doi:10.6041/j.issn.1000-1298.2026.01.027

Agri-Eval: Multi-level Large Language Model Valuation Benchmark for Agriculture

WANG Yaojun GE Mingliang XU Guowei ZHANG Qiyu BIE Yuhui

(*College of Information and Electrical Engineering, China Agricultural University, Beijing* 100083, *China*)

Abstract: Model evaluation using benchmark datasets is an important method to measure the capability of large language models (LLMs) in specific domains, and it is mainly used to assess the knowledge and reasoning abilities of LLMs. Therefore, in order to better assess the capability of LLMs in the agricultural domain, Agri-Eval was proposed as a benchmark for assessing the knowledge and reasoning ability of LLMs in agriculture. The assessment dataset used in Agri-Eval covered seven major disciplines in the agricultural domain: crop science, horticulture, plant protection, animal husbandry, forest science, aquaculture science, and grass science, and contained a total of 2 283 questions. Among domestic general-purpose LLMs, DeepSeek-R1 performed best with an accuracy rate of 75.49%. In the realm of international general-purpose LLMs, Gemini-2.0-pro-exp-02-05 standed out as the top performer, achieving an accuracy rate of 74.28%. As an LLMs in agriculture vertical, Shennong V2.0 outperformed all the LLMs in China, and the answer accuracy rate of agricultural knowledge exceeded that of all the existing general-purpose LLMs. The launch of Agri-Eval helped the LLM developers to comprehensively evaluate the model's capability in the field of agriculture through a variety of tasks and tests to promote the development of the LLMs in the field of agriculture.

Key words: large language models; assessment systems; agricultural knowledge; agricultural datasets

CLC number: TP3 **Document code**: A **Article ID**: 1000-1298(2026)01-0290-10 **OSID**:

0 Introduction

With the development of large language models, more and more models have been proposed, such as ChatGPT[1], Claude[2], LLama[3] and so on. However, the evaluation of model abilities is still at a relatively basic stage. Traditional natural language processing (NLP) evaluation benchmarks and methods can only measure the abilities of traditional language models, and are difficult to apply to evaluate the abilities of current LLMs. Today's LLMs have been equipped with some complex abilities more similar to those possessed by humans, such as reasoning and comprehension[4]. Therefore, new evaluation benchmarks needed to be proposed.

In the last two years, with the boom of LLMs, a series of assessment methods and assessment datasets emerged along with them. For example, T-Evalwas used to assess the tool-using ability of a large language model, which was refined into sub-competencies such as instruction following, planning, reasoning, etc.[5] MMLU was used to assess the ability of a large language model in understanding world knowledge and problem solving, which contained multiple-choice questions on 57 topics such as humanities, social sciences, etc.[6] While LongBench was used to assess the ability of a large language model to understand long text sequences[7]. These evaluation benchmarks aimed to assemble a variety of NLP tasks to comprehensively assess the abilities of a large language model. Other evaluation benchmarks shifted the goal of evaluation to some specific abilities of a large language model, such as GSM8K for evaluating a model's mathematical reasoning ability[8] and HumanEval for evaluating a

Received date: 2024-09-21 Accepted date: 2025-11-05

Foundation item: National Key R&D Program (2024YFD2000805)

First author: WANG Yaojun, associate professor. E-mail: wangyaojun@cau.edu.cn

model's ability to generate computer program code on demand[9].

The current popular assessment benchmark C-Eval, which encompassed science, technology, engineering, mathematics education, social sciences, and so on, focused on the model's knowledge and reasoning ability in a wide range of domains and failed to meet the needs of specific industries. The issue of food has always been crucial for the world[10]. In order to better leverage LLMs to assist agricultural production and enhance productivity, corresponding evaluation benchmarks should be developed. Although it is already possible to assess the abilities of LLMs in plant protection in C-Eval[11], considering plant protection alone is not comprehensive enough. When assessing the capabilities of LLMS in the agricultural field, emphasis should be placed on core tasks in practical agricultural production scenarios, such as pest and disease control, fertilization management, etc. Different tasks should be graded according to their importance to actual agricultural production. When formulating evaluation indicators, the weight of tasks at different grades should be taken into consideration.

In order to better promote the development of LLMs in agriculture, the Agri-Eval evaluation benchmark was proposed, which was the first evaluation system to comprehensively assess the capability of LLMs in agriculture. Agri-Eval's assessment dataset comprised 2 283 questions with multiple choice, fill-in-the-blank, judgement, and reasoning questions. It covered seven major agricultural disciplines: crop science, horticulture, plant protection, animal husbandry, forestry, aquaculture, and grass science. An experimental evaluation of the general-purpose LLMs that ranked high in the open compass list and the LLMs Shennong V2.0 in the agriculture was conducted.

1　Agri-Eval evaluation system

1.1　Agri-Eval design principles

The design goal of Agri-Eval was to help developers and users understand the abilities of the LLMs of interest to them in the agricultural domain so that they can target their model improvements and choices. Therefore, it was focused on developing datasets that evaluated the more advanced abilities of LLMs in the agricultural domain. In agricultural production, which facesd a variety of risks, including natural disasters, pests, and diseases, LLMs should have the ability to provide effective solutions in the form of questions and answers to cope with these risks and help agricultural production[12].

In order to effectively assess the capabilities of LLMs in the agricultural field, a needs analysis of the agricultural industry to identify the key issues in this domain was conducted. Taking apple tree cultivation as an example, the planting process was divided into seven stages: seed selection, seedling cultivation, transplanting, growth period, harvesting, and sales.

Subsequently, based on the actual production scenarios, each stage was further divided. For instance, in the growth period stage, it was divided into fertilization management and pest and disease control.

After that, according to the importance of each task to the production scenarios, it was classified into three levels in total. For example, in the growth period stage, the task of pest and disease control had a decisive impact on the yield and quality of fruit trees. The name of the pesticide, the dosage, and the application method must not be wrong. Therefore, it was classified as level one.

In the seedling cultivation stage, for the task of pruning seedlings, pruning can promote the branching and plant architecture optimization of seedlings, which is helpful to improve the photosynthetic efficiency and ventilation and light-transmittance, and has an important impact on the healthy growth of seedlings and the subsequent yield. However, it is not a decisive factor. Therefore, it was classified as level two.

In the harvesting stage, for the task of fruit turning, it was to increase the fruit surface's exposure to sunlight and make the coloring uniform. Compared with the task of pest and disease control, which directly affected the production quality, its importance was relatively lower. Therefore, it was classified as level three.

Through needs analysis, the key issues in the agricultural field was identified. Subsequently, following the disciplinary divisions in agronomy, the core disciplines related to agriculture were selected. Then, the issues of these disciplines were graded

according to their importance to actual agricultural production, dividing them into three levels in total. Finally, drawing on the construction method of human exams, the Agri-Eval evaluation system was established to assess the key capabilities of LLMs in the agricultural field. The Agri-Eval assessment dataset covered four main types of questions: multiple-choice, fill-in-the-blank, judgement, and reasoning. In the multiple-choice section, there were four choices for each question, but only one was correct; in the fill-in-the-blanks section, there was a single correct answer for each blank; in the judgement section, there was a clear right or wrong answer for each question; and in the reasoning section, there was a path to the answer for each question. The question format of Agri-Eval covered seven different disciplines and the statistical data for each discipline was shown in Table 1.

Table 1 Statistical table of Agri-Eval data

Agronomic classification	Number
Crop science	402
Horticulture	439
Plant protection	301
Animal husbandry	263
Forest science	321
Aquaculture science	208
Grass science	349
Total	2 283

In order to achieve fairness in the assessment, the content in the reduced question set was directly presented in the training corpus of the existing large language model. Data sources from small-scale exams were chosen, such as one of the tests or end-of-course exams at a particular university. Compared with large-scale exams such as national vocational exams and postgraduate entrance exams, data from these small-scale exams were difficult to easily accesses. In addition, some of Agri-Eval's question type data came from paper-based problem sets stored in libraries that had not been electronically published and distributed. The documents and knowledge related to the question types were collated, proofread, annotated and formatted in consultation with agronomy experts to obtain the standard single-choice question types. This process ensured the quality of the data and the standardization of the question types, making it possible to achieve a comprehensive, fair and reasonable assessment of the abilities of the large language model in the agricultural domain.

1.2 Agri-Eval evaluation dataset construction

Based on the disciplinary division of agronomy, the Agri-Eval reference graduate admissions network ultimately contained seven major disciplines: crop science, horticulture, plant protection, animal husbandry, forestry science, aquaculture science, and grass science (Fig. 1). The collection of topics for each discipline was based on the consideration of the core undergraduate curriculum of the corresponding discipline as well as the specialized subjects of the graduate entrance examination (GEE). Key knowledge on the characteristics of various types of crops, planting techniques, soil management, proper fertilizer application and the use of plant growth regulators were included in crop science. Horticulture is a comprehensive discipline that encompasses a full range of knowledge from breeding and cultivation to post-harvest handling. The discipline is not only concerned with how to breed varieties with excellent characteristics, but also endeavors to study the growth habit and environmental adaptations of crops to achieve high yield and quality. Plant protection, on the other hand, includes basic areas of plant morphology, structure, reproductive mode, physiological functions, and classification, such as the phenomenon of double fertilization, root-tip water-sucking region, distribution of vesicular cells, and plant taxonomic units. In addition, it explored plant-environment interactions, including plant water and mineral nutrient uptake, photosynthesis, respiration, ecological factor effects, biome structure and function, and ecosystem services. Animal husbandry is a comprehensive discipline that involves several areas such as veterinary qualifying examinations, drug reconciliation, animal anatomy, physiological systems, pathology, microbiology, immunology and pharmacology. It aimed to improve animal health and productivity while ensuring food

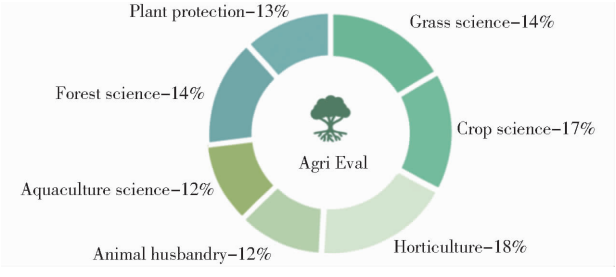

Fig. 1 Agri-Eval discipline distribution map

safety and public health. Forest science, on the other hand, encompasses a wide range of knowledge on disease diagnosis, plant growing conditions, plant protection measures, plant growth and reproduction, functional structure of plant organs, and plant nutrition and management. This discipline was dedicated to the conservation and rational use of forest resources and the promotion of ecological balance. Aquaculture science, on the other hand, covers crop species, fish habits, culture techniques, feed selection, disease control, and physiological characteristics of aquatic animals. It was concerned with the culture and conservation of aquatic organisms for the sustainable development of aquatic resources. Grass science, on the other hand, focuses on the structure and function of herbaceous plant cells, the role of hormones, response to the environment, regulation of growth and development, photosynthesis and respiration, water and nutrient uptake, as well as morphology, function and developmental processes of plant organs. This discipline was important for grassland management and herbaceous plant utilization.

During the construction of the Agri-Eval assessment dataset, firstly, for multiple-choice questions, all questions were reformatted to include four options; if a question originally had fewer than four options, it was excluded; and in cases where there were more than four options, the wrong option was randomly removed to ensure that four options were ultimately retained for each question. Fill-in-the-blank questions were designed to ensure that each blank had a unique correct answer. For judgement questions, each question had a clear correct or incorrect decision. As for the reasoning questions, each question had a clear answer. Then all the data were de-duplicated to ensure that they do not contain redundant questions. Next, the questions of each subject were divided according to their importance to the actual agricultural production scenarios, with three levels in total. Level one was the most important, as it was the factor that determined agricultural production, and so on. Finally, it was divided into three subsets according to the purpose of the assessment: development set, validation set and test set. In the development set, each question had a corresponding answer and parse, which was used to set up five shot experiments; in the validation set, each question had a corresponding answer, which was used to measure the accuracy of the assessment; in the test set, there were only questions without answers, and this dataset would be open out to maintain the fairness of the Agri-Eval assessment.

Explanatorydata generation: 'Chain-of-thought (COT) reasoning' was proposed by Kojima et al. and Wei et al.[13] The method can guide LLMs to generate specific sequences of reasoned text and can give final answers. This method has been very successful on many inference tasks. In contrast to Zero Shot, Few Shot is able to motivate the model to answer questions. To facilitate the potential use of Few Shot COT in Agri-Eval, an interpretation of the data in the development set was made, which was given by an expert in the field of agronomy. Example data were shown in Fig. 2.

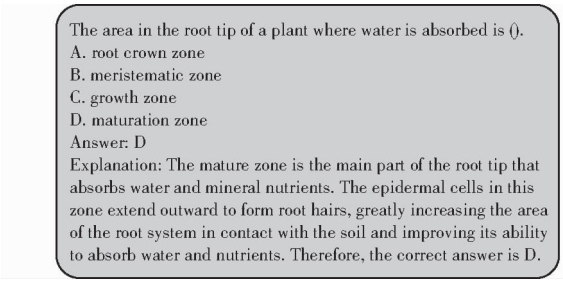
The area in the root tip of a plant where water is absorbed is ().
A. root crown zone
B. meristematic zone
C. growth zone
D. maturation zone
Answer: D
Explanation: The mature zone is the main part of the root tip that absorbs water and mineral nutrients. The epidermal cells in this zone extend outward to form root hairs, greatly increasing the area of the root system in contact with the soil and improving its ability to absorb water and nutrients. Therefore, the correct answer is D.

Fig. 2　Example data of the development set

1.3　Agri-Eval evaluation methodology

In the Agri-Eval evaluation methodology, there were two main approaches to the strategy of constructing Prompt. The first method was the Zero Shot approach, where a question and its options were selected directly from the validation set to build the Prompt, and the second method was the Five Shot approach, where five questions and their options and answers were selected from the development set, and then combined with a question and its options from the validation set to build a more detailed Prompt. Subsequently, the process of generating the answers was also two different routes. The first was an inferential approach, where the constructed Prompt was fed directly into a large language model, which generated the corresponding answers. The second approach relied on the calculation of perplexity, where each option was combined with a question to form different sequences, and the reasonableness of these sequences was evaluated by calculating their perplexity, and the answer was chosen to be the one corresponding to the sequence with the lowest

perplexity. Finally, the model's performance evaluation was based on the answers on the validation set. The overall performance of the model was measured by the proportion of the number of questions answered correctly by the model to the total number of questions, that was the accuracy rate. There were two ways of evaluation. One was to directly calculate the accuracy rate according to the subject classification. The other was to first calculate the accuracy rate according to the results of the subject grading, and then calculate the weighted average based on the weight of each grade. The specific calculation method was as follows: firstly determining the weight of each grade, the weight of level one was 0.5, denoted as W_1, the weight of level two was 0.3, denoted as W_2, and the weight of level three was 0.2, denoted as W_3, with the sum of the weights being 1. Then the weighted average was calculated, denoted as S. For example, if the score of level one questions was $X_1\%$, the score of level two questions was $X_2\%$, and the score of level three questions was $X_3\%$, the weighted average was calculated by using formula (1).

$$S = \sum_{i=1}^{n=2} X_i W_i \tag{1}$$

2 Experiments

2.1 Experimental setup

On the Agri-Eval evaluation dataset, Zero Shot[14] and Five Shot[15] experiments were set up to evaluate existing popular Chinese LLMs. The experiments used inference and perplexity to evaluate the capability of LLMs under Zero Shot and Five Shot conditions, respectively. In the Five Shot experiment, five samples were taken from the development set. In the answer extraction session, regular expressions were used to ensure that the answers were obtained. The samples were dynamically deleted and adjusted to fit the input window of the model, as the length of the input may exceed the input window of the model under the Five Shot condition. Finally, for the graded questions, under the Zero Shot condition, the accuracy rate was calculated by using reasoning, and the final score was calculated by using formula (1).

2.2 Modelling of participation reviews

In order to measure the current domestic LLMs in agriculture capability, eight open-source general-purpose LLMs were evaluated, all of which were at the top of the list of various Chinese LLMs and varied in their sizes, as shown in Table 2. Among these, due to the constraints of the API models, only Zero Shot and reasoning experiments were conducted for the API models.

Table 2 Model details

Model	Creator	Parameter/B	Access
Baichuan2 - 7B	Baichuan	7	Weights
Baichuan2 - 13B	Baichuan	13	Weights
DeepSeek - V3	DeepSeek	671	API
DeepSeek - R1	DeepSeek	671	API
Qwen1.5 - 14B	Alibaba	14	Weights
Qwen1.5 - 72B	Alibaba	72	Weights
Qwen2.5 - 72B	Alibaba	72	API
QWQ - 32B	Alibaba	32	API
Yi - 34B	01. AI	34	Weights
Internlm2 - 7B	Shanghai AI Lab	7	Weights
Internlm2 - 20B	Shanghai AI Lab	20	Weights
ChatGLM3 - 6B	ZhipuAI	6	Weights
LongAlign - 13B	ZhipuAI	13	Weights
LLama2 - 7B	Meta	7	Weights
LLama2 - 13B	Meta	13	Weights
LLama2 - 70B	Meta	70	Weights
LLama3 - 8B	Meta	8	Weights
LLama3 - 70B	Meta	70	Weights
Gemini - 2.0 - Pro	DeepMind	-	API
Claude 3.7 Sonnet	Anthropic	-	API
ChatGPT - 4o	OpenAI	-	API
Shennong V2.0	LALM	72	API

Qwen was a large language model implemented by Alibaba in 2023 based on the Transformer encoder, using a large amount of data for training and supervised fine-tuning and direct preference optimisation of trained models[16]. Currently, there were Qwen1.5 and Qwen2.5 versions of the Qwen family, which were available in seven scale sizes: 0.5B, 1.8B, 4B, 7B, 14B, 32B, and 72B. In this experiment, Qwen1.5 - 14B, Qwen1.5 - 72B, Qwen2.5 - 72B, and QWQ - 32B (based on Qwen2.5 - 32B fine-tuning) were involved in the evaluation.

Yi was a large language model developed by Zero One Everything in 2023, which aimed at bilingual language modelling, trained on 3 trillion tokens multi-language corpus. There were two versions of the Yi series, namely 34B and 6B, and Yi - 34B[17] was chosen to participate in this experiment for the evaluation.

Internlm was a large language model released in 2023 by Shanghai Artificial Intelligence Laboratory (Shanghai AI Lab) and SenseTime in conjunction with Chinese University of Hong Kong, Fudan University and Shanghai Jiaotong University, which was a large model trained on 1.5 trillion tokens and had a strong comprehensive capability. There were currently two versions, 7B and 20B[18]. In this experiment, Internlm - 7B and Internlm - 20B were selected to participate in the evaluation.

Baichuan was an open source large language model introduced by Baichuan Intelligence, trained on a high-quality corpus of 2.6 trillion tokens, and currently had two versions, 7B and 13B[19]. In this experiment, Baichuan2 - 7B and Baichuan2 - 13B were selected to participate in the evaluation.

ChatGLM was a large language model jointly released by Chip AI and KEG Lab at Tsinghua University, which used a large amount of data for pre-training[20]. In this experiment, ChatGLM3 - 6B and LongAlign - 13B were selected to participate in the evaluation.

DeepSeek was a large language model released by Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd. The DeepSeek series had currently been updated to two versions: DeepSeek - V3 and DeepSeek - R1, with DeepSeek - R1 demonstrating performance comparable to world-leading closed-source models like GPT-4o and Claude 3.5. In this experiment, DeepSeek - V3 and DeepSeek - R1 were selected to participate in the evaluation[21].

Shennong was an industry LLMs for the agricultural vertical launched by the Lab for Agriculture Large Model in 2024. The current released version of Shennong was Shennong V2.0.

LLama was a large language model released by Meta, trained on hundreds of billions of tokens datasets[22]. Currently, there were three versions of LLama series: LLama1, LLama2 and LLama3[23]. In this experiment, LLama2 - 7B, LLama2 - 13B, LLama2 - 70B, LLama3 - 8B and LLama3 - 70B were selected to participate in the evaluation.

ChatGPT was a large language model developed by OpenAI. The ChatGPT - 4o version demonstrated leading performance in reasoning, mathematics, and other domains. In this experiment, ChatGPT - 4o were selected to participate in the evaluation.

Claude was a large language model released by Anthropic. The latest version, Claude 3.7 Sonnet, exceled in response speed and reasoning capabilities. For this experiment, Claude 3.7 Sonnet were selected to participate in the evaluation.

Gemini was a large language model developed by Google DeepMind. Among its versions, Gemini - 2.0 - Pro ranked highly on the OpenCompass benchmark. In this experiment, Gemini - 2.0 - Pro - exp - 02 - 05 were selected to participate in the evaluation.

2.3 Experimental results

2.3.1 Overall reasoning ability assessment of model

As shown in Tables 3 and 4, Shennong V2.0 demonstrated superior performance in the Zero Shot and reasoning experiments, achieving an average accuracy of 92.08%. Notably, its accuracy exceeded 90% across disciplines such as crop science, plant protection, animal husbandry, aquaculture science, and grass science, with an impressive 98.23% accuracy in crop science. These results highlighted Shennong V2.0's significant advantage as a large model specialized in the agricultural vertical domain. The Chinese general-purpose large model DeepSeek - R1 showed the best performance with an average accuracy of 75.49%, followed by QWQ - 32B with an average accuracy of 73.92%, ranking second. LLama2 - 7B, on the other hand, performed relatively poorly with an average accuracy of 16.65%, ranking the lowest. The performance of each model also varied on different subject areas, for example, Qwen2.5 - 72B scored higher on horticulture, while Gemini - 2.0 - Pro - exp -02 - 05 performed better on animal husbandry. Secondly, in the Five Shot and reasoning experimental conditions, the Yi - 34B model demonstrated the best performance, achieving an average accuracy of approximately 70.24%, while under Zero Shot and reasoning conditions, its average accuracy was 61.68%. These results indicated that providing a small number of learning examples can significantly improve the model's answer accuracy. On the other hand, it exhibited stronger Zero Shot learning ability. In addition, only Internlm2 - 7B performed the best for models with 10B scale parameters, where the accuracy

Table 3 Zero Shot and reasoning %

Model	Crop science	Horticulture	Plant protection	Animal husbandry	Forest science	Aquaculture science	Grass science	Average accuracy
Baichuan2 - 7B	42.93	39.49	54.58	52.71	44.76	41.29	41.11	45.27
Baichuan2 - 13B	48.99	41.80	55.25	58.53	45.40	47.26	43.15	48.63
DeepSeek - V3	76.01	54.73	85.08	78.29	66.35	69.65	81.34	73.06
DeepSeek - R1	77.18	56.35	88.17	78.68	70.79	72.14	85.13	75.49
Qwen1.5 - 14B	54.04	35.57	62.71	60.85	50.16	41.29	46.94	50.22
Qwen1.5 - 72B	61.11	45.27	64.41	58.52	60.00	52.74	61.22	57.61
Qwen2.5 - 72B	73.48	57.27	87.46	79.07	65.40	65.67	83.09	73.06
QWQ - 32B	75.00	56.58	88.14	79.84	68.57	65.67	83.67	73.92
Yi - 34B	56.57	48.50	75.93	68.22	60.00	50.25	72.30	61.68
Internlm2 - 7B	51.77	46.42	60.68	60.08	48.57	52.74	51.90	53.17
Internlm2 - 20B	49.49	44.34	60.00	61.63	47.30	49.25	51.90	51.99
ChatGLM3 - 6B	44.19	36.49	57.63	50.78	41.90	39.30	40.52	44.40
LongAlign - 13B	11.87	18.24	21.36	19.38	14.92	15.22	16.27	16.75
LLama2 - 7B	20.20	15.47	14.58	12.02	19.05	22.83	12.43	16.65
LLama2 - 13B	21.21	14.09	12.20	19.77	19.68	11.41	16.57	30.90
LLama2 - 70B	19.19	21.25	36.61	27.91	29.21	19.57	26.92	25.81
LLama3 - 8B	19.70	20.79	38.64	25.97	21.90	22.28	23.37	24.66
LLama3 - 70B	13.89	16.40	30.17	12.79	13.02	20.65	21.89	18.40
ChatGPT - 4o	55.30	48.50	70.17	68.22	51.75	55.72	61.22	58.69
Claude 3.7 Sonnet	71.86	53.35	82.37	72.87	61.59	62.19	79.88	69.16
Gemini - 2.0 - Pro	76.01	54.50	88.14	80.23	67.62	68.66	84.84	74.29
Shennong V2.0	98.23	81.29	95.93	92.25	86.03	97.01	93.88	92.08

Table 4 Five Shot and reasoning %

Model	Crop science	Horticulture	Plant protection	Animal husbandry	Forest science	Aquaculture science	Grass science	Average accuracy
Baichuan2 - 7B	49.49	43.42	58.64	62.40	48.89	47.76	47.52	51.16
Baichuan2 - 13B	42.68	40.18	57.29	55.81	44.76	44.28	46.06	47.29
Qwen1.5 - 14B	60.35	47.81	78.98	67.83	55.87	48.76	62.39	60.28
Qwen1.5 - 72B	70.20	54.73	88.81	73.64	65.08	61.19	80.47	70.59
Yi - 34B	67.68	54.27	88.14	74.42	63.81	63.18	80.17	70.24
Internlm2 - 7B	54.29	47.81	61.69	58.53	51.11	52.24	54.23	54.27
Internlm2 - 20B	55.05	47.34	65.76	63.57	50.48	48.76	53.64	54.94
ChatGLM3 - 6B	41.16	38.11	60.34	50.00	42.86	42.29	42.27	45.29
LongAlign - 13B	39.65	33.95	38.98	34.50	37.14	38.59	37.28	37.16
LLama2 - 7B	35.10	23.09	24.41	28.29	32.38	33.15	32.54	29.85
LLama2 - 13B	34.85	27.48	35.25	33.72	29.21	38.59	34.02	33.3
LLama2 - 70B	24.49	24.48	38.98	20.93	29.84	33.70	23.37	27.97
LLama3 - 8B	44.44	39.26	56.95	47.67	35.24	40.76	44.67	44.14
LLama3 - 70B	61.11	45.73	75.25	70.93	52.38	57.07	63.02	60.78

was 53% in the Zero Shot with inference condition and 54% in the Five Shot with inference condition.

As can be seen from the comparison between Table 3 and Table 5, Shennong V2.0 performed outstandingly under both experimental conditions. In Table 3, its average accuracy was 92.08%, while in Table 5, the average accuracy was increased to 91.82%. There were changes in the scores in various subject areas, such as crop science from 98.23% to 95.2%, horticulture from 81.29% to 98%, and aquaculture science from 97.01% to 84.1%.

2.3.2 Perplexity assessment of model results

Interpreting the results from Table 3 and Table 6, most of the model accuracies were improved when converting from an inference-based approach to a perplexity calculation approach. Among them, the most

Table 5　Zero Shot and reasoning under graded conditions　　%

Model	Crop science	Horticulture	Plant protection	Animal husbandry	Forest science	Aquaculture science	Grass science	Average
Baichuan2－7B	41.80	39.40	49.80	65.50	46.60	35.50	38.30	45.27
Baichuan2－13B	44.50	44.40	51.10	57.00	43.60	48.40	41.50	47.21
Qwen1.5－14B	57.90	31.20	68.70	55.50	54.00	44.80	46.90	51.28
Qwen1.5－72B	56.70	44.60	68.40	61.70	60.90	51.40	61.00	57.81
Yi－34B	56.10	52.60	74.30	71.20	60.00	53.60	70.80	62.65
Internlm2－7B	53.10	41.30	56.20	61.10	45.20	46.90	49.50	50.47
Internlm2－20B	53.40	46.50	61.40	60.20	46.00	49.30	50.20	52.42
ChatGLM3－6B	47.50	37.20	53.90	57.30	37.30	40.10	43.90	45.31
LongAlign－13B	10.80	18.10	20.20	12.10	13.00	15.70	17.30	15.31
LLama2－7B	20.50	17.50	14.60	9.60	20.80	21.70	12.40	16.72
LLama2－13B	18.40	11.80	15.90	17.80	17.70	10.40	13.10	15.01
LLama2－70B	19.60	16.50	32.20	33.10	23.60	16.00	22.40	23.34
LLama3－8B	16.60	20.90	37.20	27.50	26.50	16.80	18.00	23.35
LLama3－70B	13.60	18.30	31.40	14.00	11.50	24.30	20.30	19.05
Shennong V2.0	95.20	98.00	79.00	98.20	96.40	84.10	91.90	91.82

significant improvement was Qwen1.5－72B, which was increased by approximately 13 percentage points, from 57.61% to 71.15%. In addition, Qwen1.5－14B was improved from 50.22% to 60.29%, an improvement of about 10 percentage points, while Yi－34B was improved from 61.68% to 67.19%, an improvement of about 6 percentage points. The increase in accuracy was due to the use of the perplexity calculation. Perplexity was calculated by concatenating each option with the question, calculating the loss and selecting the one with the least loss as the answer to the question. This approach helped to avoid the large model illusion problem[21] and allowed for a more direct measure of the model's true ability. Thus, the accuracy of a model under this computational approach reflected its more realistic performance. However, not all models' accuracies were improved. For example, the accuracy of Interlm2－7B was not improved, but the loss it incurred was almost negligible.

3.3.3　Small sample learning rate review of models

Interpreted in terms of the results shown in Table 6 and Table 7, most of the models improved in accuracy when samples before entering the problem were given. Among them, the Yi－34B model was improved by 3 percentage points, and the rest of the models was improved to varying degrees.

Table 6　Zero Shot and perplexity　　%

Model	Crop science	Horticulture	Plant protection	Animal husbandry	Forest science	Aquaculture science	Grass science	Average
Baichuan2－7B	42.93	40.18	53.90	51.94	44.13	43.28	42.86	45.60
Baichuan2－13B	47.98	42.26	53.56	59.69	46.35	49.25	44.31	49.06
Qwen1.5－14B	58.59	46.65	76.61	71.32	56.19	50.25	62.39	60.29
Qwen1.5－72B	70.20	55.66	86.78	74.81	64.44	63.68	82.51	71.15
Yi－34B	61.87	52.89	85.76	72.09	61.90	59.70	76.09	67.19
Internlm2－7B	52.53	42.73	61.36	60.85	46.98	51.74	51.90	52.58
Internlm2－20B	52.02	47.11	66.10	65.50	47.94	47.26	53.94	54.27
ChatGLM3－6B	44.95	39.49	60.00	51.55	45.71	47.76	47.23	48.10
LongAlign－13B	33.59	30.48	41.36	32.95	33.97	37.50	36.98	35.26
LLama2－7B	32.07	25.87	27.46	27.91	34.60	33.15	35.21	30.90
LLama2－13B	33.59	27.94	30.85	32.56	33.65	39.13	31.95	32.81
LLama2－70B	38.38	26.56	36.61	38.37	32.38	37.50	33.43	34.75
LLama3－8B	40.66	40.88	50.85	53.49	36.19	45.11	40.83	44.00
LLama3－70B	60.35	46.19	73.90	67.44	52.70	56.52	61.24	59.76

Table 7 Five Shot and perplexity %

Model	Crop science	Horticulture	Plant protection	Animal husbandry	Forest science	Aquaculture science	Grass science	Average
Baichuan2 - 7B	40.66	38.11	56.27	56.20	44.13	46.27	44.90	46.65
Baichuan2 - 13B	52.02	43.88	57.97	62.40	48.89	47.26	46.94	51.34
Qwen1.5 - 14B	57.32	57.58	81.26	68.60	54.60	48.26	62.68	61.47
Qwen1.5 - 72B	73.48	54.04	87.46	73.64	66.35	58.71	81.63	70.76
Yi - 34B	67.42	53.81	90.51	73.26	64.76	64.68	79.88	70.62
Internlm2 - 7B	52.78	46.42	63.73	60.47	51.75	53.23	52.48	54.41
Internlm2 - 20B	53.03	44.80	66.10	64.73	48.89	49.25	54.52	54.47
ChatGLM3 - 6B	47.22	39.26	61.02	53.10	49.84	41.29	46.94	48.38
LongAlign - 13B	38.64	31.87	38.98	38.37	36.19	34.24	37.87	36.59
LLama2 - 7B	36.11	25.17	31.86	29.46	33.65	32.61	31.36	31.46
LLama2 - 13B	38.64	33.95	39.32	31.78	31.43	39.67	34.91	35.67
LLama2 - 70B	45.96	36.72	43.73	39.92	35.24	36.96	36.69	39.32
LLama3 - 8B	43.94	38.80	55.59	50.00	38.10	45.11	46.15	45.38
LLama3 - 70B	59.85	45.50	73.90	68.60	53.02	55.43	64.79	60.16

3 Conclusion

(1) Agri-Eval, designed as a benchmark specifically for the agricultural field, featured a comprehensive evaluation dataset comprising 2 283 questions that spaned seven major agricultural disciplines: crop science, horticulture, plant protection, animal husbandry, aquaculture science, grass science, and forestry science. Within each discipline, questions were categorized into three levels based on their importance to actual agricultural production, along with a corresponding scoring mechanism. This approach provided a holistic test of models' agricultural knowledge and reasoning abilities.

(2) In the evaluation, general-purpose LLMs with hundreds of billions of parameters, such as DeepSeek - R1, achieved an accuracy rate as high as 75% without using the graded scoring mechanism. Qwen1.5 - 14B, with less than 20 billion parameters, also surpassed an accuracy rate of 60%. It was particularly worth noting that Shennong, the agricultural-specific large model, outperformed all general-purpose LLMs in terms of agricultural knowledge answering accuracy, regardless of whether the graded scoring mechanism was used. This highlighted the distinct advantages of vertical-domain models in specific fields. Moreover, after employing the graded scoring mechanism, the scores for each model varied. This demonstrated that the evaluation of agricultural LLMs should take into account the actual production scenarios and their relative importance.

(3) The launch of Agri-Eval not only helped R&D developers to comprehensively evaluate and improve the application capability of LLMs in the agriculture domain, but also promoted the further LLMs in the field of agricultural science research and development. In the future, with the advancement of agricultural science and the innovation of language modelling technology, Agri-Eval should also be continuously updated to meet new assessment needs and promote interdisciplinary research.

References

[1] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv Preprint, arXiv:2303.18223, 2023.

[2] ZHAO H, CHEN H, YANG F, et al. Explainability for large language models: a survey[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(2): 1 - 38.

[3] CHANG Y, WANG X, WANG J, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1 - 45.

[4] YIN S, FU C, ZHAO S, et al. A survey on multimodal large language models[J]. arXiv Preprint, arXiv:2306.13549, 2023.

[5] CHEN Z, DU W, ZHANG W, et al. T-eval: evaluating the tool utilization capability step by step[J]. arXiv Preprint, arXiv: 2312.14033, 2023

[6] ONO K, MORITA A. Evaluating large language models: ChatGPT - 4, Mistral 8x7B, and Google Gemini benchmarked against

MMLU[J/OL]. Authorea Preprints, 2024. DOI: 10.36227/techrxiv.170956672.21573677/v1.
[7] BAI Y, LV X, ZHANG J, et al. Longbench: a bilingual, multitask benchmark for long context understanding[J]. arXiv Preprint, arXiv:2308.14508, 2023.
[8] ZHONG Q, WANG K, XU Z, et al. Achieving > 97% on GSM8K: deeply understanding the problems makes LLMs perfect reasoners[J]. arXiv Preprint, arXiv:2404.14963, 2024.
[9] LI D, MURR L. HumanEval on latest GPT models—2024[J]. arXiv Preprint, arXiv:2402.14852, 2024.
[10] VAN DER HEIJDEN J. Unravelling programme success and complex causation in agricultural research for development (AR4D): a systematic and comprehensive literature review[J]. Agricultural Systems, 2024, 215: 103851.
[11] HUANG Y, BAI Y, ZHU Z, et al. C-Eval: a multi-level multi-discipline Chinese evaluation suite for foundation models[J]. Advances in Neural Information Processing Systems, 2024, 36: 62991 - 63010.
[12] POHANKOVÁ E, HLAVINKA P, KERSEBAUM K C, et al. Expected effects of climate change on the soil organic matter content related to contrasting agricultural management practices based on a crop model ensemble for locations in Czechia[J]. European Journal of Agronomy, 2024, 156: 127165.
[13] FENG G, ZHANG B, GU Y, et al. Towards revealing the mystery behind chain of thought: a theoretical perspective[J]. Advances in Neural Information Processing Systems, 2024, 36: 70757 - 70798.
[14] TANG B, ZHANG J, YAN L, et al. Data-free generalized zero-shot learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 5108 - 5117.
[15] SUCHOLUTSKY I, GRIFFITHS T. Alignment with human representations supports robust few-shot learning[J]. Advances in Neural Information Processing Systems, 2024, 36: 73464 - 73479.
[16] YU C, ZANG L, WANG J, et al. Token-free LLMs can generate Chinese classical poetry with more accurate format[J]. arXiv Preprint, arXiv:2401.03512, 2024.
[17] YOUNG A, CHEN B, LI C, et al. Yi: open foundation models by 01. AI[J]. arXiv Preprint, arXiv:2403.04652, 2024.
[18] DONG X, ZHANG P, ZANG Y, et al. InternLM - XComposer2: mastering free-form text-image composition and comprehension in vision-language large model[J]. arXiv Preprint, arXiv:2401.16420, 2024.
[19] XIA X, DONG S. Optimizing inference abilities in Chinese NLP: a study on lightweight generative language models for knowledge question answering[C]// 2024 4th International Conference on Neural Networks, Information and Communication (NNICE). IEEE, 2024: 359 - 363.
[20] MUA T. Chatglm-6b-pl: application of parallel reinforcement models for LLMs in unbalanced medical small sample datasets [M]// Intelligent computing technology and automation. Amsterdam: IOS Press, 2024.
[21] VERMA M, BHAMBRI S, KAMBHAMPATI S. Theory of mind abilities of large language models in human-robot interaction: an illusion? [C]// Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. 2024: 36 - 45.
[22] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: open and efficient foundation language models[J]. arXiv Preprint, arXiv:2302.13971, 2023.
[23] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[J]. arXiv Preprint, arXiv:2307.09288, 2023.

Agri-Eval:农业领域大语言模型多层次评估基准

王耀君　葛明亮　徐国威　张齐豫　别宇辉
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 利用基准数据集对模型进行评估,是衡量大语言模型(LLMs)在特定领域能力的重要方法,主要用于评测其知识水平与推理能力。为更好地评估大语言模型在农业领域的能力,本文提出了 Agri-Eval:一个用于评估农业领域大语言模型知识与推理能力的基准。Agri-Eval 的评测数据集涵盖农业领域 7 个主要学科:作物科学、园艺学、植物保护学、畜牧学、林学、水产科学和草业科学,共包含 2 283 道试题。在国内通用大语言模型中,DeepSeek - R1 表现最佳,准确率达 75.49%;在国际通用大模型中,Gemini - 2.0 - Pro - exp - 02 - 05 以 74.28% 的准确率位居首位。作为农业垂直领域大模型,神农 V2.0(Shennong V2.0)的综合表现超越了所有国内通用大模型,其在农业知识问答的准确率亦优于所有现有的通用模型。Agri-Eval 的发布有助于开发者通过多样化任务与测试,全面评估模型在农业领域的综合能力,从而推动农业领域大语言模型的发展。

关键词: 大语言模型; 评估体系; 农业知识; 农业数据集

基金项目: 国家重点研发计划项目(2024YFD2000805)