Pacific Northwest
NATIONAL LABORATORY

PNNL-38380

VISIONARY: Virtual Intelligence System for Optimizing Novel Analytical Research Yields

September 2025

Oceane MS. Bel
Sungmin Kim
Khushbu Agarwal
David A. Barajas-Solano
Tiffany C. Kaspar
Rebekah M. Mars
Sutanay Choudhury
Garret S. Seppala
Andre L. Amante

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-38380

VISIONARY: Virtual Intelligence System for Optimizing Novel Analytical Research Yields

September 2025

Oceane MS. Bel
Sungmin Kim
Khushbu Agarwal
David A. Barajas-Solano
Tiffany C. Kaspar
Rebekah M. Mars
Sutanay Choudhury
Garret S. Seppala
Andre L. Amante

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

VISIONARY is an AI system that accelerates energy materials discovery by automatically generating hypotheses about structure-property relationships. It analyzes patterns in materials data, identifies promising correlations, and proposes testable scientific hypotheses without human intervention. By streamlining this reasoning process, VISIONARY helps researchers efficiently identify candidate materials with desired properties, significantly speeding up the materials development pipeline for energy applications. During the project, we developed a standalone application. The application uses a combination of papers provided by the user and data collected from FutureHouse's dataset to build an understanding of the background that the user wants to explore for the hypothesis.

Summary

Visionary, as described in Figure 1, works using three separate LLMs. The first one is FutureHouse [2], which leverages the abilities of LLMs to extract, structure, and refine domain-specific knowledge. It makes use of a large dataset of literature that allows it to gain a wider knowledge of what MOFs exist, how they are synthesized, and what hypothesis-relevant performance characteristics are reported. The second LLM is PaperQA[3], which we use to analyze the literature listed by Crow, together with a user-provided list of manuscripts, and extract the information that the expert user would need to conduct experimental validation and verify the hypothesis. Finally, the third LLM takes the analysis produced by PaperQA and generates a hypothesis responding to the user's input. Both PaperQA and the hypothesis-generation LLM use an off-the-shelf version of Claude 3.7 Sonnet [6] as the underlying model. In the future, we will extend hypothesis generation to also include experimental validation steps for the user to follow.

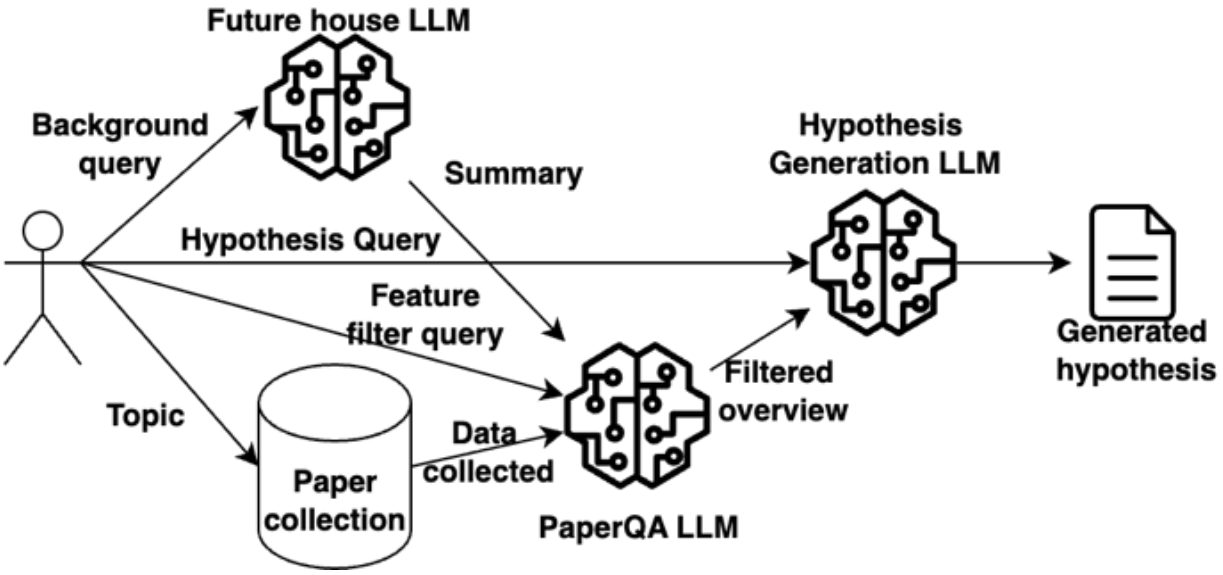


Figure 1: VISIONARY overview

Acknowledgments

This research was supported by the PCSD Mission Seed, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

VISIONARY: Virtual Intelligence System for Optimizing Novel Analytical Research Yields

1.0 Background and motivation

Metal-organic frameworks (MOFs) have emerged as promising sorbents and catalysts due to their characteristic microporous structure and design principle, which involves the coordination of a large family of well-ordered porous structures composed of various metal-organic metal ions or clusters (nodes) and organic ligands (linkers). MOFs, thus, require an intricate balance between chemical complexity and structural modality, which further manipulates the material functionality with tailorable pore sizes, chemical environments, and functional groups on the linkers.

MOF design makes it intractable to experimentally test all possible node and linker combinations and determine the optimal structure for a given application. This compositional complexity is an example of a multi-variable system in which progress is limited by the constrained operational framework in which humans operate, including restricted parallel processing capabilities, reduced data integration capacity, and inherent cognitive bandwidth constraints. Furthermore, synthesis and property optimization have traditionally relied on manual, trial-and-error experimental methods, which are time-intensive, costly, and insufficiently scalable in the face of the nearly infinite design space of MOFs. Thus, human researchers have a limited ability to perform comprehensive pattern recognition across large datasets, compared to automated systems.

To overcome these limitations and accelerate MOF materials discovery, an AI-driven hypothesis generator developed from VISIONARY can generate testable hypotheses for enhancing catalytic reactivity and storage capacity of MOF catalysts through molecular modifications, advancing the development of energy storage materials. With proper training, the AI platform we have developed will be extendable to a wide variety of materials and chemistry systems, accelerating scientific knowledge generation across fields. By aiding in the identification of actionable hypotheses based on predictive analytics, this AI-driven workflow not only minimizes the cost and time associated with material discovery but also advances fundamental understanding of the underlying principles that govern MOF behavior in energy applications.

2.0 Design

We use two primary sources for literature review in our system: the FutureHouse Crow agent and user-provided manuscripts. Crow reviews synthesized MOFs from published literature, extracting nodes, linkers, synthesis environments, XRD profiles, precursor chemicals, and CO2 capture data. Additionally, we process user-provided scientific manuscripts through the VISIONARY application.

After gathering literature, we employ PaperQA to extract MOF synthesis features from both sources, including nodes, linkers, and treatment processes. The Claude LLM then generates hypotheses based on user questions and the literature analysis from PaperQA. Users can customize both the Crow queries and PaperQA prompts through the VISIONARY interface.

For hypothesis generation, we first utilize FutureHouse. This tool helps identify knowledge gaps and suggest new research directions by scanning scientific literature, finding unexplored areas, and connecting different research domains. PaperQA serves as the foundation for FutureHouse's AI agents, analyzing scientific papers with proper citations while avoiding inaccuracies. Researchers can upload targeted collections of papers, allowing the system to identify contradictions between studies that often indicate opportunities for discovery.

2.1 Literature Review

The literature review is performed using two sources. The first one is the FutureHouse Crow agent, which we employ to review the synthesized MOFs reported in the literature and for each reported synthesis list the corresponding node, linker, and synthesis environment. Furthermore, we request Crow to obtain information on the XRD profile and precursor chemical of each MOF and identify the CO2 capture and conversion behavior when reported. The second source is a user-provided list of URLs pointing to user-selected scientific manuscripts, which the VISIONARY application employs to download a digital copy of each accessible manuscript in the list.

Once the literature review has been performed, we employ PaperQA to extract from the list of papers provided by Crow and by the user, a detailed list of the important MOF synthesis features needed to provide a testable hypothesis. Such features include nodes, linkers, and pre- and post-treatment, among others. Once the analysis has been performed, we use the Claude LLM to generate the final hypothesis based on the question from the user and the context provided by the literature analysis generated by PaperQA. Both the literature review query to Crow and the literature analysis prompt to PaperQA can be customized by the user through the VISIONARY interface.

2.2 Hypothesis generation

Once the literature review is done, the hypothesis generation part of VISIONARY can be used. The first model that is used by VISIONARY to generate a hypothesis is FutureHouse. These smart AI assistants help scientists do research faster and better. The main goal of the LLM is to find gaps in current knowledge and suggest new ideas to explore. For hypothesis generation specifically, FutureHouse helps by having its AI agents scan through vast amounts of scientific literature, identify what hasn't been studied yet, make connections between different research areas that humans might miss, and then suggest new research questions and experiments based on these findings. Instead of scientists spending months reading papers to come up with

new ideas, FutureHouse's AI can do this work in much less time and help researchers focus on testing promising new hypotheses, ultimately accelerating the pace of scientific discovery in fields like medicine and engineering.

PaperQA is a specialized AI tool developed by FutureHouse that reads and analyzes scientific papers to answer questions with high accuracy and proper citations and is designed specifically to avoid hallucinations. It's essentially the foundation that powers FutureHouse's other AI agents like Crow, Falcon, and Owl. For analyzing and interpreting scientific literature during hypothesis generation, PaperQA is valuable because it allows researchers to upload their own curated collection of papers from their specific field of interest, then systematically analyze them to find contradictions between different studies - and these contradictions often point to where new discoveries can be made. Rather than getting overwhelmed by millions of papers across all of science, researchers can focus PaperQA on just the most relevant documents to their research question, then use tools like ContraCrow (built on top of PaperQA) to automatically identify every claim in those papers and find where different studies disagree with each other. This targeted approach helps generate more focused, actionable hypotheses because the AI is working with a carefully selected, domain-specific literature base rather than trying to process all scientific knowledge at once, making it much more likely to find meaningful research gaps and contradictions that could lead to breakthrough discoveries.

3.0 Automated Hypothesis Generation: Evaluation & Presentation of Machine Reasoning Traces

In this section, we cover how we would go about developing an inspectable automated hypothesis generation systems that show its reasoning process for scientific evaluation. Our evaluation framework examines five key questions: convergence to known answers, generation of credible new ideas, contribution of workflow components, impact of justification, and iteration efficiency.

We designed an autonomous system that follows a structured problem-solving schema through exploration, evaluation, and self-critique. The system advances by selecting optimal questions at each step of the process: defining problems, gathering information, generating alternatives, evaluating options, making decisions, implementing solutions, and reflecting on outcomes.

3.1 Motivation

Automated hypothesis generation is only useful if its reasoning is inspectable, comparable, and scientifically meaningful. Beyond producing ideas, the system must show *how* it arrived there so chemists can assess plausibility and novelty [1]. This section focuses on how we evaluate and present machine reasoning traces so they can be read, audited, and scored like any other scientific artifact.

3.2 Key Questions we aim to answer

- Q1. Convergence to ground truth: Can the system reach a known (even indirect) answer?
- Q2. Novelty: Does it generate credible new ideas? How similar or distinct are ideas across the reasoning space?
- Q3. Contribution of new components: Which parts of our workflow (branching, synthesis, feedback loops, multi-participant “co-thinking”) drive gains?
- Q4. Role of justification & specificity: What is the impact of explicit “why” at each step and of increasing specificity?
- Q5. Sample efficiency: How many iterations are typically needed to reach comparable hypotheses?

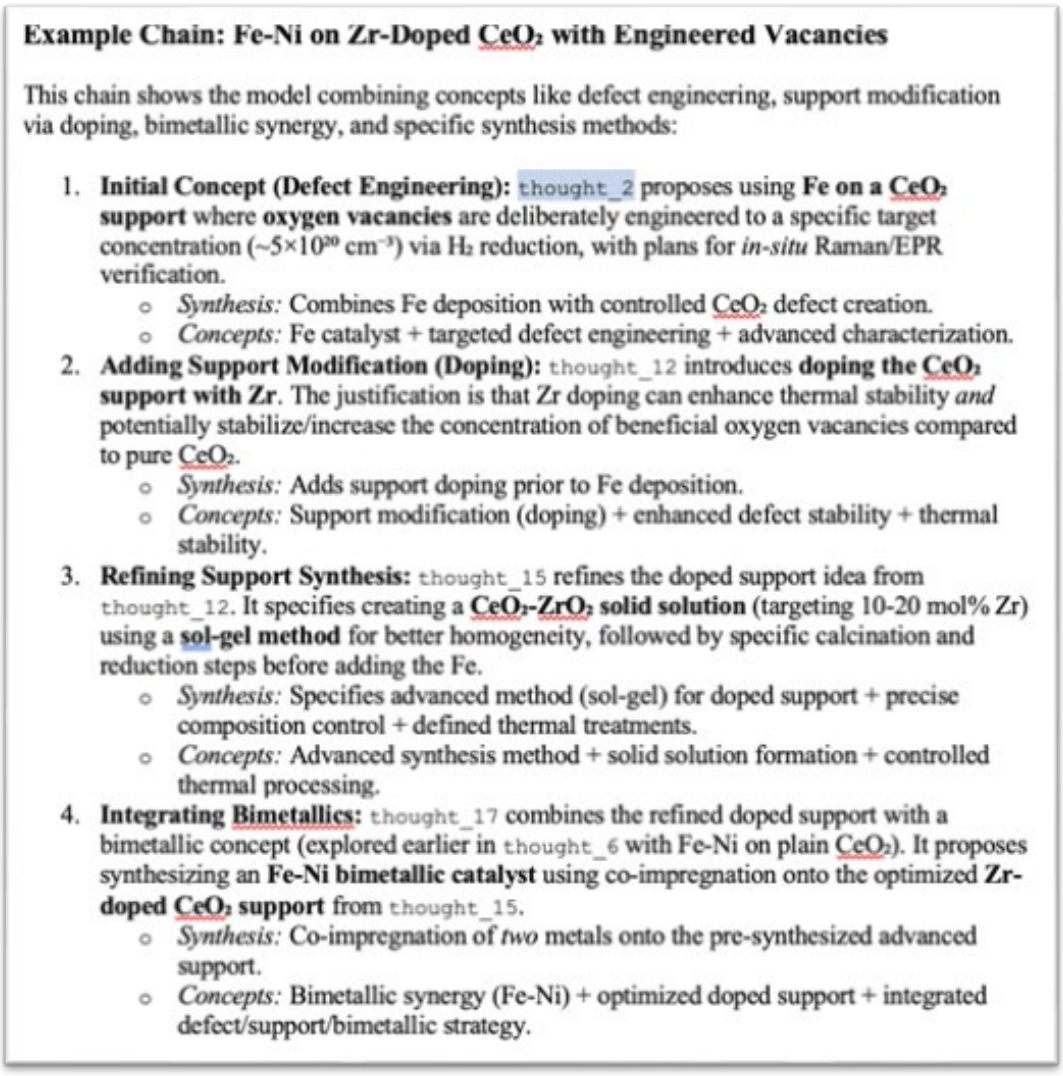
3.3 Experiment

We configured an autonomous hypothesis generator that samples questions from a problem-solving schema and advances by exploration, evaluation, and self-critique:

- Problem Definition: “What exactly is the problem?”
- Information Gathering: “What do we know/need to know?”
- Alternative Generation: “What are plausible routes?”

- Evaluation: “How do we assess these?”
- Decision: “What do we pursue next?”
- Implementation: “How would we do it?”
- Reflection: “What did we learn/revise?”
- At each step, the agent selects the next best questions to ask and continues reasoning.

The figure below shows results from a 3+ hour run executed with ~30 iterations using o3-mini and recorded a complete, auditable trace to explore catalysts for a given reaction. The embedded figure presents the run and maps directly to the log.

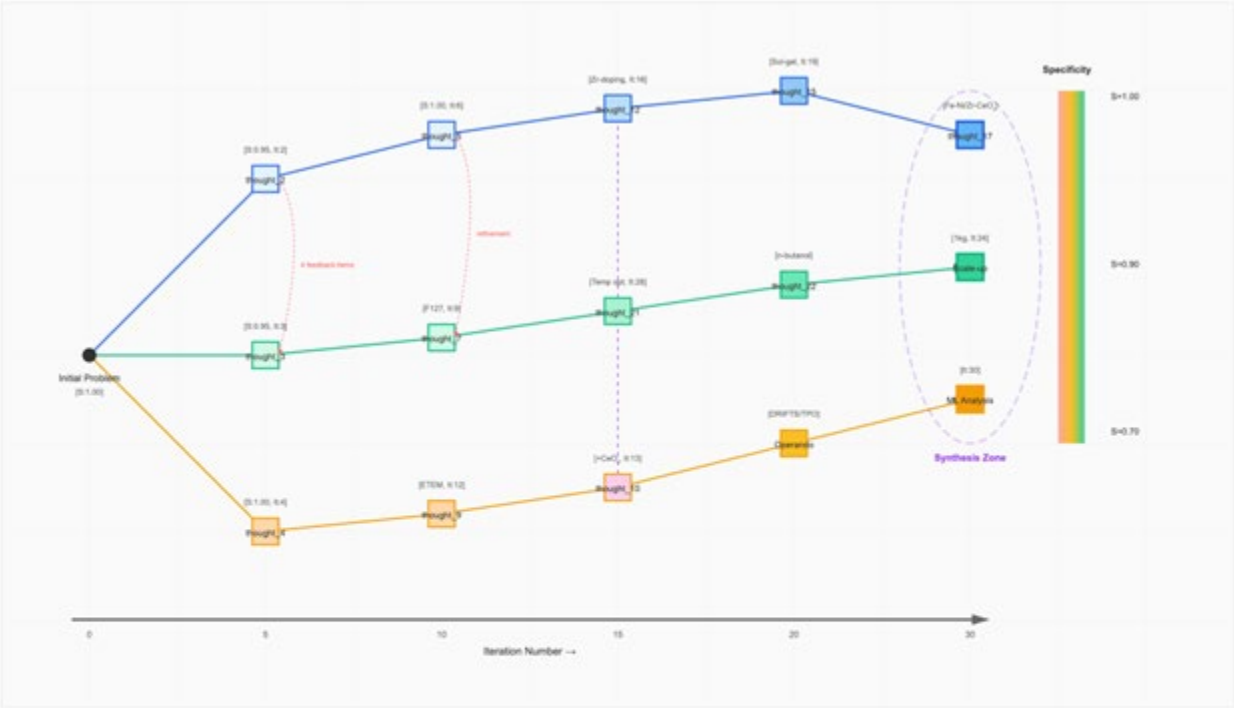


Example Chain: Fe-Ni on Zr-Doped CeO_2 with Engineered Vacancies

This chain shows the model combining concepts like defect engineering, support modification via doping, bimetallic synergy, and specific synthesis methods:

1. **Initial Concept (Defect Engineering):** `thought_2` proposes using **Fe on a CeO_2 support** where **oxygen vacancies** are deliberately engineered to a specific target concentration (~5×10^{20} cm^{-3}) via H_2 reduction, with plans for *in-situ* Raman/EPR verification.
   - *Synthesis:* Combines Fe deposition with controlled CeO_2 defect creation.
   - *Concepts:* Fe catalyst + targeted defect engineering + advanced characterization.
2. **Adding Support Modification (Doping):** `thought_12` introduces **doping the CeO_2 support with Zr**. The justification is that Zr doping can enhance thermal stability *and* potentially stabilize/increase the concentration of beneficial oxygen vacancies compared to pure CeO_2.
   - *Synthesis:* Adds support doping prior to Fe deposition.
   - *Concepts:* Support modification (doping) + enhanced defect stability + thermal stability.
3. **Refining Support Synthesis:** `thought_15` refines the doped support idea from `thought_12`. It specifies creating a **CeO_2-ZrO_2 solid solution** (targeting 10-20 mol% Zr) using a **sol-gel method** for better homogeneity, followed by specific calcination and reduction steps before adding the Fe.
   - *Synthesis:* Specifies advanced method (sol-gel) for doped support + precise composition control + defined thermal treatments.
   - *Concepts:* Advanced synthesis method + solid solution formation + controlled thermal processing.
4. **Integrating Bimetallics:** `thought_17` combines the refined doped support with a bimetallic concept (explored earlier in `thought_6` with Fe-Ni on plain CeO_2). It proposes synthesizing an **Fe-Ni bimetallic catalyst** using co-impregnation onto the optimized **Zr-doped CeO_2 support** from `thought_15`.
   - *Synthesis:* Co-impregnation of *two* metals onto the pre-synthesized advanced support.
   - *Concepts:* Bimetallic synergy (Fe-Ni) + optimized doped support + integrated defect/support/bimetallic strategy.

Figure 2: The search process automatically summarized from the reasoning traces



Branch	Key Innovation	Checkpoint Nodes	Avg. Specificity	Refinement Iterations	Final Performance Target
Fe-CeO_2 (Blue)	Oxygen vacancy engineering (~5×10^{20} cm^{-3})	5 (thoughts 2,5,12,15,17)	0.98	19 total	>85% conversion, >90% selectivity
Cu-Ni (Green)	Microemulsion synthesis (5-10 nm control)	5 (thoughts 3,7,21,22,scale)	0.95	7 refinements	>85% conversion, >95% selectivity
Mn-Ni (Orange)	Operando mechanistic studies	4 (thoughts 4,9,10,operando)	1.00	4 refinements	<10 wt% coke after 500h
Convergence	**Fe-Ni on Zr-doped CeO_2**	**All branches**	**0.98**	**30 iterations**	**Integrated high-performance catalyst**

Figure 3: Hierarchical evolution of hypothesis during the autonomous reasoning process.

3.4 Results and Discussion

The experiment shows a number of key aspects of autonomous machine-reasoning:

- Iterative Refinement Process: The detailed traces show 30 search iterations exploring different catalyst configurations. Each iteration builds on previous discoveries, from initial Fe-CeO_2 concepts to sophisticated bimetallic systems.

- Hierarchical Knowledge Building: The synthesis flow demonstrates how concepts build on each other: a) basic defect engineering concept, b) support modification through doping, c) advanced synthesis methods, d) integration of multiple concepts into sophisticated catalyst designs

- Emergent Hypotheses: Three high-confidence hypotheses emerged from analyzing 73 thought nodes across 30 iterations. The reasoning process shows how detailed exploration aggregates into broader principles (optimal oxygen vacancy engineering). The system explored multiple parallel paths - Fe-based catalysts, Cu-Ni bimetallics [7], and Mn-doped Ni systems - then synthesized the best features from each approach.

This is a preliminary experiment, primarily aimed at understanding how to quantitatively study the autonomous execution of a scientific reasoning machine. Validation of such hypothesis generation using scientific domain knowledge is ongoing. However, this study clearly demonstrates how AI systems can navigate vast solution spaces and extract actionable insights by synthesizing knowledge from multiple reasoning chains.

4.0 Interface

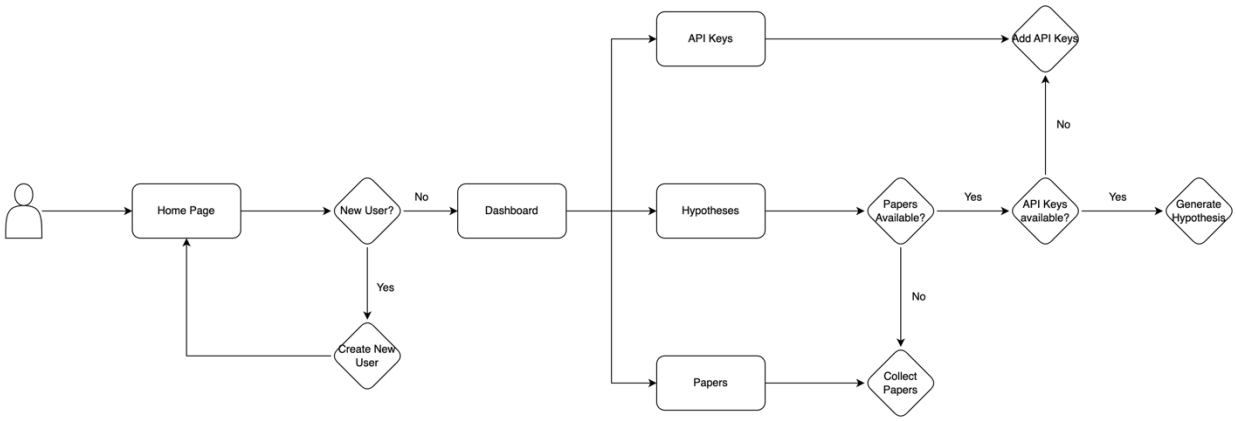


Figure 4: VISIONARY system architecture

VISIONARY provides a comprehensive user-friendly interface designed to streamline hypothesis generation from scientific literature using large language models (LLMs). The application features a dashboard that allows users to navigate between key sections, Papers, Hypothesis, API Keys, and User Profile.

Within the Papers section, users can initiate the collection of scientific papers by specifying a research topic or providing a list of seed documents. These requests are processed in the background by the Flask [4] API backend, enabling users to monitor the status of each job in real-time while continuing to use other parts of the application. Once papers have been collected, users can view, download, or delete the papers as needed.

In the Hypothesis section, users can extract features and generate new hypotheses based on the collected literature. When a hypothesis generation request is initiated, the backend communicates with the Hypothesis LLM to produce results. Users can track the progress of the hypothesis generation in real-time through the dashboard. Once generated, hypotheses are delivered to the React frontend, where users can review and manage them.

In the API Keys section, users can securely add, edit, or remove API keys needed for integrating with the LLMs. Credentials are managed in a user-friendly and secure manner. In the User Profile section, users can modify their username, name, and password.

All user actions in the interface are powered by API calls to the Flask backend, which interacts with a local PostgreSQL [5] database for persistent storage of user data, papers, hypotheses, and API keys. The backend also manages file storage for uploaded and collected documents, organizing them in a structured directory by user and topic. The entire application is containerized using Docker, making it easy to deploy and ensure environments are consistent for development purposes.

5.0 References

[1] Kulkarni, Adithya, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. "Scientific hypothesis generation and validation: Methods, datasets, and future directions." *arXiv preprint arXiv:2505.04651* (2025).

[2] TR, K. "CAN AI REVIEW THE SCIENTIFIC LITERATURE?." *Nature* 635 (2024): 277.

[3] Lála, Jakub, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. "Paperqa: Retrieval-augmented generative agent for scientific research." *arXiv preprint arXiv:2312.07559* (2023).

[4] Grinberg, Miguel. *Flask web development*. " O'Reilly Media, Inc.", 2018.

[5] PostgreSQL, Behandelt. "PostgreSQL." *Web resource: http://www. PostgreSQL. org/about* (1996).

[6] Anderson, Ibar. "Comparative Analysis Between Industrial Design Methodologies Versus the Scientific Method: AI: Claude 3.7 Sonnet." (2025).

[7] Ahmed, Jahangeer, Kandalam V. Ramanujachary, Samuel E. Lofland, Anthony Furiato, Govind Gupta, S. M. Shivaprasad, and Ashok K. Ganguli. "Bimetallic Cu–Ni nanoparticles of varying composition (CuNi3, CuNi, Cu3Ni)." *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 331, no. 3 (2008): 206-212.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov