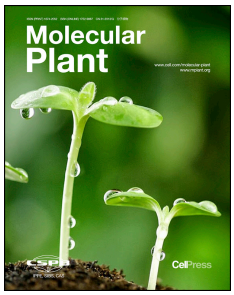
Journal Pre-proof



SeedLLM·Rice: A large language model integrated with rice biological knowledge graph

Fan Yang, Huanjun Kong, Jie Ying, Zihong Chen, Tao Luo, Wanli Jiang, Zhonghang Yuan, Zhefan Wang, Zhaona Ma, Shikuan Wang, Wanfeng Ma, Xiaoyi Wang, Xiaoying Li, Zhengyin Hu, Xiaodong Ma, Minguo Liu, Xi-Qing Wang, Fan Chen, Nanqing Dong

PII: S1674-2052(25)00172-8

DOI: https://doi.org/10.1016/j.molp.2025.05.013

Reference: MOLP 1916

To appear in: *MOLECULAR PLANT*

Received Date: 21 February 2025

Revised Date: 8 May 2025

Accepted Date: 23 May 2025

Please cite this article as: Yang F., Kong H., Ying J., Chen Z., Luo T., Jiang W., Yuan Z., Wang Z., Ma Z., Wang S., Ma W., Wang X., Li X., Hu Z., Ma X., Liu M., Wang X.-Q., Chen F., and Dong N. (2025). SeedLLM·Rice: A large language model integrated with rice biological knowledge graph. Mol. Plant. doi: https://doi.org/10.1016/j.molp.2025.05.013.

SeedLLM·Rice: A large language model integrated with rice biological knowledge graph

Fan Yang[a,1,2], Huanjun Kong[b], Jie Ying[b], Zihong Chen[b], Tao Luo[a], Wanli Jiang[b], Zhonghang Yuan[b], Zhefan Wang[b], Zhaona Ma[a], Shikuan Wang[a], Wanfeng Ma[a], Xiaoyi Wang[a], Xiaoying Li[a], Zhengyin Hu[e], Xiaodong Ma[c], Minguo Liu[f], Xi-Qing Wang[c,g,1], Fan Chen[a], and Nanqing Dong[b,d,1,2]

[a]Yazhouwan National Laboratory, Sanya, 572025, China.
[b]Shanghai Artificial Intelligence Laboratory, Shanghai, 200232, China.
[c]State Key Laboratory of Plant Environmental Resilience, College of Biological Sciences, China Agricultural University, Beijing 100193, China.
[d]Shanghai Innovation Institute, Shanghai, 200231, China.
[e]National Science Library (Chengdu), Chengdu, 610299, China
[f]College of Grassland Agriculture, Northwest A & F University, Shaanxi, 712100, China.
[g]Frontier Technology Research Institute, China Agricultural University, Shenzhen, 518000, China.
[1]Correspondence to: dongnanqing@pjlab.org.cn; yangfan@yzwlab.cn; wangxq21@cau.edu.cn
[2]These authors contributed equally.

ABSTRACT

Rice biology research involves complex decision-making, requiring researchers to navigate a vast and growing body of knowledge that includes extensive literature and multiomics data. The exponential increase in biological data and scientific publications has posed significant challenges in efficiently extracting meaningful insights. While large language models (LLMs) show promise for knowledge retrieval, their application to rice-specific research is hindered by the absence of specialized models and the challenge of synthesizing multimodal data integral to the field. Moreover, the lack of standardized evaluation frameworks for domain-specific tasks impedes the assessment of model performance in this area. To address these challenges, we introduce SeedLLM·Rice (SeedLLM), a 7-billion-parameter model trained using 1.4 million rice-related publications, which represent nearly 98.24% of global rice research. Additionally, we present a novel human evaluation framework designed to assess LLM performance in rice biology tasks. Initial evaluations of rice-specific tasks demonstrate that SeedLLM outperforms general-purpose models such as OpenAI GPT-4o1 and DeepSeek-R1, achieving win rates ranging from 57% to 88%. Furthermore, SeedLLM is integrated with the rice biological knowledge graph (RBKG), which consolidates genome annotations for *Nipponbare* and large-scale synthesis of transcriptomic and proteomic information from over 1,800 studies. This integration enhances the ability of SeedLLM to address complex research questions requiring the fusion of textual and multiomics data. To facilitate global collaboration, we provide free access to SeedLLM and the RBKG via an interactive web portal (https://seedllm.org.cn/). SeedLLM represents a transformative tool for rice biology research, facilitating unprecedented discoveries in crop improvement and climate adaptation through its advanced reasoning capabilities and comprehensive data integration.

KEY WORDS

LLM, Knowledge Graph, Multiomics Data Integration, GPT, DeepSeek

SUMMARY

SeedLLM·Rice, a 7-billion-parameter language model trained on 1.4 million rice-related publications, outperforms general-purpose models such as GPT-4o and DeepSeek-R1 in rice-specific tasks. Through integration with a rice biological knowledge graph, it demonstrates

superior capacity for multiomics data synthesis, positioning it as a robust tool for AI-enabled crop genomics and systems biology, with continued validation anticipated to broaden its applicability.

INTRODUCTION

Paddy rice (*Oryza sativa*) is a crucial staple crop that supports nearly half of the global population (Shi et al., 2023). Research on rice biology is inherently complex, requiring a comprehensive understanding of literature, experimental data, and hypothesis formulation to advance knowledge in the field (Majumdar et al., 2017). This iterative process often challenges or refines established theories, leading to new insights. For example, investigating the function of uncharacterized genes in rice requires a thorough understanding of related biological pathways, as documented in prior studies (Consortium, 2009; Warde-Farley et al., 2010). Similarly, crop breeding efforts rely heavily on genetic and phenotypic data to select optimal parental lines (Huang et al., 2022b). Therefore, a deep understanding of existing knowledge and empirical data is fundamental to advancing rice research.

However, the rapid expansion of biological data, coupled with an overwhelming volume of research literature, presents significant challenges (Bornmann and Mutz, 2015). The sequencing of the rice genome in 2002 marked a key turning point, and subsequent advancements in high-throughput technologies have accelerated the generation of vast datasets (Li et al., 2014; Project, 2007). Despite the increasing availability of data, extracting meaningful insights remains a labor-intensive task. Researchers require extensive academic training to navigate this expanding body of knowledge, making the discovery of novel insights increasingly difficult. This highlights the urgent need for advanced tools that are capable of efficiently navigating and extracting relevant information from the growing corpus of rice-related biological data.

A promising solution lies in the application of large language models (LLMs) (Naveed et al., 2023). However, several challenges hinder their effective use in rice biology research. First, there is a lack of standardized evaluation frameworks tailored to rice biology, making it difficult to assess the performance of general-purpose LLMs in domain-specific tasks such as comprehension, reasoning, and language generation (Lam et al., 2024). Although benchmarks exist for multilingual tasks or biomedical applications (Li et al., 2020; Singhal et al., 2023), no such tools have been developed for rice biology. Second, the absence of domain-specific training limits the effectiveness of general-purpose models in this domain (Nazi and Peng, 2024), where domain-specific LLMs have been shown outperforming performance over general-purpose LLMs. This gap stems from the limited availability of large-scale, diverse corpora needed to

train rice-specific models. Finally, rice biology studies generate complex, multiomics data (Huang et al., 2022a)—such as transcriptomics and genomic sequences—that are difficult to fully represent in textual formats. LLMs, which are predominantly trained using textual data, struggle to synthesize multimodal information, restricting their ability to address complex research questions that require an integrated approach.

To address these challenges, we introduce SeedLLM·Rice (hereafter referred to as SeedLLM), a 7-billion-parameter large language model trained using 1.4 million rice-related publications, covering nearly 98.24% of the global literature in the field. To fill this gap, SeedLLM is designed as a domain-specific LLM capable of processing and integrating diverse datasets relevant to rice biology. A novel human evaluation framework is designed to assess the performance of SeedLLM on tasks such as gene function prediction, textual integration of transcriptional and proteomic data, and variety breeding. Initial results from over human evaluations demonstrate that SeedLLM outperforms general-purpose models such as DeepSeek-R1 in rice-specific tasks, with a 88.14% win rate. Furthermore, SeedLLM is integrated with the rice biological knowledge graph (RBKG), a comprehensive resource that includes the latest genome annotations for *Nipponbare* and large-scale synthesis of transcriptional and proteomic information from over 1,800 academic publications. This integration enables SeedLLM to address complex rice biology questions by drawing from both textual data. Our results show that, when augmented with the RBKG, SeedLLM significantly outperforms all general-purpose LLMs in advanced-level rice omics tasks, despite some limitations in reasoning ability. To increase accessibility, we have developed an interactive web portal (https://seedllm.org.cn/) that allows researchers worldwide to freely access both SeedLLM and the RBKG, thus accelerating the pace and depth of rice biology research.

RESULTS

Overview of SeedLLM development

To develop a specialized LLM for rice research, we collaborated with experts in rice biology to create RiceCorpus, a rice-specific corpus designed to address the lack of a specialized dataset for rice-domain LLMs (Figure 1A). This rice-specific corpus was developed under the guidance of experts in the field of rice biology research, ensuring comprehensive coverage (Supplemental Figure 1). RiceCorpus integrates both English and Chinese data, reflecting the

primary languages used in rice research. The corpus comprises 1.4 million peer-reviewed papers containing keywords such as "rice" and "*Oryza sativa*" published over the past 40 years. This collection represents approximately 98.24% of global rice-related research in these languages during this period, providing a robust foundation for training rice-specific LLMs. Additionally, RiceCorpus includes 1,207 rice-related books. The corpus is exclusively textual and covers a wide range of disciplines in rice research, including molecular biology, plant breeding, and management practices. With a total size of 3,397.49 GB of textual data, RiceCorpus was processed through a multistep quality control pipeline (see the Methods for details), including language detection, content filtering, and deduplication, ensuring that only high-quality data were retained. This meticulous process meets the stringent requirements necessary for training large-scale LLMs.

SeedLLM construction and automated evaluation

We selected Qwen2.5-7B, a Transformer-based general-purpose LLM with 7 billion parameters (Qwen et al., 2024), as the base model (Figure 1B). This base model that had been pretrained using a large, multiphase dataset comprising 18 trillion tokens. Previous reports have demonstrated that Qwen excels in language proficiency, comprehension, reasoning, and mathematics. We pretrained the base model using RiceCorpus and GeneralCorpus, the latter being a widely used general-purpose corpus for LLM training (Huang et al., 2024; Penedo et al., 2024). To evaluate its effectiveness, we assessed the model on two rice biology datasets: MCQ-ACC, consisting of 300 single-choice questions, and Gen-QA-ACC, containing 517 short-answer questions (Figure 1C). The pretrained model outperformed the base model that had not been pretrained with RiceCorpus (Figure 1D), demonstrating that pretraining with rice-specific data enabled the LLM to acquire complex rice biology knowledge and domain-specific language patterns. Subsequently, we fine-tuned the pretrained model using RiceQA, a large annotated question–answer dataset in rice biology, along with GeneralQA, a commonly used dataset in general domains (Dong et al., 2024). This process enhanced the model's performance on rice-related tasks where preserving its general-purpose capabilities. The resulting model, SeedLLM (Figure 1B), shares the same architecture as the base model but exhibits improved understanding of rice biology and domain-specific linguistic features due to the pretraining and fine-tuning procedures (see Supplemental Methods for more details).

We also conducted an automated evaluation of SeedLLM using the Agri series dataset, which consists of 1,975 question–answer pairs across 10 subdatasets, each with various task types, such as essay-style, summary, language understanding, and multiple-choice questions. SeedLLM outperformed general-purpose LLMs, including Qwen2.5 and Llama3.1 (Grattafiori et al., 2024), across all subdatasets, achieving the highest accuracy, F1, and ROUGE scores in automated evaluations (Figure 1E-G). These results demonstrate the effectiveness of the fine-tuning process in optimizing SeedLLM for rice-specific tasks. Additionally, we assessed the generalizability of SeedLLM by fine-tuning it with GeneralQA, a benchmark dataset for general knowledge widely used to assess LLMs' abilities in general-purpose tasks. Despite being primarily pretrained and fine-tuned for rice-related tasks, SeedLLM achieved accuracy scores comparable to those of general-purpose models in various general knowledge tasks, such as mathematics problem-solving with GSM8K (Figure 1H). This finding suggests that SeedLLM retains the ability to perform general language understanding, reasoning, and mathematics tasks, making it a versatile tool for both domain-specific and general-purpose applications.

In summary, we developed SeedLLM, a 7-billion-parameter large language model specifically designed for rice research. By leveraging RiceCorpus, a comprehensive rice-specific corpus, and applying a robust pretraining and posttraining methodology, we created a model capable of outperforming general-purpose LLMs in rice-related tasks, as validated by two rounds of automated evaluations.

Human-Centric Evaluation of SeedLLM Performance

LLMs are capable of generating long, coherent, and complex responses. However, they are also prone to factual inaccuracies (Huang et al., 2023), necessitating careful verification by human experts, particularly in specialized fields such as rice biology. To fully evaluate SeedLLM’s domain-specific comprehension and knowledge retrieval, we developed a human-centric framework (Figure 2A). This framework enables the assessment of the model’s ability to generate complete and accurate answers to real-world research questions.

We began by constructing HumanDesignRiceQA, a high-quality, human-designed question-answering benchmark tailored for rice biology. This benchmark enables LLM-generated responses and facilitates subsequent assessment of their accuracy and quality. Developed by rice biology experts, the benchmark comprises 253 questions spanning 6 topics,

including rice gene function, multiomics, genome-wide association studies (GWAS), traditional breeding and molecular breeding (Figure 2B). Questions are categorized into three levels on the basis of their complexity: basic, intermediate, and advanced. Basic-level questions can be answered via publicly available information (e.g., research abstracts or online resources) without requiring prior education or experience in rice biology. In contrast, advanced-level questions require individuals to have undergone at least minimal formal training in rice biology and to synthesize information from multiple academic papers, integrating insights from scientific literature and experimental data. The benchmark design reflects the cognitive challenges encountered by individuals with varying levels of expertise when addressing rice biology problems. Additionally, input from a diverse group of experts ensured that the benchmark encompassed a broad spectrum of contemporary rice research topics.

Next, we tasked SeedLLM, along with several general-purpose LLMs, with generating responses to the questions in the HumanDesignRiceQA benchmark. To establish a baseline for comparison, undergraduate students specializing in agronomy or crop breeding who had completed relevant coursework also provided responses to the same set of questions. These undergraduate students' answers served as a representative baseline ability of human performance, allowing for a direct comparison between SeedLLM and typical human understanding of rice biology.

We assembled a panel of human evaluators to assess whether responses demonstrated correct or incorrect rice-specific reading comprehension and knowledge retrieval. Over 326 individuals with academic backgrounds in agronomy, including 83 experts in rice biology and variety development, participated in the evaluation (Figure 2C). Evaluations were conducted in a blinded manner, ensuring that the evaluators were unaware of which responses were generated by SeedLLM. Responses were rated on a scale from 0 to 100 or ranked from best to worst, using answer keys or expert experience as reference.

SeedLLM received higher human evaluation scores than all other tested LLMs in the HumanDesignRiceQA benchmark across both rounds (Figure 2D, Supplemental Figure 2). Superior performance of SeedLLM was observed across all question difficulties, as it outperformed all other tested LLMs in basic, intermediate, and advanced-level questions (Supplemental Figure 3). Evaluators also assigned that the highest number of the best responses came to SeedLLM (Figure 2E). These results demonstrate that SeedLLM outperforms general-

purpose LLMs in rice-specific question-answering tasks, as validated by human evaluators over time. This suggests that SeedLLM maintains its leading performance, even as general LLMs evolve.

In conclusion, the human-centric evaluation confirmed that SeedLLM outperforms both general-purpose LLMs and the human baseline across a broad range of rice biology questions. The model demonstrated state-of-the-art performance in tasks requiring a deep understanding of rice biology. However, SeedLLM achieved an average score of 69.98 in answering advanced-level questions, highlighting areas for improvement, particularly in multistep reasoning and integrating complex biological data into textual information.

Construction of the rice biological knowledge graph

Recent studies have shown that incorporating external knowledge graphs into LLMs enhances their reasoning and data fusion capabilities (Pan et al., 2023; Peng et al., 2024). Motivated by these findings, we developed the rice biological knowledge graph (RBKG), a multimodal graph that integrates transcriptional and proteomic data from over 1,879 papers and gene annotation information (Figure 3A and 3B). The construction of the RBKG occurred in three phases: textual integration of transcriptional and proteomic data, integration of rice genome annotation, and comprehensive data fusion.

We first identified scientific papers reporting rice transcriptome and proteome data. Through a comprehensive literature search, we identified 1,879 papers provided raw or preanalyzed transcriptomic and proteomic data. However, inconsistencies in experimental protocols and analysis methods present challenges to data standardization. To address this problem, we structured the transcriptomic and proteomic data by representing each gene's transcriptional event as a sentence, using an approach similar to that of CellAnnoation (Fang et al., 2024). This structure included gene expression levels, protein abundance, experimental attributes (e.g., rice variety, genetic background, tissue type, developmental stage), and conditions specified in the respective studies. These data were then modeled within a knowledge graph framework, with nodes representing gene IDs, transcriptional events, and experimental attributes and edges denoting relationships among them (Figure 3A). This approach facilitated the standardization and harmonization of diverse rice omics data, resulting in a cohesive rice omics knowledge graph. Next, we integrated gene annotation data, including Gene Ontology

(GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations, for all genes identified in the transcriptomic and proteomic datasets. Annotation nodes were created and linked to the corresponding gene IDs, ensuring that functional annotations were directly connected to the underlying transcriptomic and proteomic data.

Finally, we fused the individual knowledge graphs—transcriptomic, proteomic, and gene annotation—into a unified RBKG using graph matching algorithms to reconcile discrepancies and resolve conflicts from overlapping data sources. The resulting RBKG comprises 401,094 nodes and 1,573,258 edges, representing transcriptional, translational, and genomic data for 33,599 rice genes (Figure 3B). To our knowledge, the RBKG is the first knowledge graph that integrates both transcriptional and proteomic data for rice, providing a comprehensive and structured framework for rice multiomics retrieval and reasoning.

SeedLLM integrates knowledge graphs for complex biological queries

To investigate whether the RBKG could enhance the ability of SeedLLM to answer complex biological questions related to rice, we developed a framework that integrates the RBKG into the response generation process of SeedLLM (Figure 3C), hereafter referred to as SeedLLM-KG. This framework comprises three key steps: query decomposition, entity grouping, and knowledge augmentation. In the first step, SeedLLM-KG decomposes user queries into essential entities and attributes, identifying critical components for subsequent processing. The second step involves constructing entity groups by linking the queried entity with relevant concepts from the knowledge graph. Finally, in the knowledge augmentation phase, SeedLLM-KG uses its parametric knowledge base to establish connections between the queried entity and pertinent concepts, thereby enriching the context of the query.

For example, given the following query: “Does the rice gene AGIS_Os06g035130 respond to various environmental conditions?” SeedLLM-KG initially fails to locate AGIS_Os06g035130 in its textual knowledge base (Figure 4A and 4B). It then queries the RBKG for the node representing AGIS_Os06g035130, expanding the search to identify connected nodes that include gene annotations, functional descriptions, and multiomics data. The system retrieves all relevant nodes and edges corresponding to transcriptomic and translational data, synthesizing this structured knowledge into a coherent, human-readable response. Notably, SeedLLM-KG consistently provided correct responses when the AGIS ID of the rice gene was

converted to the corresponding MSU ID and RAP-DB ID in the queries (Supplemental Figures 4 and 5).

SeedLLM-KG integrates multiomic and literature data for complex rice biology questions

To evaluate the performance of SeedLLM-KG in addressing complex rice biology questions, we conducted human-centric assessments via advanced-level questions from the HumanDesignRiceQA dataset. Human experts evaluated the correctness of responses from SeedLLM-KG, awarding an median score of 85, which was significantly higher than the score of 67 awarded to SeedLLM alone (Supplemental Figure 6). This improvement stems from SeedLLM-KG's ability to retrieve and synthesize transcriptional, proteomic and specialized rice research data, which are absent from general-purpose text corpora used to train conventional LLMs (Figure 4A; Supplemental Figures 4 and 5). We also included state-of-the-art general-purpose LLMs, such as DeepSeek-R1 (DeepSeek-AI et al., 2025), DeepSeek-V3 (DeepSeek-AI et al., 2024), GPT-4o1 (OpenAI et al., 2023), and GPT-4o3-mini in our evaluation. SeedLLM-KG demonstrated a unique ability to integrate heterogeneous biological data sources, outperforming all tested general-purpose LLMs (Figure 3D), with win rates ranging from 57.63% to 88.14% (Figure 3E). Notably, it surpassed DeepSeek-R1 in rice-specific task, a leading LLM at the time of writing.

Interestingly, human evaluators assigned similar scores in assessing SeedLLM's reasoning ability, which were not statistically different from other LLMs (Figure 3F). This result was unexpected, as other LLMs, such as DeepSeek-V3 with its 671 billion parameters, were expected to exhibit superior reasoning capabilities, especially considering that SeedLLM-KG is based on SeedLLM, which has only a 7-billion-parameter architecture and lacks reinforcement learning in the post-training stage. The results suggest that human evaluators considered SeedLLM-KG to exhibit similar reasoning abilities compared to LLMs with larger architectures in rice-specific advanced-level tasks.

DISCUSSION

LLMs have revolutionized the field of AI, particularly in general-purpose content generation. However, their application in specialized fields such as rice biology remains limited because of a lack of domain-specific training data. Here, we presented SeedLLM, a 7-billion-

parameter model developed from scratch using 1.4 million rice-related research publications, representing nearly 98.24% of the global literature on rice biology. To our knowledge, SeedLLM is the first LLM specifically designed for this domain. Its potential to advance both rice biology and broader plant biology research is substantial, addressing critical gaps in the current research landscape.

While general-purpose LLMs have garnered considerable interest, their application in specialized fields such as rice biology raises an important question: can these models effectively contribute to scientific discovery in this domain? An ideal model for rice biology must be capable of mastering a vast body of knowledge, enabling efficient information retrieval and fostering scientific breakthroughs. However, a comprehensive framework for evaluating LLMs in rice biology is lacking. To address this gap, we propose a novel evaluation framework featuring a robust question-answer dataset of rice biology, comprising 253 human-designed questions and 1,975 additional automatically generated questions. Our evaluations—both subjective human-led assessments and objective automated evaluations—demonstrate that general-purpose LLMs underperform relative to SeedLLM across multiple rice biology tasks. These findings underscore the necessity for domain-specific models to increase research efficiency and accuracy. The performance of SeedLLM is significantly enhanced by the integration of a knowledge graph, which consolidates transcriptional, proteomic data from over 1,800 research papers—data typically inaccessible to general-purpose models. This knowledge graph is a key factor driving the superior performance of the SeedLLM in rice biology tasks, highlighting the importance of specialized knowledge sources in optimizing model output.

Despite these advances, several limitations must be addressed. First, although SeedLLM is specifically designed to improve rice biology knowledge, it still suffers from hallucinations—incorrect or fabricated responses—particularly in specialized tasks. During our evaluation, distinguishing between hallucinations and information retrieval failures proved challenging, as both can lead to inaccurate outputs. This issue is exacerbated by the absence of a more comprehensive knowledge graph, which could serve as a structured repository of accurate data to guide model outputs. Expanding and refining this domain-specific knowledge graph in future iterations of SeedLLM could reduce hallucinations and enhance model reliability.

Second, while scaling laws suggest that larger models exhibit improved performance with increased training data and model complexity (Kaplan et al., 2020), the current version of

SeedLLM—comprising 7 billion parameters—faces limitations owing to available computational resources and human expertise. Increasing the model size—such as for 14B, 32B, or even 72B parameters—may lead to performance gains. However, it remains uncertain whether such increases significantly impact the discovery of novel phenomena in rice biology that the current configuration of SeedLLM has not yet captured. As such, we propose that future studies should prioritize improving data diversity and integrating cutting-edge rice biology datasets rather than solely focusing on scaling the model.

Third, the RiceQA dataset used for supervised fine-tuning, while instrumental to SeedLLM development, has limitations in terms of data quality and diversity. Although RiceQA is comprehensive, it could better represent underexplored areas in rice biology and more diverse experimental conditions. Expanding the dataset to cover a broader range of topics (Han et al., 2023) and ensuring higher accuracy in question design would likely improve the performance of SeedLLM, especially in addressing more complex or nuanced queries. Additionally, a higher-quality dataset could help mitigate issues such as hallucinations, making model outputs more reliable.

Rice biology research has traditionally been labor intensive, requiring scientists to manually process vast amounts of textual and biological data. SeedLLM represents a substantial advancement in research efficiency, enabling scientists to interact with the model via natural language queries and obtain critical information much faster than traditional search engines and databases. This is exemplified by our recent integration of the high-quality RiceData database (www.ricedata.cn) into SeedLLM's response pipeline (Supplemental Figure 7). This capability is expected to significantly accelerate the pace of discovery and analysis in rice biology.

We look forward to further enhancing the performance of SeedLLM. Insights gained from this study—particularly regarding the importance of knowledge graphs—will guide the development of an expanded, more sophisticated rice knowledge graph. This graph integrates recent advances in rice biology, including single-cell sequencing and spatial transcriptomics, and transforms these complex, high-dimensional datasets into a format that SeedLLM can process more effectively. Traditional academic literature often fails to capture the intricacies of such cutting-edge datasets, creating a challenge for researchers. By enabling SeedLLM to incorporate and analyze these advanced datasets, we aim to create a more comprehensive and accurate model of rice biology. Additionally, recent studies indicate that LLM agents, which utilize domain-

specific external tools to autonomously execute tasks, provide significant advantages over traditional chatbots (Kapoor et al., 2024). We propose integrating tools like RiceNavi (Wei et al., 2021) with SeedLLM to streamline breeding tasks such as target and parental line selection, thereby significantly enhancing the practical utility of SeedLLM. Although SeedLLM currently answers queries by retrieving rice epistatic QTL pairs (Wei et al., 2024) that serve as input data for RiceNavi pipeline, this integration would further elevate its capabilities (Supplemental Figure 8).

Our long-term vision for SeedLLM is to evolve into a global, comprehensive knowledge atlas that will provide researchers with unprecedented access to insights previously hidden owing to reliance on isolated data points. This vision aligns with the broader trend of AI-driven knowledge synthesis, where models such as SeedLLM will enable new discoveries by integrating diverse and complex data sources. Additionally, as we develop specialized LLMs for other crops, we anticipate that cross-species knowledge reasoning will become increasingly feasible, empowering researchers with more powerful tools for AI-assisted seed design in the future.

METHODS

Comprehensive search and retrieval of rice-related publications. The dataset used in this study comprises scientific publications in both English and Chinese. English-language publications were retrieved from the Web of Science using a search query that included various rice-related terms, resulting in 1,148,299 publications. Chinese-language publications were sourced from the China National Knowledge Infrastructure using a corresponding set of rice-related terms, yielding 232,445 publications. All publications used for the construction of RiceCorpus are up to December 31, 2024.

Construction of RiceCorpus. To generate a high-quality, rice-specific corpus, we developed a reusable, high-granularity data cleaning pipeline consisting of four primary stages. First, raw PDF documents of rice publication were converted to text using the MinerU tool (Wang et al., 2024a), with the Layout model accurately recognizing document sections (e.g., titles, abstracts) to ensure semantic coherence. Text extraction, formula recognition, and table conversion were performed, while PaddleOCR assisted in optical character recognition (Du et al., 2020). Post-

processing with regular expressions optimized the identification of rice-related terms, gene names, numbers, and punctuation. Second, heuristic cleaning was applied to address redundancy and irrelevant content in the resulting TXT files. Statistical analysis of rice literature informed the development of regular expression-based rules to filter low-quality text, retaining 72.82% high-quality documents, thereby enhancing the corpus' rice knowledge density. To further improve model training efficiency and reduce overfitting, we performed sentence-level deduplication using MinHash (Broder, 1997), which calculates n-gram similarity between text pairs. After replacing MinHash's tokenizer with SeedLLM's, deduplication was conducted at the sentence level, removing 1,834,317 sentences (25% of the data). Lastly, model-based filtering was used to eliminate non-rice content by applying the IndustryCorpus2_Classifier (Wang et al., 2024b), which classified text into 31 domain categories, retaining only agricultural, biological, and chemical content. The CCI3-HQ-Classifier provided quality scores (Wang et al., 2024b), filtering out segments with scores below 2. These combined methods resulted in a defined corpus termed as RiceCorpus consisting of 1.1 billion tokens, ensuring high relevance and quality for subsequent model training.

Model Pretraining. To enhance rice domain capabilities without compromising generalization, we pretrained base model Qwen-2.5-7B with RiceCorpus along with GeneralCorpus. The GeneralCorpus included downsampled Fineweb-Edu (English) (Penedo *et al.*, 2024), Fineweb-Edu-Chinese-V2.1 (Chinese) (Huang *et al.*, 2024), as well as code-mathematical corpus such as Opc-Fineweb-Math-Corpus and Opc-Fineweb-Code-Corpus. All model-related hyperparameters of pre-training matching those of Qwen-2.5-7B (Qwen *et al.*, 2024). The AdamW optimizer was used with β_1=0.9, β_2=0.999, weight decay=0.0, and a maximum context length of 4K tokens. The learning rate followed a linear increase from 8e-10 to 6e-07 for the first 10% of training steps, then decayed to 2e-14 following a cosine curve. Pretraining was conducted on 16 NVIDIA H100 GPUs with a batch size of 128K tokens. An ablation experiment using only the RiceQA demonstrated the model's ability to avoid catastrophic forgetting, enhancing both cross-domain generalization and rice-domain capabilities, as evidenced by a substantial improvement in the BBH metric compared to the baseline.

Model Post-training. GraphGen generates high-quality synthetic data for LLM fine-tuning by leveraging knowledge graphs. The process begins with knowledge construction, where text is segmented into semantically coherent chunks, and a synthesizer model M_{synth} extracts entities and relationships, which are merged to form a structured knowledge graph (Ibrahim et al., 2024). This enables effective long-text processing and reduces content hallucination. In the comprehension evaluation phase, M_{synth} generates paraphrased statements and negations to assess the model's understanding, with a comprehension loss computed based on the model's confidence scores. Graph organization follows, where subgraphs are extracted using methods like k-hop and selection strategies to balance complexity and relevance. Finally, question-answer pairs generation is performed for various scenarios, including atomic, aggregated, and multi-hop question-answer pairs, based on the subgraphs. For the SFT phase, RiceQA dataset is created by categorizing the Infinity Instruct dataset into six categories, selecting the top 300k question-answer pairs based on vector similarity, and augmenting the data with the synthetic pairs via AutoIF (Dong *et al.*, 2024). The final dataset of question-answer pairs integrates domain-specific knowledge with general capabilities while maintaining data quality through expert curation and overlap removal.

Training Configuration. During the training phase, we utilized XTuner as the training framework, based on the Transformer architecture and optimized using the AdamW optimizer. The learning rate followed a linear schedule with a warm-up phase, and gradient clipping was applied to stabilize training. The model training employed several key parameters to optimize performance. The maximum sequence length was set to 2048, defining the maximum input sequence size. A learning rate of 2e-5 was used to control the step size for weight updates, while a weight decay of 0.1 helped mitigate overfitting by penalizing large weights. Gradient clipping was applied with a threshold of 1 to stabilize training and prevent gradient explosion. A batch size of 64 (16×4) was selected, determining the number of samples processed per optimization step. The AdamW optimizer, a variant of Adam with decoupled weight decay, was used for improved generalization, with β1 and β2 values set at 0.9 and 0.999, respectively, for exponential decay of moment estimates. To gradually increase the learning rate, a warm-up ratio of 0.03 was applied for the initial fraction of the total training steps, and the model was trained for 2 epochs, completing two full passes over the dataset.

Automated evaluations of LLM performance. To assess model performance, we use accuracy for classification tasks and perplexity (PPL) for language modeling (Hu et al., 2024). The pre-trained model was evaluated on PPL-MCQ-ACC and Gen-QA-ACC datasets. The supervised fine-tuned model was evaluated on several general-purpose benchmarks CMMLU (Li et al., 2023), GSM8K (Cobbe et al., 2021), BBH (Srivastava et al., 2022), MMLU (Hendrycks et al., 2020) and Agri series dataset, which a domain-specific rice dataset from SeedBench (Ying et al., 2025), focusing on accuracy for multiple-choice tasks and PPL for fill-in-the-blank tasks.

Human-mediated evaluation of LLM performance. To evaluate LLM within the domain of rice biology, we adapted a structured human evaluation framework inspired by the methodology proposed by Petrov et al. (2025). With domain-specific modifications, we developed HumanDesignRiceQA, a curated benchmark comprising 253 expert-authored questions spanning six major topics: gene function, multi-omics, genome-wide association studies (GWAS), traditional breeding, molecular breeding, and gene editing. Each question was classified into one of three complexity tiers—basic, intermediate, and advanced—based on the depth of biological knowledge and reasoning required. Reference answers were derived from peer-reviewed literature, and evaluation rubrics were constructed by biological science experts with Master's-level training. Responses generated by LLMs, including SeedLLM, were assessed alongside those written by students through a blinded review process conducted by 326 human evaluators, of whom 83 were domain experts in rice biology. Each response was independently scored by three evaluators using a 0–100 scale and ranked relative to other answers based on predefined criteria encompassing factual accuracy, logical structure, and clarity. This evaluation framework constitutes the first domain-adapted, human-mediated assessment pipeline in plant science, establishing a rigorous benchmark for comparing LLM outputs in rice biology. The results demonstrate that SeedLLM consistently outperforms peer models across all levels of question complexity.

Construction of Rice Biological Knowledge Graph. We developed a Python pipeline to identify publications on rice transcriptomics and proteomics by searching for relevant keywords in titles and abstracts. After filtering potential papers, each was manually reviewed to confirm

focus on rice transcriptomics or proteomics, and whether raw or processed data were available. From these studies, we curated lists of the most upregulated and downregulated genes and proteins, standardizing them to rice AGIS IDs (Shang et al., 2023). Experimental metadata, including genotype, tissue or organ, growth stage, and treatments, were extracted and used to generate structured annotations (e.g., "Genotype_X under treatment_Y shows differential expression of gene AGIS_ID_1 in organ_Z at growth stage_W"). These annotations were converted into triples (subject, relation, object) to represent transcriptional and translational events, as well as experimental conditions. Each AGIS_ID was cross-referenced with databases like RAP-DB (Sakai et al., 2013) and Gramene (Jaiswal, 2011) to obtain functional annotations and subcellular localization information. These data were incorporated into the knowledge graph, linking them to the corresponding gene or protein nodes.

Visualization of Rice Biological Knowledge Graph. For basic visualization, we used the networkx library, assigning node and edge styles based on entity types (e.g., proteins in blue, growth stages in green). The spring layout algorithm optimized node positioning for clarity. For advanced visualizations, the graph was exported in GraphML format and imported into Gephi and Cytoscape. These tools enabled customization, such as adjusting node size by protein expression magnitude and edge thickness by interaction strength, allowing for a more detailed exploration of the data.

Construction of HumanDesignRiceQA. We developed a benchmark for rice-specific knowledge based on academic papers by creating three question levels—basic, intermediate, and advanced—covering five major research areas: gene function, transcriptomics, proteomics, traditional breeding, and molecular breeding. The levels are distinguished by the complexity of reasoning and knowledge integration, not by specific topics. Basic questions rely on readily accessible information, such as abstracts or general knowledge from search engines. Intermediate questions require a deeper understanding of rice biology, focusing on the paper's results with experimental details. Advanced questions demand specialized expertise in rice research, with the ability to analyze biological data within a broader biological context. The levels differ in analysis depth, knowledge integration, and reasoning complexity. Academic papers from each research

area were randomly selected, reviewed, and used to design the questions and corresponding answer sheets.

Graph-based Retrieval-Augmented Generation (GraphRAG). To enhance the performance of SeedLLM, we developed a framework that integrates external knowledge through a structured retrieval process to guide SeedLLM's response generation. This approach combines graph structures with dense indexing methods to represent relationships between knowledge fragments, facilitating the retrieval of relevant information for LLM-generated responses (Peng et al., 2024). The GraphRAG framework consists of three main components graph-guided indexing, retrieval, and text generation. In the graph-guided indexing phase, data is preprocessed into manageable chunks, followed by entity and relation extraction to form "entity, relation, description" tuples, which are organized into a knowledge graph. Persistence ensures that these relationships are embedded and stored in a database for efficient access. In the graph-guided retrieval phase, logic form method and dual-level method is conducted for rice-related queries and for general queries, respectively. The logic form method decomposes the query into operators and parameters, generating sub-queries whose results are merged to form a retrieval context. Pre- and post-checks verify whether the context sufficiently supports the LLM's response. If the logic form method is insufficient, the dual-level method decomposes the query into high-level semantic representations and low-level entity-based components, using fuzzy matching to identify relevant nodes and relationships. The results are merged to create a comprehensive retrieval context. Finally, in graph-guided text generation, the retrieved context and the original query are input into the LLM, which generates the final output based on the enriched context. This integrated approach enables LLMs to leverage structured external knowledge, leading to more accurate and contextually relevant responses.

Code Availability. SeedLLM will continue to grow and improve through version control. Currently, SeedLLM·Rice, version 0.6a is available via an interactive web portal https://seedllm.org.cn/.

SUPPLEMENTAL INFORMATION. Supplemental Information is available at Molecular Plant Online.

FUNDING. SeedLLM research and development was supported by Yazhouwan National Laboratory Project (grant no. 2310CF01) and Shanghai Artificial Intelligence Laboratory. Corpus preparation was supported by the Hainan Yazhou Bay Seed Laboratory Project (grant no. B21HJ0001). Human evaluation of LLM performance was supported by Biological Breeding-National Science and Technology Major Project (2023ZD04076).

AUTHOR CONTRIBUTIONS. F.Y. and N.Q.D. conceptualized the idea, designed the experiments, wrote the manuscript, supervised the project, and acquired fundings. F.Y., N.Q.D., and H.J.K designed the methodology. H.J.K, J.Y, Z.H.C,W.L.J, Z.H.Y, and Z.F.W trained the SeedLLM·Rice and SeedLLM·Rice-KG models, conducted automated model evaluations, and developed the web portal for user access to SeedLLM·Rice-KG. T.L., Z.N.M., S.K.W., X.Y.W., W.F.M., and X.Y.L. prepared the RiceCorpus and HumanDesignRiceQA datasets and performed the data analysis in human evaluations. X.D.M., M.G.L., and X.Q.W. organized human-centric model evaluations. Z.Y.H. conducted comprehensive searching for rice-related publications.

ACKNOWLEDGEMENTS. We thank members of the Innovation Platform for Seed Design Team and Hongqing Ling in YNL fruitful discussions. We thank Xinquan Yang for inviting human evaluators. We thank YZBSTCACC for support with GPU resources and Pengcheng Cloud Brain II for data storage support. We thank SiliconFlow for support with GPU resources for SeedLLM deployment on the cloud. We thank Gaojun Fan in Pengcheng Laboratory for OpenI Community support. No conflict of interest is declared.

REFERENCES

Bornmann, L., and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66:2215-2222. https://doi.org/10.1002/asi.23329.

Broder, A.Z. (1997). On the resemblance and containment of documents. Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training Verifiers to Solve Math Word Problems.

Consortium, T.U. (2009). The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research 38:D142-D148. 10.1093/nar/gkp846.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., et al. (2024). DeepSeek-V3 Technical Report.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Dong, G., Lu, K., Li, C., Xia, T., Yu, B., Zhou, C., and Zhou, J. (2024). Self-play with Execution Feedback: Improving Instruction-following Capabilities of Large Language Models.

Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., et al. (2020). PP-OCR: A Practical Ultra Lightweight OCR System.

Fang, C., Wang, Y., Song, Y., Long, Q., Lu, W., Chen, L., Wang, P., Feng, G., Zhou, Y., and Li, X. (2024). How do Large Language Models understand Genes and Cells. bioRxiv:2024.2003.2023.586383. 10.1101/2024.03.23.586383.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The Llama 3 Herd of Models.

Han, L., Zhong, W., Qian, J., Jin, M., Tian, P., Zhu, W., Zhang, H., Sun, Y., Feng, J.-W., Liu, X., et al. (2023). A multi-omics integrative network map of maize. Nature Genetics 55:144-153. 10.1038/s41588-022-01262-1.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding.

Hu, Y., Huang, Q., Tao, M., Zhang, C., and Feng, Y. (2024). Can Perplexity Reflect Large Language Model's Ability in Long Text Understanding?

Huang, F., Jiang, Y., Chen, T., Li, H., Fu, M., Wang, Y., Xu, Y., Li, Y., Zhou, Z., Jia, L., et al. (2022a). New Data and New Features of the FunRiceGenes (Functionally Characterized Rice Genes) Database: 2021 Update. Rice 15:23. 10.1186/s12284-022-00569-1.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.

Huang, S., Cheng, T., Liu, J.K., Hao, J., Song, L., Xu, Y., Yang, J., Liu, J.H., Zhang, C., Chai, L., et al. (2024). OpenCoder: The Open Cookbook for Top-Tier Code Large Language Models.

Huang, X., Huang, S., Han, B., and Li, J. (2022b). The integrated genomics of crop domestication and breeding. Cell 185:2828-2839. 10.1016/j.cell.2022.04.036.

Ibrahim, N., Aboulela, S., Ibrahim, A., and Kashef, R. (2024). A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. Discover Artificial Intelligence 4:76. 10.1007/s44163-024-00175-8.

Jaiswal, P. (2011). Gramene database: a hub for comparative plant genomics. Methods Mol Biol 678:247-275. 10.1007/978-1-60761-682-5_18.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models.

Kapoor, S., Stroebl, B., Siegel, Z.S., Nadgir, N., and Narayanan, A. (2024). AI Agents That Matter.

Lam, H.Y.I., Ong, X.E., and Mutwil, M. (2024). Large language models in plant biology. Trends in Plant Science 29:1145-1155. https://doi.org/10.1016/j.tplants.2024.04.013.

Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., and Mehdad, Y. (2020). MTOP: A Comprehensive Multilingual Task-Oriented Semantic Parsing Benchmark.

Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., and Baldwin, T. (2023). CMMLU: Measuring massive multitask language understanding in Chinese.

Li, J.-Y., Wang, J., and Zeigler, R.S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience 310.1186/2047-217x-3-8.

Majumdar, J., Naraseeyappa, S., and Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data 4:20. 10.1186/s40537-017-0077-4.

637 Naveed, H., Ullah Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N.,
638 and Mian, A. (2023). A Comprehensive Overview of Large Language Models.
639 Nazi, Z.A., and Peng, W. (2024). Large Language Models in Healthcare and Medical Domain: A
640 Review. Informatics 11:57.
641 OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Leoni Aleman, F., Almeida,
642 D., Altenschmidt, J., Altman, S., et al. (2023). GPT-4 Technical Report.
643 Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2023). Unifying Large Language
644 Models and Knowledge Graphs: A Roadmap.
645 Penedo, G., Kydlíček, H., Ben allal, L., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., and
646 Wolf, T. (2024). The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale.
647 Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., and Tang, S. (2024). Graph
648 Retrieval-Augmented Generation: A Survey.
649 Project, R.A. (2007). The Rice Annotation Project Database (RAP-DB): 2008 update *. Nucleic
650 Acids Research 36:D1028-D1033. 10.1093/nar/gkm978.
651 Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., et al.
652 (2024). Qwen2.5 Technical Report.
653 Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C.,
654 Iwamoto, M., Abe, T., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative
655 and interactive database for rice genomics. Plant Cell Physiol 54:e6. 10.1093/pcp/pcs183.
656 Shang, L., He, W., Wang, T., Yang, Y., Xu, Q., Zhao, X., Yang, L., Zhang, H., Li, X., Lv, Y., et al.
657 (2023). A complete assembly of the rice Nipponbare reference genome. Molecular Plant
658 16:1232-1236. 10.1016/j.molp.2023.08.003.
659 Shi, J., An, G., Weber, A.P.M., and Zhang, D. (2023). Prospects for rice in 2050. Plant, Cell &
660 Environment 46:1037-1045. https://doi.org/10.1111/pce.14565.
661 Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-
662 Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature
663 620:172-180. 10.1038/s41586-023-06291-2.
664 Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro,
665 A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the Imitation Game: Quantifying and
666 extrapolating the capabilities of language models.

667 Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et
668 al. (2024a). MinerU: An Open-Source Solution for Precise Document Content Extraction.
669 Wang, L., Zhang, B.-W., Wu, C., Zhao, H., Shi, X., Gu, S., Li, J., Ma, Q., Pan, T., and Liu, G.
670 (2024b). CCI3.0-HQ: a large-scale Chinese dataset of high quality designed for pre-training large
671 language models.
672 Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M.,
673 Grouios, C., Kazi, F., Lopes, C.T., et al. (2010). The GeneMANIA prediction server: biological
674 network integration for gene prioritization and predicting gene function. Nucleic Acids Research
675 38:W214-W220. 10.1093/nar/gkq537.
676 Wei, X., Qiu, J., Yong, K., Fan, J., Zhang, Q., Hua, H., Liu, J., Wang, Q., Olsen, K.M., Han, B.,
677 et al. (2021). A quantitative genomics map of rice provides genetic insights and guides breeding.
678 Nat Genet 53:243-253. 10.1038/s41588-020-00769-9.
679 Wei, X., Chen, M., Zhang, Q., Gong, J., Liu, J., Yong, K., Wang, Q., Fan, J., Chen, S., Hua, H., et
680 al. (2024). Genomic investigation of 18,421 lines reveals the genetic architecture of rice. Science
681 385:eadm8762. doi:10.1126/science.adm8762.
682 Ying, J., Chen, Z., Wang, Z., Jiang, W., Wang, C., Yuan, Z., Su, H., Kong, H., Yang, F., and
683 Dong, N. (2025). SeedBench: A Multi-task Benchmark for Evaluating Large Language Models
684 in Seed Science.
685

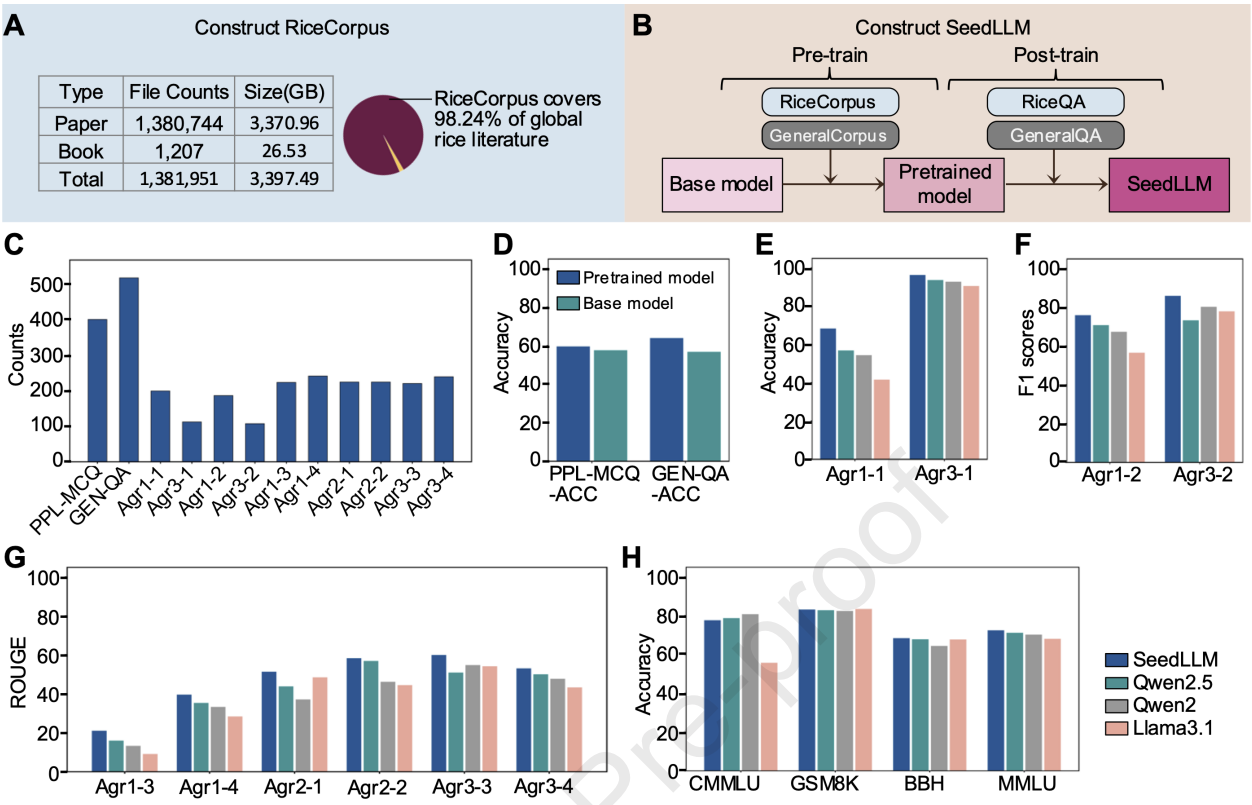
Figure 1. Overview of SeedLLM development and automated evaluation. (A) The RiceCorpus is a comprehensive dataset of scientific publications and books related to rice, encompassing 1.38 million academic papers in both Chinese and English. **(B)** The RiceCorpus was used to train the base model, Qwen2.5, which was fine-tuned with the riceQA dataset to specialize SeedLLM for rice-specific questions. GeneralCorpus, consisting of datasets like Curated FineWeb-Edu and Curated OpenCoder, was used for general language model training. RiceQA, designed for rice biology, includes Key-info QA (extracting key data from RiceCorpus), Bad-case QA (addressing difficult scenarios), and Graph QA (using a knowledge graph for relevant questions). GeneralQA, which includes Curated Infinity Instruct and Curated AutoIF, further fine-tunes the model for instruction-following and automated inference tasks. **(C)** The pre-trained model was evaluated on two tasks: Gen-QA-ACC for open-ended questions and PPL-MCQ-ACC for multiple-choice questions. SeedLLM's performance was assessed using rice-specific datasets from the Agr series, which included single-choice, multiple-choice, and fill-in-the-blank questions. The number of questions across these datasets was also recorded. **(D)** The pre-trained model outperforms the baseline on both Gen-QA-ACC and PPL-MCQ-ACC, confirming the effectiveness of the pretraining process. **(E-G)** SeedLLM demonstrates its superiority over other LLMs in rice-specific tasks in terms of accuracy, F1 score, and ROUGE. Accuracy measures the proportion of correct predictions, F1 score balances precision and recall, and ROUGE evaluates the overlap between model-generated outputs and reference texts. **(H)** SeedLLM exhibited robust performance on general-purpose tasks across multiple datasets, including CMMLU (Chinese multitask understanding), GSM8K (grade school math), BBH (beyond current model capabilities), and MMLU (general multitask language understanding). All evaluations in (D-H) utilized an automated pipeline to extract model responses and compare them to a reference key for correctness.

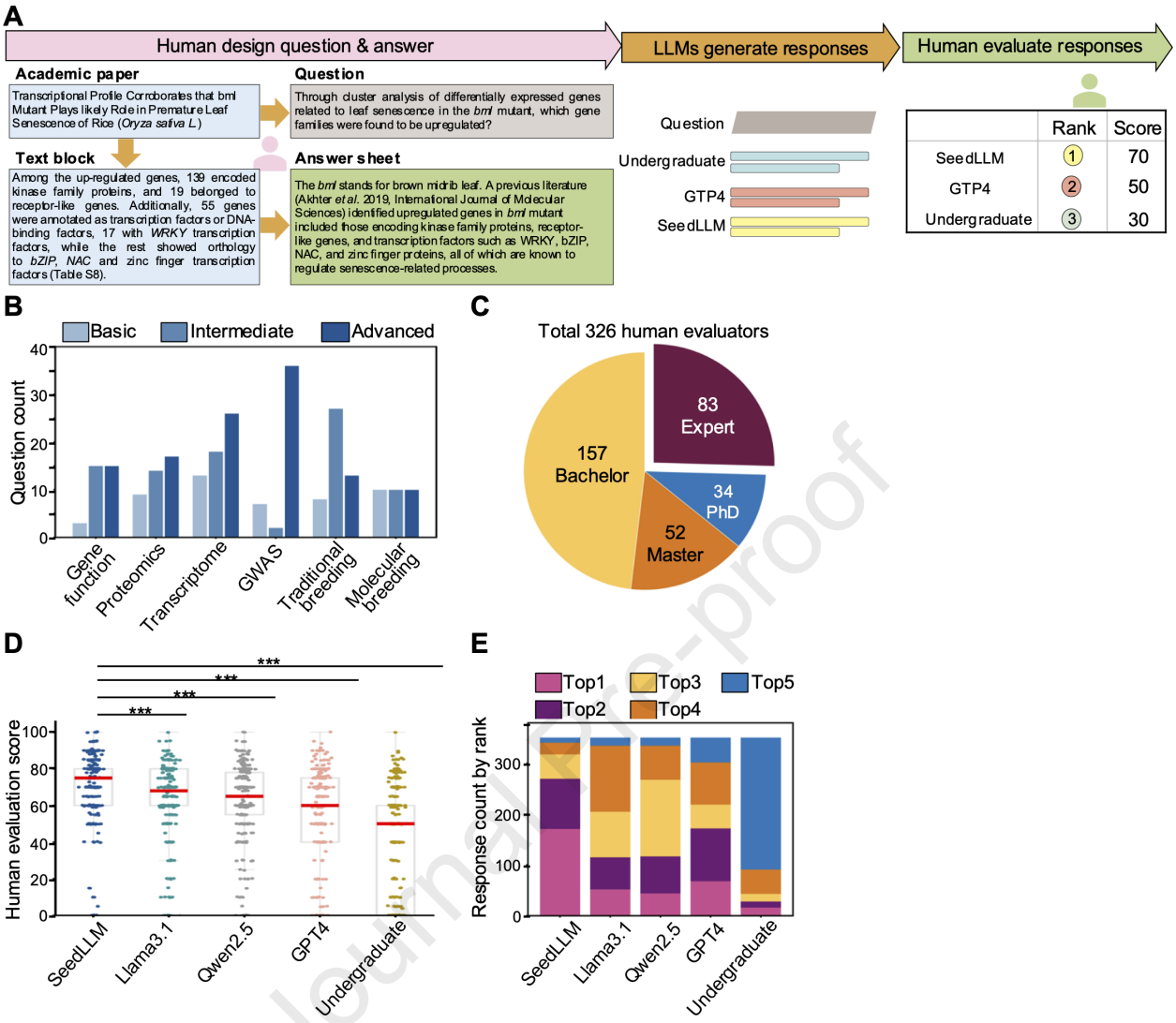
Figure 2. Human-centric evaluation confirms SeedLLM's superior performance compared to general-purpose LLMs in HumanDesignRiceQA. (A) Overview of human-centric evaluation. HumanDesignRiceQA, a dataset comprising question-answering pairs with plant biology expertise derived from academic publications, was constructed. SeedLLM, other general-purpose LLMs, and undergraduate responses were tasked with answering questions from this dataset. A panel of evaluators, experts in rice biology, ranked the quality of responses for each question from best to worst. Evaluators also assigned grades based either on the provided answer key or their own expertise. **(B)** Distribution of questions in HumanDesignRiceQA is categorized into three difficulty levels basic, intermediate, and advanced. **(C)** Distribution of evaluator educational backgrounds. All evaluators possess academic training in agronomy, with degrees ranging from Bachelor's to PhD. Experts are defined as individuals who have published research, filed patents, or contributed to rice variety development in the past five years. **(D)** Human evaluation scores evaluated in the HumanDesignRiceQA dataset. **(E)** Human-assigned rankings from top1 to top5 for responses generated by SeedLLM and other LLMs.

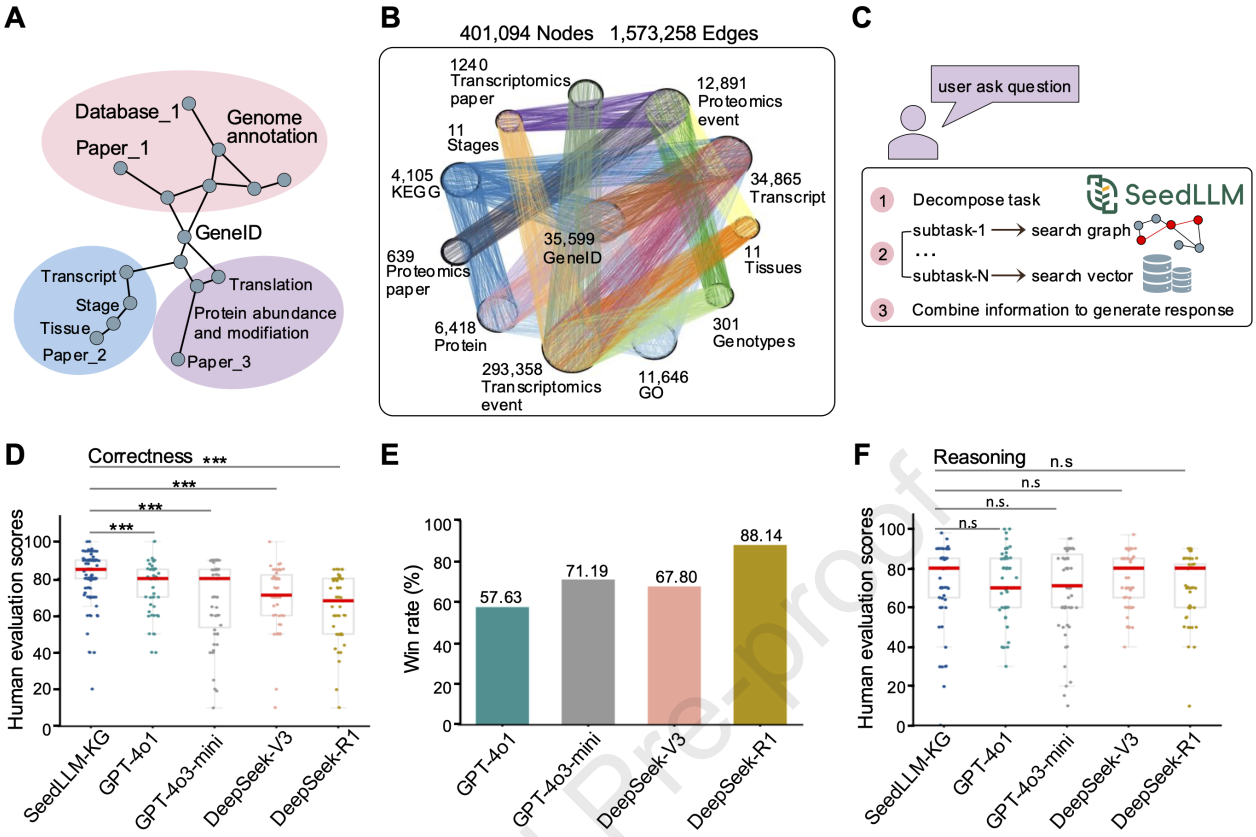
Figure 3. Encoding multiomics data into the rice biological knowledge graph for SeedLLM-based agricultural query responses. (A) Schematic overview for encoding transcriptomic, proteomic, and genome annotation data as graph structures. Transcriptional and proteomic events for each gene are represented as sentences and converted into triples. These triples are transformed into nodes (representing rice gene IDs, transcriptional and translational events, and experimental metadata) and edges (depicting relationships). Gene ID nodes are linked to genome annotation attributes such GO terms and KEGG. **(B)** Overview of the Rice Biological Knowledge Graph. Node types are color-coded to represent different data categories. The graph includes 35,599 unique rice AGIS IDs identified from rice transcriptomic and proteomic studies, shown with their relationships as edges. **(C)** SeedLLM-KG working mechanism. SeedLLM processes agricultural queries by decomposing them into sub-queries, retrieving answers from either the LLM's database or the knowledge graph. These sub-answers are integrated to form a comprehensive response. Non-agricultural queries are handled directly by SeedLLM. **(D)** Correctness of SeedLLM-KG and other LLMs were assessed by human evaluators using advanced-level questions of the HumanDesignRiceQA dataset. Asterisks indicate statistical significance between comparisons (t-test, $P < 0.001$). **(E)** Win rate of SeedLLM-KG against various LLMs. SeedLLM-KG is considered to win if it receives a higher human evaluation scores than the other model on the same questions from HumanDesignRiceQA dataset. The win rate represents the percentage of questions where SeedLLM-KG outperforms each model, as labeled in the plot. **(F)** Reasoning ability of SeedLLM-KG and other LLMs were assessed by human evaluators using advanced-level questions of the HumanDesignRiceQA dataset. n.s. indicates no significant difference between comparisons (t-test, $P < 0.05$).

752

753 **Figure 4. Comparison of response quality between SeedLLM and other LLMs. (A)** Models
754 were tasked with generating responses to the query: "Does the rice gene AGIS_Os06g035130
755 respond to various environmental conditions?" The models tested include SeedLLM-KG,
756 DeepSeek-R1, and OpenAI GPT-4o1. Response quality was assessed by verifying content
757 against RiceCorpus literature. Correct content is highlighted in blue, model reasoning in green,
758 and inability to answer in red. **(B)** A list of references identified in the literature search that
759 corroborate the responses. Note that none of the LLMs generated references post-response.
760







A

Does the rice gene AGIS_Os06g035130 respond to various environmental conditions?

Correct Answer
Reasoning
Unable to Answer

SeedLLM

Yes, the rice gene AGIS_Os06g035130 does respond to various environmental conditions. It shows changes in transcriptional levels in response to different environmental stimuli. This gene's expression is measured under a wide range of experimental conditions, including drought stress, anaerobic germination, early chilling stress, and osmotic stress. Specifically, AGIS_Os06g035130's transcript expression is upregulated during the seedling stage under certain conditions, such as osmotic stress, and in the panicle during specific developmental stages under drought stress. Additionally, its expression is studied in different rice varieties and growth stages, highlighting its dynamic regulatory mechanisms in response to various stress conditions.

deepseek

Okay, so I need to figure out if the rice gene AGIS_Os06g035130 responds to various environmental conditions and whether its transcriptional or translational levels change under different environmental stimuli. Let me start by breaking down the problem.
First, I need to confirm the correct identifier for the gene. The user provided AGIS_Os06g035130. Wait, sometimes gene IDs can have different formats depending on the database. For example, Rice Genome Annotation Project (RGAP) uses LOC_Os06g035130, where "LOC" is a common prefix. The user's ID starts with AGIS, which might be a typo or from a different database. Let me check that first because if the ID is incorrect, any further steps could be wrong.
Assuming AGIS_Os06g035130 is correct, but maybe it's supposed to be LOC_Os06g035130. Maybe "AGIS" is part of another naming convention? Alternatively, perhaps it's from the AgBase database? Hmm. Alternatively, maybe it's a typo and should be RAP-DB's ID, which might be Os06g0135200 or similar. Wait, RGAP and RAP-DB are two major databases for rice genes.
Let me cross-reference this. The format Os06g035130 would correspond to chromosome 6, locus 035130. But RGAP uses LOC_Os06g035130. For example, LOC_Os06g035130 is a known gene encoding a protein kinase. If AGIS is part of another system, maybe I need to check that.
Alternatively, maybe the user intended to write "LOC" but wrote "AGIS". Let me proceed with LOC_Os06g035130 as a possible correct ID and see what information I can find.
Now, moving on to the main question: does this gene respond to environmental conditions? Changes in transcription or translation levels?
I know that plant genes often respond to environmental stimuli such as drought, salinity, cold, heat, pathogen attack, etc. Transcription factors or kinases might be involved in stress responses.
Looking up LOC_Os06g035130: According to RGAP (Rice Genome Annotation Project), this gene is annotated as a receptor-like protein kinase. Protein kinases are often involved in signal transduction pathways in response to environmental stresses.
To check expression data under different conditions: There are databases like RiceXPro (Rice Expression Profile Database), which provide expression profiles under various conditions. Also, CREP (Database of Composit).

OpenAI

The rice gene AGIS_Os06g035130, identified in the Gene Expression Omnibus (GEO) dataset GSE73609, is involved in the plant's response to various environmental conditions. This dataset examined global gene expression patterns in rice under different cultivation methods and climatic conditions, highlighting how environmental factors influence gene expression.
Additionally, research on Environmental Gene Regulatory Influence Networks (EGRINs) in rice has shown that genes like AGIS_Os06g035130 are part of complex networks that coordinate responses to environmental stimuli such as water deficit and high temperature. These networks involve interactions between transcription factors and target genes, modulating gene expression in response to environmental changes.
Furthermore, studies on stress-induced gene expression in rice have demonstrated that environmental stresses can lead to changes in DNA methylation patterns near stress-responsive genes. This suggests that genes like AGIS_Os06g035130 may undergo transcriptional and epigenetic modifications in response to environmental stimuli.
In summary, AGIS_Os06g035130 is responsive to various environmental conditions, with its expression levels and epigenetic modifications adjusting to environmental stimuli.

B

1.Kaur, S., Seem, K., Duhan, N., Kumar, S., Kaundal, R., & Mohapatra, T. (2023). Transcriptome and Physio-Biochemical Profiling Reveals Differential Responses of Rice Cultivars at Reproductive-Stage Drought Stress. *International Journal of Molecular Sciences, 24*(2), 1002. https://doi.org/10.3390/ijms24021002
2.Baldoni, E., Bagnaresi, P., Locatelli, F., Mattana, M., & Genga, A. (2016). Comparative leaf and root transcriptomic analysis of two rice japonica cultivars reveals major differences in the root early response to osmotic stress. *Rice, 9*(1), 23. https://doi.org/10.1186/s12284-016-0098-1
3. Han, B., Ma, X., Cui, D., Wang, Y., Geng, L., Cao, G., Zhang, H., & Han, L. (2019). Comprehensive evaluation and analysis of the mechanism of cold tolerance based on the transcriptome