Implementation Guidelines for LLM and Agentic Frameworks in Drug Discovery

Dr Raminderpal Singh

raminderpal@20visioneers15.com

raminderpal@hitchhikersai.org

deepseek

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

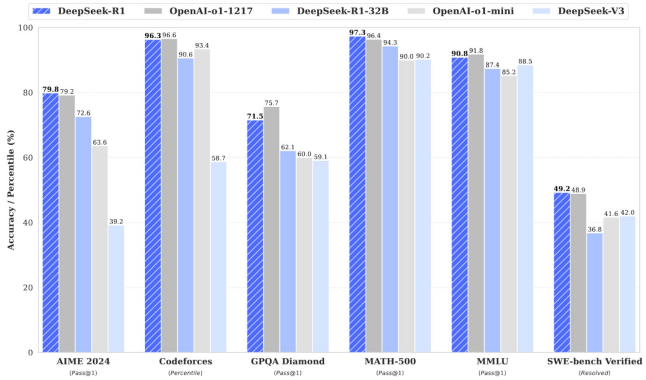


Figure 1 | Benchmark performance of DeepSeek-R1.

BEFORE WE START: THERE'S AN ELEPHANT IN THE ROOM …

Best-in-class reasoning → very effective

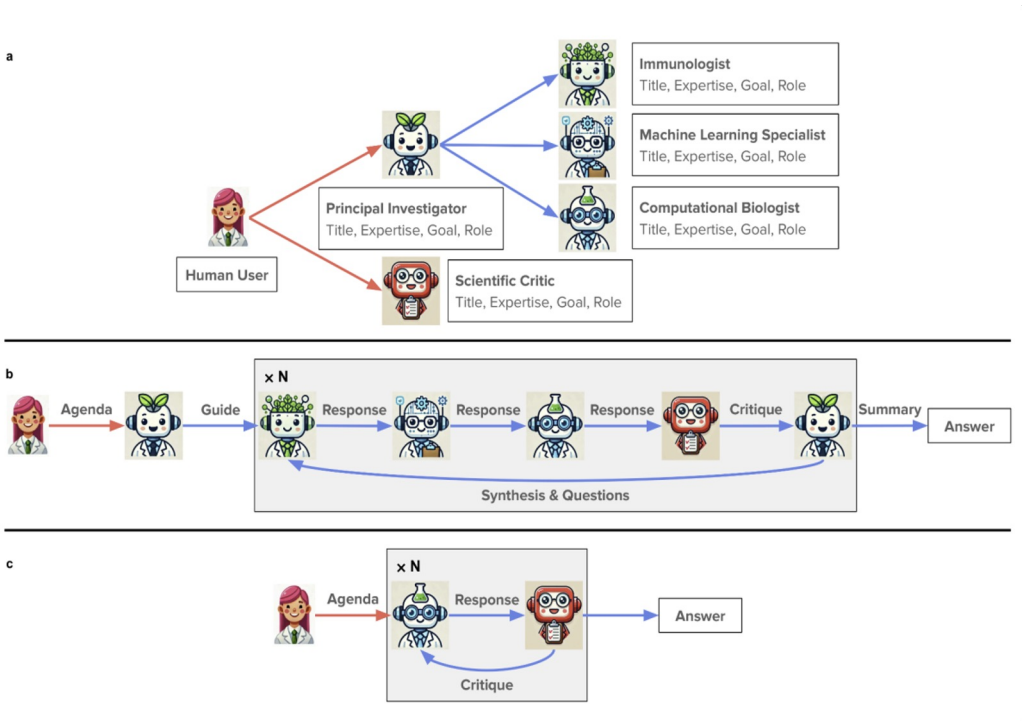
Open-source → commodity pricing

This significantly raises the pressure on commercial providers to up their game. Which in turn gives us (users) options.

- LLMs and Agents
- Specific to Drug Discovery
- Examples
- Guidelines

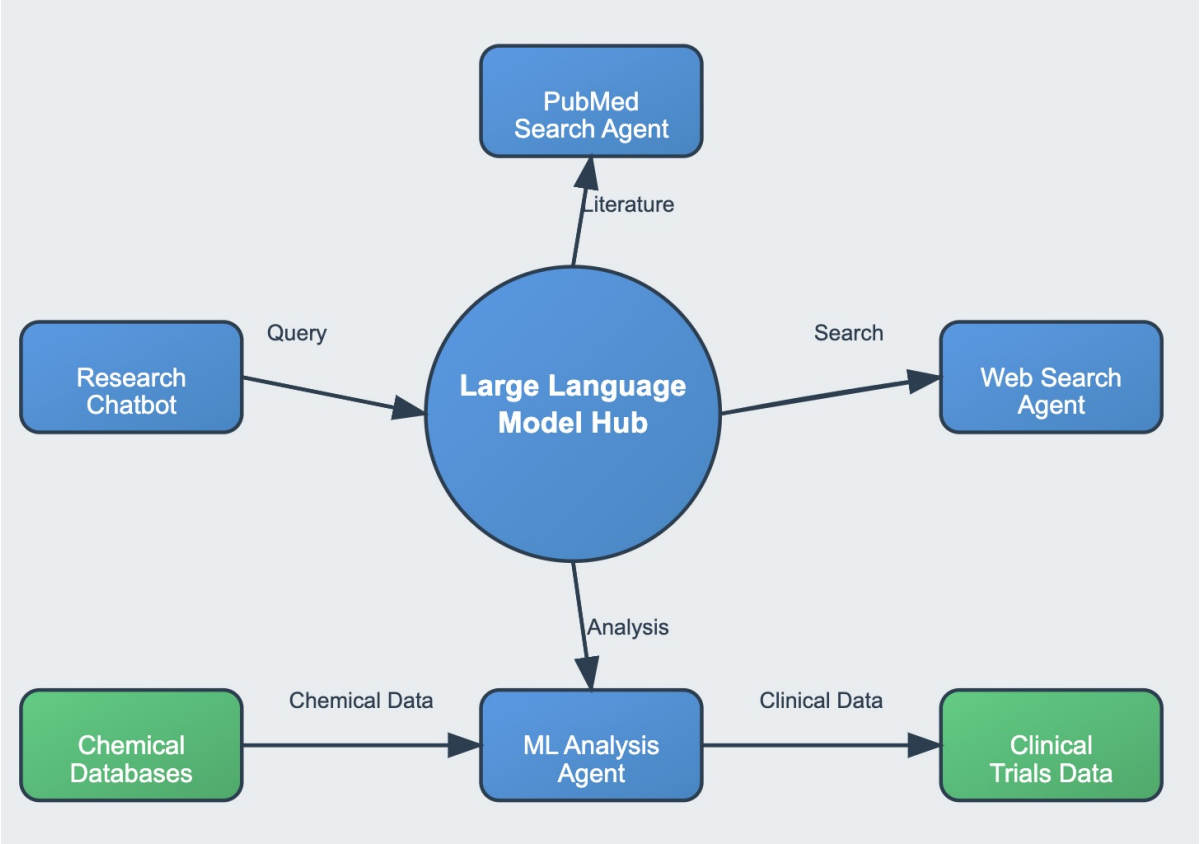
LLMs and Agents

LLM Agents are cool

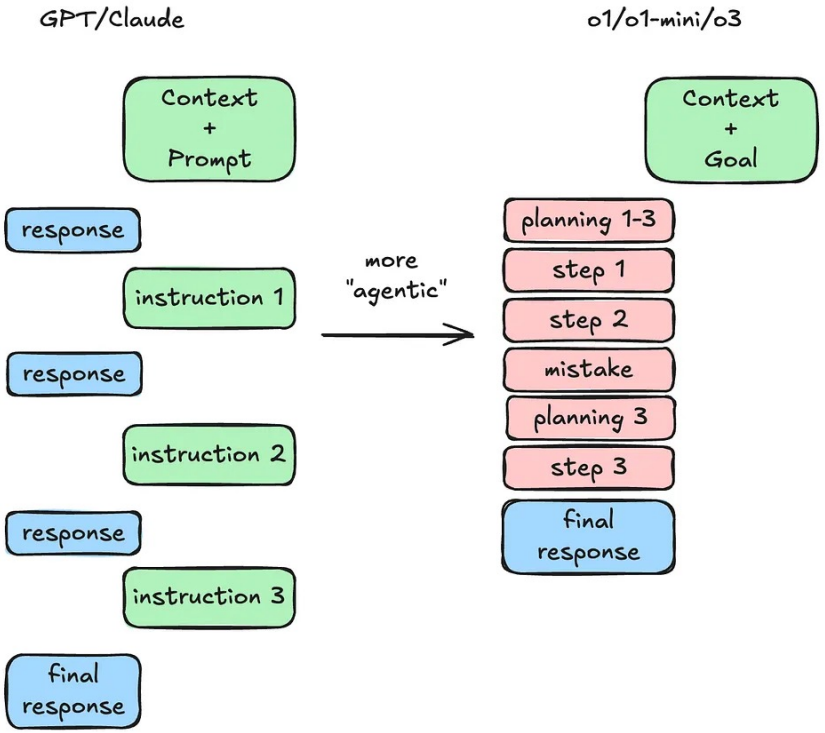


The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation
https://www.biorxiv.org/content/10.1101/2024.11.11.623004v1

LLMs sit next to traditional data processing

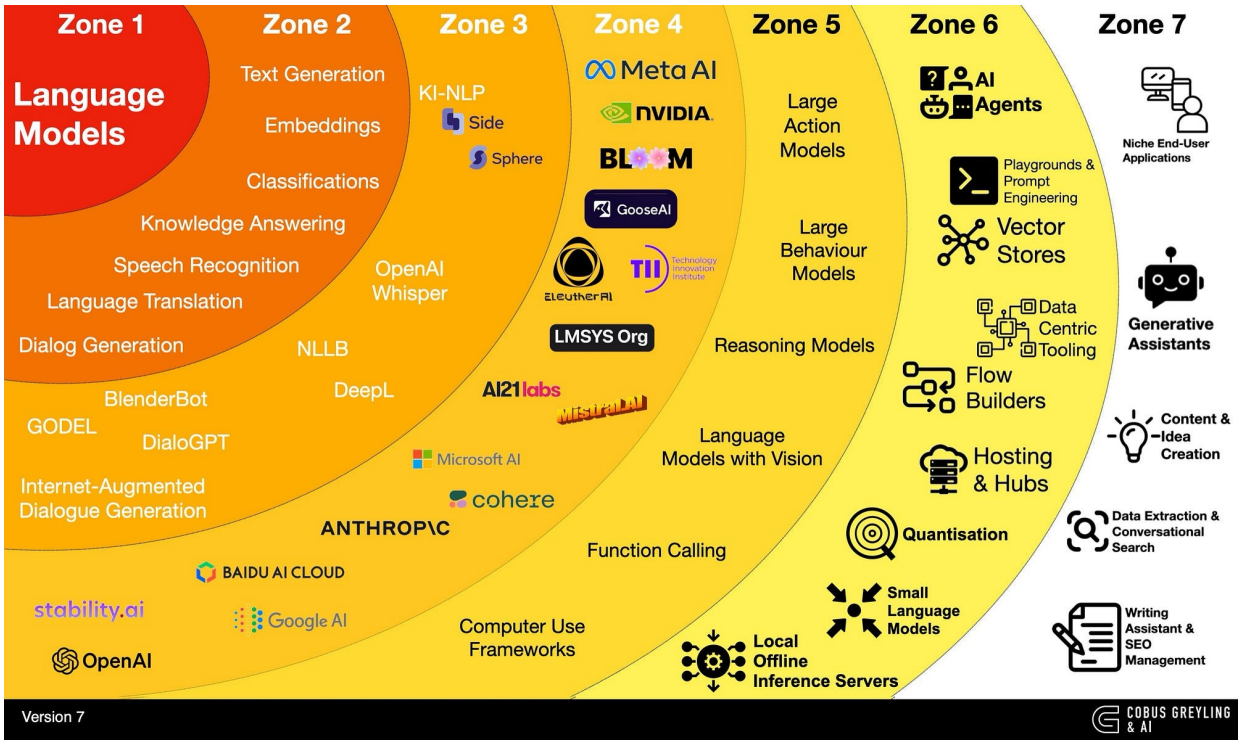


LLM themselves are becoming agentic



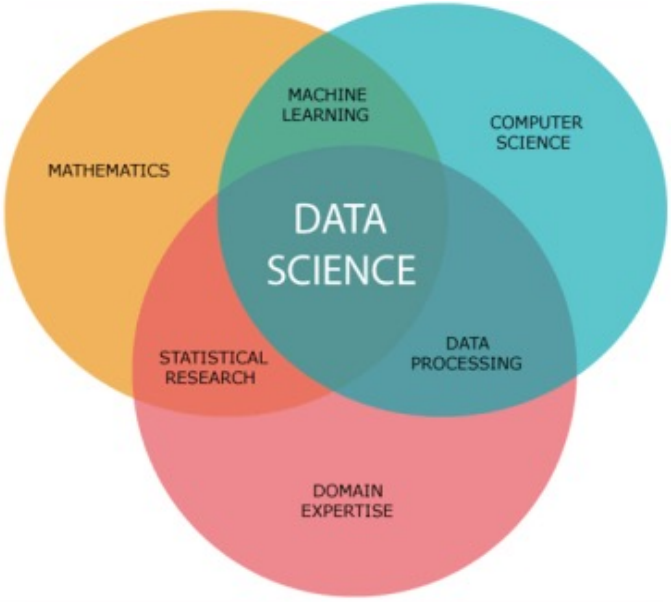
https://www.latent.space/p/o1-skill-issue

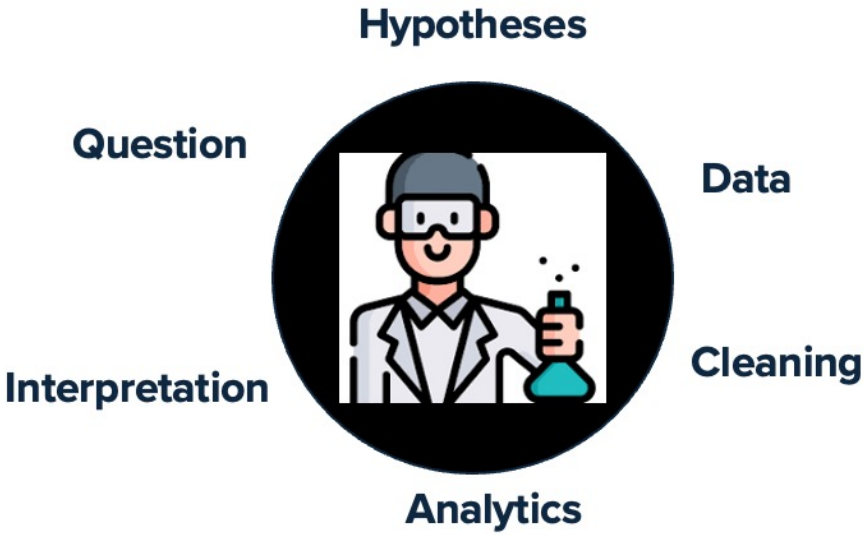
Lots going on here ...



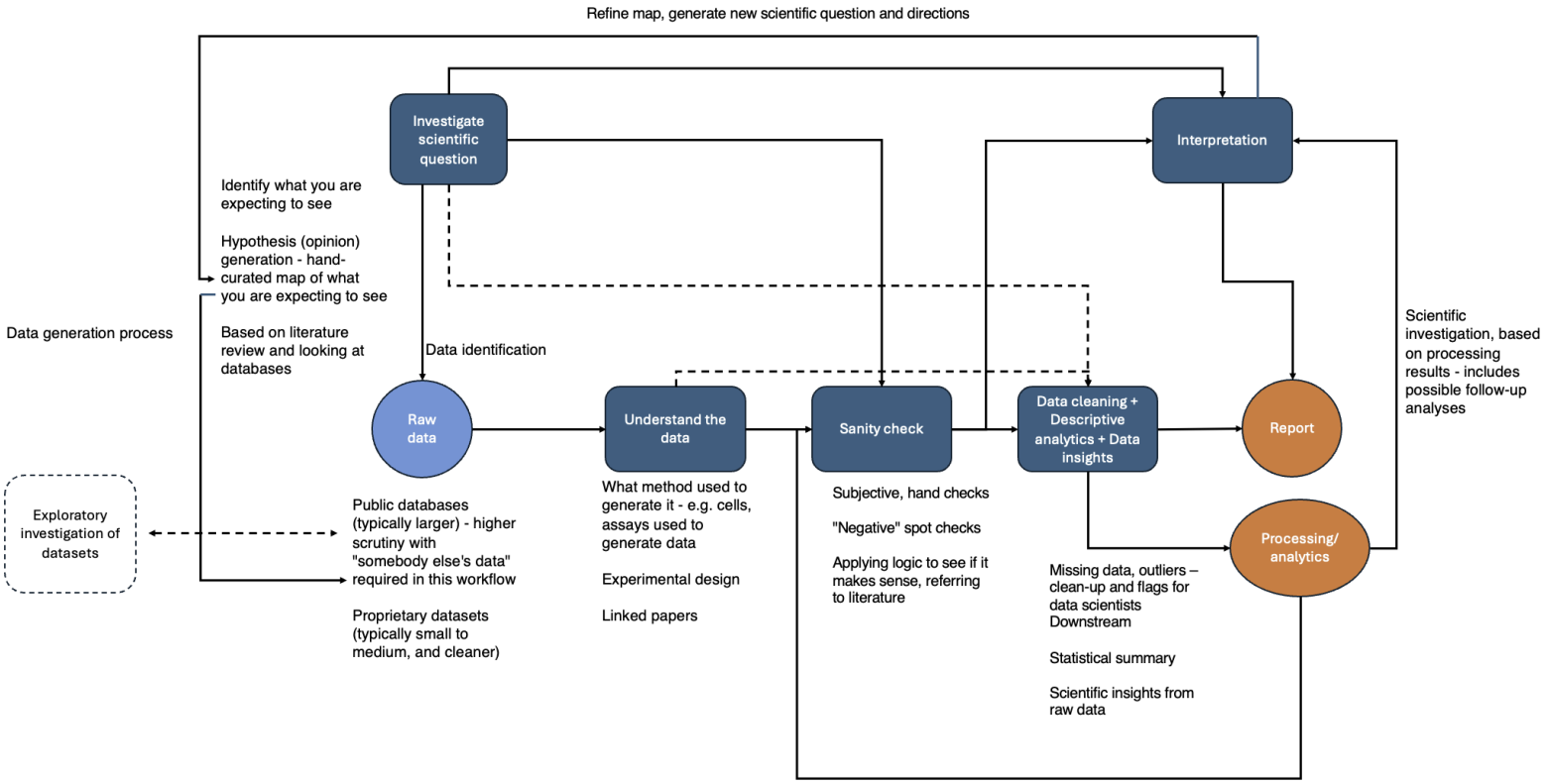
Specific to Drug Discovery

Data science in Drug Discovery is exciting





… and complex



Hypotheses

Raw data

Publications

Insights

Processed data

Reports

Challenges for LLMs

Molecular Representation: LLMs struggle with 3D molecular structures since they're designed for text sequences. Converting molecules to text formats (like SMILES) loses crucial spatial information about atomic arrangements.

Biological Complexity: Models must understand entire biological systems, including protein interactions, cell membrane penetration, and pathway effects. Context-dependent interactions make predictions challenging.

Data Limitations: Pharmaceutical data is sparse (failed experiments often unpublished), noisy (biological variability), heavily proprietary, and imbalanced across different drug types and targets.

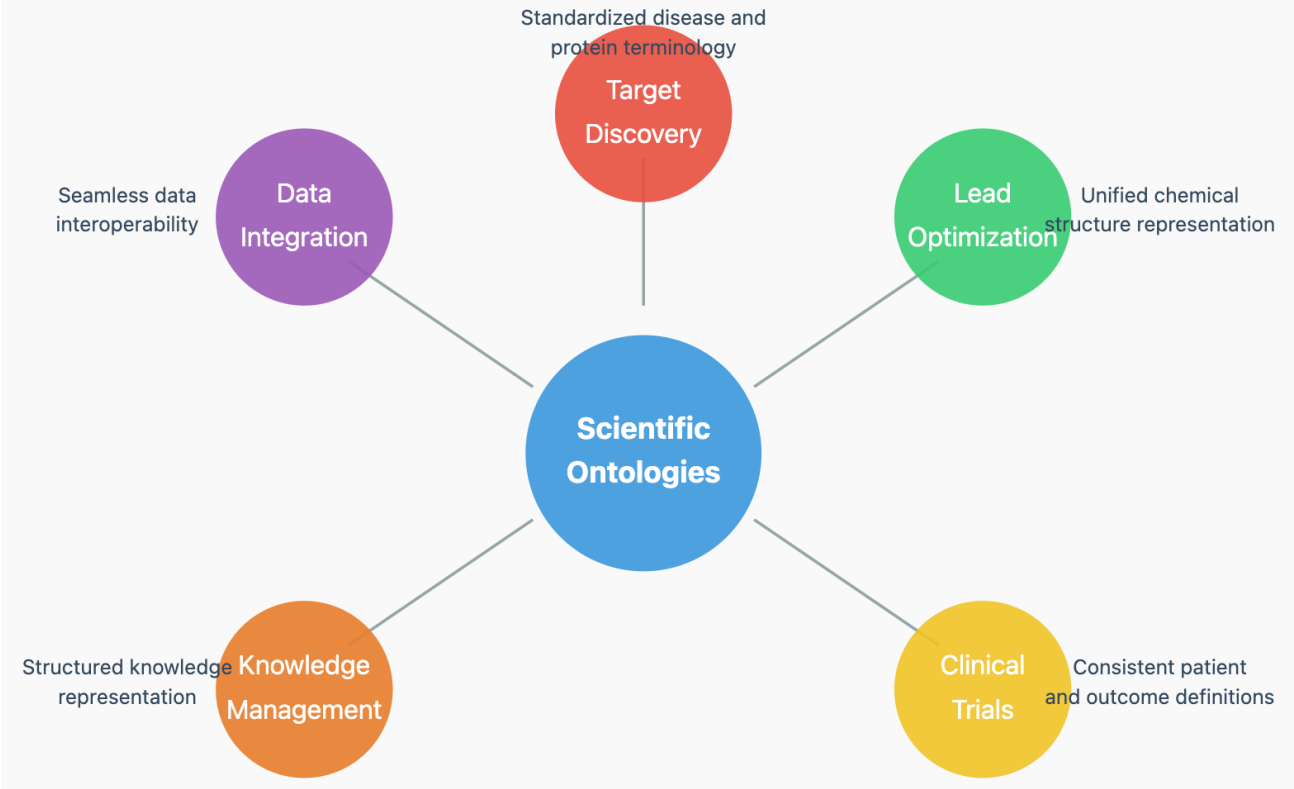
Interpretability Requirements: Drug discovery demands understanding why models make specific predictions for scientific validity, regulatory compliance, and safety assurance.

Vast Chemical Space: Models must generalize from training on a tiny fraction of possible drug-like molecules (estimated 10^{60} total possibilities) to make useful predictions about unexplored compounds.

Uncertainty Quantification: High-stakes decisions require reliable confidence estimates, especially for early-stage decisions about compound synthesis that can cost millions.

Integration Challenges: Combining structural biology data, experimental results, and text-based information into a single model architecture remains difficult.

You need ontologies to connect the pieces



While vector databases excel at straightforward question-answer matching, they have limitations when dealing with complex relationships where information needs to be connected across multiple sources.

LLMs are fundamentally limited

LLMs predict words (tokens) across large sets of unstructured data (primarily text). They predict links between entities, and extract knowledge.

LLM vector databases excel at straightforward question-answer matching.

They have limitations when dealing with complex relationships where information needs to be connected across multiple sources.

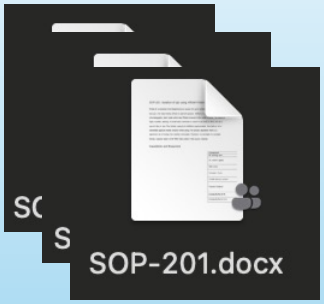
Examples

These are apps I have recently written. Please contact me for more details.

EXAMPLE 1: TAC-KNOW LAB ASSISTANT™

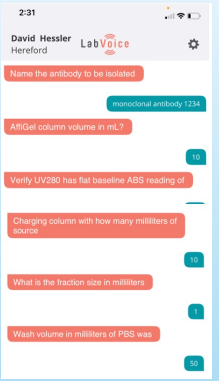
Reference docs

SOP docs



Antibody Separation

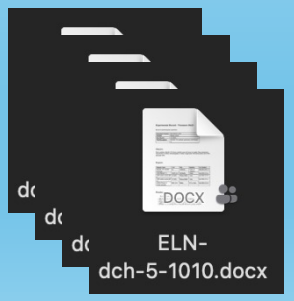
1. Charged column with 10mL of rMoAB-1234 solution
2. Fraction collection every 1.0mL
3. Wash column with 50mL of PBS
4. Wash column 10 mL of acetate ph6 buffer
5. Switch fractions to 0.25mL
6. Switched to acetate ph3.5 buffer
7. Noted second UV peak fractions are 74-79
8. Continued until UV signal baselines
9. Dialysis of pooled fractions 74-79 into PBS for 1 hour at 4C



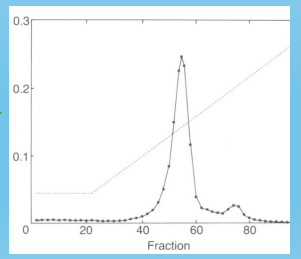
LabVoice

Tacit Knowledge Capture, leading to QC improvements

Q&A GUI

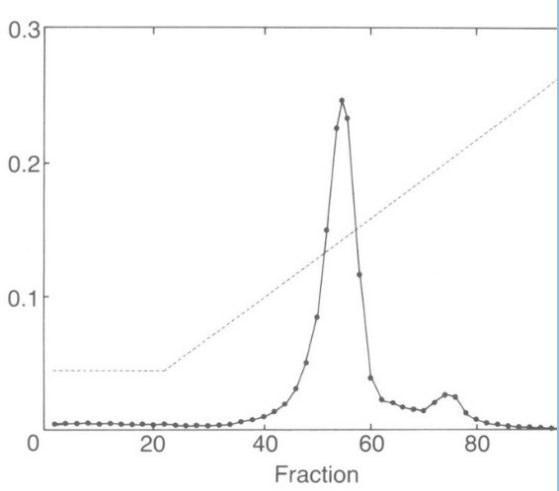


Chromatography separation system



Seamless integration into the scientist's current workflow

ELN



Conclusion

Separation of IgG from general lysate was achieved however there was insufficient separation between the waste peak and the IgG peak so the target is likely to be lower purity than desired. Will send fractions 55 (waste), and pooled 74-79 off for analytical purity analysis by PAGE.

Tac-Know db

=== Tacit Knowledge Entry ===
Date: 2025-01-07 13:37:10
Type: Lab Process Recommendation
Context: IgG Purification
Source: Expert System Analysis

Based on the laboratory conclusion, it appears that the separation of IgG from the general lysate was achieved, but there was insufficient separation between the waste peak and the IgG peak, resulting in a lower purity than desired.

To address this issue, I recommend re-evaluating the washing and elution steps in the Affinity Chromatography (LC Protein-A) process. According to the reference document, "Washing the column with PBS, and then extracting waste at pH 6, must be done with care. The UV280 detector will show when the column is truly clear of eluted proteins. Assuming that a given number of column volumes of wash will work is a mistake: use the effluent detector to indicate when a stage is complete." (Reference: IgG antibody purification strategies)

Specifically, I suggest re-washing the column with 2X column volumes of PBS or until UV280 detection baseline is achieved, as stated in SOP-201: Isolation of IgG using AffiGel Protein-A. Additionally, the pH 6 acetate buffer washing step should be repeated until the UV280 detector indicates a stable baseline, ensuring that the column is truly clear of eluted proteins.

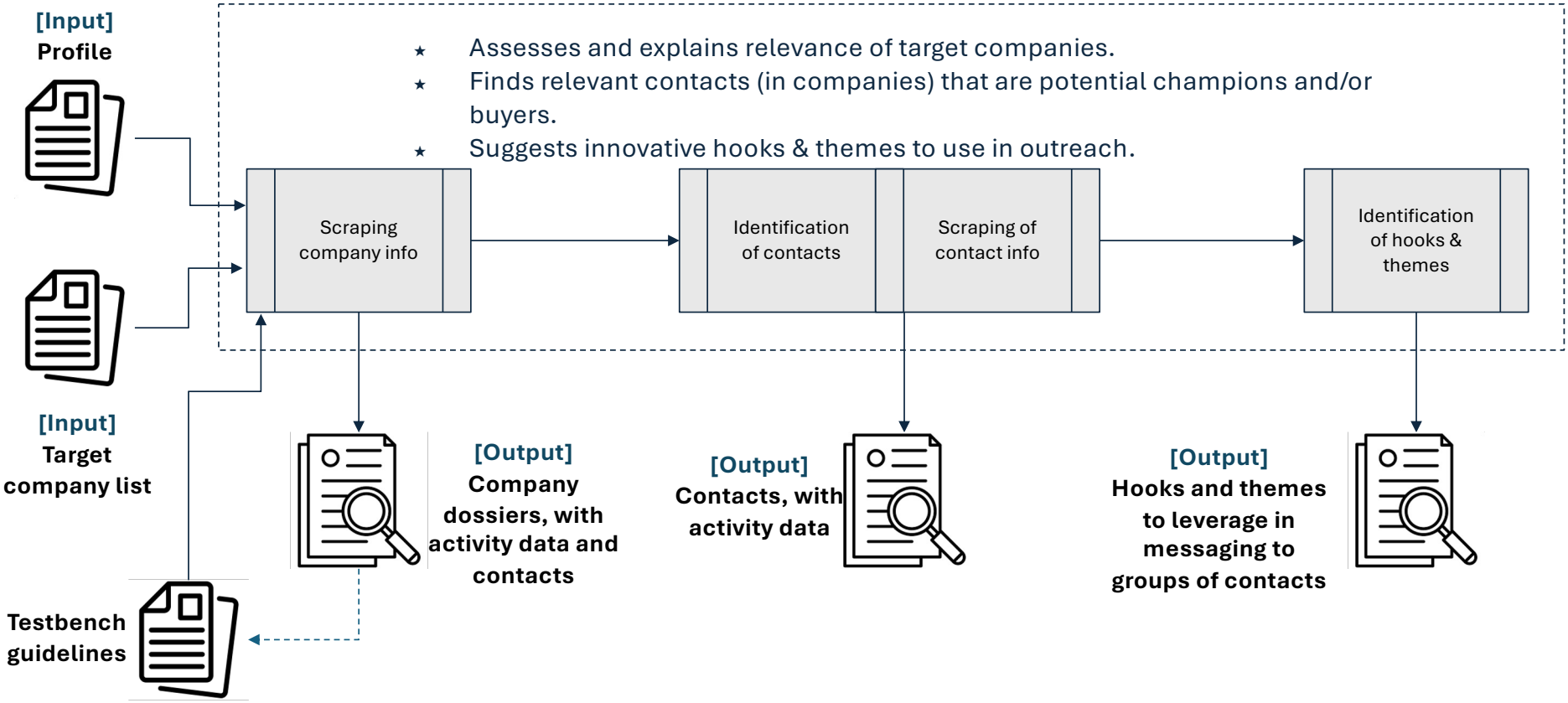
Furthermore, the reference document highlights the importance of running the column slowly enough to allow binding time for the IgG. Therefore, I recommend verifying that the column flow rate is set at no faster than 10% column volume per minute, as specified in the reference document.

If the issue persists, consider implementing additional polishing steps, such as Ion Exchange Chromatography (LC using DEAE media) or Precipitation, as described in the reference document. These methods can provide additional purification and help achieve the desired purity levels.

Lastly, it is essential to ensure that the analytical purity analysis by PAGE is conducted correctly to verify the purity of the IgG fractions. This will help determine the effectiveness of the revised purification strategy and identify any further adjustments that may be necessary.

EXAMPLE 2: LeadgenAgent

Which contacts would buy my offering? Who is the right entry point into an account? What "hooks" will work for a particular contact? What messaging themes will resonate with sub-groups of clients?



EXAMPLE 2: LeadgenAgent (con't)

Some early user feedback …

"The score for "decision-making authority" for identified contacts is very cool.

Saves lots of time with the PubMed results for companies and identified contacts.

There is a simple connection of useful tools:

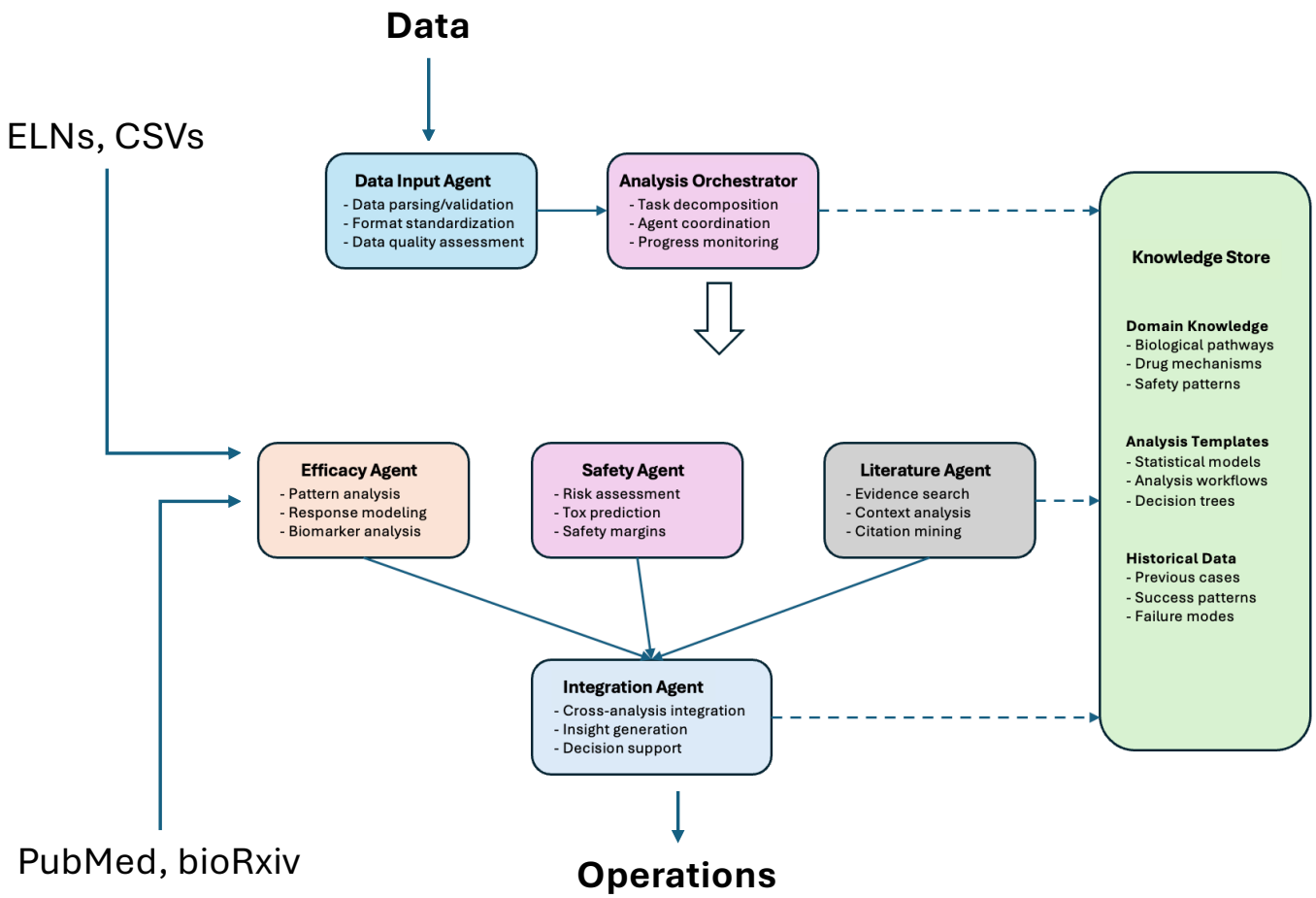
a. Find potential companies on **Google**,
b. Use **LeadgenAgent** to add relevance and find new contacts,
c. Follow-up with **Apollo.io** to learn about & message contacts on LinkedIn"

Joe Webb, CEO & Founder, Vitality Robotics

To use the tool for free: https://www.hitchhikersai.org/leadgenagent

HitchhikersAI.org: A non-profit impact community, accelerating the adoption of AI/ML and data in drug discovery & development.

EXAMPLE 3: Accelerating IND submission



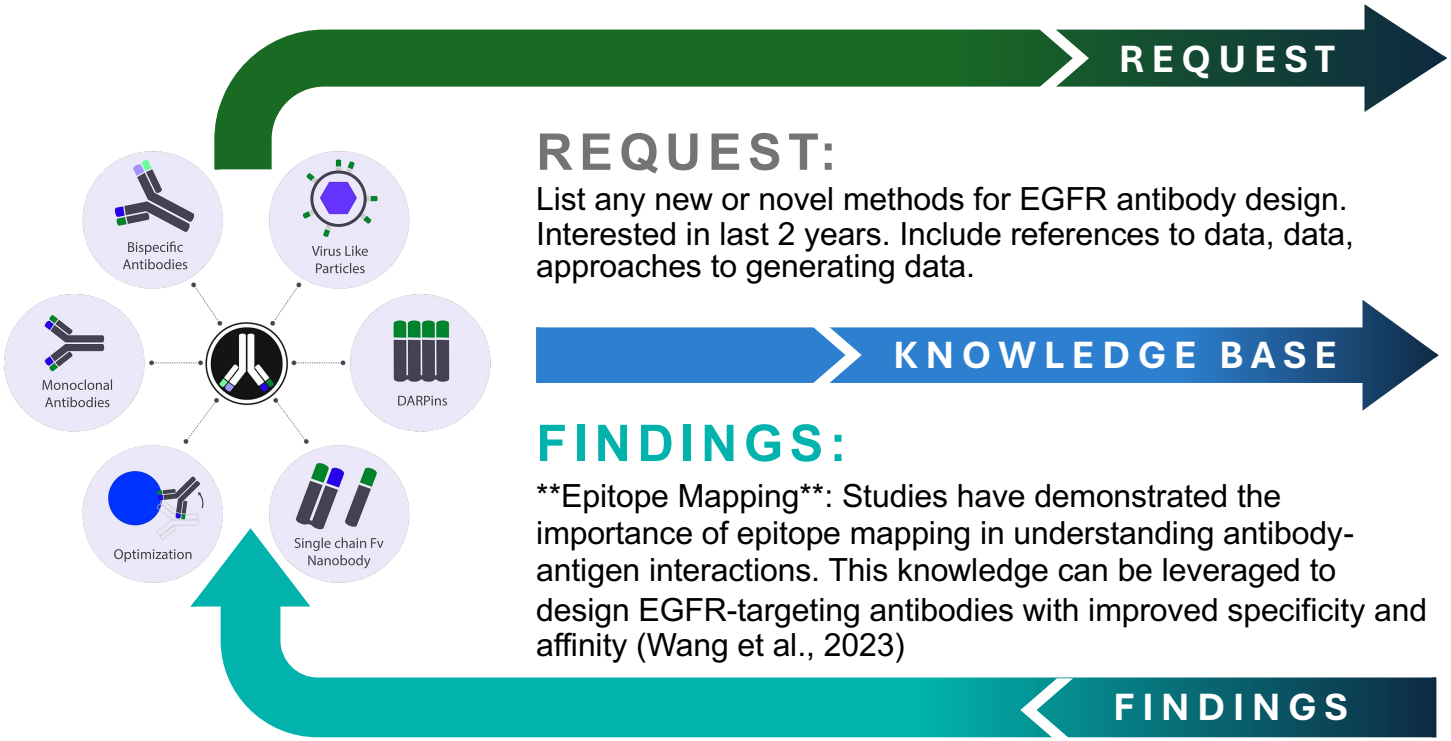
Using LLMs to discover new hypotheses.

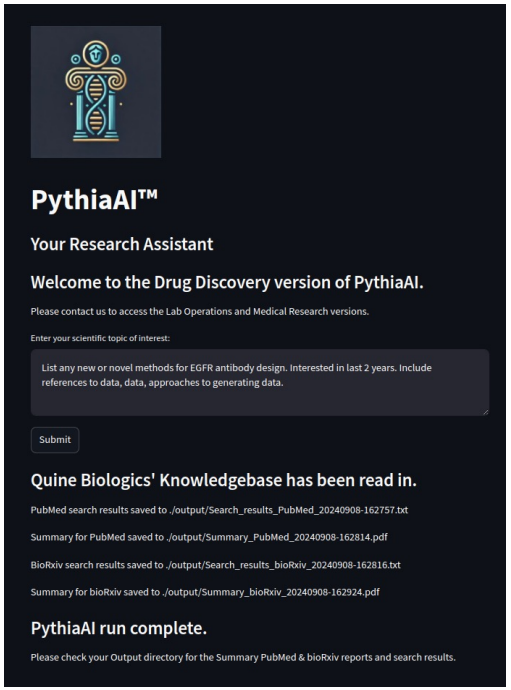
Finding supporting and explanatory insights for study data.

Addressing efficacy and toxicity.

EXAMPLE 4: PythiaAI™ Research Assistant

Quine Biologics, Inc.





For scientists to quickly discover actionable scientific insights in antibody design

Guidelines

Pragmatic guidelines to building LLM systems

Build your system using first-principles. Understand your work and data flows.

"Wait a year, and someone will offer it for free!" As with other industries, these words of wisdom stand (relatively) true. It is important to be agile and respond to changes in this balance.

Your LLM system may need to support both research-intensive activities and workflow-driven tasks, such as extracting knowledge from papers vs designing a lab experiment.

Design your system thinking about a junior employee performing the proposed LLM functions, and then swap the LLM technology in for the human once the design is ready.

Pragmatic guidelines to building LLM systems (con't)

Model Options

PRINCIPLE: Use affordable models if you can. Think about cost and reproducibility.

- Full function APIs: GPT and Claude 3.5 Sonnet are the go-to places, but can be expensive.
- Lesser function APIs: OpenAI and Anthropic offer older models at cheaper costs, but my preference is to go to open-source models through low-cost services such as GROQ CLOUD. For example, Llama3.3:70b is a great model.
- Local models: Ollama is great. It offers numerous native models as well as running GGUF files available on HuggingFace. There is a lot you can do with smaller models, e.g. Phi4 (14b) and Llama3.1 (8b).

Agentic Frameworks

PRINCIPLE: Be wary of black-box frameworks with 100ks of code in them. Things tend to break when you scale.

- Cloud providers offer them and there are several open-source. For example, within Bedrock in AWS.
- My favorite is Pydantic-AI as it is open-source, light-weight, and a natural extension of Pydantic (which many of us already use in our coding). Also, it support numerous LLM providers.

THANK YOU!

These slides are downloadable from:

https://www.hitchhikersai.org/reading