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

Charvi Jain[1(✉)], Sahar Vahdati[1,2,7], Pauline Porschitz[8], Mounir El Hankouri[6], Nandu Gopan[1,3,5], Ivo F. Sbalzarini[1,2,3,5,6], and Jens Lehmann[1,2,4]

[1] Faculty of Computer Science, Dresden University of Technology, Nöthnitzer Str. 46, 01187 Dresden, Germany
{charvi.jain,sahar.vahdati,nandu.gopan, ivo.sbalzarini,jens.lehmann}@tu-dresden.de
[2] Center for Scalable Data Analytics and Artificial Intelligence ScaDS.AI, Dresden/Leipzig, Germany
[3] Center for Systems Biology Dresden, Pfotenhauerstr. 108, 01307 Dresden, Germany
[4] Amazon, Dresden, Germany
[5] Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany
[6] Cluster of Excellence, Physics of Life, TU Dresden, Dresden, Germany
mounir.el_hankouri@tu-dresden.de
[7] Leibniz University Hanover, Welfengarten 1, 30167 Hannover, Germany
[8] Center for Regenerative Therapies Dresden (CRTD), Technische Universität (TU) Dresden, 01307 Dresden, Germany
pauline.porschitz@tu-dresden.de

Abstract. AI-based research assistants for domains with cross disciplinary and multi-resource demands, such as systems biology, require deep reasoning and integration across diverse knowledge sources. In this work, we study the application of large language model (LLM) agents for long-form scientific question answering that requires multi-source reasoning across heterogeneous biological knowledge resources. To achieve our aim, we first introduce a high-quality, human-curated dataset, ThinkSysBio, that offers a structured testbed for long-form question answering (LFQA) with multi-source, tool-augmented scientific reasoning in systems biology. Second, we devise an adaptation of the ReAct framework, namely ReAct-LF, for inference tokens efficient LFQA. In addition, we evaluate both ReAct and ReAct-LF using four state-of-the-art open-source and closed-source LLMs. Third, we perform a systematic evaluation of agentic workflows by analyzing both the intermediate reasoning traces and the generated scholarly reports. Our results reveal a practical trade-off between report quality and computational efficiency in agent-based scientific workflows, demonstrating

J. Lehmann—Work done outside of Amazon.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-37682-4_19.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
M. Baratchi et al. (Eds.): ECML PKDD 2026, LNAI 16949, pp. 324–341, 2027.
https://doi.org/10.1007/978-3-032-37682-4_19

that relevance-aware evidence filtering can substantially reduce inference token usage by an average of 61.3% across datasets while maintaining comparable report quality. **Code and dataset available at:** https://github.com/Charvijain16/LFQA_SB.

Keywords: LLMs · Long-Form Question Answering · AI for Science

1 Introduction

Systems biology is an interdisciplinary field that combines biology, mathematics, computation, and physics to understand how biological function emerges from interactions among components such as genes, proteins, pathways, cells, and tissues [3]. Research-oriented questions in systems biology are inherently cross-resource, mechanism-centered, and synthesis-heavy, and hence answering these questions rarely reduces to a single fact lookup. Instead, it often requires integrating evidence from heterogeneous resources, including biological databases, pathway repositories, and scientific literature, and synthesizing it into an expert-style scholarly report that provides mechanistic and evidence-backed explanations. In practice, this workflow is both time-consuming and cognitively demanding: researchers must reformulate searches, inspect partially overlapping evidence, reconcile different sources and assemble them into a valuable article.

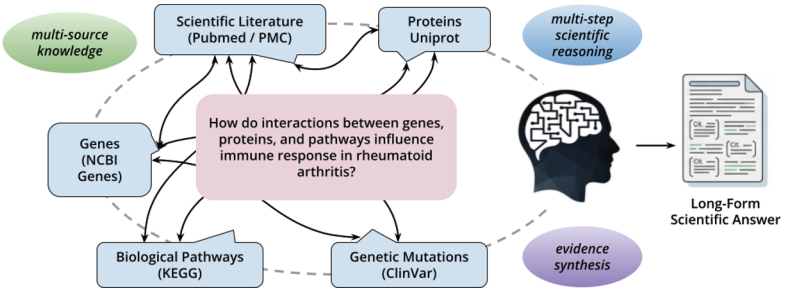


Fig. 1. Illustration of research workflow for systems biology researchers

Recent advances in AI-driven biological research have supported important but relatively isolated tasks such as protein and drug design [7] or DNA/RNA effect prediction [11]. While these efforts demonstrate the growing utility of AI in biology, they do not directly address the broader problem of multi-source scientific synthesis. Real systems biology workflows require not only retrieving information but also deciding what is relevant, connecting evidence across modalities, and producing coherent explanations. Furthermore, existing tools such as Undermind [6], Aurora [9] are valuable for literature exploration, but they focus mainly on publications and follow traditional literature-extraction processes involving a

static sequence of title filtering, abstract filtering, or/and full-text filtering. For systems biology, however, literature alone is often insufficient. Relevant information must often be gathered from iterating back and forth across different biological sources as shown in Fig. 1, rather than following a static, step-by-step pipeline. To enable such dynamic exploration, AI assistants require the ability to reason. ReAct [19] is a widely known approach for interleaving reasoning and tool use, but it has mainly been studied in short-form multi-hop question answering (QA) [14]. This is a limited fit for systems biology. Short-form QA is typically focused and specific, whereas long-form, research-like questions are broad and open-ended. While short-form QA often retrieves answers directly from readily available information, LFQA requires the selection of relevant sources and conceptual entities, because the answer should be grounded in evidence drawn from diverse and credible sources. In this work, we study the application of LLM-based agents for LFQA in systems biology and provide practical insights into how such agents should be studied and evaluated under expert-domain constraints.

First, we extend ReAct with a relevance-filtering step tailored to long-form answer generation, ReAct-LF. The motivation is to remove uninformative data from retrieved evidences before answer generation to improve efficiency and reduce noise. Hence, we compare ReAct and ReAct-LF across several state-of-the-art LLMs. To enable the assessment of this application, we curated ThinkSysBio, an evaluation dataset for LFQA in systems biology with the help of domain experts. The dataset consists of 60 doctoral-level, high-quality scholarly questions with domain-expert reasoning steps and scholarly reports. Each sample required substantial expert effort, high cognitive load, and manual evidence verification across multiple resources. Our goal is to have a controlled testbed for depth and realism, and for analyzing agent behavior in a difficult expert-domain application. We evaluate the agents along multiple axes, consisting of generated report quality, reasoning behavior, and computational efficiency. Our results show that ReAct-LF substantially reduces token usage relative to standard ReAct while maintaining broadly comparable performance quality, highlighting an important efficiency–quality trade-off for long-form scientific QA. Through this work, we provide an application framework for studying how agentic systems can support multi-source scientific reasoning and scholarly report generation in systems biology.

Problem Formulation. We formulate the problem as a multi-source, multi-step reasoning and long-form synthesis task. Given a research-style question from the field of systems biology and access to a set of external biological knowledge sources, the LLM agent must iteratively gather relevant information and integrate into a coherent scholarly report. The task requires producing (i) a step-by-step sequence of tool-augmented reasoning steps, (ii) a final synthesized report conditioned on the original question and the information collected via tool calls. The central challenge lies in orchestrating multi-step retrieval across heterogeneous sources and modalities with well-structured final synthesis.

2 Related Work

Scientific Question Answering. Scientific and biomedical QA has been shaped by benchmarks such as PubMedQA [10] and BioASQ [18], and later extended to broader scholarly settings such as SciQA [1] and literature-centric benchmarks like LitQA/LitQA2 [17]. These resources have advanced scientific QA, but they mainly emphasize short-form answers, literature-only evidence, or scholarly knowledge. In contrast, our setting targets long-form systems biology questions that require synthesis across literature, genes, proteins, pathways, and variants.

LLM Agents and Tool Use. Agentic tool use builds on ReAct, which interleaves reasoning and acting. In science, PaperQA2 [17] iteratively retrieves and synthesizes evidence from scientific papers. Biomedical agents such as Biomni [8] and STELLA [12], which extend tool use to broader biomedical workflows, are evaluated on settings such as HLE-Biomed [16] and LAB-Bench [13]. These works show the promise of tool-augmented LLMs in science, but most focus on literature workflows or short-form evaluation. Our work instead studies domain-specific, long-form systems biology QA with explicit reasoning traces and report generation.

Long-Form Scientific Reasoning. Recent work has begun to study long-context and long-form reasoning in research settings. LongBench-v2 [2] evaluates deep reasoning over realistic long-context tasks, while CURIE [4] focuses on scientific long-context understanding across disciplines. Benchmarks such as DeepResearch Bench [5] and DeepScholar-Bench [15] move closer to open-ended, citation-rich report generation and research synthesis, and related work has also emerged on long-form medical QA with expert annotations. However, these benchmarks are largely cross-domain, web-based, or literature-centered and do not target systems biology as a multi-modal scientific workflow. Our work differs in focusing on domain-specific long-form scientific reasoning grounded in structured biological resources and expert reasoning traces.

3 Methodology

The methodology is illustrated in Fig. 2. For the curation of the evaluation dataset ThinkSysBio and for inference on the systems biology research questions, the methodology remains consistent. During the dataset curation procedure, domain experts interact with a custom user interface to record their thought process and scholarly report for a given question. In case of inference, the domain expert is replaced by an LLM agent that simulates the thought process and generates a scholarly report for the question. For uniformity, we use the term *agent* to refer to either the human domain expert or the LLM, depending on context. In both cases, the agent interacts with an external environment to gather relevant information for the question and frames a scholarly report with proper references.

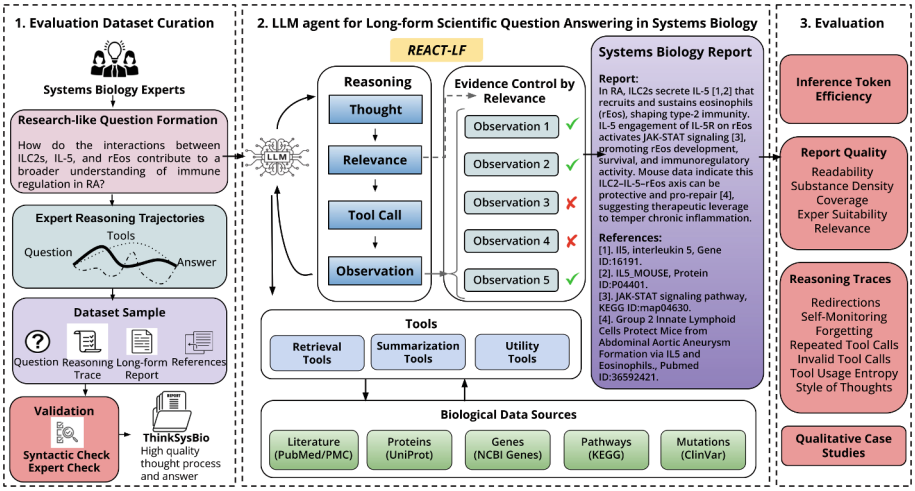


Fig. 2. End-to-End Application Framework for Systems Biology Scholarly Report.

The external environment consists of the tools listed in Table 1. For each data repository, Python tools (or functions) were created for retrieval and summarization. These tools were the only way an agent could interact with the environment to gather information. For a scholarly question Q, the tools can be categorized as retrieval or summarization tools; $\{f_{R_1}, f_{R_2}, \ldots, f_{R_m}, f_{S_1}, f_{S_2}, \ldots, f_{S_m}\} \in F$, where R stands for retrieval, S for summarization, m for data repositories, and F is the tool set.

Retrieval Tools. Given a search string q, the retrieval tool selects a ranked subset of documents from a collection $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$. Formally, the tool returns the top-k elements $f_R(q, \mathcal{D}) = \{d_{i_1}, d_{i_2}, \ldots, d_{i_k}\} where \quad k \ll N$, and the indices $\{i_1, \ldots, i_k\}$ are ordered according to a relevance function $\rho(d, Q)$. This is similar to performing title or abstract-based filtering for literature review.

Summarization Tools. For any subset of retrieved documents $G \subseteq f_R(q, \mathcal{D})$, the summarization tool generates succinct information w.r.t the given question Q as $f_S(G, Q) = \{g_1, g_2, \ldots, g_{|G|}\}$ where each g_j is a summary/gist of the corresponding document $d_j \in G$. This process is similar to full-text filtering for literature review.

Agent–Environment Interaction. For a given question Q, the agent dynamically alternates between using its internal knowledge and interacting with the external environment to gather relevant information for answering the question. We denote the internal knowledge of the agent by B and external knowledge by E. The agent's goal is to produce a final report $\mathcal{L}$ for question Q

$$\mathcal{L} = agent(Q, B, E)\,. \tag{1}$$

Table 1. Tools available to agents and their descriptions.

Tool	Description
Retrieve PubMed	Retrieve papers from PubMed using a search string
Summarize PMC	Summarize information from PubMed Central paper w.r.t question
Retrieve UniProt	Retrieve proteins from UniProt using a search string
Summarize UniProt	Summarize information on UniProt page for an ID and question
Retrieve KEGG	Retrieve pathways from KEGG using a search string
Summarize KEGG	Summarize information on KEGG page for an ID and question
Retrieve Genes	Retrieve genes from NCBI Genes using a search string
Summarize Genes	Summarize information on Gene page for an ID and question
Retrieve ClinVar	Retrieve mutations from NCBI ClinVar using a search string
Summarize ClinVar	Summarize information on ClinVar page for an ID and question
Exch-PMC-PMID	Map PubMed Central IDs to PubMed IDs and vice versa
Fetch Info Any	Gather information for a thought from any of the sources

For the original ReAct approach, the agent yields a trajectory of thoughts T, actions A, and observations O, and generates the final report from the collected thoughts and observations. In ReAct, external knowledge is defined as:

$$E = [\{T_1, A_1, O_1\}, \ldots, \{T_p, A_p, O_p\}], \tag{2}$$

where p is the number of reasoning steps.

When dealing with biological data in LFQA, titles or names are often not enough to understand the underlying concepts. For instance, only on the basis of the names of proteins, like DLG1_DANRE or DLG1L_DANRE, it is difficult to understand their biological function. This is exacerbated by the same protein in different species often having different names. Understanding this requires going through the abstract or description to decide the relevance of a search result. When an agent invokes a retrieval tool f_R, the top-k entries are fetched, out of which only a fraction are typically relevant to the question. Similarly, for a summarization tool f_S, not all possible summaries are equally relevant to the question. Hence, in line with the "garbage in, garbage out" principle, we hypothesize that inclusion of irrelevant documents obtained during the reasoning process introduces noise while generating the scholarly report, whereas filtering such information will improve token efficiency and report quality by focusing on relevant data.

The central methodological contribution of ReAct-LF is the relevance-aware filtering mechanism. Unlike standard ReAct, which carries forward the full observation history regardless of how useful it is, ReAct-LF explicitly separates observation generation from evidence retention. At each reasoning step, the agent identifies the retrieved or summarized items that are relevant to the question and records their identifiers in a relevance field. Only judged-relevant observations are retained for final report generation.

Formally, ReAct-LF introduces a relevance field I in the reasoning and a post-filtering step $O \rightarrow O'$. This extends the ReAct approach for LFQA by modifying the $\{T, A, O\}$ reasoning step to $\{T, I, A, O'\}$. External knowledge for ReAct-LF then is:

$$E' = [\{T_1, I_1, A_1, O'_1\}, .., \{T_p, I_p, A_p, O'_p\}] \tag{3}$$

After an action yields an observation, the next ReAct-LF step combines relevance and thought into a single *structured thought*. We group these together to make explicit the connection to the thought step in standard ReAct and because the agent has to jointly reason about them. Since each entry of a tool output comes with a unique identifier from the source data repository, we directly define the relevance field I over those identifiers.

To illustrate O'_t at a certain iteration t, let

$$O_t = [\, o_1, o_2, \ldots, o_k \,] \tag{4}$$

be a list of dictionaries, where each dictionary o_i has a unique identifier $\mathrm{id}(o_i)$, and let

$$I_{t+1} = [\, i_1, i_2, \ldots, i_m \,] \tag{5}$$

be a list of relevant identifiers such that

$$I_{t+1} \subseteq \{\, \mathrm{id}(o_1), \mathrm{id}(o_2), \ldots, \mathrm{id}(o_k) \,\}. \tag{6}$$

Then, the filtered list O' is defined as:

$$O'_t = [\, o \in O_t \mid \mathrm{id}(o) \in I_{t+1} \,]. \tag{7}$$

This ensures that the report is grounded in structured reasoning and distilled external information. Relevant identifiers are selected by experts during dataset generation and by LLM agent at inference time, with trajectory filtering retaining only relevant observations for final answer generation.

4 Dataset Construction

We constructed the ThinkSysBio in a collaborative, expert-driven process involving 15 researchers from a world-renowned systems biology research institute. Four lead contributors, comprising doctoral researchers and principal investigators, served as co-authors and domain experts, involved in the formulation of scholarly questions, documenting reasoning processes, and authoring detailed answers. To ensure high quality and accuracy, we conducted three group workshops and multiple one-on-one sessions with all systems biology researchers, focusing on data creation, curation, and validation. The resulting dataset represents a rigorously curated, high-quality resource that reflects substantial expert effort, deep domain reasoning, and a strong commitment to quality.

Question Formation. The questions in the ThinkSysBio dataset are formulated by doctoral researchers in systems biology. They are equally distributed

across six major categories: Definition, Mechanism, Reasoning, Comparison, Application, and Review. Table 2 details the question types for each of the categories. The categories ensure diversity and coverage of question types.

Repository Selection. To integrate multiple sources across distinct data modalities, we selected one repository per modality to build a proof of concept. The types of modalities commonly used in systems biology include: research articles, proteins, genes, biological pathways, and genetic mutations. The selected repositories (Table 3) were chosen for their practical usefulness, widespread recognition in the research community, and regular use by the contributing experts.

Table 2. Definitions of six question categories.

Category	Questions that ask for	Question format
Definition	clear definition or basic description	What is/are ...?
Mechanism	a step-by-step explanation	How does/What is the role of..?
Reasoning	the cause of a phenomenon	Why does ...?/What causes ...?
Comparison	comparison between alternatives	Which approach is better for ...?
Application	implications or practical use	How can it be applied in practice?
Review	high-level synthesis of findings/trends.	What are current trends in ...?

Validation. Two levels of quality checks were implemented to validate the constructed dataset. Syntactic validation, ensuring proper categorization of questions and the correct structure of entries, and expert validation, where domain experts evaluated the scientific correctness and usefulness of the final answer and the references.

Dataset Characteristics. The resulting dataset contained a total of 60 questions, 10 per category, each with an average of 8 structured-thought iterations. The answers are designed to be bounded by the knowledge accessible via the five sources listed in Table 3, recognizing that the complete systems biology domain may contain additional relevant data. Summary statistics are reported in Table 4.

5 Evaluation Framework

Datasets. To examine whether our analysis extends beyond the curated corpus, we also conduct evaluation on BioASQ, an established biomedical QA benchmark. For this study, we focus on the BioASQ dataset's summary questions, as they best align with our task formulation and provide reference summary answers derived from multiple sources such as PubMed and UniProt.

Models. We evaluate four strong LLMs spanning both closed-source and open-source model families: GPT-4.1, Claude 3.5 Sonnet, Gemini 2.5 Pro, and Llama

Table 3. Data modalities and their selected repositories.

Modality	Data Repo
Paper's metadata	PubMed
Paper's full text	PubMed Central
Biological pathways	KEGG
Proteins	UniProt
Genes	NCBI Gene
Mutations	NCBI ClinVar

Table 4. Summary statistics of the ThinkSysBio corpus.

Statistic	Value
Total questions	60
Question categories	6
Questions per category	10
Avg. reasoning traces	8
Evidence modalities	5

3.3 70B. All models support function calling for tool invocation, structured output generation, and context lengths of at least 128k tokens. To keep the comparison controlled, all models are prompted with the same system-level instructions, and no model-specific *deep research* mode is enabled.

Framework. We evaluate base LLMs, ReAct, and ReAct-LF. Base LLMs do not produce a reasoning chain and only a scholarly report. To ensure a fair comparison, ReAct and ReAct-LF use the same reasoning trajectories for each question, rather than independently sampling separate trajectories. This is important because LLM outputs are probabilistic, and multiple runs on the same question may produce different but valid reasoning paths. Using the same trajectory ensures that differences in the final reports are not affected by differences in sampled reasoning. While ReAct generates the final report using all observations, ReAct-LF filters observations based on the Relevance field before report generation. This setup isolates the effect of relevance-based filtering while keeping the underlying reasoning trajectory fixed.

Our evaluation centers on four research questions that focus on (i) how does evidence control impact efficiency, (ii) which agents produce high-quality scholarly reports, (iii) how do reasoning behaviors differ across models, and (iv) what do qualitative cases reveal beyond the aggregate metrics.

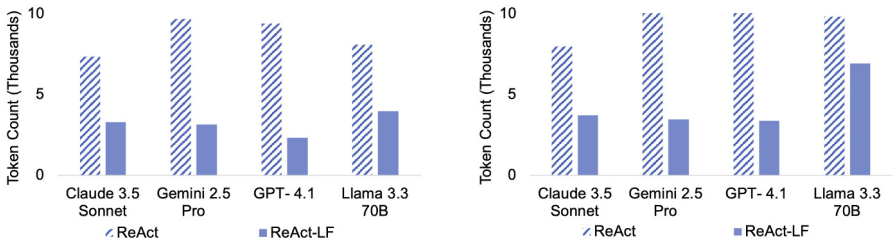
5.1 RQ1: How Does Evidence Control Impact Efficiency?

To examine the use of inference tokens in ReAct-LF and ReAct, we measured tokens with and without post-processing the observations. The Fig. 3a for BioASQ and Fig. 3b for ThinkSysBio clearly illustrate the significant drop of 63% (on average) and 59.6% (on average) in tokens, respectively, when ReAct-LF is chosen over ReAct.

5.2 RQ2: Which Agents Produce High Quality Scholarly Reports?

The scholarly report is the long-form answer produced after gathering information during the reasoning process. To qualitatively evaluate scholarly reports, we use an LLM as a judge on five different criteria: readability, substance density, coverage, expert suitability, and relevance. Their definitions are given in Table 5. For LLM judges, we use two state-of-the-art models, namely GPT-OSS 120B and QWEN3 30B, and prompt them to evaluate each report for each criterion on a Likert scale of 1–5, with 5 being the best. We perform this evaluation on the complete datasets, with each report scored using a single inference run for each judge model and criterion. The scores obtained using GPT-OSS-120B as a judge for BioASQ are shown in Fig. 4a and for ThinkSysBio in Fig. 4b.

Overall, Gemini and GPT perform best across all criteria, followed by Claude with moderate performance, while Llama performs the weakest. This consistent ranking in both figures suggests that model choice influences the quality of the final report. Across agents, score patterns are largely consistent: relevance and coverage are highest, substance density and readability are mid-range, and expert suitability is lowest. The base LLM shows the lowest expert suitability, suggest-



(a) BioASQ (b) ThinkSysBio

Fig. 3. Comparison of token usage for ReAct and ReAct-LF agents

Table 5. Report evaluation criteria and corresponding definitions used for qualitative LLM-as-judge evaluation.

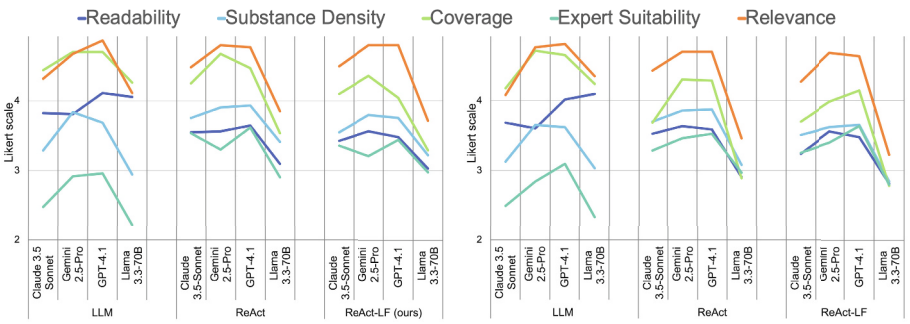
Criterion	Judge prompt
Readability	Assess how easily the report can be followed at the sentence and paragraph level based on linguistic clarity, coherence, transitions, and grammatical flow, independent of content correctness
Substance-Density	Assess whether the report prioritizes precision, concrete claims rather than filler content or loosely related information
Coverage	Assess how well the report addresses the major aspects required to answer the question, regardless of verbosity or brevity
Expert-Suitability	Assess whether the report targets a domain expert audience, assuming prior knowledge and emphasizing insights rather than basic explanations
Relevance	Assess how well the response aligns with the core intent of the question, accounting for off-topic material without evaluating writing quality

ing a trade-off toward layman-friendly explanations over technical depth. Expert suitability increases in ReAct and ReAct-LF in comparison to the base LLM.

Comparing ReAct and ReAct-LF, we observe a slight drop in scores for ReAct-LF. A plausible explanation is that ReAct-LF restricts content to strictly relevant information, excluding background and contextual details. Although the core facts remain similar across ReAct and ReAct-LF, the additional contextual information in ReAct may support clearer explanations and conceptual links, resulting in slightly higher scores. Additionally, we measure the alignment between the two judges using quadratic weighted kappa (46%) and Krippendorff's alpha (44%), indicating moderate reliability; while exact agreement is about 50%, the very high within-one agreement (94%) and low MAE (0.56) indicate that most disagreements are minor (±1 point), which are potentially due to the open-ended and subjective nature of doctoral-level responses.

5.3 RQ3: How Do Reasoning Behaviors Differ Across Models?

The reasoning trajectories are generated by the agents before producing the final report. Analyzing these sequences helps us understand how reasoning influences the final output. We examine reasoning behaviors, tool-use distributions, and the style of generated thoughts. For ThinkSysBio, we compare the reasoning trajectories of ReAct-LF with expert-driven trajectories. The baseline LLM is excluded because it does not generate reasoning traces. For BioASQ, reasoning trajectories for ReAct-LF cannot be compared with expert or ground-truth trajectories, as the dataset lacks them.



(a) BioASQ (b) ThinkSysBio

Fig. 4. Qualitative evaluation of model-generated scholarly reports using GPT-OSS-120B 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.

Table 6. Reasoning behaviors of models on BioASQ and ThinkSysBio, evaluated by GPT-OSS. Values indicate the average percentage of intermediate steps associated with each behavior. ↓ Green = lower/↑ Orange = higher within each agent block. Forgetting and Invalid Tool Calls: lower is better.

	Agent	Model	Redirections	Self Monitoring	Forgetting	Repeated Tool Calls	Invalid Tool Calls
BioASQ	ReAct-LF	Claude 3.5 Sonnet	30.79%	26.88%	2.15%	6.49%	5.01%
		Gemini 2.5 Pro	27.81%	21.58%	0%	0.30%	1.36%
		GPT 4.1	35%	25.02%	0.22%	0.50%	1.76%
		Llama 3.3 70B	31.14%	21.61%	1.16%	1.15%	1.65%
ThinkSysBio	ReAct-LF	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%

Reasoning Behaviors. We evaluate the reasoning behaviors with the help of the same LLM judges as earlier, GPT-OSS 120B and QWEN3 30B on the complete datasets, along five dimensions: (1) thought redirections, defined as explicit strategic pivots after dead ends, ambiguity, or missing information (e.g., "this approach failed; I'll try X instead"); (2) self-monitoring, capturing meta-reasoning where the agent verifies or reflects on its own inferences (e.g., "let me double-check" or "this seems inconsistent"); (3) forgetting incidents, where previously observed information is not used later, such as repeating searches or ignoring constraints it found earlier; (4) invalid tool invocations, including incorrect tool names, malformed queries, or using a tool in a way that shows a misunderstanding of how it works; and (5) repeated tool calls, where the agent reissues the same query without modification, reflecting a failure to learn from prior results. The judges were given clear definitions and examples for each category and asked to respond with integer counts for each dimension and for every solution process. These checks show how often agents change their strategy, think about their own reasoning, or lose important context, and quantify weaknesses in how agents execute their plans. This helps measure how consistent their planning is and how well they manage their thought process, beyond the scholarly report.

The scores for all five phenomena using GPT-OSS-120B for both datasets are shown in Table 6. Domain experts exhibit fewer redirections and self-monitoring and almost no forgetting or repeated tool calls, indicating that human reasoning is more targeted and implicit. Gemini 2.5 Pro features minimal forgetting, few repeated tool calls, and comparatively stable reasoning behavior, representing the most balanced model across both datasets. GPT-4.1 shows high redirections and self-monitoring but almost no forgetting, suggesting strong retention despite a broader and more exploratory strategy. Llama is typically in the middle, but

Claude shows weak memory retention and poor understanding of tool use as demonstrated by forgetting and repetitive or invalid tool calls, respectively.

Tool-Use Behavior. To understand how different models, across different agents, use tools, we compute the Shannon entropy of the distribution of tool calls. The entropies are shown in Table 7. Across datasets, LLMs exhibit higher tool entropy than humans, indicating more diverse and exploratory tool usage patterns. GPT consistently shows the highest entropy, while Claude demonstrates low entropy.

Table 7. Tool entropies across models. Lower is better.

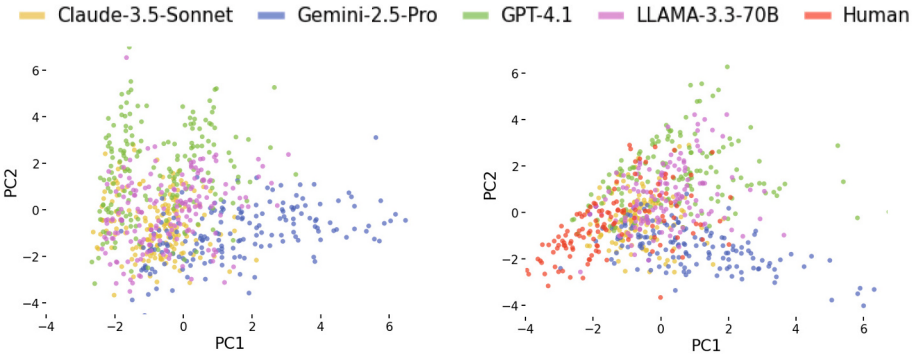
Dataset	Agent	Claude 3.5 Sonnet	Gemini 2.5 Pro	GPT 4.1	Llama 3.3 70B	Expert
BioASQ	ReAct-LF	1.65	2.14	**2.46**	2.33	NA
ThinkSysBio	ReAct-LF	1.58	2.05	**2.45**	2.33	1.46

Stylometric Analysis of Thoughts. In the reasoning process, thoughts are the foundation from which actions and observations emerge, so we examine how humans and different models structure them. We sample three randomly selected thoughts from the reasoning trajectory of each data sample. In total, we thus collect 150 thoughts for each model and original dataset (human-written). For each thought, we extract a suite of stylometric features designed to capture linguistic and narrative structure: token and bigram counts, stopword usage, part-of-speech ratios (nouns, adjectives, verbs), presence of discourse markers indicative of reasoning flow (e.g., "first", "next", "finally"), sentence complexity, and information density. The resulting feature vectors were standardized and projected into two-dimensional space via principal component analysis (PCA) to visualize aggregate differences and overlaps between human and model reasoning.

The PC1-PC2 scatter plot for both datasets for ReAct/ReAct-LF is shown in Fig. 5. The plots exhibit similar overall patterns across the two datasets. The BioASQ plot does not include red dots, as this dataset does not contain original or human-written reasoning traces. Overall, it is evident from the figures how human and model-generated thinking differ. Model-generated traces go toward longer, more structured, and denser reasoning, whereas human traces tend to cluster around shorter, simpler, and less explicit concepts. This implies that LLMs are more likely than humans to overtly externalize their reasoning. GPT and Llama are the models that produce the longest and most complicated ideas, Gemini is comparatively long but less complex, and Claude is yet more succinct and more human-like.

5.4 What Do Qualitative Cases Reveal Beyond Aggregate Metrics?

Aggregate metrics summarize overall performance but often hide important differences in how models gather evidence and construct reports. To better understand these dynamics, we conducted qualitative case studies with domain experts



(a) BioASQ (b) ThinkSysBio

Fig. 5. Stylometric PCA of randomly sampled thoughts from BioASQ and ThinkSysBio. PC1 captures length and lexical-structural variety, while PC2 reflects sentence-level complexity and information density.

on two representative doctoral-level questions from the ThinkSysBio for all four models under both ReAct and ReAct-LF settings. For one of them, a detailed transcript is provided in the Supplementary Material, while Tables 8 and 9 summarize the observations from all case studies, based on reasoning trajectories and final reports.

Table 8. Qualitative comparison of reasoning behavior across models

Aspect	Claude3.5 Sonnet	Gemini2.5 Pro	GPT 4.1	Llama3.3 70B
Reasoning Pattern	Degradation in longer chains	Methodical	Exploratory, Coherent	Exploratory, Inconsistent
Memory	Forgetful	Strong	Strong	Weak
Tool use	Incorrect, unnecessary	Correct and coherent	Broad tool exploration	Weakly justified
Evidence	Weak	Particular	Effective	Weak

Reasoning. Domain expert's reasoning is strong and implicit rather than step-by-step articulation. This affirms the observation in reasoning behavior in Sect. 5.3 and Table 6 where experts show fewer redirections and self-monitoring and almost no forgetting or repeated tool calls. LLM agents, on the other hand, externalize their reasoning to varying degrees. GPT models exhibit coherent reasoning behavior, strong memory retention, and effectively selecting relevant evidence. Their strategy tends to be breadthwise, they explore multiple sources

before converging on the most relevant evidence. Gemini shows similar powerful reasoning, but it does it in a more organized and thorough way, going over previous steps in the reasoning process and explicitly ruling out evidences that are not relevant. Although this increases focus, it sometimes results in an unnecessarily narrow selection of evidences, discarding information that is vaguely related yet possibly helpful. Claude starts with rational paths, but they get worse over longer chains, which leads to repeated or unnecessary tool calls. This indicates poor retention of prior reasoning context. Llama shows the least stable behavior, with exploratory reasoning behaviors and inconsistent relevance selection that often result in incomplete evidence collection.

Scholarly Reports. The quality of the reports produced is directly influenced by the reasoning processes used. Under the standard ReAct setting, GPT and Gemini produce detailed and comprehensive answers that closely follow the question structure, although GPT sometimes supplements retrieved information with internal model knowledge and hallucinations. Claude and Llama, by contrast, rely heavily on the completeness of retrieved evidence, which often leads to fragmented or incomplete reports when reasoning trajectories degrade. Under ReAct-LF, since the final synthesis stage only uses filtered observations, all models produce reports that are more structured and straightforward. This filtering gets rid of noise in the reasoning trace and makes the reports more factually grounded. However, the reduced evidence set can also remove contextual scaffolding that sometimes contributes to richer explanations, resulting in focused but occasionally less broad reports.

Table 9. Qualitative comparison of reports under ReAct and ReAct-LF

Aspect	Claude3.5 Sonnet	Gemini2.5 Pro	GPT 4.1	Llama3.3 70B
ReAct report	Rely only on reasoning trace for information	Verbose, mechanistic, and well-structured; minor factual issues	Well-structured; uses internal knowledge; few hallucinations	Weakly structured, limited mechanistic explanation
ReAct-LF report	Improved focus and structured; internal knowledge presents ideas	Precise and straightforward; retains key points with improved references	More direct and better grounded, with fewer inconsistencies	Slightly better organized, but still incomplete due to dependence on reasoning trace

Overall, the qualitative cases reveal that differences in reasoning style, memory retention, tool use, and relevance selection play an important role in determining the quality of long-form scientific answers. These distinctions are not always captured by aggregate report metrics alone, highlighting the importance of qualitative evaluation when evaluating agentic systems for open-ended tasks in domain-specific fields.

6 Discussion

Our results provide several practical insights for building AI systems that support long-form scientific question answering in systems biology and require navigating heterogeneous biological resources, performing multi-step reasoning, and synthesizing evidence into coherent reports.

1. **LFQA Reasoning:** Agent-based frameworks greatly enhance the capacity of LLMs to address research-oriented inquiries in systems biology by facilitating the integration and reasoning of information from external biological sources, such as genes, proteins, literature, biological pathways, and genetic variants, instead of solely depending on the model's internal knowledge. We also saw how unchecked observations during ReAct reasoning might raise the cost of tokens without making the answers any better.
2. **Evidence Control:** In systems biology procedures, the evidence that is collected can often be too wide, redundant, or only vaguely relevant to the topic. ReAct-LF is a valuable adaptation since it filters observations by relevance before producing the report. This makes the final synthesis more focused and much cheaper, even if some redundant context is lost.
3. **Model Architectures:** The efficacy of agentic systems relies not solely on the reasoning methodology but also on the model's capacity to manage long context and structured reasoning without losing track of earlier evidence. This matters in systems biology because multi-source questions often require following biological relationships over several steps before a coherent explanation can be formed.

7 Conclusion

We presented an application of LLM-based agents to long-form question answering in systems biology, where useful answers require integrating evidence across literature, genes, proteins, pathways, and variants to form mechanistic, expert-style explanations. To study this setting, we introduced ThinkSysBio, an expert-curated dataset designed to capture not only challenging research-style questions but also the reasoning processes that underlie scientific synthesis. We proposed ReAct-LF, a relevance-aware variant of ReAct for LFQA that filters intermediate observations before report generation. Across models and datasets, ReAct-LF yields token savings of 61.3% on average while maintaining comparable report quality. Overall, our findings highlight the promise of agentic systems for supporting multi-source reasoning and long-form scientific synthesis in systems biology.

Limitations: Our work has a few limitations. The ThinkSysBio contains 60 curated samples. While this limits the dataset size, it reflects a deliberate emphasis on high-quality expert reasoning traces rather than large-scale automatic generation. The primary intellectual value of the dataset lies in exposing domain

expert's thought processes, rather than providing static final answers that may evolve as scientific knowledge progresses. We did not compute aggregate false-positive/false-negative rates of the relevance field in ReAct-LF over the full dataset, since this would require expert annotations for all intermediate observations. Instead, we provided a qualitative analysis through two representative case studies in Sect. 5.4 and in the Supplementary Material. Our current setup uses only one representative repository for each biological modality. This choice keeps the environment controlled and interpretable for studying multi-source reasoning. Similarly, each question is associated with a single expert reasoning trajectory. In open-ended scientific discovery, there may be multiple valid reasoning paths, but capturing all such alternatives would require extensive additional expert effort.

Acknowledgments. This work was supported by BMFTR (Federal Ministry of Research, Technology and Space) in DAAD project 57616814 (SECAI, School of Embedded Composite AI) as part of the program Konrad Zuse Schools of Excellence in Artificial Intelligence. We thank all participants of the workshop at the Center for Systems Biology Dresden for their inputs during discussions.

References

1. Auer, S., et al.: The SciQA scientific question answering benchmark for scholarly knowledge. **13**(1), 7240. https://doi.org/10.1038/s41598-023-33607-z, https://www.nature.com/articles/s41598-023-33607-z
2. Bai, Y., et al.: LongBench v2: towards deeper understanding and reasoning on realistic long-context multitasks (2025). https://arxiv.org/abs/2412.15204
3. Bruggeman, F.J., Westerhoff, H.V.: The nature of systems biology. Trends Microbiol. **15**(1), 45–50 (2007)
4. Cui, H., Shamsi, Z., Cheon, G., et al: Curie: evaluating LLMs on multitask scientific long context understanding and reasoning (2025). https://arxiv.org/abs/2503.13517
5. Du, M., Xu, B., Zhu, C., Wang, X., Mao, Z.: DeepResearch bench: a comprehensive benchmark for deep research agents (2025). https://arxiv.org/abs/2506.11763
6. Hartke, T., Ramette, J.: Benchmarking the undermind search assistant (2024)
7. Hayes, T., et al.: Simulating 500 million years of evolution with a language model. Science **387**(6736), 850–858 (2025)
8. Huang, K., et al.: Biomni: a general-purpose biomedical AI agent. biorxiv (2025)
9. Jain, C., Vahdati, S., Gopan, N., Sbalzarini, I.F., Lehmann, J.: Evaluating large language model literature reviews in interdisciplinary science: a systems biology perspective (2024). https://ceur-ws.org/Vol-3967/PD_paper_173.pdf
10. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: PubMedQA: a dataset for biomedical research question answering (2019). https://arxiv.org/abs/1909.06146
11. Jin, Q., Yang, Y., Chen, Q., Lu, Z.: GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics **40**(2) (2024). https://doi.org/10.1093/bioinformatics/btae075
12. Jin, R., Zhang, Z., Wang, M., Cong, L.: Stella: self-evolving LLM agent for biomedical research. arXiv preprint arXiv:2507.02004 (2025)

13. Laurent, J.M., et al.: Lab-bench: measuring capabilities of language models for biology research (2024). https://arxiv.org/abs/2407.10362
14. Lehmann, J., Bhandiwad, D., Gattogi, P., Vahdati, S.: Beyond boundaries: a human-like approach for question answering over structured and unstructured information sources. Trans. Assoc. Comput. Linguist. **12**, 786–802 (2024). https://doi.org/10.1162/tacl_a_00671, https://doi.org/10.1162/tacl_a_00671
15. Patel, L., et al.: Deepscholar-bench: a live benchmark and automated evaluation for generative research synthesis (2026). https://arxiv.org/abs/2508.20033
16. Phan, L., et al.: A benchmark of expert-level academic questions to assess AI capabilities. Nature **649**(8099), 1139–1146 (2026). https://doi.org/10.1038/s41586-025-09962-4
17. Skarlinski, M.D., et al.: Language agents achieve superhuman synthesis of scientific knowledge (2024). https://arxiv.org/abs/2409.13740
18. Tsatsaronis, G., et al.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. **16**(1), 138 (2015)
19. Yao, S., et al.: React: synergizing reasoning and acting in language models (2023). https://arxiv.org/abs/2210.03629