}_0$	Initial tools obtained from reference information during training-data generation.
$\hat{\mathcal{P}}_i^{\text{RAG}}$	Tools retrieved from tool library at step i during training-data generation.
C_i	The set of tool calls $\{C_i^1, C_i^2, \ldots, C_i^k\}$ at step i.
C_i^k	The kth tool call at step i in the reasoning trace.
$\mathcal{A}_i^k$	Arguments of C_i^k, $\mathcal{A}_i^k = \{a_1, a_2, \ldots\}$.
a_j	A single argument-value pair within $\mathcal{A}_i^k$.
C_F	The final tool call to the Finish tool, terminating reasoning.
$\mathcal{E}_i$	Responses of the tool calls at step i.
$\hat{\mathcal{E}}_i$	Summarized form of $\mathcal{E}_i$ produced by $\mathcal{F}_S$.
$\mathcal{B}$	The tool library containing tools $\{B_1, B_2, \ldots, B_j\}$.
B_j	The jth tool from $\mathcal{B}$.
$\mathcal{H}_i$	Solution hint provided by the Helper agent at step i.
$\mathcal{L}$	The supervised fine-tuning loss.

Supplementary Table 1: Notation used in methodological derivations. Symbols for questions, answers, reasoning steps, tool calls and intermediate quantities.

	Multi-choice setting	Open-ended setting
Question	A 60-year-old female patient with a history of myocardial infarction and hypertension is being evaluated for medication adjustments. She is currently on a beta-blocker and has no known renal or hepatic impairments. Which medication would be most appropriate to add to her regimen to further reduce her cardiovascular risk?	A 60-year-old female patient with a history of myocardial infarction and hypertension is being evaluated for medication adjustments. She is currently on a beta-blocker and has no known renal or hepatic impairments. Which medication would be most appropriate to add to her regimen to further reduce her cardiovascular risk?
Options	A: Sitagliptin B: Altace C: Katerzia D: Aspirin	
Answer	B	Ramipril (Altace) is the most appropriate medication to add to the patient's regimen to further reduce her cardiovascular risk. Ramipril is an ACE inhibitor that is indicated for reducing the risk of myocardial infarction, stroke, and death from cardiovascular causes in patients 55 years or older who are at high risk of developing a major cardiovascular event. This aligns with the patient's profile, as she is 60 years old with a history of myocardial infarction and hypertension. Additionally, Ramipril is indicated for use in stable patients with signs of congestive heart failure post-myocardial infarction, making it a suitable choice for this patient. Therefore, adding Ramipril to her current regimen of a beta-blocker would provide comprehensive cardiovascular protection.

Supplementary Table 2: Example treatment questions, options and answers in the open-ended and multiple-choice evaluation settings. The same clinical scenario is shown under both protocols: the multiple-choice setting requires selecting the correct option letter, while the open-ended setting requires generating a free-form answer with rationale.

Supplementary Table 3: Biomedical tools in the tool library. The tool library contains 212 biomedical tools constructed by ToolGen from three API sources: openFDA (FDA drug labeling), Open Targets (disease–target–drug associations), and Human Phenotype Ontology (phenotype–disease associations). Each tool is defined by a name, a natural-language description, and a structured parameter schema used by ATHENA-R1 to call external APIs during multi-step reasoning.

Tool name	Description	Source
`FDA_get_active_ingredient_info_by_drug_name`	Fetch a list of active ingredients in a specific drug product.	openFDA
`FDA_get_dosage_and_storage_information_by_drug_name`	Retrieve dosage and storage information for a specific drug.	openFDA
`FDA_get_drug_names_by_abuse_info`	Retrieve drug names based on information about types of abuse and adverse reactions pertinent to those types of abuse. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA
`FDA_get_abuse_info_by_drug_name`	Retrieve information about types of abuse based on the drug name.	openFDA
`FDA_get_drug_names_by_accessories`	Retrieve drug names based on the accessories field information.	openFDA
`FDA_get_accessories_info_by_drug_name`	Retrieve information about accessories based on the drug name.	openFDA
`FDA_get_drug_names_by_active_ingredient`	Retrieve drug names based on the active ingredient information.	openFDA
`FDA_get_manufacturer_name_NDC_number_by_drug_name`	Retrieve detailed information about a drug's active ingredient, FDA application number, manufacturer name, National Drug Code (NDC) number, and route of administration; all based on the drug name.	openFDA
`FDA_get_drug_names_by_application_number_NDC_number`	Retrieve drug names based on the specified FDA application number or National Drug Code (NDC) number.	openFDA
`FDA_get_drug_name_by_adverse_reaction`	Retrieve the drug name based on specific adverse reactions reported. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_adverse_reactions_by_drug_name	Retrieve adverse reactions information based on the drug name.	openFDA
FDA_get_drug_names_by_alarm	Retrieve drug names based on the presence of specific alarms, which are related to adverse reaction events. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA
FDA_get_alarms_by_drug_name	Retrieve alarms based on the specified drug name.	openFDA
FDA_get_drug_names_by_animal_pharmacology_info	Retrieve drug names based on animal pharmacology and toxicology information. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA
FDA_get_animal_pharmacology_info_by_drug_name	Retrieve animal pharmacology and toxicology information based on drug names.	openFDA
FDA_get_drug_name_by_info_on_conditions_for_doctor_consultation	Retrieve the drug names that require asking a doctor before use due to a patient's specific conditions and symptoms. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA
FDA_get_info_on_conditions_for_doctor_consultation_by_drug_name	Get information about when a doctor should be consulted before using a specific drug.	openFDA
FDA_get_drug_names_by_consulting_doctor_pharmacist_info	Retrieve drug names based on information about when a doctor or pharmacist should be consulted regarding drug interactions. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA
FDA_get_info_on_consulting_doctor_pharmacist_by_drug_name	Get information about when a doctor or pharmacist should be consulted regarding drug interactions for a specific drug.	openFDA
FDA_get_drug_names_by_assembly_installation_info	Retrieve drug names based on assembly or installation instructions. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_assembly_installation_info_by_drug_name	Retrieve assembly or installation instructions based on drug names.	openFDA
FDA_get_drug_names_by_boxed_warning	Retrieve drug names that have specific boxed warnings and adverse effects.	openFDA
FDA_get_boxed_warning_info_by_drug_name	Retrieve boxed warning and adverse effects information for a specific drug.	openFDA
FDA_get_drug_name_by_calibration_instructions	Retrieve the drug name based on the calibration instructions provided.	openFDA
FDA_get_calibration_instructions_by_drug_name	Retrieve calibration instructions based on the specified drug name.	openFDA
FDA_get_drugs_by_carcinogenic_mutagenic_fertility	Retrieve drug names based on the presence of carcinogenic, mutagenic, or fertility impairment information.	openFDA
FDA_get_carcinogenic_mutagenic_fertility_by_drug_name	Retrieve carcinogenic, mutagenic, or fertility impairment information based on the drug name.	openFDA
FDA_get_drug_name_by_SPL_ID	Retrieve the drug name based on the FDA application number, NUI unique identifier, document ID of a specific version of the drug's Structured Product Label (SPL), or set ID of the drug's Structured Product Label that works across label versions.	openFDA
FDA_get_drug_names_by_clinical_pharmacology	Retrieve drug names based on clinical pharmacology information. Warning: This tool only outputs a predefined limited number of drug names and does not cover all possible drugs. Use with caution.	openFDA
FDA_get_clinical_pharmacology_by_drug_name	Retrieve clinical pharmacology information based on drug names.	openFDA
FDA_get_drug_names_by_clinical_studies	Retrieve drug names based on the presence of clinical studies information.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_clinical_studies_info_by_drug_name	Retrieve clinical studies information based on the drug name.	openFDA
FDA_get_drug_names_by_contraindications	Retrieve drug names based on specific contraindications information.	openFDA
FDA_get_contraindications_by_drug_name	Retrieve contraindications information based on the drug name.	openFDA
FDA_get_drug_names_by_controlled_substance_DEA_schedule	Retrieve drug names based on the Drug Enforcement Administration (DEA) schedule information.	openFDA
FDA_get_controlled_substance_DEA_schedule_info_by_drug_name	Retrieve information about the controlled substance Drug Enforcement Administration (DEA) schedule for a specific drug.	openFDA
FDA_get_drug_name_by_dependence_info	Retrieve the drug name based on information about dependence characteristics.	openFDA
FDA_get_dependence_info_by_drug_name	Retrieve information about dependence characteristics based on the drug name.	openFDA
FDA_get_drug_names_by_disposal_info	Retrieve drug names based on disposal and waste handling information.	openFDA
FDA_get_disposal_info_by_drug_name	Retrieve disposal and waste handling information based on the drug name.	openFDA
FDA_get_drug_name_by_dosage_info	Retrieve the drug name based on dosage and administration information.	openFDA
FDA_get_drug_names_by_dosage_forms_and_strengths_info	Retrieve drug names based on specific dosage forms and strengths information.	openFDA
FDA_get_dosage_forms_and_strengths_by_drug_name	Retrieve dosage forms and strengths information based on the drug name.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_drug_names_by_abuse_dependence_info	Retrieve the drug name based on information about drug abuse and dependence, including whether the drug is a controlled substance, the types of possible abuse, and adverse reactions relevant to those abuse types.	openFDA
FDA_get_abuse_dependence_info_by_drug_name	Get information about drug abuse and dependence based on the drug name, specifically information on whether the drug is a controlled substance, the types of possible abuse, and adverse reactions relevant to those abuse types.	openFDA
FDA_get_drug_names_by_lab_test_interference	Retrieve drug names that have known interference with laboratory tests.	openFDA
FDA_get_lab_test_interference_info_by_drug_name	Retrieve information about laboratory test interferences for a specific drug.	openFDA
FDA_get_drug_names_by_drug_interactions	Retrieve a list of drug names that have the specified drug interactions.	openFDA
FDA_get_drug_interactions_by_drug_name	Retrieve drug interactions based on the specified drug name.	openFDA
FDA_get_drug_names_by_effective_time	Retrieve drug names based on the effective time of the labeling document.	openFDA
FDA_get_effective_time_by_drug_name	Retrieve effective time of the labeling document based on the drug name.	openFDA
FDA_get_drug_name_by_environmental_warning	Retrieve the drug name based on the specified environmental warnings.	openFDA
FDA_get_environmental_warning_by_drug_name	Fetch environmental warnings for a specific drug based on its name.	openFDA
FDA_get_drug_names_by_food_safety_warnings	Retrieve drug names based on specific food safety warnings.	openFDA
FDA_get_drug_names_by_general_precautions	Retrieve drug names based on specific general precautions information.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_general_precautions_by_drug_name	Retrieve general precautions information based on the drug name.	openFDA
FDA_get_drug_names_by_geriatric_use	Retrieve drug names that have specific information about geriatric use.	openFDA
FDA_get_geriatric_use_info_by_drug_name	Retrieve information about geriatric use based on the drug name.	openFDA
FDA_get_dear_health_care_provider_letter_info_by_drug_name	Fetch information about dear health care provider letters for a specific drug. The letters are sent by drug manufacturers to provide new or updated information about the drug.	openFDA
FDA_get_drug_names_by_dear_health_care_provider_letter_info	Fetch drug names based on information about dear health care provider letters. The letters are sent by drug manufacturers to provide new or updated information about the drug.	openFDA
FDA_get_drug_names_by_health_claim	Retrieve drug names based on specific health claims.	openFDA
FDA_get_health_claims_by_drug_name	Retrieve health claims associated with a specific drug name.	openFDA
FDA_get_drug_name_by_document_id	Retrieve the drug name based on the document ID.	openFDA
FDA_get_document_id_by_drug_name	Retrieve the document ID based on the drug name.	openFDA
FDA_get_drug_name_by_inactive_ingredient	Retrieve the drug name based on the inactive ingredient information.	openFDA
FDA_get_inactive_ingredient_info_by_drug_name	Fetch a list of inactive ingredients in a specific drug product based on the drug name.	openFDA
FDA_get_drug_names_by_indication	Retrieve a list of drug names based on a specific indication or usage.	openFDA
FDA_get_indications_by_drug_name	Retrieve indications and usage information based on a specific drug name.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_drug_names_by_information_for_owners_or_caregivers	Retrieve drug names based on information for owners or caregivers.	openFDA
FDA_get_information_for_owners_or_caregivers_by_drug_name	Retrieve specific information for owners or caregivers based on the drug name.	openFDA
FDA_get_info_for_patients_by_drug_name	Fetch information for patients based on the drug name.	openFDA
FDA_get_drug_names_by_instructions_for_use	Retrieve drug names based on specific instructions for use.	openFDA
FDA_get_instructions_for_use_by_drug_name	Retrieve instructions for use information based on the drug name.	openFDA
FDA_retrieve_drug_name_by_device_use	Retrieve the drug name based on the intended use of the device.	openFDA
FDA_retrieve_device_use_by_drug_name	Retrieve the intended use of the device based on the drug name.	openFDA
FDA_get_drug_names_by_child_safety_info	Retrieve drug names based on whether the product should be kept out of the reach of children and instructions about what to do in the case of accidental contact or ingestion.	openFDA
FDA_get_child_safety_info_by_drug_name	Retrieve child safety information for a specific drug based on its name.	openFDA
FDA_get_drug_name_by_labor_and_delivery_info	Retrieve the drug name based on information about the drug's use during labor or delivery.	openFDA
FDA_get_labor_and_delivery_info_by_drug_name	Retrieve information about the drug's use during labor or delivery based on the drug name.	openFDA
FDA_get_drug_names_by_lab_tests	Retrieve drug names based on laboratory tests information.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_lab_tests_by_drug_name	Retrieve laboratory tests information based on drug names.	openFDA
FDA_get_mechanism_of_action_by_drug_name	Retrieve the mechanism of action information for a specific drug.	openFDA
FDA_get_drug_names_by_mechanism_of_action	Retrieve drug names based on the specified mechanism of action information.	openFDA
FDA_get_drug_name_by_microbiology	Retrieve the drug name based on microbiology field information.	openFDA
FDA_get_microbiology_info_by_drug_name	Retrieve microbiology information based on the drug name.	openFDA
FDA_get_drug_names_by_nonclinical_toxicology_info	Retrieve drug names based on nonclinical toxicology information.	openFDA
FDA_get_nonclinical_toxicology_info_by_drug_name	Retrieve nonclinical toxicology information based on drug names.	openFDA
FDA_get_drug_names_by_nonteratogenic_effects	Retrieve drug names based on the presence of nonteratogenic effects information.	openFDA
FDA_get_nonteratogenic_effects_by_drug_name	Retrieve information about nonteratogenic effects based on the drug name.	openFDA
FDA_get_drug_names_by_info_for_nursing_mothers	Retrieve drug names based on information related to nursing mothers.	openFDA
FDA_get_info_for_nursing_mothers_by_drug_name	Retrieve information about nursing mothers for a specific drug.	openFDA
FDA_get_drug_name_by_other_safety_info	Retrieve the drug name based on the provided safety information. This tool looks through safety information that may not be specified in other fields.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_other_safety_info_by_drug_name	Retrieve safety information that may not be specified in other fields based on the provided drug name.	openFDA
FDA_get_drug_names_by_overdosage_info	Retrieve drug names based on information about signs, symptoms, and laboratory findings of acute overdosage.	openFDA
FDA_get_overdosage_info_by_drug_name	Retrieve information about signs, symptoms, and laboratory findings of acute overdosage based on the drug name.	openFDA
FDA_get_drug_name_by_principal_display_panel	Retrieve the drug name based on the content of the principal display panel of the product package.	openFDA
FDA_get_principal_display_panel_by_drug_name	Retrieve the content of the principal display panel of the product package based on the drug name.	openFDA
FDA_retrieve_drug_names_by_patient_medication_info	Retrieve drug names based on patient medication information, which is about safe use of the drug.	openFDA
FDA_retrieve_patient_medication_info_by_drug_name	Retrieve patient medication information (which is about safe use of the drug) based on drug names.	openFDA
FDA_get_drug_names_by_pediatric_use	Retrieve drug names based on pediatric use information.	openFDA
FDA_get_pediatric_use_info_by_drug_name	Retrieve pediatric use information based on drug names.	openFDA
FDA_get_drug_name_by_pharmacodynamics	Retrieve the drug name based on pharmacodynamics information.	openFDA
FDA_get_pharmacodynamics_by_drug_name	Retrieve pharmacodynamics information based on the drug name.	openFDA
FDA_get_drug_name_by_pharmacogenomics	Retrieve the drug name based on pharmacogenomics field information.	openFDA
FDA_get_pharmacogenomics_info_by_drug_name	Retrieve pharmacogenomics information based on the drug name.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_drug_names_by_pharmacokinetics	Retrieve drug names based on specific pharmacokinetics information, such as absorption, distribution, elimination, metabolism, drug interactions, and specific patient populations.	openFDA
FDA_get_pharmacokinetics_by_drug_name	Retrieve pharmacokinetics information (e.g. absorption, distribution, elimination, metabolism, drug interactions, and specific patient populations) for a specific drug based on its name.	openFDA
FDA_get_drug_name_by_precautions	Retrieve the drug name based on the precautions field information.	openFDA
FDA_get_precautions_by_drug_name	Retrieve precautions information based on the drug name.	openFDA
FDA_get_drug_names_by_pregnancy_effects_info	Retrieve drug names based on information about effects the drug may have on pregnant women or on a fetus.	openFDA
FDA_get_pregnancy_effects_info_by_drug_name	Retrieve information about the effects on pregnancy for a specific drug.	openFDA
FDA_get_drug_name_by_pregnancy_or_breastfeeding_info	Retrieve the drug names based on pregnancy or breastfeeding information.	openFDA
FDA_get_pregnancy_or_breastfeeding_info_by_drug_name	Retrieve the pregnancy or breastfeeding information based on the specified drug name.	openFDA
FDA_get_contact_for_questions_info_by_drug_name	Retrieve information on who to contact with questions about the drug based on the provided drug name.	openFDA
FDA_get_recent_changes_by_drug_name	Retrieve recent major changes in labeling for a specific drug.	openFDA
FDA_get_drug_name_by_reference	Retrieve the drug name based on the reference information provided in the drug labeling.	openFDA
FDA_get_reference_info_by_drug_name	Retrieve reference information based on the drug name provided.	openFDA
FDA_get_drug_names_by_residue_warning	Retrieve drug names based on the presence of residue warnings.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_residue_warning_by_drug_name	Retrieve the residue warning based on drug name.	openFDA
FDA_get_drug_names_by_risk	Retrieve drug names based on specific risk information, especially regarding pregnancy or breastfeeding.	openFDA
FDA_get_risk_info_by_drug_name	Retrieve risk information (especially regarding pregnancy or breastfeeding) based on the drug name.	openFDA
FDA_get_drug_names_by_route	Retrieve the drug names based on the route of administration.	openFDA
FDA_get_route_info_by_drug_name	Retrieve the route of administration information based on the drug name.	openFDA
FDA_get_drug_names_by_safe_handling_warning	Retrieve drug names that have specific safe handling warnings.	openFDA
FDA_get_safe_handling_warnings_by_drug_name	Retrieve safe handling warnings for a specific drug based on its name.	openFDA
FDA_get_drug_name_by_set_id	Retrieve the drug name based on the Set ID of the labeling.	openFDA
FDA_get_drug_names_by_spl_indexing_data_elements	Retrieve drug names based on Structured Product Labeling (SPL) indexing data elements.	openFDA
FDA_get_spl_indexing_data_elements_by_drug_name	Retrieve Structured Product Labeling (SPL) indexing data elements based on drug names.	openFDA
FDA_get_drug_names_by_medication_guide	Retrieve drug names based on the presence of specific information in the medication guide.	openFDA
FDA_get_medication_guide_info_by_drug_name	Retrieve medication guide information based on the drug name.	openFDA
FDA_get_drug_name_from_patient_package_insert	Retrieve the drug name based on the information provided in the patient package insert.	openFDA
FDA_get_patient_package_insert_from_drug_name	Retrieve the patient package insert information based on the drug name.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
FDA_get_drug_names_by_ingredient	Retrieve drug names based on a specific ingredient present in the drug product.	openFDA
FDA_get_ingredients_by_drug_name	Retrieve a list of drug ingredients based on the drug name.	openFDA
FDA_get_spl_unclassified_section_by_drug_name	Retrieve the SPL unclassified section information based on the drug name.	openFDA
FDA_get_drug_name_by_stop_use_info	Retrieve the drug name based on the stop use information provided.	openFDA
FDA_get_stop_use_info_by_drug_name	Retrieve stop use information based on the drug name provided.	openFDA
FDA_get_drug_name_by_storage_and_handling_info	Retrieve the drug name based on storage and handling information.	openFDA
FDA_get_storage_and_handling_info_by_drug_name	Retrieve storage and handling information based on the drug name.	openFDA
FDA_get_drug_names_by_safety_summary	Retrieve drug names based on the summary of safety and effectiveness information.	openFDA
FDA_get_safety_summary_by_drug_name	Retrieve a summary of safety and effectiveness information based on the drug name.	openFDA
FDA_get_drug_names_by_teratogenic_effects	Retrieve drug names based on specific teratogenic effects categories.	openFDA
FDA_get_teratogenic_effects_by_drug_name	Retrieve teratogenic effects information based on the drug name.	openFDA
FDA_get_drug_names_by_population_use	Retrieve drug names based on their use in specific populations, such as pregnant women, nursing mothers, pediatric patients, and geriatric patients.	openFDA
FDA_get_population_use_info_by_drug_name	Retrieve information about the use of a drug in specific populations based on the drug name.	openFDA

(continued on next page)

(continued from previous page)

Tool name	Description	Source
`FDA_get_user_safety_warning_by_drug_names`	Retrieve specific user safety warnings based on drug names.	openFDA
`FDA_get_drug_names_by_user_safety_warning`	Retrieve drug names that have specific user safety warnings.	openFDA
`FDA_get_drug_name_by_warnings`	Retrieve the drug names based on specific warning information.	openFDA
`FDA_get_warnings_by_drug_name`	Retrieve warning information based on the drug name.	openFDA
`FDA_get_warnings_and_cautions_by_drug_name`	Retrieve warnings and cautions information for a specific drug based on its name.	openFDA
`FDA_get_drug_names_by_warnings_and_cautions`	Retrieve drug names based on specific warnings and cautions information.	openFDA
`FDA_get_when_using_info`	Retrieve information about side effects and substances or activities to avoid while using a specific drug.	openFDA
`FDA_get_brand_name_generic_name`	Retrieve the brand name and generic name from generic name or brand name of a drug.	openFDA
`FDA_get_do_not_use_info_by_drug_name`	Retrieve information about all contraindications for use based on the drug name.	openFDA
`FDA_get_purpose_info_by_drug_name`	Retrieve information about the drug product's indications for use based on the drug name.	openFDA
`FDA_get_drug_generic_name`	Get the drug's generic name based on the drug's generic or brand name.	openFDA
`OpenTargets_get_associated_targets_by_disease_efoId`	Find targets associated with a specific disease or phenotype based on efoId.	Open Targets
`OpenTargets_get_diseases_phenotypes_by_target_ensembl`	Find diseases or phenotypes associated with a specific target using ensemblId.	Open Targets
`OpenTargets_target_disease_evidence`	Explore evidence that supports a specific target-disease association. Input is disease efoId and target ensemblID.	Open Targets

(continued on next page)

(continued from previous page)

Tool name	Description	Source
`OpenTargets_get_drug_warnings_by_chemblId`	Retrieve warnings for a specific drug using ChEMBL ID.	Open Targets
`OpenTargets_get_drug_mechanisms_of_action_by_chemblId`	Retrieve the mechanisms of action associated with a specific drug using chemblId.	Open Targets
`OpenTargets_get_associated_drugs_by_disease_efoId`	Retrieve known drugs associated with a specific disease by disease efoId.	Open Targets
`OpenTargets_get_similar_entities_by_disease_efoId`	Retrieve similar entities for a given disease efoId using a model trained with PubMed.	Open Targets
`OpenTargets_get_similar_entities_by_drug_chemblId`	Retrieve similar entities for a given drug chemblId using a model trained with PubMed.	Open Targets
`OpenTargets_get_similar_entities_by_target_ensemblID`	Retrieve similar entities for a given target ensemblID using a model trained with PubMed.	Open Targets
`OpenTargets_get_associated_phenotypes_by_disease_efoId`	Find HPO phenotypes associated with the specified disease efoId.	Open Targets
`OpenTargets_get_drug_withdrawn_blackbox_status_by_chemblId`	Find withdrawn and black-box warning statuses for a specific drug by chemblId.	Open Targets
`OpenTargets_search_category_counts_by_query_string`	Get the count of entries in each entity category (disease, target, drug) based on a query string.	Open Targets
`OpenTargets_get_disease_id_description_by_name`	Retrieve the efoId and additional details of a disease based on its name.	Open Targets
`OpenTargets_get_drug_id_description_by_name`	Fetch the drug chemblId and description based on the drug generic name.	Open Targets

(continued on next page)

(continued from previous page)

Tool name	Description	Source
`OpenTargets_get_drug_chemblId_by_generic_name`	Fetch the drug chemblId and description based on the drug generic name.	Open Targets
`OpenTargets_get_drug_indications_by_chemblId`	Fetch indications (treatable phenotypes/diseases) for a given drug chemblId.	Open Targets
`OpenTargets_get_target_gene_ontology_by_ensemblID`	Retrieve Gene Ontology annotations for a specific target by Ensembl ID.	Open Targets
`OpenTargets_get_target_homologues_by_ensemblID`	Fetch homologues for a specific target by Ensembl ID.	Open Targets
`OpenTargets_get_target_safety_profile_by_ensemblID`	Retrieve known target safety liabilities for a specific target Ensembl ID.	Open Targets
`OpenTargets_get_biological_mouse_models_by_ensemblID`	Retrieve biological mouse models, including allelic compositions and genetic backgrounds, for a specific target.	Open Targets
`OpenTargets_get_target_genomic_location_by_ensemblID`	Retrieve genomic location data for a specific target, including chromosome, start, end, and strand.	Open Targets
`OpenTargets_get_target_subcellular_locations_by_ensemblID`	Retrieve information about subcellular locations for a specific target ensemblID.	Open Targets
`OpenTargets_get_target_synonyms_by_ensemblID`	Retrieve synonyms for specified target, including alternative names and symbols, using given ensemblID.	Open Targets
`OpenTargets_get_target_tractability_by_ensemblID`	Retrieve tractability assessments, including modality and values, for a specific target ensembl ID.	Open Targets
`OpenTargets_get_target_classes_by_ensemblID`	Retrieve the target classes associated with a specific target ensemblID.	Open Targets

(continued on next page)

(continued from previous page)

Tool name	Description	Source
`OpenTargets_get_target_enabling_packages_by_ensemblID`	Retrieve the Target Enabling Packages (TEP) associated with a specific target ensemblID.	Open Targets
`OpenTargets_get_target_interactions_by_ensemblID`	Retrieve interaction data for a specific target ensemblID, including interaction partners and evidence.	Open Targets
`OpenTargets_get_disease_ancestors_parents_by_efoId`	Retrieve the disease ancestors and parents in the ontology using the disease EFO ID.	Open Targets
`OpenTargets_get_disease_descendants_children_by_efoId`	Retrieve the disease descendants and children in the ontology using the disease EFO ID.	Open Targets
`OpenTargets_get_disease_locations_by_efoId`	Retrieve the disease's direct location and indirect location disease terms and IDs using the disease EFO ID.	Open Targets
`OpenTargets_get_disease_synonyms_by_efoId`	Retrieve disease synonyms by its EFO ID.	Open Targets
`OpenTargets_get_disease_description_by_efoId`	Retrieve disease description, name, database cross references, obsolete terms, and whether it's a therapeutic area, all using the specified efoId.	Open Targets
`OpenTargets_get_disease_therapeutic_areas_by_efoId`	Retrieve the therapeutic areas associated with a specific disease efoId.	Open Targets
`OpenTargets_get_drug_adverse_events_by_chemblId`	Retrieve significant adverse events reported for a specific drug chemblId.	Open Targets
`OpenTargets_get_known_drugs_by_drug_chemblId`	Get a list of known drugs and associated information using the specified chemblId.	Open Targets
`OpenTargets_get_parent_child_molecules_by_drug_chembl_ID`	Get parent and child molecules of specified drug chemblId.	Open Targets

(continued on next page)

(continued from previous page)

Tool name	Description	Source
`OpenTargets_get_approved_indications_by_drug_chemblId`	Retrieve detailed information about multiple drugs using a list of ChEMBL IDs.	Open Targets
`OpenTargets_get_drug_description_by_chemblId`	Get drug name, year of first approval, type, cross references, and max clinical trial phase based on specified chemblId.	Open Targets
`OpenTargets_get_drug_synonyms_by_chemblId`	Retrieve the synonyms associated with a specific drug chemblId.	Open Targets
`OpenTargets_get_drug_trade_names_by_chemblId`	Retrieve the trade names associated with a specific drug chemblId.	Open Targets
`OpenTargets_get_drug_approval_status_by_chemblId`	Retrieve the approval status of a specific drug chemblId.	Open Targets
`OpenTargets_get_chemical_probes_by_target_ensemblID`	Retrieve chemical probes associated with a specific target using its ensemblID.	Open Targets
`OpenTargets_drug_pharmacogenomics_data`	Retrieve pharmacogenomics data for a specific drug, including evidence levels and genotype annotations.	Open Targets
`OpenTargets_get_associated_drugs_by_target_ensemblID`	Get known drugs and information (e.g. id, name, MoA) associated with a specific target ensemblID, including clinical trial phase and mechanism of action of the drugs.	Open Targets
`OpenTargets_get_associated_diseases_by_drug_chemblId`	Retrieve the list of diseases associated with a specific drug chemblId based on clinical trial data or post-marketed drugs.	Open Targets
`OpenTargets_get_associated_targets_by_drug_chemblId`	Retrieve the list of targets linked to a specific drug chemblId based on its mechanism of action.	Open Targets
`OpenTargets_multi_entity_search_by_query_string`	Perform a multi-entity search based on a query string, filtering by entity names and pagination settings.	Open Targets

(continued on next page)

(continued from previous page)

Tool name	Description	Source
`OpenTargets_get_gene_ontology_terms_by_goID`	Retrieve Gene Ontology terms based on a list of GO IDs.	Open Targets
`OpenTargets_get_target_constraint_info_by_ensemblID`	Retrieve genetic constraint information for a specific target ensemblID, including expected and observed values, and scores.	Open Targets
`OpenTargets_get_publications_by_disease_efoId`	Retrieve publications related to a disease efoId, including PubMed IDs and publication dates.	Open Targets
`OpenTargets_get_publications_by_target_ensemblID`	Retrieve publications related to a target ensemblID, including PubMed IDs and publication dates.	Open Targets
`OpenTargets_get_publications_by_drug_chemblId`	Retrieve publications related to a drug chemblId, including PubMed IDs and publication dates.	Open Targets
`OpenTargets_get_target_id_description_by_name`	Get the ensemblId and description based on the target name.	Open Targets
`get_joint_associated_diseases_by_HPO_ID_list`	Retrieve diseases associated with a list of phenotypes or symptoms by a list of HPO IDs.	HPO
`get_phenotype_by_HPO_ID`	Retrieve a phenotype or symptom by its HPO ID.	HPO
`get_HPO_ID_by_phenotype`	Retrieve the HPO ID of a phenotype or symptom.	HPO

Supplementary Table 4: ATHENA-R1 Evaluation Consortium: rare disease experts contributing to the blinded treatment reasoning evaluation. The consortium comprises 29 experts representing 28 patient-led and clinical organizations focused on rare diseases, recruited through the Chan Zuckerberg Initiative Rare As One network and collaborating institutions. Each expert provided disease-specific context used by QUESTIONGEN to construct treatment cases for the blinded, arena-based evaluation reported in Figure 3 and Supplementary Note 4. Two experts (Ali Rosenberg, Dianne Mitchell) hold appointments at two organizations.

Name	Organization
Ali Rosenberg	KCNT1 Epilepsy Foundation; Cure KCNH1 Foundation
Andrew Longenecker	PBD Project
Arnaud Monteil	Channeling Hope Foundation
Brooke Babineau	KCNQ2 Cure Alliance
Carlos Guerrero	Hairy Cell Leukemia Foundation
Crystal Kaya	Mission: Cure
Curt Ginder	Harvard University
Desiree Magee	CureARS
Dianne Mitchell	CAMK2 Therapeutics Network; Global DARE Foundation
Emma Rybalka	Cure ADSSL1
Gabrielle Rushing	CSNK2A1 Foundation
Jaime Lopez	Centenario Hospital Miguel Hidalgo
Jeremy Tanner	Channeling Hope Foundation
Katrin Ericson	RUNX1 Research Program
Kellan Weston	CTNNB1 Connect and Cure
Kurt Losert	Fibrolamellar Cancer Foundation
Lex Cowsert	DADA2 Foundation
Mackenzie Gamble	NKH Crusaders
Melissa Chaikof	Usher 1F Collaborative
Nathan Guo	ZTTK SON-Shine Foundation
Necia Sabin	Heterotaxy Connection
Priyanka Kakkar	Cure ADSSL1
Rachel Heilmann	The Rory Belle Foundation
Raquel Miralles	International SCN8A Alliance
Sharon Halperin	Lipodystrophy United
Shayanne Martin	Channeling Hope Foundation
Thomas Wagner	End AxD
Timothy Ogden	Bardet Biedl Syndrome Foundation
Zollie Yavarow	Cure VCP Disease

References

1. Gallifant, J. *et al. Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks* in *Findings of the Association for Computational Linguistics: EMNLP 2024* (Association for Computational Linguistics, Miami, Florida, USA, 2024), 12448–12465. doi:10.18653/v1/2024.findings-emnlp.726.
2. Chen, S. *et al.* When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine* **8,** 605 (2025).
3. Henry, K. *et al.* Drug or Pokémon? An analysis of the ability of large language models to discern fabricated medications. *medRxiv* (2026).
4. Yang, A. *et al.* Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
5. Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
6. Singh, A. *et al.* OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267* (2025).
7. Bedi, S. *et al.* Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine,* 1–9 (2026).
8. Kerby, D. S. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology* **3,** 1–9. doi:10.2466/11.IT.3.1 (2014).
9. Kass-Hout, T. A. *et al.* OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. *Journal of the American Medical Informatics Association* **23,** 596–600 (2016).
10. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research* **53,** D1467–D1475 (2025).