ISPRS Journal of Photogrammetry and Remote Sensing 225 (2025) 492–513

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Multimodal large language model for wheat breeding: A new exploration of smart breeding

Guofeng Yang [a], Yu Li [b], Yong He [a], Zhenjiang Zhou [a], Lingzhen Ye [c], Hui Fang [a], Yiqi Luo [d], Xuping Feng [a,*]

[a] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058 Zhejiang, China
[b] Zhejiang Society of Agricultural Machinery, Hangzhou 310003 Zhejiang, China
[c] College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058 Zhejiang, China
[d] Soil and Crop Sciences Section, School of Integrative Plant Science, Cornell University, Ithaca 14853 NY, USA

ARTICLE INFO

Keywords:
Data fusion
Remote sensing
Multimodal large language model
Cross-domain knowledge
Smart breeding

ABSTRACT

Unmanned aerial vehicle remote sensing technology has become a key technology in crop breeding, which can achieve high-throughput and non-destructive collection of crop phenotyping data. However, the multidisciplinary nature of breeding has brought technical barriers and efficiency challenges to knowledge mining. Therefore, it is important to develop a smart breeding goal tool to mine cross-domain multimodal data. Based on different pre-trained open-source multimodal large language models (MLLMs) (e.g., Qwen-VL, InternVL, Deepseek-VL), this study used supervised fine-tuning (SFT), retrieval-augmented generation (RAG), and reinforcement learning from human feedback (RLHF) technologies to inject cross-domain knowledge into MLLMs, thereby constructing multiple multimodal large language models for wheat breeding (WBLMs). The above WBLMs were evaluated using the newly created evaluation benchmark in this study. The results showed that the WBLM constructed using SFT, RAG and RLHF technologies and InternVL2-8B has leading performance. Then, subsequent experiments were conducted using the WBLM. Ablation experiments indicated that the combination of SFT, RAG, and RLHF technologies can improve the overall generation performance, enhance the generated quality, balance the timeliness and adaptability of the generated answer, and reduce hallucinations and biases. The WBLM performed best in wheat yield prediction using cross-domain data (remote sensing, phenotyping, weather, and germplasm) simultaneously, with R^2 and RMSE of 0.821 and 489.254 kg/ha, respectively. Furthermore, the WBLM can generate professional decision support answers for phenotyping estimation, environmental stress assessment, target germplasm screening, cultivation technique recommendation, and seed price query tasks. This study aims to improve the application of remote sensing in crop breeding by enabling precise assessment and prediction of wheat germplasm breeding materials in alignment with breeding goals, thereby accelerating the selection of superior varieties and better supporting the breeding decisions.

1. Introduction

As one of the basic food crops for mankind, wheat breeding is facing unprecedented challenges and opportunities due to the global food crisis and the promotion of sustainable agricultural development (Cheng et al., 2024; Senapati et al., 2022). Although traditional breeding methods have achieved remarkable results, their efficiency and accuracy have gradually revealed their limitations in the face of complex and changing climatic conditions, increasingly serious threats from pests and diseases, and constantly upgrading consumer demands (Hu et al., 2022; Xiong et al., 2021). With the rapid development of unmanned aerial vehicle (UAV) technology, its application in the agricultural field has become more and more extensive, especially UAVs have shown great potential in intelligent breeding (Das et al., 2021; Fei et al., 2023; Jiang et al., 2021). With its advantages of high efficiency, precision and flexibility, it is becoming an indispensable tool in smart breeding and can provide important monitoring data for smart breeding. Remote sensing-based vegetation indexes (VIs) are widely used in crop breeding for phenotypic identification, genotype screening, and breeding efficiency enhancement (Sun et al., 2022; Tao et al., 2022). However, wheat

* Corresponding author.
E-mail address: fengxp@zju.edu.cn (X. Feng).

https://doi.org/10.1016/j.isprsjprs.2025.03.027
Received 25 September 2024; Received in revised form 14 February 2025; Accepted 29 March 2025
Available online 15 May 2025

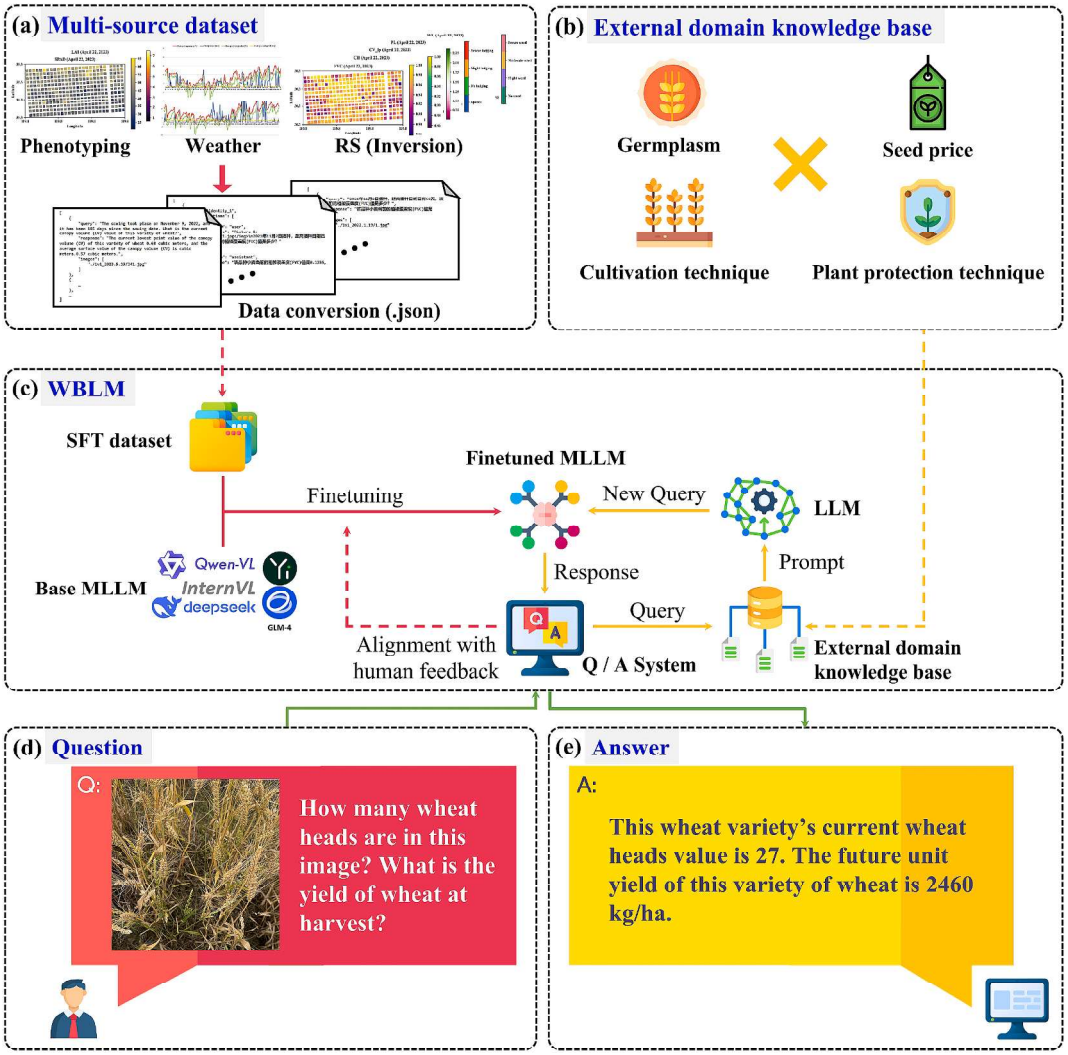

Fig. 1. A workflow for the construction and application of a multimodal large language model for wheat breeding (WBLM). (a) Multi-source dataset construction. (b) External domain knowledge base construction. (c) The WBLM with domain knowledge is constructed using supervised fine-tuning, retrieval augmented generation, and reinforcement learning from human feedback. (d) The question (image and text) of the user is sent to WBLM. (e) The WBLM answers the question.

breeding involves the intersection of multiple disciplines such as biology, genetics, weather, and soil science, professionals have to cross literature and data from many fields when engaged in breeding work and even need to write code to access data, which greatly limits their work efficiency (Kaur et al., 2021).

Smart breeding, as an innovative mode that integrates modern information technology, biotechnology, and agricultural science, is gradually becoming a key path to solving this problem (Li et al., 2024c; Xu et al., 2022). Image recognition technology can accurately extract wheat growth indicators from remote sensing (RS) images, such as chlorophyll content (Feng et al., 2023; Zhang et al., 2024) and leaf area index (Chen et al., 2022; Du et al., 2023); text analysis technology can extract valuable germplasm descriptions (Liu et al., 2024a; Sansaloni et al., 2020), breeding optimization strategies (Xu et al., 2023a; Yao et al., 2022) and market trends (Garg et al., 2022; Padhy et al., 2024) from research manuscripts, reports and databases. The complementarity of image and text information not only enhances the credibility and interpretability of the data, but also promotes the integration of interdisciplinary knowledge, opening up new avenues for wheat breeding research. Due to the diversity of wheat varieties, breeding information has long lacked a unified tool, and data knowledge has shown an "isolated" distribution, which has created barriers to the learning of wheat breeding knowledge (Bhat et al., 2023). At present, there is an urgent need to integrate cross-domain data (RS data, phenotyping data, environment data, germplasm data, cultivation data, and price data), use artificial intelligence algorithms and big data technology to build a cross-scale and cross-domain comprehensive analysis tool for wheat breeding, accelerate the screening and optimization of wheat target varieties, and improve the efficiency and accuracy of wheat breeding.

In recent years, the large language model (LLM) has attracted much attention due to their powerful text generation and understanding capabilities (Huang et al., 2024b; Lin et al., 2024). After fine-tuning, the LLM has shown strong interactive capabilities and the potential to improve productivity (Min et al., 2024). However, since the LLM can only process plain text and cannot process images, voice, and video, their application scope is limited (Wang et al., 2023). In the context of cross-domain, multi-modal data fusion applications, several multimodal large language models (MLLM) have been developed to enhance the ability of LLM to perceive and understand visual signals (01.AI et al., 2024; Bai et al., 2023; Chen et al., 2024b; DeepSeek-AI et al., 2024; GLM et al., 2024; Li et al., 2024b). Although a lot of study has been done to explore the limitations and effectiveness of MLLM, the current open-source general MLLM still has the problem of insufficient accuracy in professional applications (Huang et al., 2024a; Li et al., 2024a), such as wheat breeding selection. When faced with wheat breeding query, general MLLM often evades the question or gives irrelevant answers due

to lack of breeding knowledge. This hinders the further exploration and application of MLLM in the field of wheat breeding.

However, the research and application of MLLM in crop breeding is still in its early stages and faces many challenges and opportunities (Zhu et al., 2024). The MLLM needs to integrate more complex data from different sources, including but not limited to RS data, phenotyping data, and environment data during the crop growth period. Data from these sources vary significantly in format, dimensionality, and sparsity. How to efficiently and accurately integrate these cross-domain heterogeneous data is a difficulty in current research (Kuska et al., 2024). MLLM needs to have strong learning and generalization capabilities to adapt to the needs of smart breeding and predict the performance of crops under different environmental conditions, which places extremely high demands on algorithm design and computing resources of the model. Additionally, professional knowledge and experience in the field of crop breeding are crucial for the development and application of the model (Li et al., 2024d). How to integrate professional knowledge into model design to make the model more in line with actual breeding needs, while discovering new breeding strategies and optimization solutions through the model, is another issue that needs in-depth exploration. MLLM is expected to reveal the deep mechanisms of crop trait formation by integrating cross-modal information and accelerate the screening and breeding process of new varieties.

This study aims to innovatively develop an MLLM for wheat breeding (WBLM) centered on remote sensing through cross-domain data fusion and artificial intelligence technologies, and explore its potential in wheat breeding goals. The purposes of this study are to: (i) evaluate the contribution of the integrated application of domain knowledge technologies to achieving wheat breeding goals, and analyze the performance of cross-domain data fusion in wheat yield prediction; (ii) explore the response of WBLM in coping with multidimensional breeding goals, and generate personalized decision support from the aspects of phenotyping estimation, environmental stress assessment, target germplasm screening, cultivation technique (CT) recommendation, seed price (SP) query; (iii) release the study dataset to promote research and application innovation in this field.

2. Methodology

We construct a WBLM with domain knowledge using the supervised fine-tuning (SFT), retrieval augmented generation (RAG) and reinforcement learning from human feedback (RLHF) technologies based on different MLLMs and cross-domain knowledge base. WBLM consists of two parts. The part 1 builds a RAG component based on the LlamaIndex large model application framework (https://www.llamaindex.ai/) and the Milvus vector database (https://milvus.io/) (Fig. 1c, yellow line). The part 2 uses STF and RLHF methods to obtain a WBLM that follows breeding preferences (Fig. 1c, red line). We regard the input of part 1 as a query (Fig. 1d), and the original query and the output of part 1 as a new query. The part 2 receives the new query, processes it, and then outputs the result to the Q/A system (Fig. 1e).

To make WBLM better adapt to breeding tasks, we use RAG technology to conveniently and efficiently supplement the knowledge not covered by STF and RLHF methods. First, we convert the data of the files in the external domain knowledge base, use BGE-M3 to generate embedding vectors (Chen et al., 2024a), and then import the data into the Milvus vector database. Then, LlamaIndex receives user input and initiates a retrieval request to the Milvus database based on the input, using a hybrid retrieval (BGE-M3) and re-ranking (BGE-Reranker-V2-M3) pipeline. LlamaIndex merges the retrieval results with the input to form a new prompt. After that, LlamaIndex inputs the new prompt into the LLM (InternLM2.5-7B-Chat) to perform reasoning and answer generation based on the retrieved knowledge (Cai et al., 2024). InternLM2.5-7B-Chat is employed as the LLM backbone in both the RAG component and subsequent experiments.

The training process using the STF and RLHF methods is divided into three stages. In the first stage, we collect demonstration data and train a supervised policy. Based on the cross-domain knowledge base, we artificially construct a question-answering dataset that we hope the model generates aligned answers to SFT the pre-trained MLLM. For questions we want the model to answer well, we collect the answers we want the model to output to increase the probability of the model generating the expected answers. We choose currently popular and competitive open-source MLLMs for model construction and fine-tuning, including Qwen-VL (Qwen-VL-Chat), InternVL (InternVL2-8B and InternVL2-2B), Yi-VL (Yi-VL-6B), Deepseek-VL (DeepSeek-VL-7B-chat and DeepSeek-VL-1.3B-chat) and GLM-4 (GLM-4V-9B), all of which support English and Chinese. It cannot be denied that we mainly provide a method to build an MLLM suitable for wheat breeding, rather than having to choose a specific open-source MLLM, as models with better performance are constantly emerging. We used Scalable lightWeight Infrastructure for Fine-Tuning (SWIFT) to fine-tune the MLLM (The ModelScope Team, 2024). We refer to the SWIFT recommended training script, in which the LoRA (Low-Rank Adaptation) fine-tuning method is selected, the number of epochs is set to 5, and no further hyperparameter adjustment is performed. All fine-tuning and inference were performed on four V100 NVIDIA GPUs with 32 GB. Specifically, dataset D consists of the model input prompt and the answer that the breeder expects the model to output, denoted by x and y respectively. The model is denoted by π_θ, and the length of y is T. When prompt x is given, the probability that the model generates answer y can be expressed as formula (1). Where $\pi_\theta(y_t|x,y_{1:t-1})$ denotes the probability of the model outputting the t-th token given the input before the t-th token.

$$\pi_\theta(y|x) = \left[\prod_{t=1}^{T} \pi_\theta(y_t|x,\ y_{1:t-1})\right] \tag{1}$$

The SFT stage uses the following loss to train the model. For simplicity, we call the model trained in this stage the SFT model.

$$L_{\mathrm{SFT}}\left(\pi_\theta\right) = -E_{(x,y)\sim D}[\log\pi_\theta(y|x)\,] = -E_{(x,y)\sim D}\left[\sum_{t=1}^{T} \log\pi_\theta(y_t|x,y_{1:t-1})\right] \tag{2}$$

The second stage collects comparative data consisting of questions and different answers. Labels indicate their preferred answer given the input. A reward model (RM) is then trained to predict the preferred answer of the breeder. Specifically, the model takes prompt and answer as input and outputs a scalar value. We use r_ϕ to denote the RM, and $r_\phi(x,y)$ denotes the scalar output of the RM given prompt x and answer y. The dataset D used to train the RM consists of the model input prompt, the answer that the breeder wants the model to output, and the answer that the breeder does not want the model to output, denotes by x , y_w, and y_t respectively. We are given data (x, y_w, y_t), and use the maximum likelihood estimation loss to train the RM. σ denotes the sigmoid function. During training, K answers are selected each time and combined in pairs.

$$L_{\mathrm{RM}}(r_\phi) = -E_{\left(x,y_w,y_t\right)\sim D}\left[\log\sigma(r_\phi(x,y_w) - r_\phi(x,\ y_t))\,\right] \tag{3}$$

In the third stage, proximal policy optimization (PPO) is used to optimize the policy for the RM. We use the output of the RM as a scalar reward and use the PPO algorithm to fine-tune the SFT model to optimize this reward (Schulman et al., 2017). Among them, the second and third stages can be carried out iteratively. We collect more comparison data on the current best policy, use it to train a new RM, and then train a new policy. In practice, most of our comparison data comes from supervised policies, and some of it comes from PPO policies. After training the RM, we have obtained an approximate reward function. Consider the model as a policy in the reinforcement learning (RL) problem, and the input and output of the model can be regarded as the state and action of RL respectively. We can use RL algorithms to train the model and train the model to output the answer with the highest reward. However, the RL training process is very unstable, so the SFT model is used as a

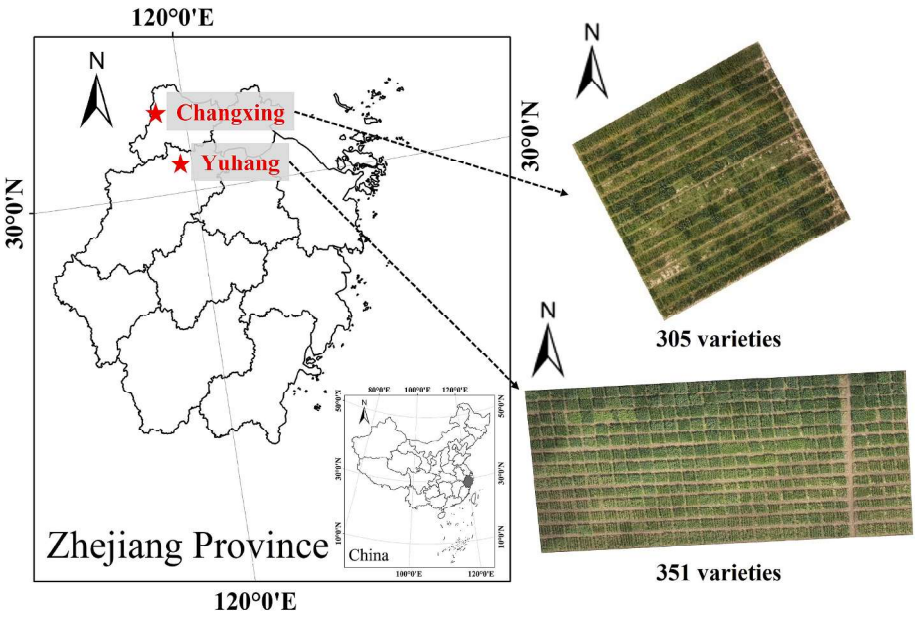

Fig. 2. Study area.

Table 1
Definition of features extracted from different sensors.

Sensor	Features	Formulation	References
MS (Spectral information)	Normalized Difference Vegetation Index	$NDVI = \frac{(NIR - R)}{(NIR + R)}$	(Tucker, 1979)
	Soil-Adjusted Vegetation Index	$SAVI = \frac{(1 + \mathrm{L}) \times (NIR - R)}{(NIR + R + \mathrm{L})}$, L = 0.5	(Huete, 1988; Richardson and Everitt, 1992)
	kernel Normalized Difference Vegetation Index	$kNDVI = tanh\left(\left(\frac{NIR - R}{2\sigma}\right)^2\right)$, where σ is a tunable length-scale parameter intended to capture nonlinear sensitivity of $NDVI$ to vegetation density. If $\sigma = 0.5 \times (NIR + R)$, which simplifies to $kNDVI = tanh((NDVI)^2)$.	(Camps-Valls et al., 2021)
	Near-Infrared Reflectance of Vegetation	$NIR_v = NIR \times NDVI$	(Badgley et al., 2017)
	Plant Senescence Reflectance Index	$PSRI = \frac{(R - G)}{NIR}$, R, G, and NIR denote the reflectance of the red, green, and near-infrared bands from the 5-band Phantom 4 multispectral, respectively.	(Cao et al., 2019)
HS (Spectral information)	Normalized Difference Vegetation Index	$NDVI = \frac{(NIR - R)}{(NIR + R)}$	(Tucker, 1979)
	Soil-Adjusted Vegetation Index	$SAVI = \frac{(1 + \mathrm{L}) \times (NIR - R)}{(NIR + R + \mathrm{L})}$, L = 0.5	(Huete, 1988; Richardson and Everitt, 1992)
	kernel Normalized Difference Vegetation Index	$kNDVI = tanh\left(\left(\frac{NIR - R}{2\sigma}\right)^2\right)$, where σ is a tunable length-scale parameter intended to capture nonlinear sensitivity of $NDVI$ to vegetation density. If $\sigma = 0.5 \times (NIR + R)$, which simplifies to $kNDVI = tanh((NDVI)^2)$.	(Camps-Valls et al., 2021)
	Near-Infrared Reflectance of Vegetation	$NIR_v = NIR \times NDVI$	(Badgley et al., 2017)
	Plant Senescence Reflectance Index	$PSRI = \frac{(R_{680} - R_{500})}{R_{750}}$, R_{680}, R_{500} and R_{750} denote the reflectance at 680 nm, 500 nm, and 750 nm respectively.	(Cao et al., 2019)
LiDAR (Structural information)	Canopy Height (CH)	CH = DSM − DEM	(Maimaitijiang et al., 2020)
LiDAR (Structural information)	Canopy Volume (CV)	CV = Excavated volume + Filled volume, DJI Terra software (DJI)	https://enterprise.dji.com/dji-terra
RGB image (Structural information)	Plant Lodging (PL)	$PL = \frac{\text{Number of lodging pixels in the plot}}{\text{Total number of the plot pixels}} \times 100\%$	(Zhang et al., 2023)
RGB video (Reproduction information)	Wheat Head (WH)	WH = The number of wheat head detection	(Zhu et al., 2022)
RGB image (Environmental information)	Weed Level (WL)	$WL = \frac{\text{Number of weed pixels in the plot and specific area}}{\text{Total number of the plot and specific area pixels}} \times 100\%$	(Anderegg et al., 2023)
MS (Environmental information)	Fractional Vegetation Coverage (FVC)	$FVC = \frac{\text{Number of crop pixels in the plot}}{\text{Total number of the plot pixels}} \times 100\%$	(Cui et al., 2023)

* R, G and NIR are the reflectance of red, green and near-infrared bands, respectively.

reference model, denoted by π_{ref}. The Kullback-Leibler (KL) divergence regularization term with the reference model is added to the objective function (4) to limit the update range of the model. Among them, the hyperparameter β controls the degree of deviation from the reference model.

$$J_{r_\phi}(\pi_\theta) = E_{x\sim D, y\sim\pi_\theta}\left[r_\phi(x,\ y)\right] - \beta D_{KL}\left(\pi_\theta(y|x)\middle|\pi_{\text{ref}}(y|x)\right) \quad (4)$$

According to the objective function (4), the KL divergence term is expanded, and the final comprehensive reward is:

$$r(x,y) \ = r_\phi(x,\ y) - \beta(\log\pi_\theta(y|x) - \log\pi_{\text{ref}}(y|x)) \quad (5)$$

The objective function is:

$$J_{r_\phi}(\pi_\theta) = E_{x\sim D, y\sim\pi_\theta}[r(x,\ y)] \quad (6)$$

We created three different datasets: (1) The SFT dataset (75 k) is created using multi-source datasets from the field, which contains questions and answers for training the SFT model. 80 % of this dataset is used for training, and 20 % is used for accuracy testing of wheat breeding model evaluation benchmark (phenotyping estimation task and environmental stress assessment task). (2) The RM dataset, with breeder rankings of model outputs, is used to train the RM. (3) The PPO dataset, without any breeder labels, which are used as inputs for RLHF fine-tuning. The external domain knowledge base is mainly used by RAG technology. The dataset format refers to SWIFT (Table S1 in Supplementary material 1).

3. Experiment

3.1. Study area

The experiment was conducted for two years (2021–2022 and 2022–2023), with 305 and 351 varieties, respectively. The varieties in the second year included the varieties in the first year and 46 varieties were added (Table S2 in Supplementary material 1). The experimental fields were set up in Changxing and Yuhang Agricultural Experimental Bases of Zhejiang University Agricultural Experiment Station in Zhejiang Province, China (Fig. 2; Table S3 in Supplementary material 1). Specific planting information is shown in Table S4 (Supplementary material 1). The experimental fields were not sprayed with pesticides and herbicides, and weeds were not controlled manually or mechanically. Maintaining the natural growth environment is to develop competitive wheat varieties that have a greater competitive advantage than wheat weeds. The water required for wheat growth comes from natural rainfall. Both places have a subtropical monsoon climate.

3.2. Data description

The cross-domain knowledge base (in Chinese and English) used in this study consists of a field multi-source dataset (Fig. 1a) and an external domain knowledge base (Fig. 1b).

In this study, we conducted a full growth cycle analysis of breeding materials from diverse wheat germplasm resources over two years. Specifically, in the first year, we collected data six times from 305 germplasms, while in the second year, data were collected nine times from 351 germplasms. The multi-source dataset collected in the field includes: (1) RS data, including hyperspectral (HS), light detection and ranging (LiDAR), multispectral (MS), and RGB data acquired using UAVs. We process remote sensing data to obtain VI, canopy height (CH), canopy volume (CV), plant lodging (PL), wheat head (WH), weed level (WL), and fractional vegetation coverage (FVC). (2) phenotypic data, including manually collected values of SPAD, leaf area index (LAI), CH, and yield. (3) environmental data, including continuously recorded meteorological data as well as manually collected and measured soil data. Detailed data acquisition and processing procedures are provided in Supplementary Material 2, with definitions of the features extracted from different sensors presented in Table 1.

The external domain knowledge base includes: (1) wheat germplasm data (22k), mainly including variety name, place of origin, nutritional quality, resistance and agricultural traits. In terms of nutritional quality, key parameters such as crude protein, lysine, and sedimentation value are involved. In terms of resistance, it covers important characteristics such as disease resistance (stripe rust, leaf rust, powdery mildew, etc.), drought resistance, and cold resistance. In addition, agronomic traits such as maturity, plant height, thousand grain weight, and grain hardness are also recorded. (2) wheat cultivation technique data (10k), covers the entire process from pre-planting preparation to post-harvest storage, mainly including the selection of suitable varieties, soil preparation, sowing, fertilizer management, irrigation and pest and disease control techniques. (3) wheat plant protection technique data (10k), mainly including disease prevention and control, pest prevention and control, weed management, as well as new techniques such as precision plant protection based on artificial intelligence and RS, agricultural UAV flight control and green prevention and control. (4) wheat seed price data (20k), mainly including observation point, variety name, price, specification, planting area and time, etc. For observation points, the economic level and planting demand of different regions lead to price differences; for varieties, different varieties have different characteristics and application scenarios, and high-quality varieties are usually more expensive; different specifications of packaging and planting areas affect the final selling price; prices change over time due to changes in market supply and demand, especially natural disasters or policy adjustments.

The wheat germplasm data comes from the Chinese Crop Germplasm Information Network (https://www.cgris.net/). The data on wheat CT and wheat plant protection technique (PPT) come from search engines (Google, Bing, Baidu), and the search terms include “wheat cultivation technique, wheat plant protection technique, 小麦栽培技术, 小麦植保技术”. The wheat SP data comes from the National Seed Market Monitoring Information Release Platform (http://202.127.45.18/). All data from this study have been made publicly available (https://doi.org/10.5281/zenodo.13710825).

3.3. Evaluation benchmark

In order to comprehensively evaluate the performance of MLLM in scientific breeding work, it is necessary to build a standardized machine and human evaluation benchmark. This benchmark aims to integrate challenges from various wheat breeding tasks. This benchmark aims to integrate challenges from various wheat breeding tasks. We have specifically designed many professional Chinese and English questions on wheat breeding and corresponding standard answers, covering five tasks: phenotyping estimation, environmental stress assessment, target germplasm screening, CT recommendation, and SP query. The phenotyping estimation includes seven subtasks (Yield, SPAD, LAI, CH, CV, WH and PL), the environmental stress assessment includes two subtasks (WL and FVC), the target germplasm screening includes five subtasks: high quality (HQ), disease resistance (DS), drought resistance (DR), maturity period (MP), adapt to mechanized (AM), the CT recommendation includes two subtasks: CT and PPT, and the SP query includes one subtask SP query. Through objective evaluation indicators and manual scoring and sorting, the evaluation team conducted a detailed evaluation of the answers of multiple MLLMs including WBLM, covering three aspects: accuracy, stability, and reasoning.

(1) Accuracy

We evaluate different subtasks under five tasks, respectively. For the phenotyping estimation task and the environmental stress assessment task, the root mean square error (RMSE) and coefficient of determination (R^2) evaluation indicators are selected for regression tasks (estimating the values of FVC, SPAD, LAI, CH, CV, WH, and Yield), and the accuracy evaluation indicator is selected for the classification task (estimating the categories of WL and PL). For each subtask under the

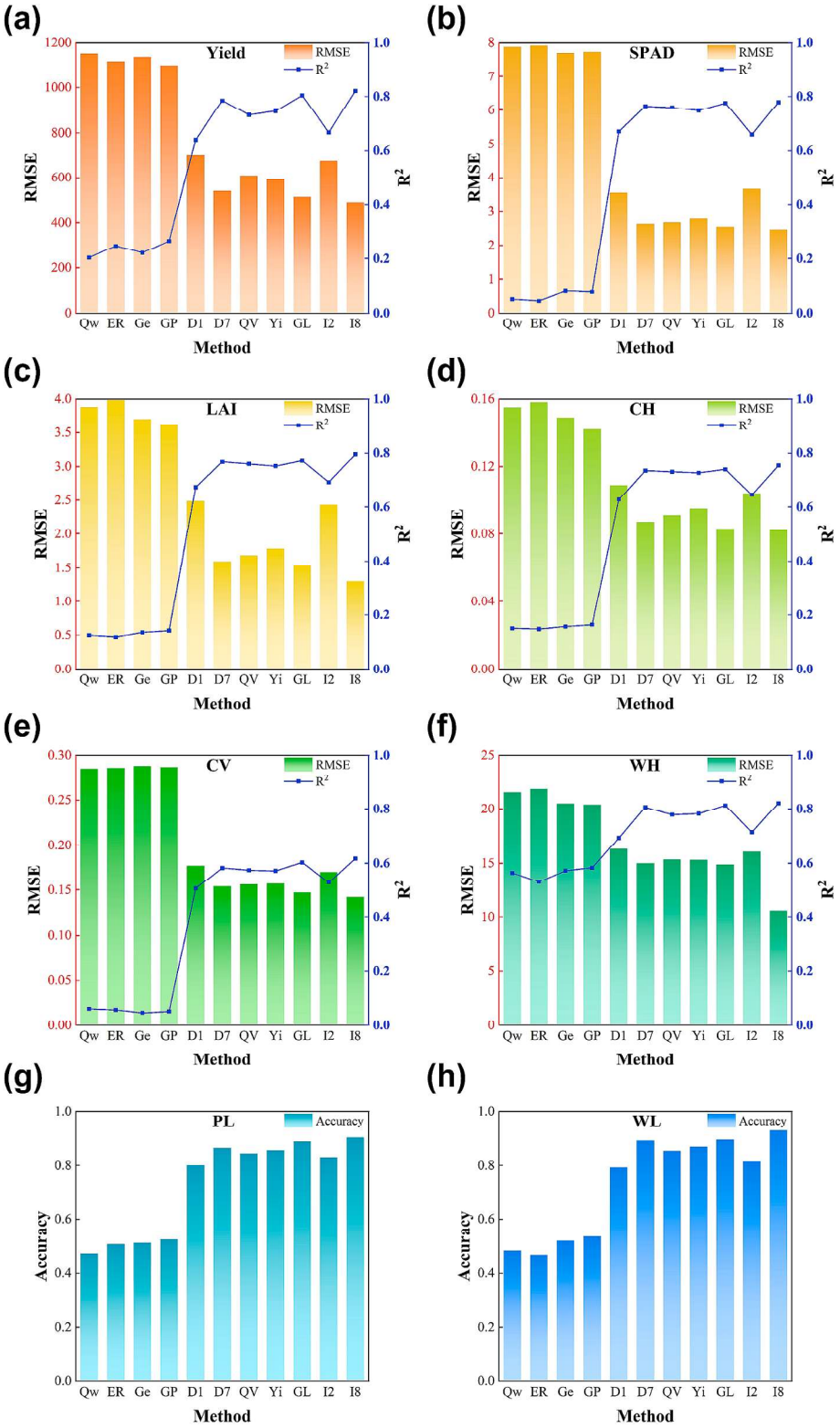

Fig. 3. Comparison of different MLLMs on the evaluation benchmark (accuracy).

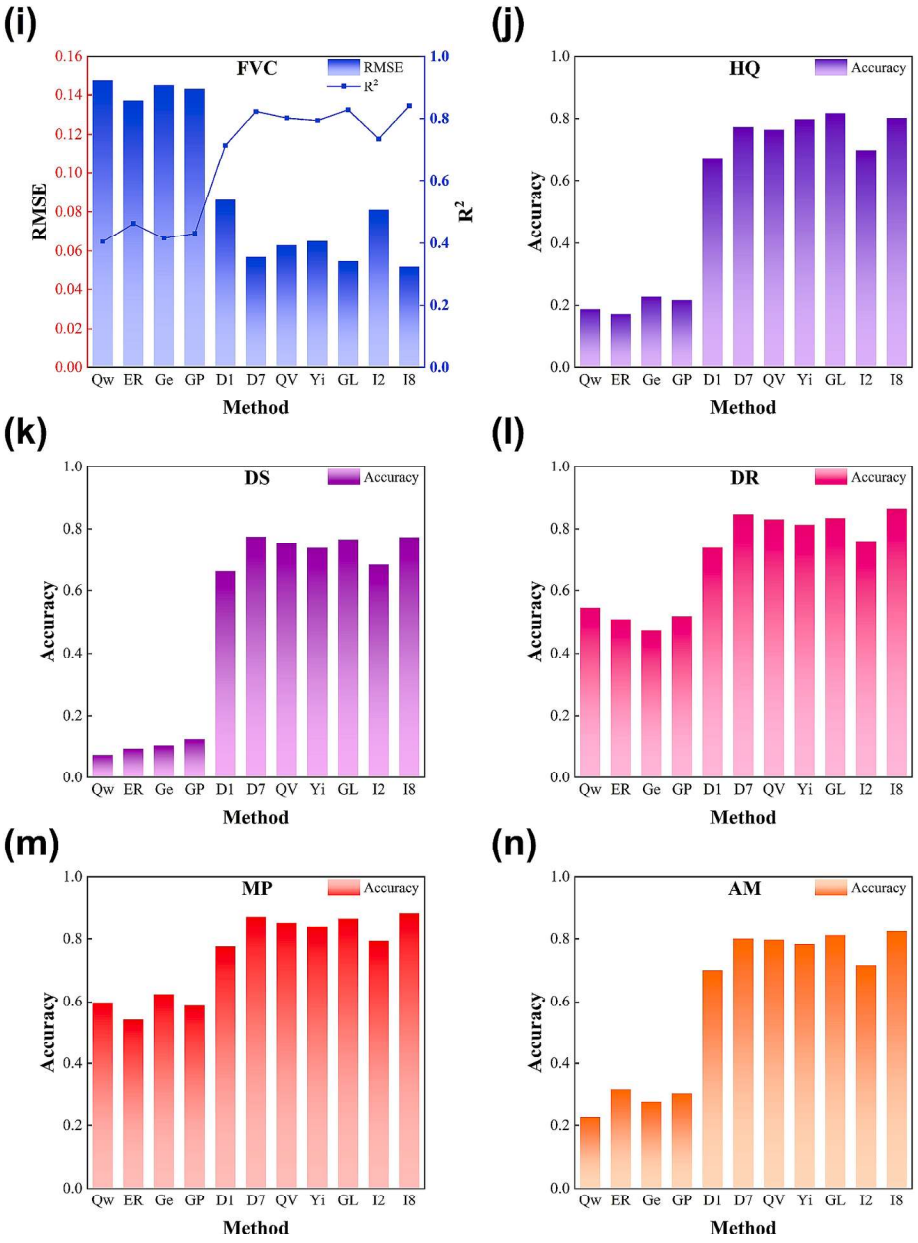

Fig. 3. (*continued*).

target germplasm screening task and CT recommendation task, each MLLM is tested multiple times (1k), and each test is manually judged whether the answer is correct. Then the proportion of the number of correct answers to the total number of tests is used as the evaluation indicator. For the SP query task, we counted whether each result after multiple (1k) tests was consistent with the data in the knowledge base (price $\pm$ 10 %), and took the proportion of the total number of consistent data to the total number of tests as the evaluation indicator.

(2) Stability

We evaluate the consistency and robustness of the answers. The evaluation indicator of stability is the proportion of times that satisfy the stability requirements in the total number of tests (1k), where the stability requirement is that the deviation of the numerical value in the answer is within $\pm$10 %, and whether the text satisfies the requirement is manually judged. Consistency: Since cross-domain data are related, the answers given after inputting different data from cross-domain data into the model should remain basically consistent. Robustness: The answer or performance of the model should not change significantly when faced with small changes in input data. Robustness requires that the model should have a certain degree of fault tolerance.

(3) Reasoning

We evaluate the logical deduction, inductive reasoning, and explanation of the answers. The evaluation indicator of reasoning is the proportion of the total reasoning score of a single model after multiple tests (1k) to all scores. After one test, x MLLM models output x answers. After manually analyzing x answers, we select a score from 1 to x points for each answer and the score can only be selected once until all answers are scored. We sum up the answer scores of different models after multiple tests as the total reasoning score of the corresponding model. Then calculate the proportion of the total reasoning score of different models to all scores ($(1 + x) \times x/2$). Logical deduction: The model can deduce new conclusions or answers through logical reasoning based on known facts, rules and conditions. This requires the model to understand the logical relationships in the question and accurately apply

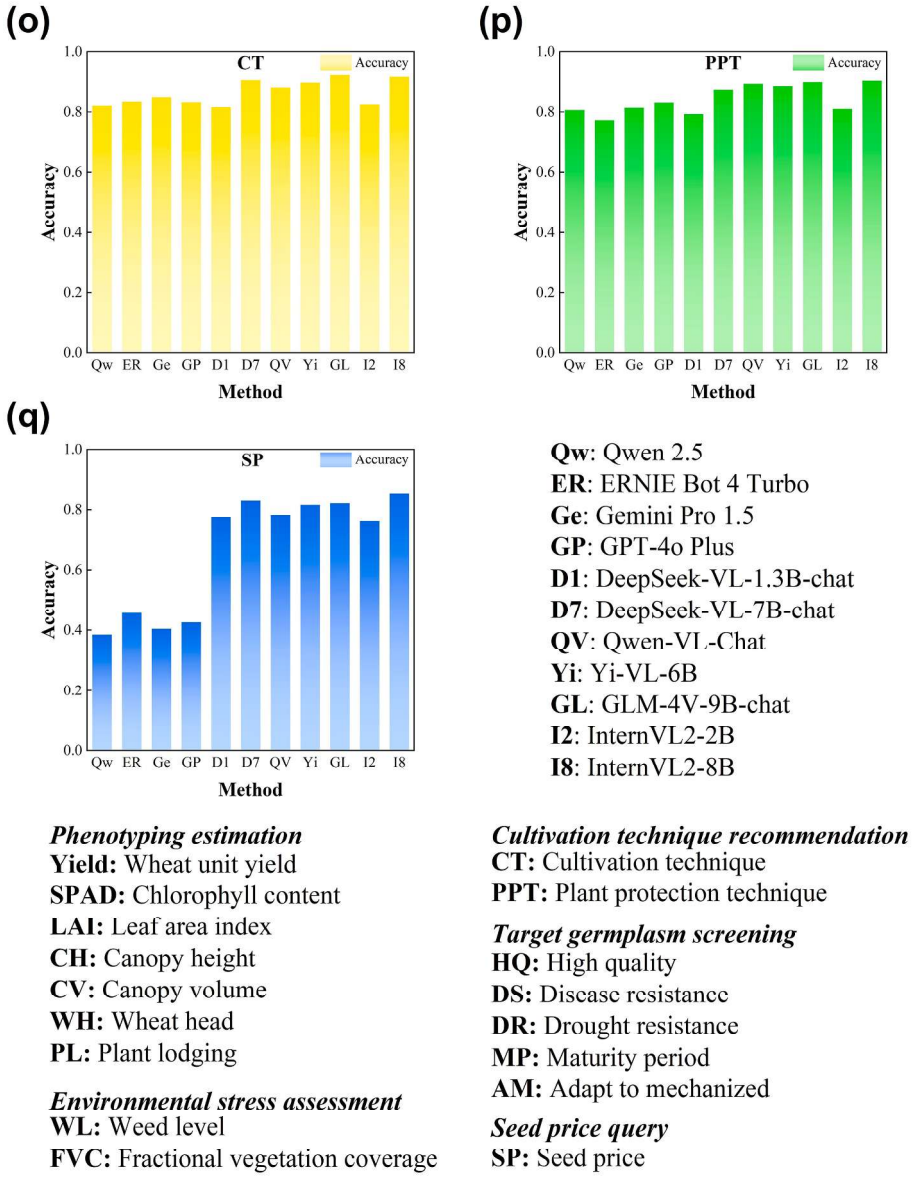

Fig. 3. (*continued*).

these relationships to deduce answers. Inductive reasoning: The model can summarize general rules or patterns from a series of specific examples or observations and give reasonable answers based on them. Inductive reasoning requires the model to have the ability to abstract and generalize information. Explanation: The reasoning process is explained or visualized to a certain extent to help user understand the source and basis of the answer.

4. Results and discussion

4.1. Evaluation of WBLM on the new benchmark

The newly constructed wheat breeding model evaluation benchmark was used to evaluate closed-source MLLMs (Qwen 2.5, ERNIE Bot 4 Turbo, GPT-4o Plus and Gemini Pro 1.5, abbreviated as Qwen, ERNIE Bot, ChatGPT, Gemini) and WBLMs built based on different open-source MLLMs (Qwen-VL, InternVL, Yi-VL, Deepseek-VL and GLM-4) (Figs. 3-5; Table S5 in Supplementary material 1).

Accuracy evaluation: On various tasks of the evaluation benchmark, the comprehensive performance of WBLM based on InternVL2-8B is better than that of WBLM built on other open-source MLLMs, and is significantly better than closed-source ChatGPT, Gemini, and Qwen. Among them, for LAI evaluation, the R^2 based on InternVL2-8B is 0.795 and the RMSE is 1.302. The R^2 and RMSE of other open-source MLLMs range from [0.674, 0.772] and [1.542, 2.486], and the R^2 and RMSE of closed-source MLLMs range from [0.118, 0.142] and [3.61, 3.985], respectively. ChatGPT shows leading performance among open-source models, although still far behind the WBLM based on InternVL2-8B. In addition, we find that consistent with existing study (Chen et al., 2024b), the performance of MLLMs with small parameters in the same series is lower than that of MLLMs with larger parameters, such as InternVL2-8B and InternVL2-2B. We attribute this modest decline to the smaller size of the MLLMs.

Stability evaluation: Compared with closed-source MLLMs, WBLMs based on open-source models outperform in stability tests. Specifically, the WBLM based on InternVL2-8B achieved the best performance in consistency, with a stability score of 0.811, while the stability score of the lowest-performing Qwen 2.5 was only 0.053. The robustness scores of closed-sources MLLM ([0.829, 0.895]) are higher than those of WBLM based on open-source MLLM ([0.735, 0.802]), because closed-source

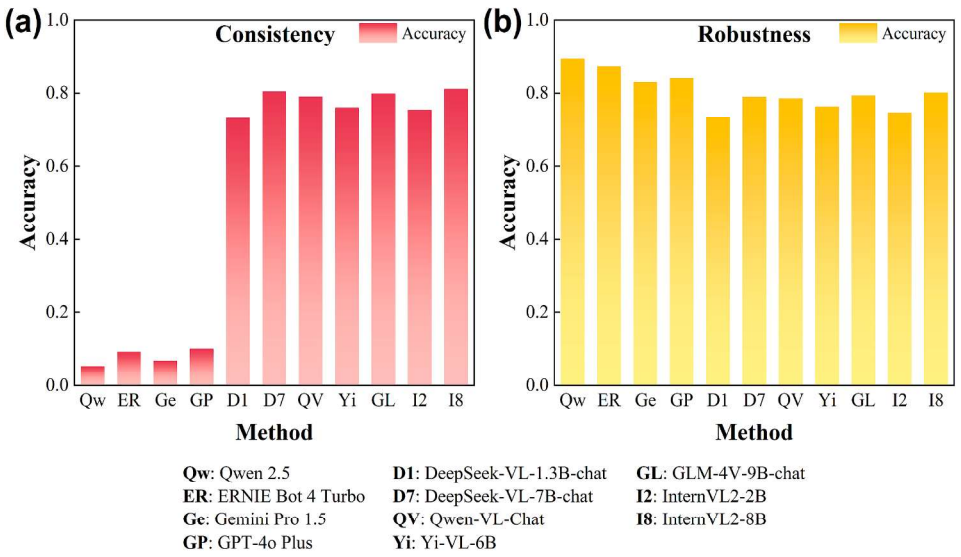

Fig. 4. Comparison of different MLLMs on the evaluation benchmark (stability), which includes consistency and robustness.

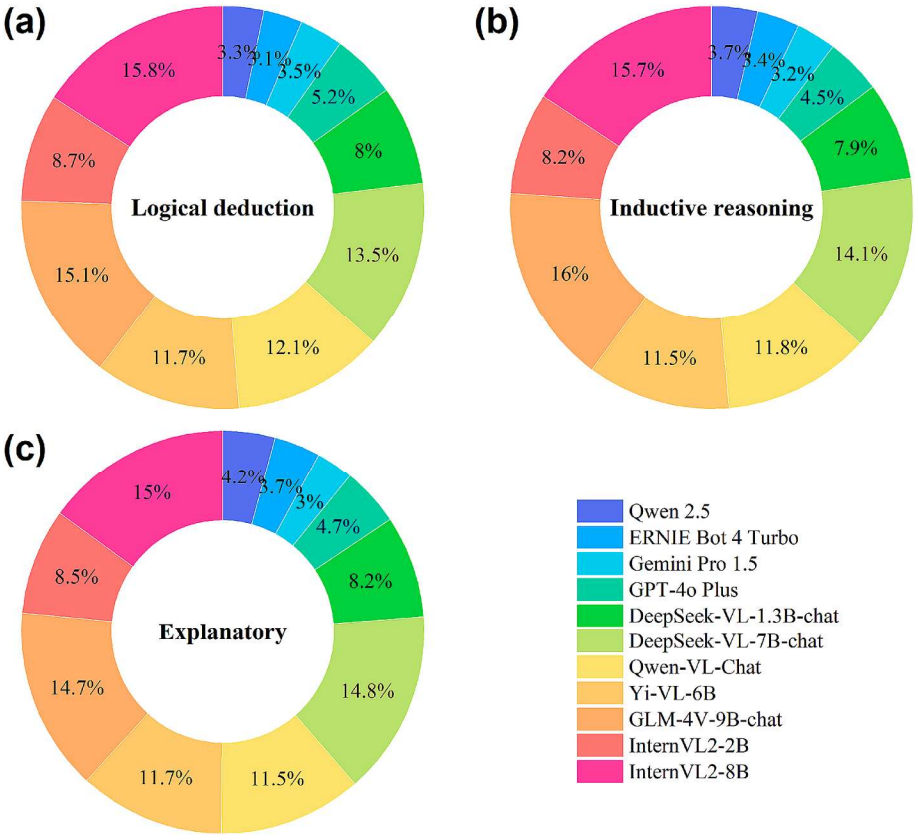

Fig. 5. Comparison of different MLLMs on the evaluation benchmark (reasoning). This figure shows the total reasoning score as a proportion of all scores after multiple tests of a single MLLM.

MLLM cannot accurately answer relevant breeding questions in most tasks, and the answers are still wrong and similar when faced with slight changes in input data. However, among multiple WBLMs based on open-source MLLM, WBLM based on InternVL2-8B has a higher robustness score.

Reasoning evaluation: The WBLM based on InternVL2-8B outperforms other MLLMs in logical deduction and explanation evaluation, with the highest scores of 0.158 and 0.15, respectively. The WBLM based on GLM-4 V-9B-chat outperforms other MLLMs in inductive reasoning evaluation, with the highest score of 0.16. Compared with accuracy and stability evaluation, reasoning evaluation tests the breeding decision support and problem-solving capabilities of the model more. Excellent reasoning ability can enhance users' trust in MLLM and improve user satisfaction (Li et al., 2024d).

InternVL2-8B based WBLM shows leading performance in most benchmark tests (Figs. 3-5; Table S5 in Supplementary material 1). However, the closed-source commercial MLLMs used in this study encountered significant difficulties in the wheat breeding evaluation