ECML PKDD 2026

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

Dresden University of Technology, Germany

Abstract: Systems biology questions often require integrating evidence across heterogeneous biological resources like literature, genes, proteins, pathways, and variants to build mechanistic, long-form explanations. We study LLM agents for this multi-source, multi-step reasoning setting and introduce ReAct-LF (ReAct for Long-form QA), a relevance-aware approach that filters intermediate evidence before scientific synthesis.

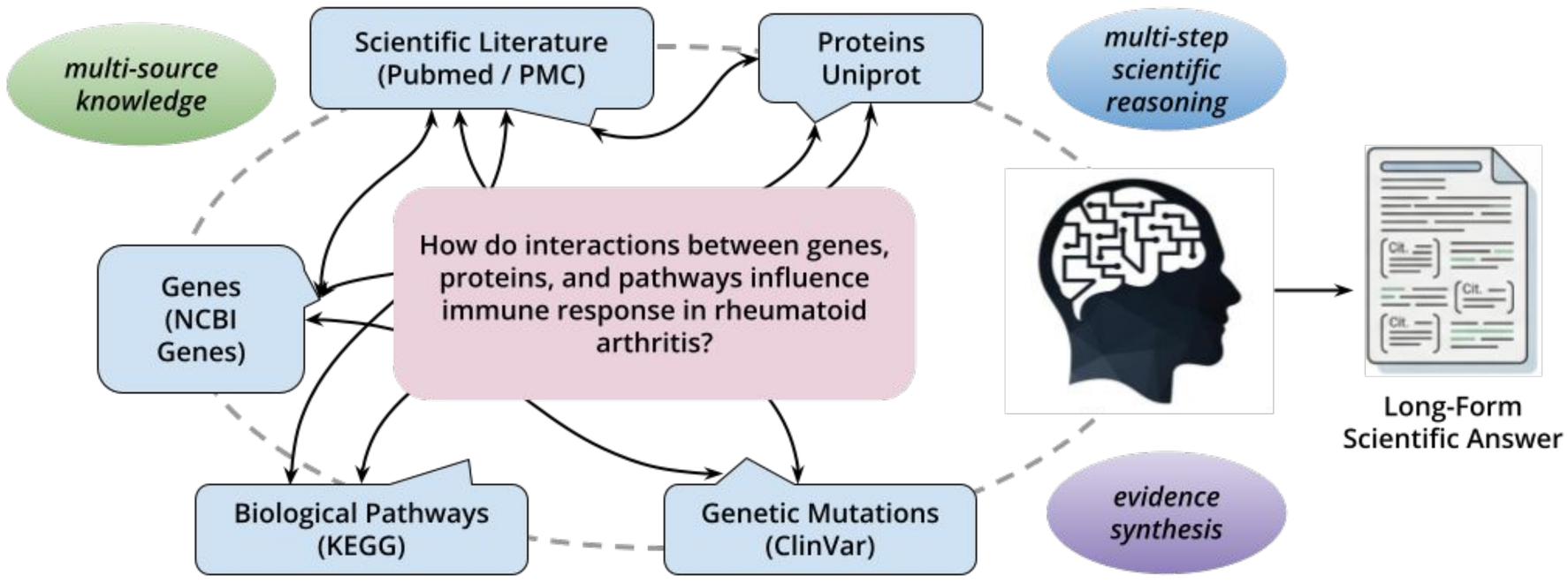
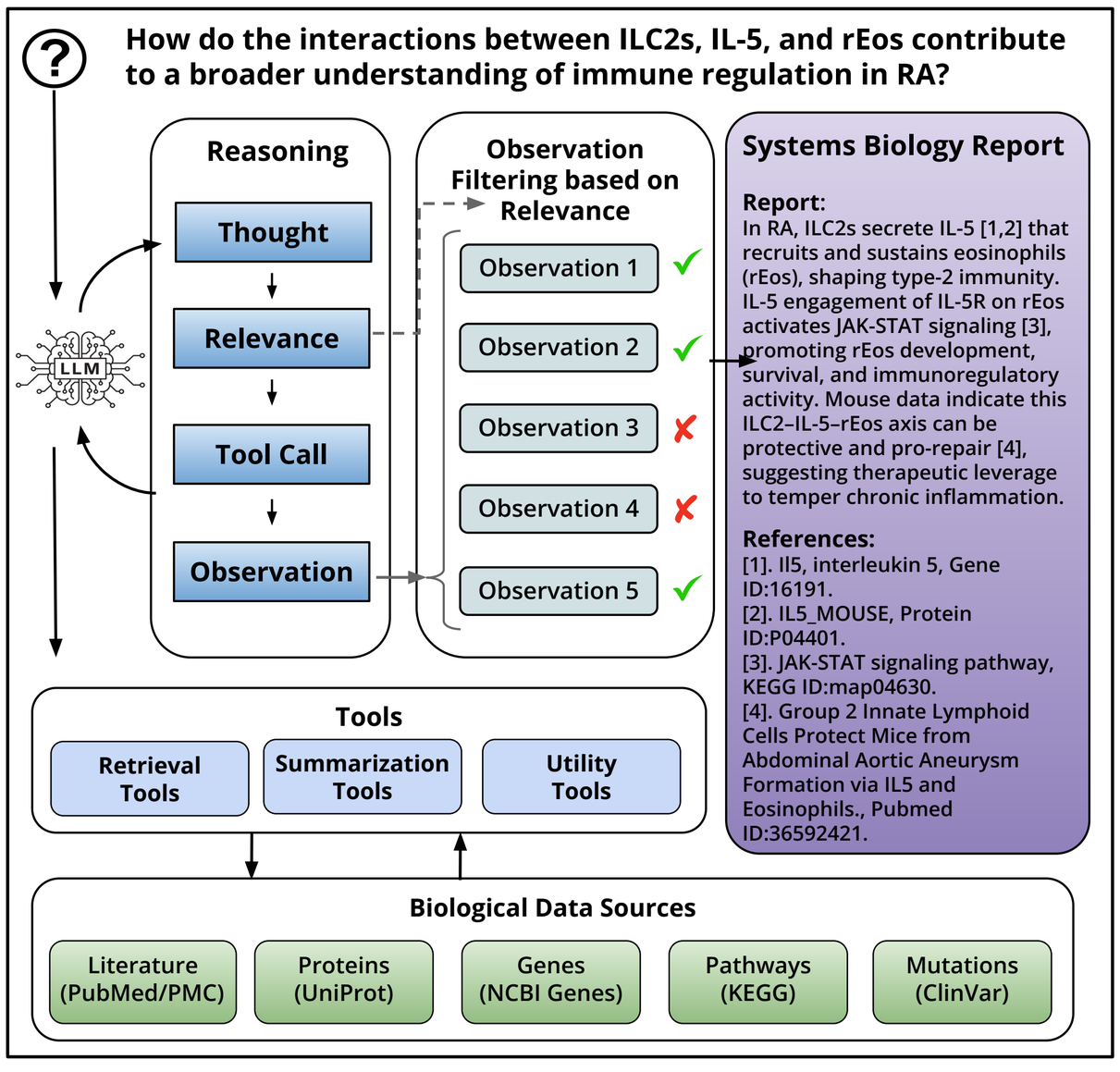


Illustration of research workflow for systems biology researchers

Methodology

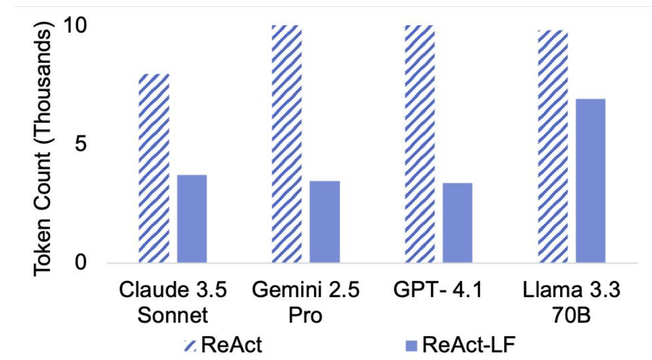
ReAct-LF : *Reason broadly, synthesize selectively.*



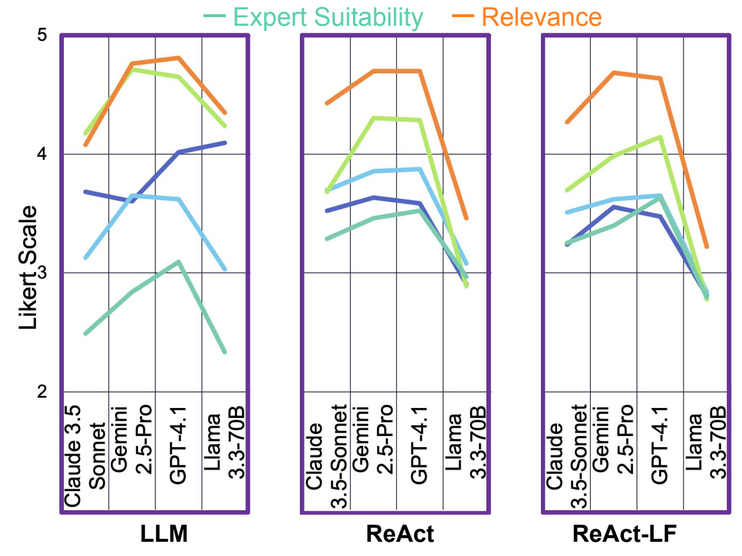
ReAct-LF : LLM agent for Long-form Scientific Question Answering in Systems Biology

Evaluation Results

Token Efficiency : ReAct-LF uses up to 60% fewer inference tokens than ReAct, making long-form question answering substantially more efficient.



Report Quality: ReAct-LF maintains broadly comparable report quality, but filtering can remove useful contextual scaffolding, leading to more focused reports.



Reasoning Behaviour : GPT/Gemini are stable, however, Claude/Llama struggle more with long reasoning chains.

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%
Domain Expert	9.77%	7.41%	0%	0.19%	1.83%

Evaluation Testbed - ThinkSysBio

ThinkSysBio is an expert-curated testbed for long-form systems biology QA. Each data sample contains doctoral-level questions, multi-source evidence, expert reasoning trajectories, and long-form scholarly reports.

4 Systems Biology researchers	**60 doctoral-level questions**
5 biological modalities	**6 question categories**

Takeaways for Systems Biology

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

GitHub: https://github.com/Charvijain16/LFQA_SB

Contact: charvi.jain@tu-dresden.de

Dresden University of Technology
SECAI School of Embedded Composite Artificial Intelligence
SPONSORED BY THE Federal Ministry of Education and Research
DAAD Zuse Schools Konrad Zuse Schools of Excellence in Artificial Intelligence