禹冰,文纪威,张霄鹏,等.基于畜牧业LIVED基准的大语言模型评估[J].华中农业大学学报,2026,45(3):137-147.
DOI:10.13300/j.cnki.hnlkxb.2026.03.012

基于畜牧业LIVED基准的大语言模型评估

禹冰[1],文纪威[1],张霄鹏[1],杜小勇[1],李国亮[1],刘丑生[2],邱小田[2],方亚平[1]

1.作物遗传改良全国重点实验室/湖北洪山实验室/农业农村部智慧养殖技术重点实验室/
农业智能技术教育部工程研究中心/湖北省农业大数据工程技术研究中心/
华中农业大学信息学院,武汉 430070; 2.全国畜牧总站,北京 100125

摘要 为促进大语言模型更好地服务于畜牧业,系统评估了11个通用大语言模型和3个畜牧领域专用模型在专业化考试基准数据集中的表现。构建的LIVED(livestock knowledge evaluation dataset)数据集包含11 037道中文畜牧业知识评估数据集,涵盖选择题、判断题、填空题与问答题4种类型。评估结果显示,各大语言模型性能差异显著,综合性能排名前3的模型分别是DeepSeek-R1(准确率0.700)、Kimi K2(准确率0.681)和DeepSeek-V3.1(准确率0.675)。DeepSeek-R1的选择题准确率最高(0.868),GPT-5的判断题准确率最高(0.842),Kimi K2的填空题准确率最高(0.651),问答题Qwen3-235B-A22B的准确率最高(0.504)。畜牧业专用模型中,神农百晓大模型整体性能最佳(0.630),但仍落后DeepSeek-R1 7百分点。错题分析显示,当前大语言模型的错误主要富集在畜牧经济与管理、畜牧生产管理等方面。结果表明,大语言模型在专业知识准确性与推理能力方面仍存在明显不足,未来垂直大模型研发需优先解决专业知识与推理性能的短板。

关键词 大语言模型; 畜牧领域; 问答对; 基准测试; 评估

中图分类号 TP18 **文献标识码** A **文章编号** 1000-2421(2026)03-0137-11

大语言模型(large language models,LLMs)已在医疗、法律与金融等关键领域展现出重要价值。依据功能特性,大语言模型可分为多模态(如GPT-5)[1]、工具增强(如WebGPT)[2]及领域专用(如FinGPT)[3]3类。它们在推理、代码生成等任务上各具优势,但也面临领域知识缺失和泛化能力不足的挑战。

为推动模型能力的发展,学术界已构建多种评测基准,覆盖医学(PubMedQA[4]和CMExam[5])、法律(LAiW)[6]、金融(FinEval)[7]等多个专业领域。现有通用基准(如MMLU[8])难以针对性地评测模型在特定专业领域的表现。由于缺乏针对大语言模型在畜牧领域专业能力的系统评估,相关技术的应用落地与深化研究均面临挑战。

为评估大语言模型在畜牧领域的知识水平与应用潜力,本研究基于自建的高质量中文畜牧业问答数据集(livestock knowledge evaluation dataset,LIVED),对11个通用及3个领域垂直模型进行综合性能评估,分别比较这14个大模型在选择题、判断题、填空题和问答题上的准确性,系统考察各模型在畜牧领域背景下的适用性与性能差异,旨在为大语言模型在畜牧领域的实际应用提供客观依据,同时为后续模型在专业知识覆盖与领域适配方面的优化提供方向。

1 材料与方法

1.1 数据获取

为确保畜牧知识的全面性与层次性,从标准化考试的角度收集了职业院校、本科专业、执业兽医3个层级共147套考试试卷及答案,具体信息如表1所示。首先是13套职业院校考试试题,难度最低,侧重于考查对畜牧领域的基础知识掌握与初步认知。其

收稿日期: 2025-09-30
基金项目:国家重点研发计划项目(2023ZD04061);中央高校基本科研业务费专项(2662025PY018;2662025DKPY005);湖北省技术创新计划项目(2024BBA001);种猪平台项目(221107)
禹冰,E-mail: yb_20020727@webmail.hzau.edu.cn
通信作者: 方亚平,E-mail:ypfang@mail.hzau.edu.cn

次为中等难度的本科专业试题，旨在评估对畜牧领域系统化理论的掌握程度。最后是难度最高的22套执业兽医资格考试题，重点考查在贴近真实临床场景下的复杂问题解决、推理与决策能力。

表1 试题来源及题型分布

Table 1 Sources of test items and distribution of question types

试题来源 Sources of test items	试卷数量 Number of tests	题目数量 Number of questions	题目类型 Question types			
			选择题 Multiple-choice	判断题 Judgment	填空题 Completion	问答题 Open-ended
职业院校考试 Vocational college examinations	13	930	550	145	132	103
本科专业考试 Undergraduate major examinations	112	11 918	5 830	2 289	1 968	1 831
执业兽医资格考试 Veterinary licensing examinations	22	3 618	3 045	398	96	79
合计 Total	147	16 466	9 425	2 832	2 196	2 013

最终整合得到的语料库共包含147套试卷，16 466个问答对。其中选择题占比最高(57.24%)，其次为判断题(17.20%)、填空题(13.34%)和问答题(12.22%)。

1.2 数据处理

原始试题和答案组成的语料库经过4步处理流程转化为结构化数据集LIVED。

1)文本提取。对图像等不可编辑文件，使用Pytesseract包进行光学字符识别(OCR)[9]。

2)初步筛选。人工匹配试题与答案，并剔除1 779道与主题无关、不完整或没有参考答案的题目。

3)冗余去除。利用BERT模型生成文本嵌入向量，移除余弦相似度高于0.95的3 650道重复问题，旨在去除语义上完全重复的题目，并保留语义上有价值的表述变体。

4)格式校验。将所有问答对统一转换为JSON格式，并经人工最终校验，保留11 037个有效问答对。

1.3 评估指标

LIVED数据集包含的4种题型可归为2类：客观题(选择题、判断题)和主观题(填空题、问答题)。每类题目都使用准确率作为评估指标，2类题型的计算方式略有不同。对于答案明确的客观题，直接计算准确率(accuracy，A)，其定义如公式(1)所示：

$$A=\frac{n_{\text{correct}}}{N_{\text{total}}} \tag{1}$$

式(1)中，n_{correct}为正确作答的题目数，N_{total}为题目总数。该指标能有效衡量模型对固定知识点的理解程度，其值域为[0,1]，值越大代表性能越优。

对于答案多样的主观题，首先采用余弦相似度(cosine similarity，cos$S$)评估模型生成答案与标准答案的语义吻合度，接着设置余弦相似度阈值(T)来计算准确率。计算公式如公式(2)和公式(3)所示。

$$\cos S=\frac{E_{\text{ref}}\cdot E_{\text{cand}}}{||E_{\text{ref}}||\cdot||E_{\text{cand}}||} \tag{2}$$

$$A=\frac{n_{\cos S\geqslant t}}{N_{\text{total}}} \tag{3}$$

式(2)中，E_{ref}和E_{cand}分别对应标准答案与模型生成答案的嵌入向量，由paraphrase-multilingual-MiniLM-L12-v2模型[10]编码生成。其得分区间为[−1,1]，值越大则语义相似度越高。式(3)中，$n_{\cos S\geqslant T}$表示余弦相似度大于等于阈值的题目数量，T代表阈值，N_{total}为题目总数。为评估模型整体性能，分别采用0.7、0.8与0.9三个余弦相似度阈值计算主观题准确率(相似度≥阈值判定为正确)，结果表明，尽管阈值升高导致准确率普遍下降，但所有模型的整体性能排名保持稳定，本研究统一采用0.8作为阈值。

为评估不同模型在主观题上作答表现的一致性，采用皮尔逊相关系数(Pearson correlation coefficient，r)[11]来量化模型间的表现关联度。计算公式如下式(4)所示：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

式(4)中，x_i与y_i分别表示2个不同模型在第i道题目上的余弦相似度，$\bar{x}$和$\bar{y}$则分别代表各自模型在所有题目上得分的均值。r可以揭示不同模型在执行相同任务时，其表现之间关联的强度与方向，取值范围为−1～1，越接近1，表明模型间一致性越强；值接近0，则表示弱相关或无相关；若为负值，则说明两者呈负相关关系。

2　结果与分析

为系统性评估通用领域模型与畜牧领域专用模型的性能表现，基于LIVED数据集对14个模型进行综合评估。首先，收集3类试卷4种题型构建包含11 037个问答对的数据集。其次，针对不同题型的问答对构建答题准确性评估指标。最后，比较14种大语言模型答题的准确性[12]，衡量其在特定题型上的性能。

2.1　数据集统计

经数据收集、处理、筛选及人工校对后，初始的16 466道试题被精炼为包含11 037个高质量问答对的LIVED数据集。该数据集中选择题占比最大，共6 592道（59.7%）；其次为判断题1 820道（16.5%）、填空题1 750道（15.9%）和问答题875道（7.9%）。具体数据格式如图1所示。该数据集规模可观且题型多样，能够支持从多角度全面评估大语言模型在知识掌握、理解、推理与生成等方面的能力。

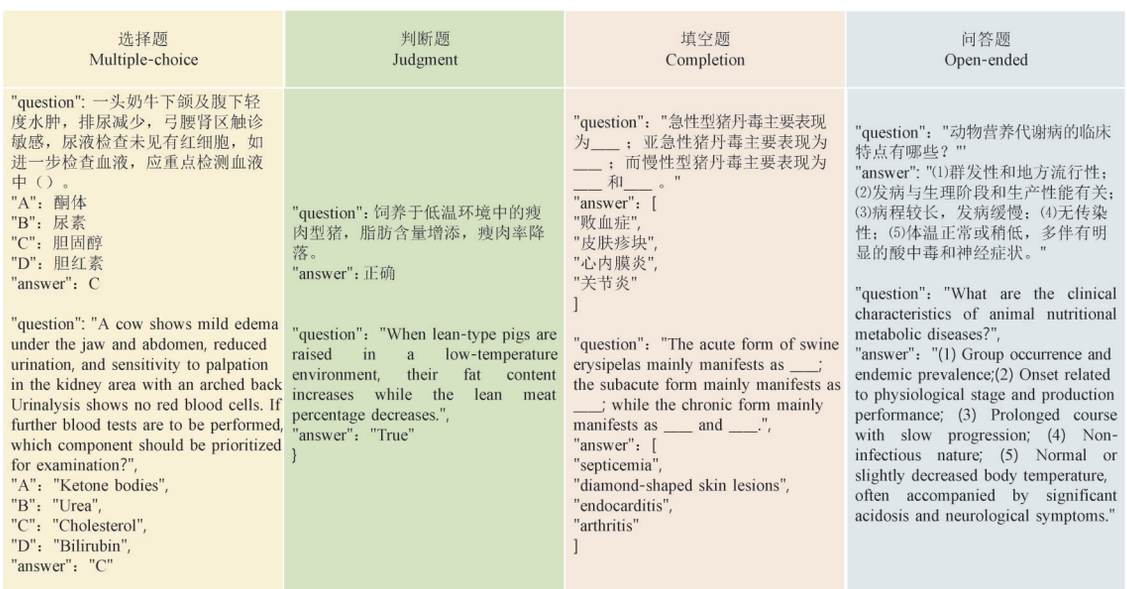


图1　不同问题类型的格式示例

Fig.1　Example formats of different question types

2.2　模型选择

选取11个广泛使用的通用领域大模型及3个畜牧领域大模型（表2），评估它们在LIVED数据集上的性能。在畜牧领域大模型的选择过程中，排除了专精于某类农业而非畜牧领域的模型，以及部分未开放公开测试的畜牧专用模型，以确保评估的可行性与针对性。通用领域大模型包括国际知名模型，如GPT系列（OpenAI）[13]、Llama3.3（Meta）[14]，以及国内领先模型，如Qwen系列（阿里云）[15]和DeepSeek系列（DeepSeek）[16]。

畜牧领域模型包括由雄安新区农业农村局研发的雄小农模型（该模型融合DeepSeek-R1以增强推理与决策能力，https://www.xiongan.gov.cn/20250228/8bad33753ede42cf8048ff71119748ba/c.html，）、由哈尔滨工业大学基于Qwen1.5-14B-Chat开发的KwooLa模型（https://github.com/HIT-Kwoo/KwooLa）以及由中国农业大学研发的神农百晓模型（https://news.cau.edu.cn/zhxwnew/79fe95cb93ec4983b2b6035c5e2f5271.htm）。这些模型基于领域专用数据训练微调，专为畜牧业知识体系构建。

1）通用领域模型评估。11种通用领域模型评估结果如表3所示，总体上，国内大模型在LIVED中文畜牧数据集上的表现优于国外大模型。尤其是DeepSeek-R1在此次评估中取得了最优的平均性能（0.700），在所有评估模型中排名第1。在选择题（0.868）、判断题（0.839）中准确率均超过0.8，表明该模型在处理结构化任务时，能够准确提取并解读关键信息。然而，该模型在问答题上的表现相对欠佳，准确率仅为0.462，反映出其在综合分析与自然语言生成方面存在局限。

Kimi K2综合排名第2，在填空题上性能最佳（0.651）。DeepSeek-V3.1模型在本次评估中排名第3，但是在4种题型上的性能均不突出，展现了其在处理不同问题上的均衡能力。Qwen3-235B-A22B综

表2 待评估的大语言模型列表

Table 2 Large language models for evaluation

	模型 Models	开发者 Developer	参数量 Number of parameters/B	网址 Websites
通用大模型 General-purpose LLMs	GPT-4o	OpenAI	200	https://openai.com
	GPT-5	OpenAI	\	https://openai.com
	Gemini-2.5-flash	Google	\	https://ai.google.dev
	Llama3.3*	Meta	70	https://huggingface.co meta-llama/Llama-3.3-70B-Instruct
	DeepSeek-R1*	DeepSeek	671	https://www.deepseek.com
	DeepSeek-V3.1*	DeepSeek	685	https://www.deepseek.com
	ERNIE-3.5	百度 Baidu	\	https://developer.baidu.com
	Kimi K2*	Moonshot AI	1 000	https://platform.moonshot.cn
	Moonshot-V1	Moonshot AI	\	https://platform.moonshot.cn
	Qwen2.5-72B*	阿里云 Alibaba Cloud	72	https://www.aliyun.com
	Qwen3-235B-A22B*	阿里云 Alibaba Cloud	235	https://www.aliyun.com
领域专用模型 Domain-specific models	雄小农 Xiongxiaonong	雄安新区农业农村局 Xiong'an New Area Agriculture and Rural Affairs Bureau	\	https://www.xaxxn.cn
	天工开悟 KwooLa	哈尔滨工业大学 Harbin Institute of Technology	\	https://www.tgkwai.com
	神农百晓 Shennong Baixiao	中国农业大学 China Agricultural University	\	https://shennong.cc/welcome

注：符号 \ 表示参数量未公开，符号 * 代表该模型为开源模型。下同。Note: The symbol \ represents that the number of parameters is unavailable. The superscript * indicates an open source model. The same below.

表3 通用领域模型的评估结果

Table 3 Evaluation results of general-purpose models

	模型 Models	选择题 Multiple-choice	判断题 Judgment	填空题 Completion	问答题 Open-ended	平均值 Average
国外大模型 International LLMs	GPT-4o	0.773	0.751	0.499	0.465	0.622
	GPT-5	0.849	**0.842**	0.577	0.406	0.669
	Gemini-2.5-flash	0.827	0.810	0.539	0.379	0.639
	Llama3.3	0.706	0.696	0.351	0.398	0.538
国内大模型 Domestic LLMs	DeepSeek-R1	**0.868**	0.839	0.631	0.462	**0.700**
	DeepSeek-V3.1	0.814	0.779	0.637	0.471	0.675
	ERNIE-3.5	0.730	0.737	0.527	0.433	0.607
	Kimi K2	0.856	0.821	**0.651**	0.395	0.681
	Moonshot-V1	0.664	0.711	0.495	0.479	0.587
	Qwen2.5-72B	0.806	0.771	0.502	0.441	0.630
	Qwen3-235B-A22B	0.831	0.782	0.575	**0.504**	0.673

注：粗体表示最优性能。Accuracy作为4种题型的评估指标。平均值指Accuracy的算术平均值。Note: Bold font indicates superior performance. Accuracy is used as the evaluation metric for the four question types. Average refers to the arithmetic mean of the accuracies.

合排名第4，在问答题上取得最高准确率(0.504)。国外大模型中GPT-5表现最好但综合排名第5，在判断题上取得最高准确率(0.842)。Moonshot-V1和Llama3.3的整体表现相对较弱，其平均性能分别为0.587和0.538。其中Moonshot-V1在选择题中准确率为0.664，但在问答题中的准确率达到0.479。这说明尽管该模型在处理结构化任务时表现不佳，却在语言生成与上下文理解方面具有较好能力。上述结果表明，不同模型各具优势，可根据具体任务需求进行针对性选用。同时DeepSeek-R1和DeepSeek-V3.1评估结果说明，模型的版本更新未必会提升模型性能。

2)领域专用模型评估。对特定领域开展评估能够更准确地衡量模型的领域专业知识、增强其实际应用的可靠性,并为领域适应性优化提供关键见解。为此,选取3个畜牧业专用模型(雄小农、天工开悟与神农百晓)进行评估。如表4所示,这3个模型在畜牧业领域中展现出明显的性能差异。神农百晓模型性能最佳(0.630),选择题的准确率为0.798,判断题准确率为0.769,填空题上准确率为0.546,领先于其他2个模型。天工开悟模型在问答题上取得了最高的准确率(0.422),体现了其在长文本生成与知识整合方面的优势。然而,该模型在填空题上的性能(0.307)相对较弱,暴露出其在术语抽取与精确信息保留方面存在局限。值得关注的是,雄小农模型在测试中存在生成无效响应的现象,主要分为空白响应与错误响应2类。此类无效响应会显著影响模型性能表现。空白响应指模型返回内容为空,错误响应指模型返回“我是农业技术问答助手……”这类无关输出。为排除这一干扰,需在剔除无效响应后重新评估雄小农模型的性能。剔除后其选择题准确率从0.399显著提升至0.689,其他题型表现亦有不同程度改善。

表4　领域专用模型的评估结果(准确率)

Table 4　Evaluation results (accuracy) of domain-specific models

模型 Models	选择题 Multiple-choice	判断题 Judgment	填空题 Completion	问答题 Open-ended	平均值 Average
雄小农 Xiongxiaonong	0.399(0.689)	0.631(0.690)	0.304 (0.305)	0.417 (0.441)	0.438(0.531)
天工开悟 KwooLa	0.674	0.678	0.307	**0.422**	0.520
神农百晓 Shennong Baixiao	**0.798**	**0.769**	**0.546**	0.408	**0.630**

注:括号内数值表示雄小农模型排除无效回答后的性能表现。Note: Values in parentheses indicate the performance of Xiongxiaonong after excluding invalid responses.

以上结果表明,不同模型在各任务类型上的表现存在显著差异。各模型在客观题上的性能明显优于主观题。此外,尽管雄小农能够生成比天工开悟更高质量的响应,但其整体性能受频繁产生的无效响应所制约,凸显了建立有效响应验证机制的必要性。

2.3　影响大语言模型性能的因素

1)模型参数对性能的影响。为深入探讨模型参数量对大语言模型(LLMs)性能[17]的影响,本研究选取Qwen2.5与DeepSeek-R1及其在不同参数配置下的变体进行分析,结果如图2所示。对Qwen2.5系列模型进行分析,发现选择、判断和填空题上模型性能通常随参数规模增大而稳定提升。然而,问答题上变化趋势并非如此,图2的结果显示,问答题上

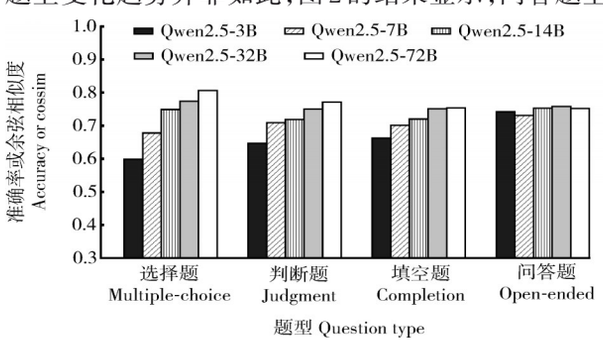


图2　不同参数量的Qwen2.5系列模型在同一题型下的性能差异

Fig.2　Performance differences among Qwen2.5 series models with different parameter sizes on the same question

模型参数增加反而会导致性能降低,如Qwen2.5-32B的余弦相似度为0.758,Qwen2.5-72B却降至0.752。为深入探究这一现象,对Qwen2.5-32B与Qwen2.5-72B生成答案的绝对长度进行定量分析,两者平均响应长度分别为76.10与79.51个词元,表明模型规模增大时,其回答趋于冗长。这种回答长度的增加可能导致模型生成答案与标准答案之间的余弦相似度下降。因此,在实际应用中选择合适的模型规模,不仅有助于保障模型性能,也能降低硬件负担,实现更优的资源调配与系统效率。

2)领域专用模型与通用领域模型的性能比较。为比较通用模型与领域专用模型在畜牧业任务中的性能差异,参照文献[18],计算2类模型在4种题型上的平均准确率,并进行统计分析。如表5所示,在选择题上,通用领域模型的平均准确率为0.793,显著高于领域专用模型的0.623。配对样本t检验表明二者差异具有统计学意义($P<0.001$),通用领域模型在该题型上平均准确率领先0.170。在判断题上,通用领域模型的平均准确率为0.776,领域专用模型为0.692,差异显著($P<9.165\times10^{-39}$),平均差距为0.084。在填空题上,通用领域模型的平均准确率为0.544,高于领域专用模型的0.386,差异显著($P<2.217\times10^{-98}$),准确率差距为0.158。在问答题上,通用领域模型的平均准确率为0.439,略高于领域专用模型的0.416,该差异同样具有统计学意义($P<5.955\times10^{-3}$),准确率差距为0.023。

表5 领域专用模型与通用领域模型的性能(准确率)对比

Table 5 Comparison results of domain-specific and general-purpose models performance(accuracy)

模型 Models	选择题 Multiple-choice	判断题 Judgment	填空题 Completion	问答题 Open-ended
通用模型 General-purpose models	0.793	0.776	0.544	0.439
领域专用模型 Domain-specific models	0.623	0.692	0.386	0.416

3)大型语言模型的一致性分析。前述研究中雄小农模型存在生成无效响应的问题,这对整体结果产生了显著影响。因此,为保证分析的有效性,后续的一致性分析将排除该模型。首先,将剩余13个模型的回答划分为3类:全部正确(所有模型均回答正确)、全部错误(所有模型均回答错误)以及部分正确(模型回答正确性不一致)。后续针对部分正确题目展开详细分析。部分正确/错误题型对应的题目数量分别为:选择题3 798道、判断题1 128道、填空题1 318道、问答题614道。

如表6所示,选择题上,DeepSeek-R1表现出最稳定的性能,其准确率从0.868小幅下降至0.821,降幅为0.047。相比之下,Moonshot-V1的性能下降幅度较大,从0.664降至0.468。值得注意的是,多个性能领先的模型,包括Kimi K2(−0.056)、GPT-5(−0.060)和Qwen3-235B-A22B(−0.073)降幅相对较小,保持了其相对优势。判断题呈现与选择题相似的变化趋势。

填空题的结果则显示出不同的模式:对部分正确的问题,大多数模型的正确率有所提升。其中Kimi K2提升最为显著,从0.651升至0.706,增幅为0.055。仅KwooLa(−0.057)和Llama3.3(−0.043)出现下降。类似地,问答题也呈现上升趋势。客观题与主观题之间变化趋势的显著差异,可能源于二者评分机制的不同。

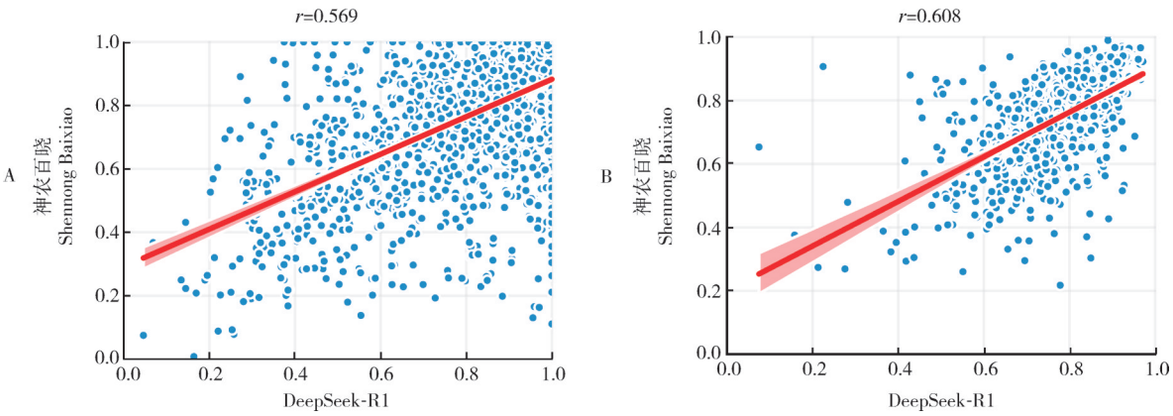
将表现最优的通用模型DeepSeek-R1与专用模型神农百晓进一步比较,结果显示,两者在客观题上回答的一致性较高;在选择题中,2个模型共同答对4 978题,共同答错593题;在判断题中,共同答对1 280题,共同答错175题。在主观题上,基于2个模型对同一问题回答的余弦相似度计算皮尔逊相关性系数,两者表现出中等程度的正相关(填空题r=0.569,问答题r=0.608),表明其回答策略存在差异(图3)。

表6 选择题和填空题上模型性能的变化

Table 6 Changes in model performance on multiple-choice and completion questions

模型 Models		选择题 Multiple-choice			填空题 Completion		
		准确率a Accuracy a	准确率b Accuracy b	差异 Difference	准确率a Accuracy a	准确率b Accuracy b	差异 Difference
国外大模型 International LLMs	GPT-4o	0.773	0.657	−0.116	0.499	0.505	0.006
	GPT-5	0.849	0.789	−0.060	0.577	0.607	0.03
	Gemini-2.5-flash	0.827	0.750	−0.077	0.539	0.558	0.019
	Llama3.3	0.706	0.541	−0.165	0.351	0.308	−0.043
国内大模型 Domestic LLMs	DeepSeek-R1	0.868	0.821	−0.047	0.631	0.680	0.049
	DeepSeek-V3.1	0.814	0.727	−0.087	0.637	0.687	0.05
	ERNIE-3.5	0.730	0.582	−0.148	0.527	0.541	0.014
	Kimi K2	0.856	0.800	−0.056	0.651	0.706	0.055
	Moonshot-V1	0.664	0.468	−0.196	0.495	0.499	0.004
	Qwen2.5-72B	0.806	0.714	−0.092	0.502	0.508	0.006
	Qwen3-235B-A22B	0.831	0.758	−0.073	0.575	0.606	0.031
	天工开悟 KwooLa	0.674	0.485	−0.189	0.307	0.250	−0.057
	神农百晓 Shennong Baixiao	0.798	0.700	−0.098	0.546	0.566	0.02

注:准确率a是大模型在相应题型上所有问题的准确率,准确率b是大模型在相应题型上有争议问题的准确率。差异定义为准确率a减去准确率b。Note:"Accuracy a" reflects the model's accuracy on all questions of the corresponding question type. "Accuracy b" refers the model's accuracy on controversial questions of the corresponding question type. "Difference" is defined as "Accuracy a" minus "Accuracy b".



A：模型在填空题回答的余弦相似度分布；B：模型在问答题回答的余弦相似度分布。A：The cosine similarity of responses for completion questions；B：The cosine similarity of responses for open-ended questions.

图 3　DeepSeek-R1 与神农百晓模型在主观题型上的回答一致性比较

Fig.3　Comparison of response consistency between the DeepSeek-R1 and Shennong Baixiao models on subjective question types

利用统计学分析 2 个模型间的性能差异。结果显示，DeepSeek-R1 的性能显著优于神农百晓。4 种题型统计学分析对应 P 值分别为 6.601×10^{-48}、2.300×10^{-11}、9.201×10^{-12} 与 1.95×10^{-3}。这表明，在所有题型上，2 个模型之间的性能差异均具有统计学意义。

2.4　模型稳定性与可复现性的比较分析

为评估模型结果的稳定性与可重复性，对国内外表现最优的模型——DeepSeek-R1 与 GPT-5 分别进行了 3 次独立重复实验，对结果的一致性进行分析。

在客观题上，2 个模型均表现出较强的稳定性，且 GPT-5 在准确率上略占优势。由表 7 可见，GPT-5 在选择题上的平均准确率为 0.834 4，略高于 DeepSeek-R1 的 0.820 0，说明其具有更优的精确度与更强的收敛性。在判断题方面，2 个模型性能相当，GPT-5 为 0.783 8，略高于 DeepSeek-R1 的 0.773 7。值得关注的是，尽管 GPT-5 的平均准确率较高，但其在判断题上的标准差显著大于 DeepSeek-R1，表明其在该类任务中的性能波动更大。

表 7　DeepSeek-R1 和 GPT-5 在客观题上 3 次重复实验结果

Table 7　Results of three repeated experiments on objective questions for DeepSeek-R1 and GPT-5

题型 Question types	模型 Models	重复实验 Repeated experiments			均值 Mean	标准差 Standard deviation
		实验 1 Experiment 1	实验 2 Experiment 2	实验 3 Experiment 3		
选择题 Multiple-choice	DeepSeek-R1	0.818 3	0.828 3	0.813 3	0.820 0	6.236×10^{-3}
	GPT-5	0.833 3	0.833 3	0.836 7	0.834 4	1.603×10^{-3}
判断题 Judgment	DeepSeek-R1	0.781 8	0.769 7	0.769 7	0.773 7	5.704×10^{-3}
	GPT-5	0.769 7	0.787 9	0.793 9	0.783 8	10.29×10^{-3}

在主观题评估方面，采用定量指标与分布图（图 4）相结合的方法衡量语义稳定性。结果显示，DeepSeek-R1 在填空题（0.823）和问答题（0.766）上的平均余弦相似度略高于 GPT-5（分别为 0.794 和 0.744）。然而，GPT-5 在这 2 项任务中均表现出更低的标准差。这一特征在图 4 的分布曲线中得到直观体现：GPT-5 的曲线通常更高、更窄，峰值更为集中。这表明，虽然 DeepSeek-R1-0528 的回答在平均意义上更接近标准答案，但 GPT-5 在多次回答同一问题时表现出更高的语义与内容一致性。

2.5　错题分析

为探究大语言模型在畜牧业领域的知识盲区，进一步分析了所有被归类为“全部错误”与部分正确的共 7 499 道题目。

LIVED 数据集中的题目被划分为 5 个主题领域：畜牧生产管理（2 761 题）、疾病防控与兽医保健（4 858 题）、育种与繁殖技术（2 289 题）、畜牧经济与管理（109 题）及其他（1 020 题）。各主题题目数量及比例分布反映出五大主题之间存在明显不均衡现象。如表 8 所示，“畜牧经济与管理”主题中“全部错

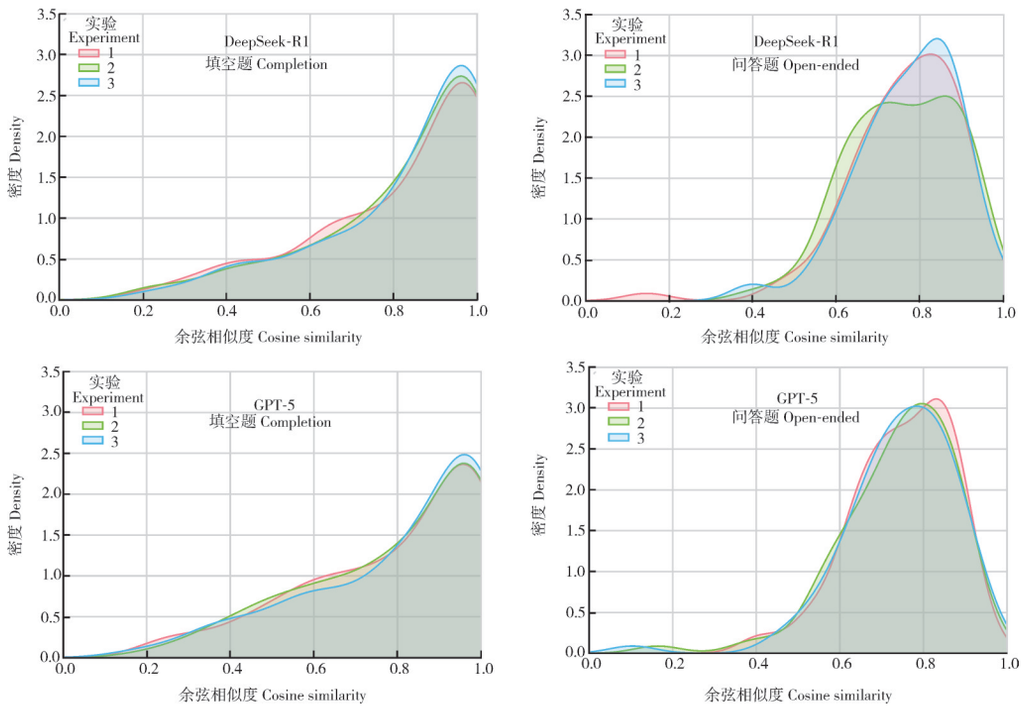


图4 DeepSeek-R1和GPT-5在主观题上3次重复实验余弦相似度分布

Fig.4 Cosine similarity distribution of three repeated experiments on subjective questions for DeepSeek-R1 and GPT-5

表8 5个主题类别的错误率

Table 8 Error rates across the five thematic categories

主题 Theme	畜牧生产管理 Livestock production management	疾病防控与兽医保健 Disease control and veterinary health	育种与繁殖技术 Breeding and reproduction technology	畜牧经济与管理 Livestock economics and management	其他 Others
原数据集 Volume of raw data	2 761	4 858	2 289	109	1 020
全部错误问题 All incorrect response (比例)(Percentage)	207 (7.50%)	241 (4.96%)	130 (5.68%)	11 (10.09%)	52 (5.10%)
有争议问题 Controversial questions (比例)(Percentage)	1 851 (67.04%)	2 722 (56.03%)	1 677 (73.26%)	76 (69.72%)	532 (52.16%)
整体错误率 Overall error rate	74.54%	60.99%	78.94%	79.81%	57.26%

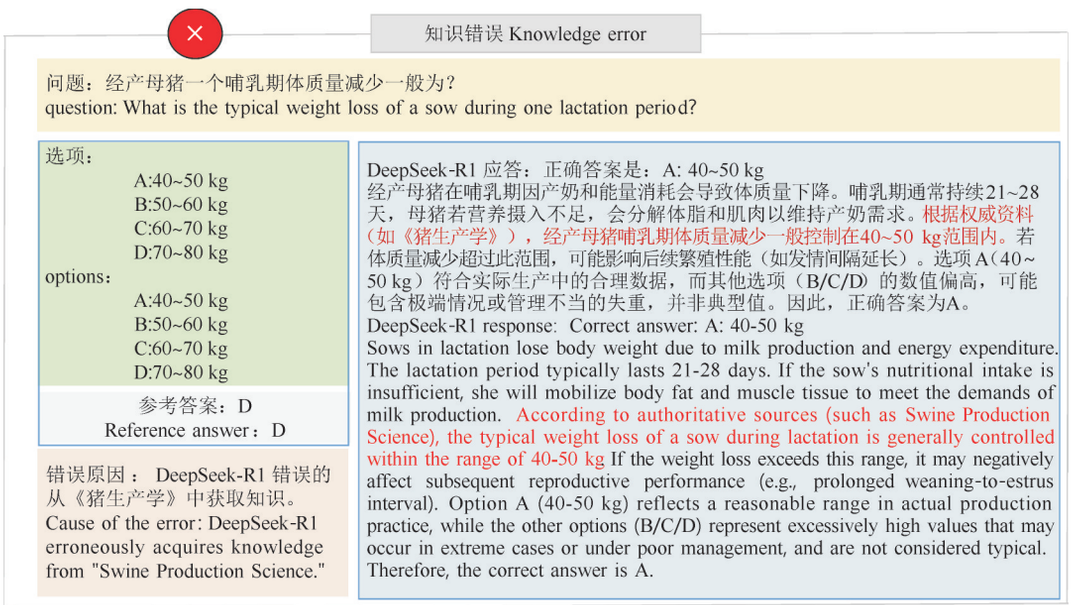
注:括号内的数值表示各主题中错误题目数量占该主题总题目数量的比例。Note: Values in parentheses represent the ratio of incorrect questions to the total questions within each topic.

误”题目占比最高(10.09%),其次为“畜牧生产管理”(7.50%),表明所有评估模型在这些领域的理解与应答能力有限,面临较大挑战。若考虑有争议问题,“畜牧经济与管理”的整体错误率最高(79.81%),其次为“育种与繁殖技术”(78.94%),说明模型在这些主题上的回答存在较大差异性。

进一步分析DeepSeek-R1回答错误的原因[19],通过对随机抽样的100道错误选择题的分析发现,知识错误最为普遍(63.0%),表明模型学习了错误信息。总结错误占22%,即模型掌握了正确知识但未能准确总结。过度推理错误占15%,表现为回答引入了不相关信息。图5为知识错误示例,DeepSeek-R1的回答表明其从《猪生产学》中检索到了错误信息。

3 讨 论

研究结果表明,当前大语言模型已能在一定程度上较准确地回答畜牧业相关问题,但整体上仍表现为客观题优于主观题。其中,选择题表现最好的是DeepSeek-R1(0.868)与Kimi K2(0.856);判断题



红色文本表示模型回答中的错误部分。The text highlighted in red color represents the erroneous part of the incorrect response.

图 5　知识错误示例

Fig.5　An example of knowledge error

为 GPT-5(0.842)和 DeepSeek-R1(0.839)；填空题为 Kimi K2(0.651)和 DeepSeek-V3.1(0.637)；问答题则以 Qwen3-235B-A22B（0.504）和 Moonshot-V1(0.479)表现相对较好。在领域专用模型中，神农百晓表现较优，选择题、判断题、填空题和问答题准确率分别为 0.798、0.769、0.546 和 0.408。但统计分析显示，领域专用模型整体仍显著弱于通用模型。主要原因如下：一是领域模型多基于旧基座模型微调，基础能力受限；二是现有畜牧业数据集覆盖范围与粒度不足；三是模型在多源知识融合与归纳推理方面存在短板，跨领域信息整合能力较弱。四是部分通用模型更新较快，其训练数据中可能已包含部分评估内容，从而影响结果。对模型共同回答错误题目的分析发现，错误主要集中在“畜牧经济与管理”和“畜牧生产管理”两类主题。这可能与不同地区管理实践与经济政策差异较大有关，因此模型难以给出普适性较强的答案。

通用模型与畜牧专用模型的对比进一步说明，后者在畜牧业任务中的整体表现暂时不及前者。这凸显了大语言模型在畜牧业应用中仍面临明显挑战。不过，以 Qwen2.5-32B 为基座模型微调得到的智慧养殖大模型，性能已明显优于基座模型，其在选择题和判断题上的准确率分别达到 0.813 和 0.777，在填空题和问答题上的相似性分别达到 0.789 和 0.776，均高于基座模型的 0.769、0.746、0.753 和 0.758。这说明，在基座模型能力相当、领域数据持续优化的前提下，专业模型仍有显著的提升空间。

通用模型强调广度，但很难在所有垂直领域同时兼顾可靠性、知识深度与更新效率。而专业模型通过聚焦，能够在成本、知识保真度、数据安全等方面构建形成优势。构建专业大模型的必要性主要体现在 3 个方面。(1)数据与计算效率。可利用高质量、小规模领域数据，通过微调和蒸馏等技术实现快速迭代，降低建模门槛。(2)模型轻量化与部署成本。可通过大模型压缩、剪枝等方法获得更轻量的小模型，便于在资源受限场景低成本部署。(3)通用大语言模型由于规模大、训练和更新成本高，而专业大模型更有利于快速吸收新数据。因此，未来更可能形成通用大模型与专业大模型互补协同的格局，而非由通用模型完全覆盖所有专业需求。

未来，知识图谱[20]、高质量畜牧业知识库和问答系统等有望推动畜牧专用模型的发展[21]。知识图谱可为畜牧业提供结构化语义表示，明确定义品种、饲料、疾病等实体及其关系[22]。通过整合教材、期刊、养殖经验等多源信息，大语言模型有望构建更实用的问答系统和实时决策系统[23]。进一步结合强化学习、图规划等算法，还可将畜牧管理建模为多目标优化与推理任务[24]，依据实时数据和用户偏好动态调整方案，支持畜牧管理策略的生成与优化，从而提升资源效率与整体收益。此外，畜牧专业领域模型的

发展还可从多方面推进，包括构建检索增强生成体系、完善贴近真实业务的评测标准、加强领域数据治理与专家参与式对齐、发展具备工具调用能力的专业智能体、推动多模态融合、增强因果与机制推理能力、提升持续学习与增量更新能力，以及加强模型可解释性、安全性和标准化生态建设等。

总体而言，当前大语言模型在知识深度、判断准确性与推理鲁棒性方面仍有不足。未来应进一步提升训练数据的多样性与质量，并充分考虑生产管理的地域差异，结合小模型响应快、部署灵活与大模型理解深、推理强的优势，采取混合建模策略，不仅有助于增强模型的实用性，也能最大限度降低硬件消耗。

本研究的数据和代码可在 https://github.com/llmlivestock/LIVED 获取。

参考文献 References

[1] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report[PP/OL]. arXiv, 2023: 2303.08774[2025-09-30]. https://arxiv.org/abs/2303.08774.

[2] MINAEE S, MIKOLOV T, NIKZAD N, et al. Large language models: a survey [PP/OL]. arXiv, 2024: 2402.06196 [2025-09-30]. https://arxiv.org/abs/2402. 06196.

[3] WANG N, YANG H Y, WANG C D. FinGPT: instruction tuning benchmark for open-source large language models in Financial Datasets [PP/OL]. arXiv, 2023: 2310.04793 [2025-09-30]. https : //doi.org/10.48550/arXiv. 2310.04793.

[4] JIN Q, DHINGRA B, LIU Z P, et al. PubMedQA: a dataset for biomedical research question answering [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong. Stroudsburg: ACL, 2019: 2567-2577.

[5] CHONG D D, GUO Z H, HUA Y N, et al. Benchmarking large language models on CMExam : a comprehensive Chinese medical exam dataset [C]//Advances in Neural Information Processing Systems 36. December 10-16, 2023. New Orleans: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 52430-52452.

[6] DAI Y F, FENG D Y, HUANG J M, et al. LAiW: a Chinese legal large language models benchmark [C]//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi: Association for Computational Linguistics, 2025: 10738-10766.

[7] GUO X, XIA H T, LIU Z W, et al. FinEval: a Chinese financial domain knowledge evaluation benchmark for large language models [C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (volume 1: Long Papers). Albuquerque, Stroudsburg: ACL, 2025: 6258-6292.

[8] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understandingar [PP/OL]. arXiv, 2020: 2009.03300 [2025-09-30]. https://doi. org/10.48550/arXiv.2009.03300.

[9] MORI S, SUEN C Y, YAMAMOTO K. Historical review of OCR research and development [J]. Proceedings of the IEEE, 1992, 80(7): 1029-1058.

[10] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China. Stroudsburg: ACL, 2019: 3980-3990.

[11] SANDMANN S, HEGSELMANN S, FUJARSKI M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making [J]. Nature medicine, 2025, 31(8): 2546-2549.

[12] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [PP/OL]. arXiv, 2020: 2005.14165 [2025-09-30]. https://doi.org/10.48550/arXiv.2005.14165.

[13] HURST A, LERER A, GOUCHER A P, et al. GPT-4o system card [PP/OL]. arXiv, 2024: 2410.21276 [2025-09-30]. https://doi.org/10.48550/arXiv.2410.21276.

[14] META LLAMA. meta-llama/llama3 [EB/OL]. GitHub, (2025-03-13) [2025-09-30]. https://github. com/meta-llama/llama3.

[15] YANG A, LI A F, YANG B S, et al. Qwen3 technical report [PP/OL]. arXiv, 2025: 2505.09388 [2025-09-30]. https://arxiv.org/abs/2505.09388.

[16] DEEPSEEK-AI, GUO D Y, YANG D J, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs *via* reinforcement learning [PP/OL]. arXiv, 2025: 2501.12948 [2025-09-30]. https://arxiv.org/abs/2501.12948.

[17] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [PP/OL]. arXiv, 2020: 2001.08361[2025-09-30]. https://arxiv.org/abs/2001.08361.

[18] WU J, WANG Z H, QIN Y F. Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese national medical licensing examination: a comparative study [J/OL]. Journal of medical systems, 2025, 49: 74 [2025-09-30]. https://doi. org/10.1007/s10916-025-02213-z.

[19] WANG P J, LI Z Z, YIN F, et al. Mv-math: evaluating multimodal math reasoning in multi-visual contexts [C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 10-17, 2025. Nashville: IEEE, 2025:

19541-19551.
[20] HOGAN A, BLOMQVIST E, COCHEZ M, et al. Knowledge graphs[J].ACM computing surveys, 2022, 54(4): 1-37.
[21] MURPHY K, SCHÖLKOPF B, VAN MOFFAERT K, et al. Multi-objective reinforcement learning using sets of Pareto dominating policies [J]. The journal of machine learning research, 2014: 3483-3512.
[22] 戴高阳, 孟小艳, 张容祯. 基于深度学习的畜牧业知识图谱构建[J]. 计算机与数字工程, 2024, 52(6): 1746-1753. DAI G Y, MENG X Y, ZHANG R Z. Construction of knowledge graph of animal husbandry based on deep learning[J]. Computer & digital engineering, 2024, 52(6): 1746-1753 (in Chinese with English abstract).
[23] POTAMITIS I. ChatGPT in the context of precision agriculture data analytics [PP/OL]. arXiv, 2023: 2311.06390 [2025-09-30]. https://arxiv.org/abs/2311.06390.
[24] 吴华瑞, 李静晨, 杨雨森. 基于大语言模型的个性化作物水肥管理智能决策方法[J]. 智慧农业(中英文), 2025, 7(1): 11-19. WU H R, LI J C, YANG Y S. Intelligent decision-making method for personalized vegetable crop water and fertilizer management based on large language models [J]. Smart agriculture, 2025, 7(1): 11-19 (in Chinese with English abstract).

Evaluation of LIVED benchmark-based large language models in animal husbandry

YU Bing[1], WEN Jiwei[1], ZHANG Xiaopeng[1], DU Xiaoyong[1], LI Guoliang[1], LIU Chousheng[2], QIU Xiaotian[2], FANG Yaping[1]

1. *National Key Laboratory of Crop Genetic Improvement/Hubei Hongshan Laboratory/Ministry of Agriculture and Rural Affairs Key Laboratory of Smart Farming Technology for Agricultural Animals/ Ministry of Education Engineering Research Center of Intelligent Technology for Agriculture/Hubei Province Engineering Technology Research Center of Agricultural Big Data/College of Informatics, Huazhong Agricultural University, Wuhan 430070, China*;
2. *National Animal Husbandry Services. Beijing 100125, China*

Abstract The performance of 11 general-purpose large language models (LLMs) and 3 specific LLMs in animal husbandry was systematically evaluated on the benchmark dataset of professional examination to promote the application of LLMs in animal husbandry. The constructed LIVED dataset included a Chinese animal husbandry knowledge assessment dataset with 11 037 questions covering multiple choice, judgment, completion, and open-ended questions. The results showed that there was significant difference in the performance of LLMs, with the top three models in terms of overall performance being DeepSeek-R1 (accuracy of 0.700), Kimi K2 (accuracy of 0.681), and DeepSeek-V3.1 (accuracy of 0.675). DeepSeek-R1, GPT-5, Kimi K2 and Qwen3-235B-A22B had the highest accuracy in multiple-choice (0.868), judgment (0.842), completion (0.651), and open-ended questions (0.504). Shennong Baixiao had the best overall performance (0.630) in the specialized livestock models, still lagging behind DeepSeek-R1 by 7.0 percentage points. The results of error analysis showed that the errors in the current LLMs were mainly concentrated in areas including the management of economy and production in animal husbandry. It is indicated that there are still significant shortcomings in the accuracy of professional knowledge and reasoning ability of LLMs. In the future, the development of vertical large models needs to prioritize addressing the shortcomings in professional knowledge and reasoning performance.

Keywords large language models (LLMs); animal husbandry; pairs of question and answer; benchmark test; evaluation

(责任编辑：张志钰)