Application of LLM Agents for Long-Form Question Answering in Systems Biology

Charvi Jain, Sahar Vahdati, Pauline Porschitz, Mounir El Hankouri, Nandu Gopan, Ivo F. Sbalzarini and Jens Lehmann

Paper number: 408



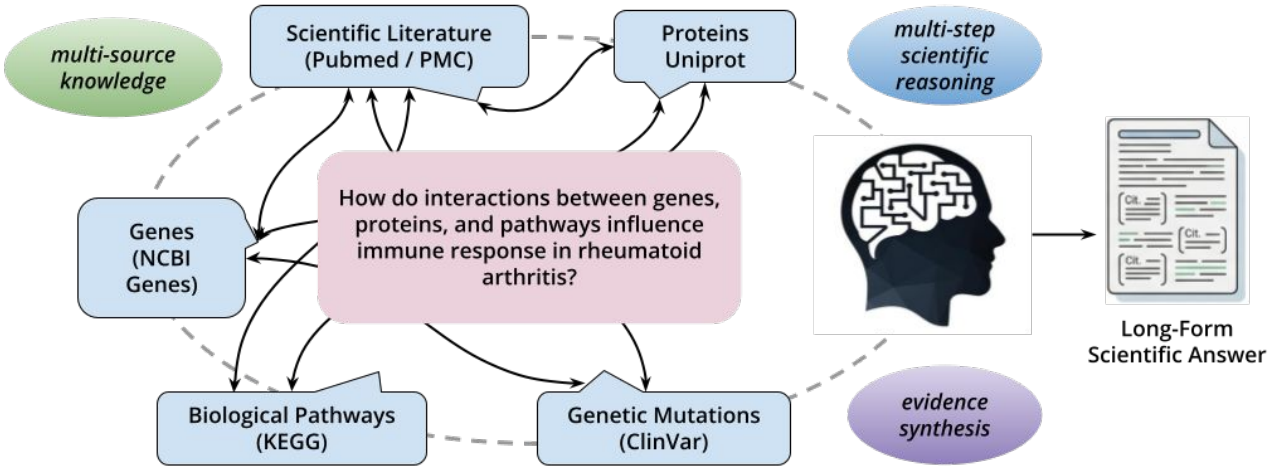




ECML-PKDD 2026 - Applied Data Science Track

Motivation

Systems Biology Question : *How do the interactions between ILC2s, IL-5, and rEos contribute to a broader understanding of immune regulation in RA?*



Systems biology questions require reconstructing mechanisms from evidence scattered across **heterogeneous biological resources**.

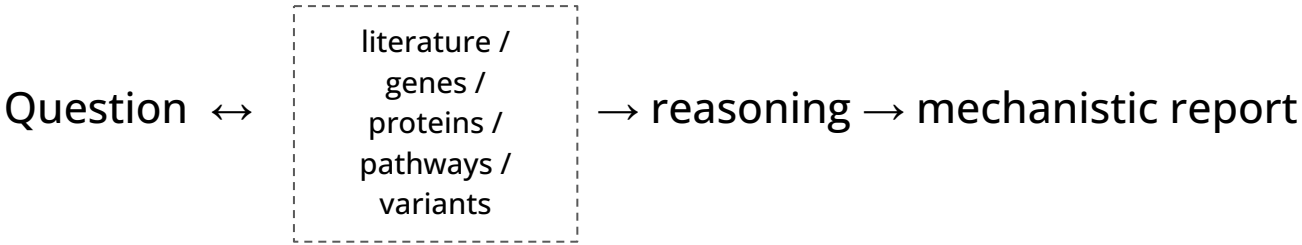
Why current QA systems are not enough?

Short-form QA

Question → retrieve fact → short answer

versus

Systems Biology Long-form QA



multi-source · multi-step · long-form

The difficult part is not retrieving one fact. It is deciding what evidence matters and connecting it into a scientific explanation.

Application Setting

- How can agents gather heterogeneous biological evidence for LFQA?
- Can relevance filtering make long reasoning more efficient?
- How do reasoning behaviors differ across models?

Evaluation Testbed for LFQA in Systems Biology

ThinkSysBio

4
systems biology
researchers

60
doctoral-level
questions

5
biological
modalities

6
question
categories

- PubMed
- NCBI Genes
- UniProt
- KEGG
- NCBI ClinVar

- Definition
- Mechanism
- Reasoning
- Comparison
- Application
- Review

Dataset Sample



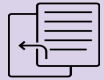
Question



Reasoning
Trace



Long-form
Report



References

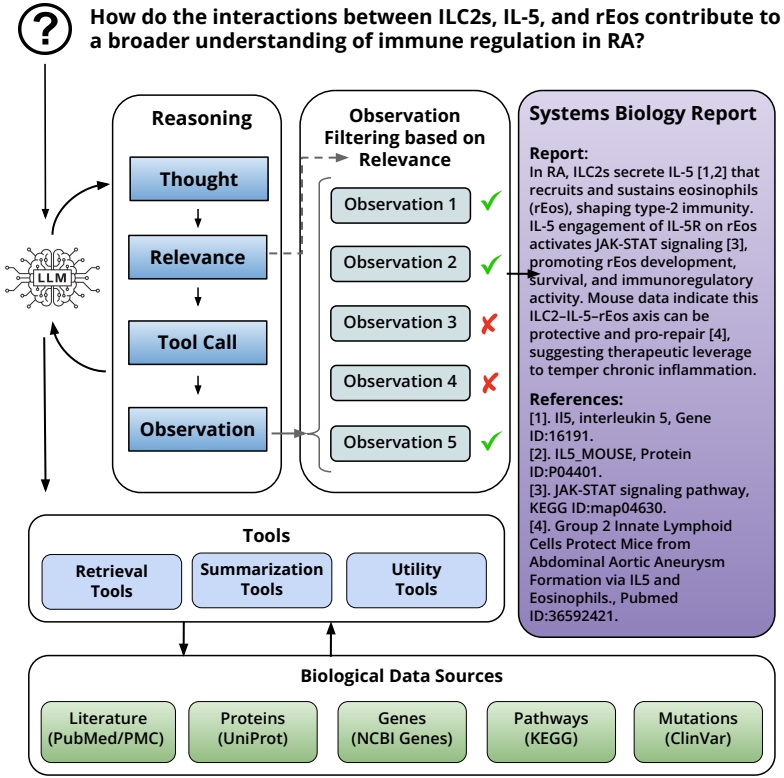
Methodology

ReAct

Reason and Write

ReAct-LF

Reason broadly,
Synthesize Selectively



LLM agent for Long-form Scientific Question Answering in Systems Biology

Experimental Setup

Evaluation Testbed Construction — Human expert reasons with biological tools
Inference — LLM agent reasons over the same environment

Models: GPT-4.1 | Gemini 2.5 Pro | Claude 3.5 Sonnet | Llama 3.3 70B

Methods: LLM | ReAct | ReAct-LF

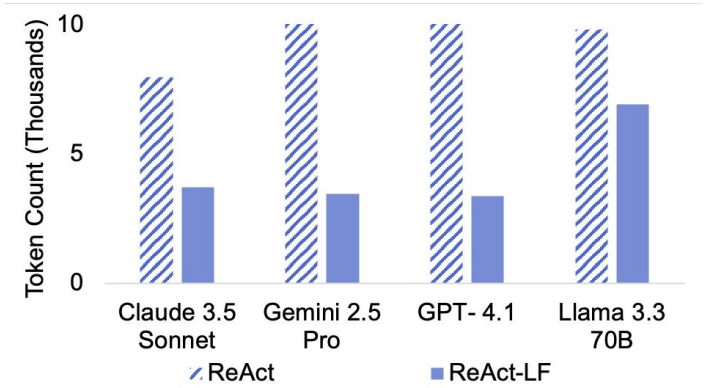
Datasets: ThinkSysBio | BioASQ

Evaluation Lens:



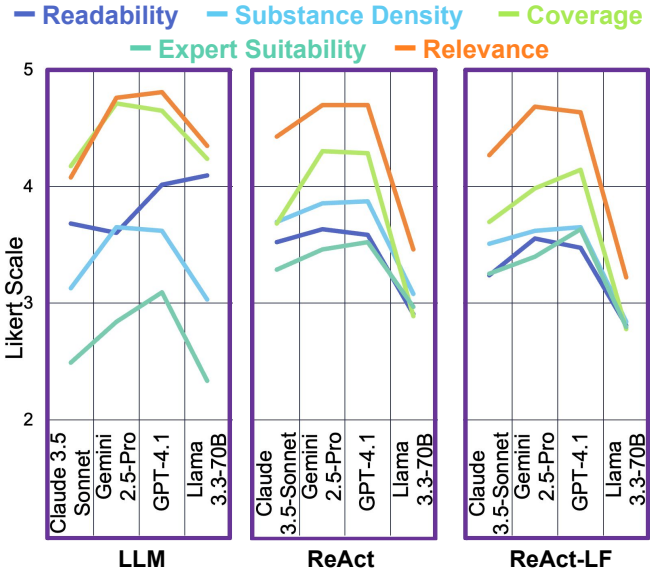
Results

Up to 60% fewer inference tokens



ThinkSysBio illustrates a significant drop of 59.6% (on average) in tokens when ReAct-LF is chosen over ReAct.

With comparable report quality



Qualitative evaluation of model-generated scholarly reports using LLM as judge, with GPT and Gemini leading, Claude mid-tier, and Llama weakest. ReAct and ReAct-LF show similar performance overall, confirming comparable result quality.

Results - Reasoning Behaviors

Frequency of Observed Reasoning Behaviors across Model Architectures

Model	Redirections	Self Monitoring	Forgetting	Repeated Tool Calls	Invalid Tool Calls
Claude 3.5 Sonnet	37.9%	36.39%	6.91%	14.4%	1.54%
Gemini 2.5 Pro	28.57%	23.88%	0.65%	1.32%	1.25%
GPT 4.1	40.85%	33.9%	0%	0%	1.05%
Llama 3.3 70B	31.71%	23.56%	1.49%	1.38%	0.79%

Reasoning behaviors of models on ThinkSysBio using LLM as judge.
Values indicate the average percentage of intermediate steps associated with each behavior, normalized by the total number of steps per example and averaged across the dataset.

↓ **Green = lower** / ↑ **Orange = higher** within each agent block.

Forgetting and Invalid Tool Calls: lower is better.

Observations from Qualitative Case Studies

Reasoning Behaviour

Model	Reasoning Pattern	Memory	Tool use	Evidence
Claude 3.5 Sonnet	Degradation in longer chains	Forgetful	Incorrect, unnecessary	Weak
Gemini 2.5 Pro	Methodical	Strong	Correct and coherent	Particular
GPT 4.1	Exploratory, Coherent	Strong	Broad tool exploration	Effective
Llama 3.3 70B	Exploratory, Inconsistent	Weak	Weakly justified	Weak

Note: The observations above, about the reasoning behaviors, are not always captured by the quantitative metrics alone, highlighting the importance of qualitative evaluation when evaluating agentic systems for open-ended tasks in domain-specific fields.

Takeaways

- **LFQA Reasoning:**
  Scientific long-form QA needs connections across multiple sources, not isolated retrieval.

- **Evidence control :**
  Relevance filtering can make multi-source reasoning substantially more efficient.

- **Model Architectures :**
  Apart from reasoning methodology, model's capacity to preserve and connect evidence across long, multi-step contexts is also important.

- **Qualitative Evaluation matters:**
  Quantitative scores alone hide important differences in how scientific evidence is gathered and retained.

Thank You