A Human-LLM Note-Taking System with Case-Based Reasoning as Framework for Scientific Discovery

Douglas B. Craig
Department of Emergency Medicine Research
Michigan Medicine, University of Michigan
Ann Arbor, Michigan USA
craigdou@med.umich.edu

Abstract

Scientific discovery is an iterative process that requires transparent reasoning, empirical validation, and structured problem-solving. This work presents a novel human-in-the-loop AI system that leverages case-based reasoning to facilitate structured scientific inquiry. The system is designed to be note-centric, using the Obsidian note-taking application as the primary interface where all components, including user inputs, system cases, and tool specifications, are represented as plain-text notes. This approach ensures that every step of the research process is visible, editable, and revisable by both the user and the AI. The system dynamically retrieves relevant cases from past experience, refines hypotheses, and structures research workflows in a transparent and iterative manner. The methodology is demonstrated through a case study investigating the role of TLR4 in sepsis, illustrating how the system supports problem framing, literature review, hypothesis formulation, and empirical validation. The results highlight the potential of AI-assisted scientific workflows to enhance research efficiency while preserving human oversight and interpretability.

1 Introduction

Large language models (LLMs) have the potential to transform scientific research. They offer broad domain knowledge and the ability to synthesize complex information. However, their application in scientific inquiry is hindered by issues such as hallucination, lack of transparency, and difficulty in tracing the reasoning process behind generated insights (Sanderson, 2023). To ensure that AI-driven research remains reliable, verifiable, and ethical, human-in-the-loop methodologies are essential.

Here we present a system that integrates case-based reasoning (CBR) (Kolodner, 1993; Watson, 1997) with a note-centric workflow to facilitate AI-assisted scientific inquiry. The system is designed around the Obsidian note-taking application (https://obsidian.md/) such that all elements of the workflow are represented as first-class plain-text notes in Obsidian. This structure provides a transparent, revisable, and interactive environment where users can inspect, modify, and refine the reasoning process at every stage.

The core workflow of the system follows a structured inquiry process. When a user poses a scientific question or problem, the system assesses whether it aligns with existing case knowledge and retrieves or adapts cases from prior solutions. Importantly, every step of a solution is documented within the note interface, including both user and LLM input, ensuring full traceability. Each step makes use of tools which can be called on explicitly, or searched for based on context.

We illustrate the potential of this approach through a case study exploring the role of TLR4[1] in sepsis. This example illustrates how the system facilitates problem framing, literature review, hypothesis generation, and data integration. The case study highlights the advantages of this structured, AI-augmented workflow.

2 Methods & Design

The system uses a human-in-the-loop approach that is note-centric. That is, all components of the system are stored as notes in the Obsidian note-taking application. All notes are plain-text documents. This includes not only user notes but all system CBR cases as well as tool specifications. This approach means that all elements of the system are transparently available to both the user and LLM as part of the workflow. This approach also means integration with the note-taking application is minimized making the system interface agnostic.

[1]Toll-like receptor 4 (TLR4) plays a central role in detecting bacterial infections. However, in some cases, it can trigger an excessive immune response, leading to sepsis.

Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities, pages 22–30

This stands in contrast with fully integrated LLM-assisted note taking applications (Suh et al., 2023) (https://notebooklm.google/).

2.1 System Workflow

Figure 1 gives an overview of the system workflow. The user interacts with Obsidian, the note-taking application. While taking notes, the user may prompt the system to answer a question or solve a problem. The system evaluates the request and searches for any applicable CBR case. A new instance of the most similar case is then created and linked to from the current user note. If no case is found, a default case is created to initiate stepwise problem solving.

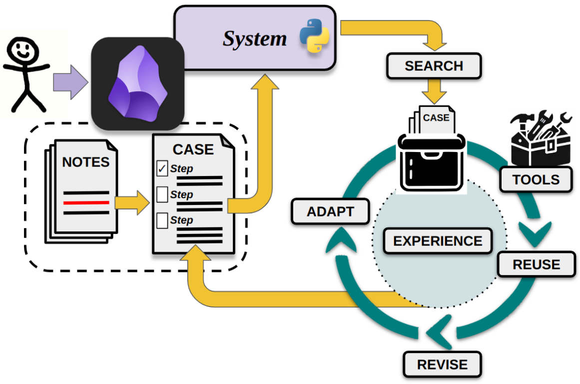


Figure 1: System workflow.

2.2 Case-based Reasoning

Case notes are structured documents that encapsulate knowledge for solving problems. Each case includes: a description of the problem, a series of steps for solving the problem and references to optional resources. Steps include an *Action* and may specify pre-conditions (*Requires*). The action is typically composed of a combination of free text instructions and references to system tools. When a tool is executed its response may be included inline in the note, or stored in a context variable. Variables may be passed to later steps. Abstractly, cases represent system experience based on previous problem solving instances. Cases may be reused, revised or adapted as new problems are encountered.

2.3 The Collaboration Process

After a case is instantiated and linked to the user's note the system begins execution of the steps. Because the case is a plain-text note, the user sees execution as it progresses. The user may pause execution to review, revise, and/or repeat steps. This keeps the user in-the-loop and makes the reasoning process interactive, transparent and traceable.

2.4 Language Enabled Tools

Tool usage and interface is specified in tool notes. As notes, this makes tools searchable both by the user and the system. This means that if a case step specifies some action, the system can search tool specifications for an appropriate tool to perform that action. Tools can also perform language functions (e.g., summarize) as well as retrieve or manipulate data (e.g., from user experiments) through a REST API. This also allows interface to any third-party database.

2.5 Implementation

The system is written in Python and interacts with the Obsidian note-taking application through notes written in plain-text markdown. For LLM-based tools Python interfaces to models (GPT4o and o1) through OpenAI's API (OpenAI, 2023). A Pinecone (https://www.pinecone.io/) serverless vector database maintains embeddings (text-embedding-ada-002 model) for all documents.

3 Case Study

Given that most available benchmarks assume significant autonomy/agency in performing knowledge discovery tasks (Liu et al., 2024; Majumder et al., 2024; Chen et al., 2024) or focus on a single correct/best answer (Rein et al., 2023; Chollet et al., 2025), we instead provide an end-to-end case study to demonstrate how a note-based system facilitates scientific inquiry through *collaboration with a user*. Specifically, the researcher initiates an exploration of how the TLR4 gene is related to sepsis. The approach supports an iterative framework that integrates user input, literature review, external database searches, hypothesis formulation and experimental results. Each step builds upon the previous, ensuring a well-documented and transparent reasoning path that is flexible, adaptable and supports a productive collaboration between human and machine.

3.1 Research Question

We begin by adding the following question to a new Obsidian note:

How is TLR4 related to sepsis?

The system must first contextualize the question/problem within the broader framework of scientific inquiry. This helps to set expectations for the nature of the insights and, more importantly, identify appropriate case-based reasoning (CBR) cases relevant to the question. A summary (Figure 2) is prepared by the *Note Change* case which assesses the original question.[2]

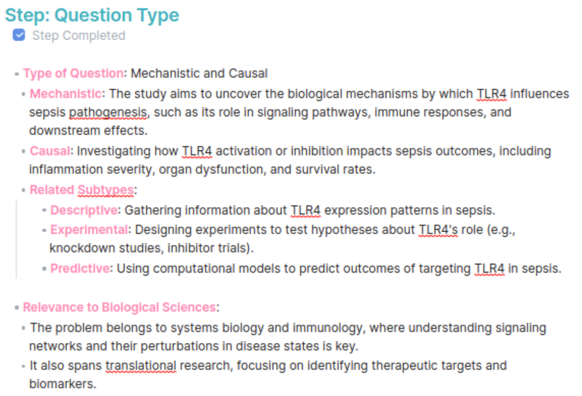


Figure 2: Defining the question type and identifying CBR cases (e.g., Mechanistic and Causal questions)

The user reviews the assessment, makes any needed changes and approves the summary (by checking "Step Completed"). This step will act as a guide for subsequent steps. That is, it informs and constrains subsequent step formulation as a part of any reasoning by the LLM. Note, especially, that this text is added to the original user note.

3.2 Initiating Case-based Reasoning

If available, an applicable "Question Type" case is chosen as a top-level starting point (if not, a "default" case formulates one). Here we have two main case types: mechanistic and causal, as well as three subtypes: descriptive, experimental and predictive. The original user note with question (and summary) will act as a top-level note with links to each subsequent reasoning step (Figure 3). The link is not to the Question Type case, but, rather, a copy which will be "populated" as each step in the case is completed and can be edited at any time by the user.

Note that steps in the following sections are specific to the above CBR case and, in fact, are only for the "Mechanistic" portion of the question as formulated in Figure 2. Though case specific, each step highlights features of the system available to

[2]Note that most figure screenshots include red spell check underlines from Obsidian.

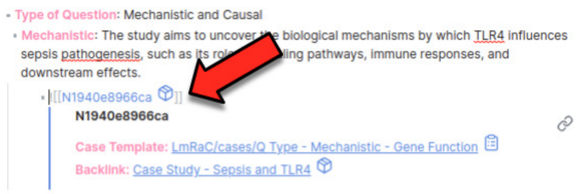


Figure 3: Instantiate intial CBR case, Mechanistic Gene Function, based on previous experience. Link this (red arrow) to the top-level user note.

any case.

3.3 Defining Scope

The first step of the 'Mechanistic - Gene Function' case is to define the scope of the problem. This requires user input. The step definition specifies this dependency in the *Requires* section with the instruction: "User input" (Figure 4). From what is known so far about the problem (which required user approval, see Figure 2) the LLM constructs a list of questions for gathering scoping information.

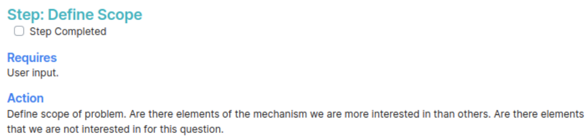


Figure 4: Case reasoning step requiring user input as part of the scope definition action.

Given the task of exploring a gene, this list (Figure 5) asks a series of questions designed to set bounds on what is to be investigated about that gene (e.g., species, interactions, relationship to disease).

Key to the system's collaborative design is that this is *not* a passive solicitation of information from the computer. Rather, the user may edit the list in any way, including using strikethrough to signal that items should be ignored. User answers are interleaved with questions and checkboxes are used to indicate the user has completed the question.

Given the user feedback, the Action portion of the step proceeds and the system proposes a working definition of scope (Figure 6). Again, the user may edit and revise as appropriate since all text is part of an Obsidian note. At this point the step is checked as completed and the next step begins.

3.4 Refining the Problem

We are now in a position to refine the original problem statement based on the agreed upon scope. In this step the Action implicitly uses the LLM to

User Input

Here's a list of questions to gather additional input before defining the scope of the problem:

1. Focus and Depth
   - [x] Are you focusing on human studies, animal models, or both?
     - human only
   - [x] Should the analysis be limited to specific tissues or organs (e.g., lungs, liver, blood)?
     - no
   - [x] Do you want to prioritize any particular stage of sepsis (e.g., early, late, or recovery phase)?
     - early (presentation to emergency department)
2. Data and Literature
   - ignore this as part of the scoping step
   - [ ] ~~Do you already have access to any datasets, or do you want guidance on identifying relevant datasets (e.g., RNA-seq, proteomics, etc.)?~~
   - [ ] ~~Should the scope include a review of the latest clinical or experimental findings on TLR4?~~
3. Mechanisms and Pathways
   - [x] Are there specific pathways or signaling cascades (e.g., NF-κB, MAPK) that you want to prioritize in the analysis?
     - no
   - [x] Should the role of TLR4 be explored in both pro-inflammatory and anti-inflammatory responses?
     - focus on pro-inflammatory in sepsis
4. Integration with Other Factors
   - [x] Do you want to explore TLR4 interactions with specific co-factors or receptors (e.g., MD-2, CD14)?
     - not sure
   - [x] Should the analysis include the microbiota or other external factors influencing TLR4 activation?
     - yes, definitely
5. Sepsis Context
   - [x] Do you want to differentiate between the role of TLR4 in bacterial, fungal, or viral sepsis?
     - focus on bacterial
   - [x] Should the analysis explore the genetic polymorphisms of TLR4 and their association with sepsis outcomes?
     - no

Figure 5: Requested user information for scope definition. User interleaves answers and then checks boxes. Users may also use strike-through to indicate information that should be ignored (i.e., is out of scope).

Scope Definition

The study will focus on understanding the biological mechanisms by which TLR4 influences early sepsis pathogenesis in humans, particularly during the presentation to the emergency department. The analysis will explore the pro-inflammatory role of TLR4 in sepsis, focusing specifically on bacterial sepsis.

The study will not be limited to specific tissues or organs, aiming for a systemic perspective of TLR4 activation. It will include external factors such as the microbiota that influence TLR4 activation and their potential role in sepsis development. While the role of TLR4 interactions with specific co-factors or receptors (e.g., MD-2, CD14) remains uncertain, this may be considered if it emerges as critical during the investigation.

The scope excludes the exploration of genetic polymorphisms of TLR4 and downstream signaling pathways (e.g., NF-κB or MAPK), focusing instead on its broader pro-inflammatory activity and systemic impact during bacterial sepsis.

Figure 6: Final system composed question scope definition.

generate a Refined Problem Statement and propose Key Questions. The Step definition and results are both given in Figure 7.

Step: Refine Problem

- [x] Step Completed

Action

Using the question scope definition: refine the original problem statement to focus on a few key questions.

Refined Problem Statement

The study aims to uncover the biological mechanisms by which TLR4 influences early sepsis pathogenesis in humans, focusing on its pro-inflammatory role during bacterial sepsis. The analysis will explore systemic effects and external factors, such as the microbiota, that influence TLR4 activation at the time of emergency department presentation.

Key Questions:

1. How does TLR4 activation contribute to the systemic pro-inflammatory response observed in early bacterial sepsis?
2. What role do external factors, including the microbiota, play in modulating TLR4 activation during bacterial sepsis?
3. Are there specific mechanisms of TLR4-mediated inflammation that can be targeted to improve outcomes in early sepsis?
4. How does the timing and magnitude of TLR4 activation influence the progression of early bacterial sepsis, and what are the critical thresholds for transitioning from a controlled to a dysregulated immune response?

Figure 7: Refined problem statement and key questions. An additional question has been added by the user (blue highlight).

Notice that the user has exercised the option of adding an additional question (blue highlight). More generally, the system also fully supports not only editing responses, but also the Action definition itself. This serves two purposes, improving responses for a particular problem step, but also providing a mechanism for system learning. Since all case instances represent experience, any changes in how a problem is approached becomes an opportunity to refine and adapt CBR cases for future problems.

3.5 Quick Review

The adoption of LLMs for scientific research has been hindered by, among other things, their propensity to fabricate both information and citations supporting those fabrications (Jones, 2025). Nevertheless, their breadth of training can make them invaluable partners *if verification is included.*

In this step the LLM is used to provide a quick (though potentially unreliable) review of the problem. The Action uses an explicit system tool call to effect a search given the previously generated Problem Statement (Figure 8).

Step: Quick Review

- [ ] Step Completed

Action

Ask GPT for a review of the Refined Problem statement.

```
res0 = tool:GPTsearch({query})
{res0["answer"]}
```

Figure 8: Explicit tool call to ask OpenAI o1 model for "quick" non-authoritative answer.

By making the call explicit the return response can be captured in a named variable, *res0*. In doing so, the response will not be included as part of the note, however, by using a substitution statement after the tool, *{res0["answer"]}*, the response (aka answer) is both included in the note text (Figure 9) and now stored in a variable for later use (see 3.7 below).

1. How does TLR4 activation contribute to the systemic pro-inflammatory response observed in early bacterial sepsis?

TLR4 detects pathogen-associated molecular patterns—especially lipopolysaccharide from Gram-negative bacteria—triggering a cascade that rapidly amplifies the release of pro-inflammatory mediators (e.g., TNF-α, IL-1β, and IL-6). This surge in cytokines and chemokines mobilizes immune cells and activates endothelial pathways, leading to widespread vascular permeability, tissue injury, and the hallmark systemic inflammatory response of sepsis.

2. What role do external factors, including the microbiota, play in modulating TLR4 activation during bacterial sepsis?

The gut microbiota can prime or dampen TLR4-mediated responses by influencing baseline immune surveillance and shaping local inflammation in barrier sites. Changes in microbial composition or permeability (as in dysbiosis) can either heighten TLR4 signaling by exposing the host to more bacterial products or restrain it by promoting beneficial commensal interactions and regulatory immune pathways.

3. Are there specific mechanisms of TLR4-mediated inflammation that can be targeted to improve outcomes in early sepsis?

Targeting downstream signaling molecules, such as MyD88 and TRIF, or blocking TLR4 directly can modulate the excessive inflammation without fully compromising pathogen clearance. Novel therapeutic strategies also focus on interrupting the downstream cytokine storm or stabilizing the endothelium, aiming to reduce the deleterious effects of hyperinflammation while preserving essential immune defenses.

4. How does the timing and magnitude of TLR4 activation influence the progression of early bacterial sepsis, and what are the critical thresholds for transitioning from a controlled to a dysregulated immune response?

Early and balanced TLR4 engagement is crucial for effective pathogen recognition and clearance. However, if TLR4 is activated too robustly or persists beyond an initial window, the immune response can become dysregulated, shifting from protective to harmful. This tipping point often involves excessive cytokine release, endothelial dysfunction, and organ injury, signifying a transition from a contained inflammatory response to full-blown septic pathology.

Figure 9: "Quick" answers to the refined problem's key questions.

The LLM (OpenAI o1 in this case) answers each Key Question in a plausible, though unverified, manner. This step is meant to offer the user a quick overview as orientation to the problem space as well as prompt revision of earlier steps if the user believes this is appropriate (e.g., answers do not support a hypothesis the user has in mind).

3.6 Literature Review

With this cursory look at the answer in mind we may now undertake a more rigorous literature review. In this step the Action explicitly requests a search of PubMed articles to answer the Key Questions (Figure 10). Key to this tool is its design to explicitly *not* answer questions using the LLM. Instead, it collects articles based on a keyword search of PubMed (e.g., genes, diseases, pathways), indexes them and uses retrieval-augmented generation (RAG) to answer the questions (Lewis et al., 2020).

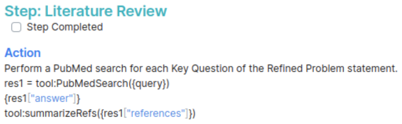


Figure 10: Initiate a literature search of PubMed on the list of Key Questions. Search results are stored in *res1* and references for search results are passed to another tool to be summarized.

Answers include paragraph-level citations to all articles used to answer the question as well as a complete linked bibliography (Figure 11).

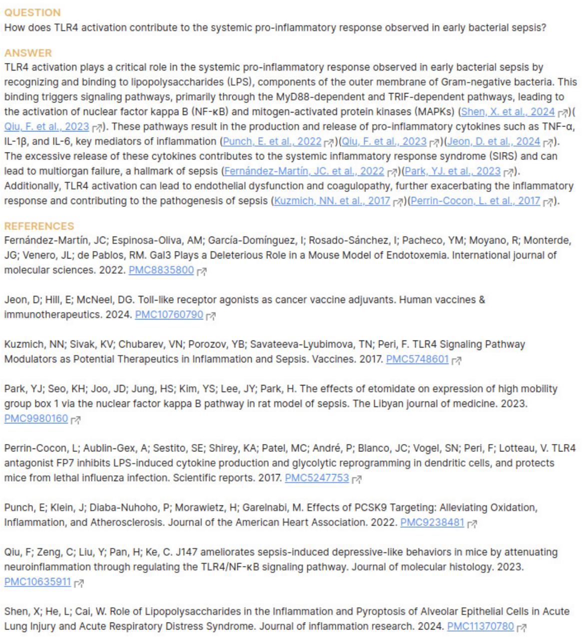


Figure 11: Literature Review answer to first Key Question. Answer is based *only* on PubMed articles with paragraph-level links to citations and bibliography (Fernández-Martín et al., 2022; Jeon et al., 2024; Kuzmich et al., 2017; Park et al., 2023; Perrin-Cocon et al., 2017; Punch et al., 2022; Qiu et al., 2023; Shen et al., 2024).

Though not shown here, the user has the option of revising any question, asking additional questions or otherwise annotating these results.

3.7 Knowledge Gaps

Having completed a first review of the Key Questions, we can now attempt to identify knowledge gaps that may warrant further investigation. This step (Figure 12) uses the same GPT tool as before but now incorporates information from previous steps (3.5 and 3.6) as part of the prompt using variable substitution (red arrows). It also explicitly specifies the structure of the response using a function prototype. Although this may border on "programming" for many users, it is shown here to demonstrate the level of control a user has over *how* the LLM answers questions.

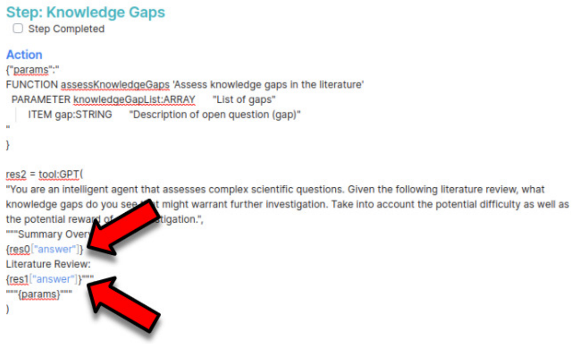


Figure 12: GPT API is used to assess knowledge gaps. An explicit prompt is provided along with results from the previous Quick Review and Literature Review. The response is structured using a function prototype parameter.

The LLM, using both reviews, then provides a list of nine potential gaps in knowledge (Figure 13 shows the first two).

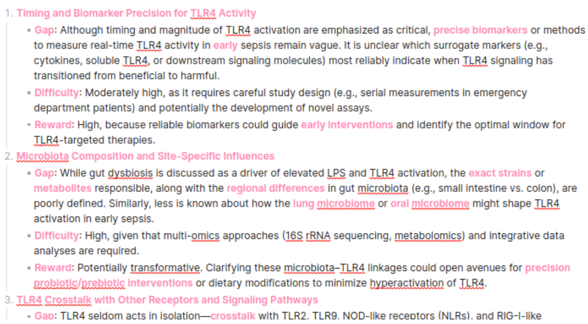


Figure 13: First two identified knowledge gaps (of nine) summarized and annotated by difficulty and reward.

Again, the strength of the LLM to identify and summarize is leveraged to provide a concise annotated summary of each potential gap. It remains up to the user to review, refine and approve the results. However, the system, by design, provides a documented, transparent path of reasoning steps to assist in this task.

3.8 Database Review

In addition to GPT and PubMed reviews, the user can also incorporate knowledge from other sources. This step (Figure 14) demonstrates the use of system tools to search, summarize and structure information from external sources, in this case: Wikipedia and GeneCards.

Step: Database Review

- [ ] Step Completed

Action

```
{"dbList:["Wikipedia","GeneCards"]}
res3 = tool:dbSearch("TLR4",{dbList})
{res3}
```

Wikipedia: TLR4

General Characteristics

- Name: Toll-like receptor 4 (TLR4), also known as CD284.
- Function: Activates innate immunity and plays a key role in bacterial infection defense.
- Structure: Transmembrane protein (~95 kDa) encoded by the TLR4 gene.
- Expression:
  - High in myeloid immune cells (e.g., monocytes, macrophages, dendritic cells).
  - Low in some non-immune cells (e.g., epithelium, endothelium, placental cells, Langerhans islet beta cells).

Activation and Ligands

- Recognizes pathogen-associated molecular patterns (PAMPs) such as lipopolysaccharides (LPS) from Gram-negative bacteria.
- Recognizes damage-associated molecular patterns (DAMPs) like HMGB1, S100 proteins, and histones, which are released during tissue damage or cell death.
- Facilitated by:
  - LPS-binding protein (LBP)
  - CD14
  - MD-2 protein (essential for receptor signaling).

Figure 14: Step definition to use two external data sources (Wikipedia and GeneCards) to summarize more information about the TLR4 gene.

The *dbSearch* tool takes a list of sources, queries them and then returns them to the user note. As before, these are added to the note, but also saved to a variable, *res3*, for later use by other functions.

3.9 Hypothesis Formulation

Given our Problem Statement; literature and database reviews; and, assessment of knowledge gaps, we can now attempt to formulate reasonable research hypotheses (Figure 15).

Step: Hypothesis Formulation

- [ ] Step Completed

Action

For the refined problem, formulate three hypotheses for further investigation using the above literature review, database review and knowledge gaps.

Figure 15: Using the results of multiple previous steps: formulate research hypotheses.

Reiterating: all work to this point has been recorded in a single user readable/editable note in Obsidian. It is available to the user, but also to the system. The Action in this step takes advantage of the accumulation of knowledge to formulate the hypotheses (Figure 16).

Three hypotheses are stated. The rationale for each is summarized and potential approaches for

Hypothesis 2: Gut Microbiota-Driven Modulation of TLR4

> ***Hypothesis***
> ***Dysbiosis*** *of the gut microbiota—characterized by* ***specific strain imbalances*** *and elevated microbial-derived lipopolysaccharide (LPS)—****amplifies TLR4 signaling*** *in the early phase of bacterial sepsis, thereby* ***exacerbating systemic inflammation*** *and worsening patient outcomes.*

Rationale

- Directly tackles the gap on microbiota composition and site-specific influences (Gap #2).
- Integrates evidence that gut-derived LPS is a major activator of TLR4, with dysbiosis increasing intestinal permeability and systemic inflammatory responses.
- Extends current knowledge on how external factors, such as diet or the microbial environment, drive TLR4-mediated sepsis.

Potential Study Approach

1. Microbiome profiling (16S rRNA sequencing) of stool samples from sepsis patients at ED admission and 24 hours later.
2. Metabolomic analysis of short-chain fatty acids and other microbial byproducts known to influence TLR4 signaling.
3. Correlation of microbiota composition and metabolite levels with TLR4-associated biomarkers (e.g., NF-κB activation, IL-6 levels).
4. Interventional arm testing whether restoring gut microbiota balance (via targeted probiotics or fecal microbiota transplantation) reduces TLR4 hyperactivation and improves clinical outcomes.

Figure 16: Proposed Hypothesis 2, including: rationale and potential stepwise approach.

their study are given. This step in particular is a starting point. A user is expected to iterate and refine a hypothesis. This may mean qualifying or constraining a given hypothesis and re-running, or it may involve returning to earlier steps to gather more information (e.g., literature review). The use of a note taking system is meant to encourage and support the dynamic collaboration that is key to a scientific workflow.

3.10 Experiments and Data Collection

Another key feature of the system is the ability to seamlessly incorporate external data into reasoning tasks. In this case study the user has indicated an interest in Hypothesis 2 which integrates gut microbiota with changes in TLR4 signaling during sepsis (Figure 16). The Action has been edited by the user to focus data source search on this hypothesis (Figure 17).

Step: Experiments / Data Collection

- [ ] Step Completed

Action

What experiments or existing data sources would be useful for the proposed hypothesis number 2? Specifically, data that would potentially help provide some preliminary data for a larger project.

Figure 17: *Implicit* search of external resources for datasets suitable for preliminary results.

This search utilizes another section of the CBR case: *Suggested Resources* (Figure 18). This gives the system an *implicit* starting point for finding relevant data. Note that initially the databases are not themselves searched, but, rather, the LLM (OpenAI o1) utilizes its own training to locate possible sources. Like the GPT Literature Review (see 3.5) this is not meant to be a final authoritative search. Rather, it quickly locates possible data as well as giving guidance to the user (not shown) on how to search the database resources (e.g., GEO and SRA).

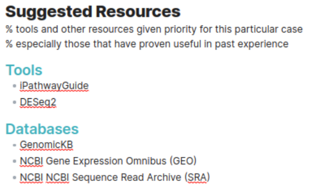


Figure 18: The CBR case includes Suggested Resources. This includes one for gene expression (GEO) and one for microbiome profiling (SRA).

Excerpted search results for both gene expression (from GEO) and microbiota profiling (SRA) are given in Figure 19. These results include accession identifiers (red arrows) as well as descriptions and relevance for Hypothesis 2 use.

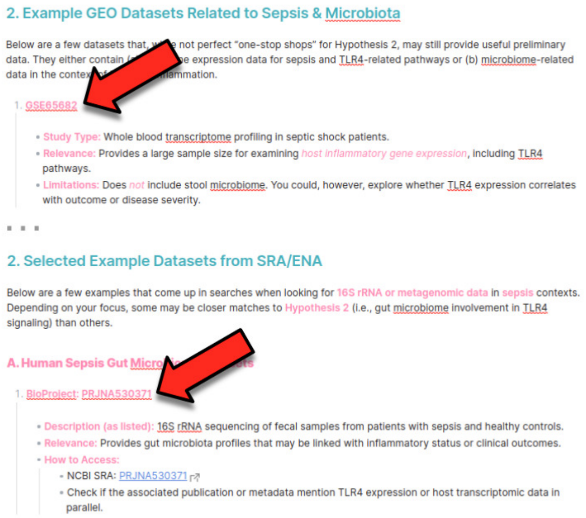


Figure 19: Excerpts from the search using OpenAI o1 model. Public datasets are identified (red arrows). Searches may also be done using tools to directly access resource APIs.

Again, this is meant as a way of using the LLM to quickly assess the availability of relevant public datasets. The user may then utilize other system tools (not shown) to search and download the actual data from GEO and SRA.

3.11 Differential Expression Analysis

Given that most data analysis, especially in the biological sciences, involves a multi-step pipeline, the advantages of initiating and monitoring a pipeline from a notes interface are limited. However, the system *does* have access, via its built-in REST API interface, for accessing the results of any analysis. What this means practically, is that these results can be incorporated into the workflow like any other text source.

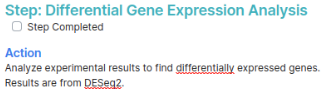


Figure 20: *Explicit* tool identification for computing differential gene expression on retrieved datasets.

Figure 20 implicitly calls a tool to interpret the results of a standard DESeq2 differential gene expression analysis. In this case the tool expects a list of genes in CSV format that includes: gene symbol (e.g., TLR4), log fold-change and significance of the change (typically, adjusted p-value). The results of the step are to summarize those genes that have been found to be significantly differentially expressed.

The user may then qualify these results relative to the workflow by posing additional questions (e.g., "Are other genes associated with an inflammatory response also up-regulated?").

3.12 Experimental Insights and Reflection

LLMs are particularly adept at summarization tasks. This CBR case takes advantage of this feature and asks the LLM in this final step to reflect on what has been discovered thus far and to suggest next steps (Figure 21).

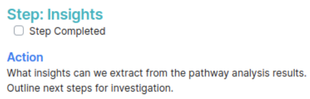
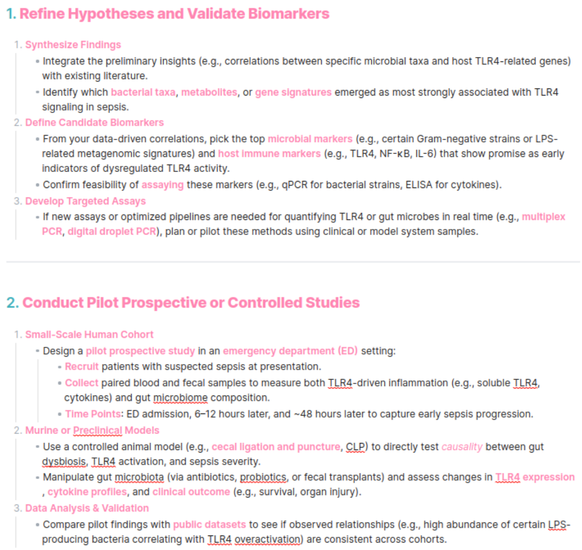


Figure 21: Preliminary results analysis and assessment of next steps sensitive to this analysis (if results are available).

Recall that each step has the accumulated context of all previous steps, so, although the Action may seem vague by asking for "insights," it is actually operating on the accumulated text of everything that has come before (nearly 20 pages in this case study).

In contrast to many autonomous reasoning systems (e.g., OpenAI's o1 model), this system is designed to support a scientific workflow that is *expected* to be open-ended and subject to continuous revision as hypotheses are generated and experiments are performed and interpreted. There is rarely, if ever, one right answer.

Figure 22 supports this workflow not only with a summary and critique of foregoing steps, but by giving guidance for refining and extending the work.



1. Refine Hypotheses and Validate Biomarkers

1. Synthesize Findings
   - Integrate the preliminary insights (e.g., correlations between specific microbial taxa and host TLR4-related genes) with existing literature.
   - Identify which bacterial taxa, metabolites, or gene signatures emerged as most strongly associated with TLR4 signaling in sepsis.
2. Define Candidate Biomarkers
   - From your data-driven correlations, pick the top microbial markers (e.g., certain Gram-negative strains or LPS-related metagenomic signatures) and host immune markers (e.g., TLR4, NF-κB, IL-6) that show promise as early indicators of dysregulated TLR4 activity.
   - Confirm feasibility of assaying these markers (e.g., qPCR for bacterial strains, ELISA for cytokines).
3. Develop Targeted Assays
   - If new assays or optimized pipelines are needed for quantifying TLR4 or gut microbes in real time (e.g., multiplex PCR, digital droplet PCR), plan or pilot these methods using clinical or model system samples.

2. Conduct Pilot Prospective or Controlled Studies

1. Small-Scale Human Cohort
   - Design a pilot prospective study in an emergency department (ED) setting:
     - Recruit patients with suspected sepsis at presentation.
     - Collect paired blood and fecal samples to measure both TLR4-driven inflammation (e.g., soluble TLR4, cytokines) and gut microbiome composition.
     - Time Points: ED admission, 6–12 hours later, and ~48 hours later to capture early sepsis progression.
2. Murine or Preclinical Models
   - Use a controlled animal model (e.g., cecal ligation and puncture, CLP) to directly test *causality* between gut dysbiosis, TLR4 activation, and sepsis severity.
   - Manipulate gut microbiota (via antibiotics, probiotics, or fecal transplants) and assess changes in TLR4 expression , cytokine profiles, and clinical outcome (e.g., survival, organ injury).
3. Data Analysis & Validation
   - Compare pilot findings with public datasets to see if observed relationships (e.g., high abundance of certain LPS-producing bacteria correlating with TLR4 overactivation) are consistent across cohorts.

Figure 22: First two (of six) suggested next steps for the investigation including, hypothesis refinement, validation and prospective controlled studies.

4 Discussion

The case study presented here illustrates how a human-in-the-loop AI system can enhance the process of scientific discovery. By structuring inquiry through case-based reasoning, the system provides a transparent, traceable, and iterative approach that naturally aligns with standard scientific workflows. A key strength of this approach is its ability to leverage LLMs as an integral tool for productive human collaboration.

A critical challenge in leveraging AI for scientific discovery is ensuring that the generated insights remain grounded in empirical evidence. LLMs are known to generate plausible yet unverified statements, which can mislead researchers if used uncritically. This system mitigates such risks by explicitly incorporating verification steps, including literature searches using PubMed and database reviews via other trusted sources. The interactive nature of the system ensures that the user remains an active partner in refining problem definitions, verifying outputs, and shaping hypotheses. This stands in stark contrast to many recent autonomous-blackbox approaches to LLM reasoning.

The foregoing case study demonstrates the value of structuring problem-solving through an evolving CBR system. Cases represent human-machine experience and as such can be reused, refined and adapted for new problems. Their implementation as first-class notes ensures transparency and encourages human collaboration as part of the reasoning process. In this example the iterative approach to scope definition, literature review, and hypothesis refinement steps serve as checkpoints, reinforcing scientific rigor while allowing for flexibility and refinement in inquiry. Using an LLM to help identify knowledge gaps and synthesize insights from multiple sources highlights the strength of this approach and demonstrates how AI can enhance, rather than replace, the natural reasoning process of scientific experts. Providing a mechanism for retrieving user experimental results further enhances the workflow by facilitating a seamless transition from hypothesis generation to empirical validation.

By embedding this approach within a human note-taking system, LLM-based tools become an integral component of the workflow, fostering a continuous cycle of learning and adaptation driven by user-machine collaboration. Furthermore, storing *all* CBR cases, tools, and generated results as user notes enhances transparency and traceability, ensuring that each step in the reasoning process remains accessible for review and refinement.

5 Conclusion

Our approach underscores the potential for human-in-the-loop AI systems to enhance scientific discovery by structuring inquiry, verifying insights, and integrating empirical data. By leveraging case-based reasoning, the approach ensures that LLM-generated outputs remain contextually relevant, empirically grounded, and are subject to a continuous step-by-step review by a collaborating human user.

The results demonstrate that while LLMs provide valuable breadth and summarization capabilities, their true scientific utility emerges when coupled with a human-in-the-loop. The interplay between user expertise and LLM-based tools creates a workflow that is not only transparent and accountable, but also adaptable to the evolving nature of all scientific inquiry. Ultimately, this approach represents a step toward AI-assisted research frameworks that align with the principles of scientific rigor and iterative discovery, paving the way for more effective collaboration between AI systems and domain experts in the pursuit of knowledge.

References

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2024. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *Preprint*, arXiv:2410.05080.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2025. ARC Prize 2024: Technical Report. *Preprint*, arXiv:2412.04604.

Juan Carlos Fernández-Martín, Ana María Espinosa-Oliva, Irene García-Domínguez, Isaac Rosado-Sánchez, Yolanda M. Pacheco, Rosario Moyano, José G. Monterde, José Luis Venero, and Rocío M. de Pablos. 2022. Gal3 Plays a Deleterious Role in a Mouse Model of Endotoxemia. *International Journal of Molecular Sciences*, 23(3):1170.

Donghwan Jeon, Ethan Hill, and Douglas G. McNeel. 2024. Toll-like receptor agonists as cancer vaccine adjuvants. *Human Vaccines & Immunotherapeutics*, 20(1):2297453.

Nicola Jones. 2025. AI hallucinations can't be stopped — but these techniques can limit their damage. *Nature*, 637(8047):778–780.

Janet Kolodner. 1993. *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, CA.

Nikolay N. Kuzmich, Konstantin V. Sivak, Vladimir N. Chubarev, Yury B. Porozov, Tatiana N. Savateeva-Lyubimova, and Francesco Peri. 2017. TLR4 Signaling Pathway Modulators as Potential Therapeutics in Inflammation and Sepsis. *Vaccines*, 5(4):34.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data. *Preprint*, arXiv:2402.17644.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. *Preprint*, arXiv:2407.01725.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Yoo Jung Park, Kwon Hui Seo, Jin Deok Joo, Hong Soo Jung, Yong Shin Kim, Ji Yung Lee, and Hunwoo Park. 2023. The effects of etomidate on expression of high mobility group box 1 via the nuclear factor kappa B pathway in rat model of sepsis. *The Libyan Journal of Medicine*, 18(1):2182683.

Laure Perrin-Cocon, Anne Aublin-Gex, Stefania E. Sestito, Kari Ann Shirey, Mira C. Patel, Patrice André, Jorge C. Blanco, Stefanie N. Vogel, Francesco Peri, and Vincent Lotteau. 2017. TLR4 antagonist FP7 inhibits LPS-induced cytokine production and glycolytic reprogramming in dendritic cells, and protects mice from lethal influenza infection. *Scientific Reports*, 7:40791.

Emily Punch, Justus Klein, Patrick Diaba-Nuhoho, Henning Morawietz, and Mahdi Garelnabi. 2022. Effects of PCSK9 Targeting: Alleviating Oxidation, Inflammation, and Atherosclerosis. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 11(3):e023328.

Fang Qiu, Changchun Zeng, Yuqiang Liu, Haobo Pan, and Changneng Ke. 2023. J147 ameliorates sepsis-induced depressive-like behaviors in mice by attenuating neuroinflammation through regulating the TLR4/NF-KB signaling pathway. *Journal of Molecular Histology*, 54(6):725–738.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *Preprint*, arXiv:2311.12022.

Katharine Sanderson. 2023. GPT-4 is here: What scientists think. *Nature*, 615(7954):773–773.

Xiao Shen, Linglin He, and Wanru Cai. 2024. Role of Lipopolysaccharides in the Inflammation and Pyroptosis of Alveolar Epithelial Cells in Acute Lung Injury and Acute Respiratory Distress Syndrome. *Journal of Inflammation Research*, 17:5855–5869.

Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, San Francisco CA USA. ACM.

Ian Watson. 1997. *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, 1st edition edition. Morgan Kaufmann, San Francisco, Calif.