基于 MQM2.0 框架的译文质量量化对比研究
——以国内大语言模型为例

刘梦祯
西安理工大学，陕西 西安 710054
DOI:10.61369/HASS.2025090002

摘　　要： 本研究基于 MQM 2.0（翻译多维质量评估）框架，以《中国农业科学》100篇中文摘要为语料，量化评估 DeepSeek-R1、Qwen2.5-Max、GLM4-Plus三款国内大语言模型（LLMs）的农业科技文本英译质量。研究采用非参数统计方法，聚焦术语、准确性、语言惯例、风格四大维度。结果显示：三款模型均满足“专业信息传递”需求，但术语、准确性、风格维度存在显著差异，语言惯例维度趋同；其中 DeepSeek-R1术语规范性最优，Qwen2.5-Max 在准确性与风格自然度上表现突出但语域判断严苛，GLM4-Plus多维度表现薄弱。研究为农业科技领域 LLM 翻译优化提供实证依据，同时指出语料单一、未纳入国际模型对比等局限。

关 键 词： 大语言大模型（LLMs）；MQM框架；农业科技文本；量化研究

A Quantitative Comparison of Translation Quality Based on the MQM 2.0 Framework — Taking Domestic Large Language Models as Examples

Liu Mengzhen
Xi'an University of Technology, Xi'an, Shaanxi 710054

Abstract: This study employs the MQM 2.0 (Multidimensional Quality Measurement for Translation) framework to quantitatively evaluate the English translation quality of agricultural science texts produced by three domestic large language models (LLMs)—DeepSeek-R1, Qwen2.5-Max, and GLM4-Plus—using 100 Chinese abstracts from the Chinese Agricultural Sciences as corpus. Nonparametric statistical methods were applied, focusing on four dimensions: terminology, accuracy, linguistic conventions, and style. Results indicate that all three models meet the “professional information transmission” requirement, yet exhibit significant differences in terminology, accuracy, and style dimensions, while showing convergence in linguistic conventions. DeepSeek-R1 demonstrates optimal terminology standardization, Qwen2.5-Max excels in accuracy and stylistic naturalness but exhibits strict register judgment, while GLM4-Plus shows weaker performance across multiple dimensions. This study provides empirical evidence for optimizing LLM translation in agricultural science and technology, while acknowledging limitations such as limited corpus diversity and the absence of international model comparisons.

Keywords: large language models (LLMs); MQM framework; agricultural science and technology texts; quantitative research

引言

随着大语言模型（LLMs）在多语言翻译任务中的规模化应用，其在专业领域文本处理中的适应性逐渐成为学界焦点。农业科技文献作为兼具术语密集性、逻辑严谨性与学科表达惯例的特殊文本类型，对 LLMs的领域知识嵌入、语义准确性及学术风格适配性提出了更高要求。然而，现有研究多聚焦通用领域翻译，对农业科技文本中专业术语的概念层级一致性、逻辑框架忠实度及学术语域规范性的系统评估仍存在明显空白。MQM框架因其多维错误分类与量化评估优势，为破解专业翻译质量分析的复杂性提供了科学工具，但其在国内 LLMs农业科技翻译场景中的系统性应用仍有待探索。本研究选取《中国农业科学》期刊中100篇中文摘要作为核心语料。该期刊作为农业科技领域的权威载体，其摘要文本涵盖作物遗传、农业生态等专业方向，具备术语专业性强、逻辑结构严密、学术规范严格的特征，能够有效检验 LLMs在专业领域的翻译能力。研究以国内领先的 DeepSeek-R1、Qwen2.5-Max和 GLM4-Plus三款 LLMs为对

作者简介：刘梦祯（2000.03—），女，汉族，陕西西安人，硕士研究生，研究方向：翻译学。

象，基于 MQM 2.0框架构建翻译质量评估体系，通过量化分析译文在术语、准确性、语言惯例及风格维度的错误分布，揭示大模型在农业科技翻译中的核心缺陷。

研究针对学术文本特性筛选维度：排除 “受众适宜性”“格式与标记” 等非核心维度，保留术语（聚焦错误术语）、准确性（含误译、过译等五类错误）、语言惯例（文本惯例）及风格（语域、不自然风格、不地道风格）四大评估维度。通过研究者与导师团队的三轮专家论证进行错误标注，并引入交叉验证机制确保评估客观性。数据分析阶段，先对各维度罚分值进行正态性检验，继而采用正态性检验及非参数 Friedman检验，量化模型间的表现差异，最后针对具有显著性差异的维度中细分错误类型进行差异性检验，进一步细化模型差异。

本研究的核心目标在于：通过错误类型量化与统计检验，揭示国内领先 LLMs在专业翻译中的优势与短板，并为农业科技领域大模型翻译能力的优化提供实证依据。与现有研究相比，其创新价值体现在：首次将 MQM 框架系统应用于国内 LLMs的农业科技翻译评估，通过微观错误类型分析与宏观统计检验的结合，为专业领域大模型翻译质量的提升提供新的研究范式。[1]

一、研究方法

（一）研究框架

鉴于农业科技文本“无跨文化适配需求、无格式排版依赖”的核心特性，本研究首先对 MQM2.0（翻译多维质量评估）原始框架进行针对性适配与优化，剔除 “受众适宜性”“格式与标记”“区域惯例” 等与农业学术文本无关的维度，最终确定四大核心评估维度，并结合专业文本的质量优先级设定错误权重，以确保评估体系贴合农业科技翻译的实际需求。其中，术语维度仅聚焦 “错误术语” 这一关键错误类型，具体指农业科技领域专属术语的误译，术语准确性是农业科技文本传递专业信息的核心基础，术语偏差会直接导致学术概念传递失真；准确性维度涵盖“误译”“过译”“欠译”“增译”“省译”五类错误，均指向核心语义的偏差或信息传递异常，“误译” 指语义与原文完全不符，“过译”为额外添加原文无相关信息，“欠译”是遗漏原文关键内容，“增译”为冗余补充非必要信息，“省译”则是缺失原文重要语义，语义准确性是确保译文学术价值的关键前提；语言惯例维度仅保留“文本惯例”类错误评估，具体包括时态统一、标点符号规范、表达逻辑连贯性等文本层面的一致性要求，由于实验过程中发现三款模型的译文均无基础语法错误或拼写错误，故排除此类错误类型；风格维度包含 “语域错误”“不自然风格”“不地道风格”三类错误，“语域错误” 指译文学术正式度与农业科技文本适配性偏差，“不自然风格”为表达生硬、不符合英文学术写作逻辑，“不地道风格” 则是非母语化表达。[2]

（二）语料设计

本研究随机选取《中国农业科学》2024年发表的100篇中文摘要作为源语文本，总字数约7.5万字，该期刊为国内农业科技领域权威核心期刊，其摘要文本具备 “术语专业性强、逻辑结构严密、学术规范严格” 的特征，能够有效检验大语言模型（LLMs）在专业领域的翻译能力。在语料处理流程上，首先进行文本预处理，剔除摘要中无关的图表、公式及注释内容，仅保留纯文本部分，随后将所有文本统一保存为 UTF-8 编码的 .txt 格式，避免因编码差异导致模型读取异常；接着进行译文生成，将预处理后的 00 篇中文摘要分别导入 DeepSeek-R1、Qwen2.5-Max、GLM4-Plus 三款国内领先 LLM 的官方接口，采用模型默认翻译参数生成英文译文，并完整保留模型原始输出结果，未进行人工干预；最后设定编码规则，为便于后续数据追踪与分析，对有效分析单元采用 “模型－章节－摘要序号” 的编码方式命名。

（三）研究工具与统计

本研究的错误标注采用 “双人标注＋交叉验证” 的三级标注机制，以确保错误分类与严重程度判断的客观性，标注人员配置上，由1名具备科技翻译经验的英语专业硕士负责初始标注，1名从事农业信息传播研究的农业科学教授与1名英文专业教授负责争议审核，标注流程为先由1名硕士独立对280个有效单元的译文，按 MQM 2.0适配框架的四大维度标注错误类型与严重程度，对于标注结果不一致的争议案例（如 “语域错误” 与 “不自然风格”的界定分歧），提交至两位教授处进行最终审核与修正，同时通过 Kappa系数验证标注一致性，最终 Kappa 系数高度一致），表明标注结果可靠。统计分析选用 SPSSPRO作为工具，按 “数据分布检验→整体差异分析→两两差异定位”的逻辑开展，首先进行正态性检验，采用 Shapiro-Wilk（S-W）检验与 Kolmogorov-Smirnov（K-S）检验判断四大评估维度罚分值的数据分布特征，结果显示所有维度罚分值均显著偏离正态分布（$p<0.01$），因此后续采用非参数检验方法进行差异分析；接着进行整体差异分析，采用多配对样本 Friedman 检验，分析三款模型在四大维度的罚分中位数差异，识别模型间存在显著差异的维度，同时计算 Cohen's f 值（整体效应量），量化差异的实际影响程度。

二、研究结果与讨论

（一）整体差异分析

对 DeepSeek-R1、Qwen2.5-Max、GLM4-Plus 三款大语言模型在农业科技文本英译任务中的译文质量，需先通过数据分布检验明确分析方法适用性，再依托非参数统计检验揭示模型在四大评估维度（术语、准确性、语言惯例、风格）的整体差异特征，为后续细分维度研究奠定基础。采用 S-W和 K-S检验对各维度罚分值进行正态性验证，结果显示所有维度罚分值均显著偏离正态分布（$p<0.01$），需采用非参数检验方法开展后续差异

分析。

采用 Friedman检验分析三款模型在四大维度的整体差异，结果如表2显示，模型表现呈现显著分化特征：在术语、准确性、风格三个维度存在显著差异（p<0.01），而语言惯例维度表现趋同（p>0.05）。其中，术语维度的统计量为199.504（p<0.001），Cohen's f值为0.357（中等效应），结合罚分值（数值越大表明模型表现越差）可知，三款模型在该维度的表现梯度为DeepSeek-R1最优、Qwen2.5-Max次之、GLM4-Plus最差，效应量表明该维度差异具有实质意义；准确性维度的统计量为87.791（p<0.001），Cohen's f 值为 0.57（大效应），表现梯度为Qwen2.5-Max最优（中位数92.5，绝对罚分总和242）、DeepSeek-R1 次之（中位数98，绝对罚分总和9930）、GLM4-Plus 最差（中位数99，绝对罚分总和 10198），且GLM4-Plus 的罚分标准差（8.507）显著高于 Qwen2.5-Max（4.522），说明其准确性表现的波动性更大；风格维度的统计量为 200（p<0.001），Cohen's f值为0.396（中等效应），表现梯度为 Qwen2.5-Max最优（中位数84，绝对罚分总和8380）、DeepSeek-R1 次之（中位数85，绝对罚分总和8480）、GLM4-Plus 最差（中位数 86，绝对罚分总和 8626）；语言惯例维度的统计量为3.211（p=0.201>0.05），Cohen's f 值为 0.097（极小效应），三款模型的中位数介于 71.5-72 之间，绝对罚分总和在7145-7190范围内，罚分分布高度集中（标准差 1.692-2.439）。

综上，三款模型在农业科技文本英译质量上的整体差异呈现“三异一同” 特征，即术语、准确性、风格维度表现分化显著，语言惯例维度能力趋同，这种分化为后续深入分析各差异维度的具体错误特征与成因提供明确方向。

表1 三组 LLMs各维度罚分值及质量得分

	DeepSeek-R1			Qwen2.5-Max			GLM4-Plus		
	绝对罚分总和	单字罚分总和	原始质量得分	绝对罚分总和	单字罚分总和	原始质量得分	绝对罚分总和	单字罚分总和	原始质量得分
术语	4929	0.066	9.34	5436	0.072	9.28	5943	0.079	9.21
准确性	9930	0.132	8.68	9242	0.123	8.77	10198	0.136	8.64
语言惯例	7145	0.095	9.05	7190	0.096	9.04	7184	0.096	9.04
风格	8480	0.113	8.87	8380	0.117	8.82	8626	0.115	8.85

表2 各错误维度多配对样本 Friedman检验结果分析表

变量名	样本量	中位数	标准差	统计量	P	Cohen’s f
术语						
DeepSeek-R1	100	48	11.591			
Qwen2.5-Max	100	53	11.648	199.504	0.000***	0.357
GLM4-Plus	100	59	11.717			
准确性						
DeepSeek-R1	100	98	7.631			
Qwen2.5-Max	100	92.5	4.522	87.791	0.000***	0.57
GLM4-Plus	100	99	8.507			
语言惯例						
DeepSeek-R1	100	71.5	2.439			
Qwen2.5-Max	100	72	2.008	3.211	0.201	0.097
GLM4-Plus	100	72	1.692			
风格						
DeepSeek-R1	100	85	2.693			
Qwen2.5-Max	100	84	2.693	200	0.000***	0.396
GLM4-Plus	100	86	2.29			

注：***、**、*分别代表1%、5%、10%的显著性水平

（二）术语维度差异分析

本研究中，术语维度的评估设计以“错误术语”为唯一核心评估对象，即仅聚焦农业科技领域专属术语的误译问题，未在该维度下进一步设置细分分类（如 “术语一致性错误”“术语概念偏差” 等），因此无需针对术语维度开展二次检验，仅基于前期整体差异分析的统计结果与文本验证，系统梳理三款模型的术语翻译表现差异及成因。

由表1来看 DeepSeek-R1 在该维度表现最优；Qwen2.5-Max 次之；GLM4-Plus 表现最差，，三者呈现清晰的梯度差异，且效应量表明该差异具有实质意义，并非统计层面的微弱波动。

农业科技文本中常涉及关联学科术语（如材料力学术语在作物品质分析中的应用），此类术语的翻译准确性直接影响学术结论的传递，以下案例可直观佐证三款模型的术语翻译差异：原文为 “产棉区原棉断裂比强度 … 显著高0.62—1.17 cN/tex”，三款模型的译文分别为：

TT-D：The breaking tenacity of raw cotton in the northern Xinjiang cotton region were significantly higher than...by 0.62 - 1.17 cN/tex.

TT-Q：The breaking strength of raw cotton in the northern Xinjiang cotton region were significantly higher than... by 0.62 - 1.17 cN/tex.

TT-G：The breaking strength of raw cotton in the northern Xinjiang cotton-growing region were significantly higher than...by 0.62-1.17 cN/tex.

需特别说明的是，“断裂强度（Breaking Strength）”与“断裂比强度（Breaking Tenacity）” 为材料力学中两个具有显著差异的核心参数：前者定义为材料断裂前承受的最大绝对力值，量纲为牛顿（N），其数值与材料几何尺寸正相关（如相同材质下粗纤维断裂强度高于细纤维）；后者则为归一化强度指标，表征为单位线密度下的断裂强度（计算公式：断裂比强度 = 断裂强度 / 线密度），量纲常为厘牛每特克斯（cN/tex），通过消除材料尺寸效应，可实现不同形态、规格纤维材料的强度可比性（王娟等，2025:184-187）。在该案例中，TT-D 与 TT-G 均将原文“断裂比强度” 误译为 “breaking strength”（断裂强度），直

接导致两个核心力学参数的概念混淆，可能引发读者对原棉强度属性的误判；而 TT-Q准确采用 “breaking tenacity” 对应 “断裂比强度”，完全符合材料科学领域的标准命名规范，也体现出其对农业科技文本中跨学科术语的精准把握。

（三）准确性维度差异分析

准确性维度主要聚焦“误译”“增译”“省译”“过译”“欠译”五类核心错误。

根据表3显示，从“误译”错误来看，三款模型的表现呈现 “统计显著但实际差异微弱” 的特征。样本量均为 100 的情况下，DeepSeek-R1 与 Qwen2.5-Max 的误译罚分中位数一致（均为 32.5），GLM4-Plus 略高（33）；Friedman 检验统计量为 42.397（p<0.001），但 Cohen's f 值仅为 0.023（小效应），表明尽管统计层面存在差异，但其实际影响程度极低。从标准差来看，三款模型差异接近（DeepSeek-R1 4.81、Qwen2.5-Max 4.516、GLM4-Plus 4.657），说明三者在误译错误的控制能力上趋同。

在 “增译” 错误上，三款模型表现分化显著且具有实际意义。Friedman 检验统计量为 17.78（p<0.001），Cohen's f=0.24（中等效应），反映出模型间的差异已对译文质量产生实质影响。具体来看，GLM4-Plus 的增译罚分中位数最低（17.5），在增译控制上表现最优；DeepSeek-R1次之（19）；Qwen2.5-Max最高（20），增译错误相对较多。值得注意的是，Qwen2.5-Max 的增译罚分标准差最小（2.494），远低于GLM4-Plus的3.573，表明其增译错误虽多，但分布高度集中，错误类型相对固定，而 GLM4-Plus虽增译较少，但标准差较大，增译内容的随机性更强。

“省译”错误是三款模型的共性短板，且无显著差异。Friedman 检验统计量仅为 0.54（p=0.763>0.05），Cohen’s f=0.052（极小效应），说明模型间的省译错误控制能力趋同。从数据来看，三款模型的省译罚分中位数集中在 25-26 之间（DeepSeek-R1 26、Qwen2.5-Max 25、GLM4-Plus 26），标准差也较为接近（3.202、3.112、3.211），反映出无论模型类型如何，均易在农业科技文本“细节信息传递”上出现缺失。

“过译”错误是三款模型差异最显著的错误类型之一，且影响程度较大。Friedman 检验统计量为 56.86（p<0.001），Cohen’s f=0.457（大效应），表明模型间的过译控制能力差异已对译文质量产生重要影响。中位数数据显示，Qwen2.5-Max 的过译罚分最低（5），在过译控制上表现最优；DeepSeek-R1 次之（8）；GLM4-Plus 最高（9），过译错误最为严重。从标准差来看，Qwen2.5-Max（2.608）远低于 DeepSeek-R1（4.075）和 GLM4-Plus（3.37），说明其过译错误不仅少，且分布集中。

“欠译”错误与过译错误类似，模型间差异显著且影响程度大。Friedman 检验统计量为 49.031（p<0.001），Cohen’s f=0.472（大效应），是准确性维度中效应量最大的错误类型。中位数对比显示，Qwen2.5-Max的欠译罚分最低（9），对原文信息的保留最完整；DeepSeek-R1 次之（13）；GLM4-Plus 最高（15），信息遗漏问题最为突出。从标准差来看，GLM4-Plus（7.809）显著高于其他两款模型（DeepSeek-R1 6.704、Qwen2.5-Max 5.021），表明其欠译错误不仅多，且波动性强。此类欠译会直接导致译文学术信息不完整，影响读者对原文研究结论的准确理解。[3]

表3 准确性维度各错误类型罚分值的多配对样本 Friedman检验

变量名	样本量	中位数	标准差	统计量	P	Cohen’s f
			误译			
DeepSeek-R1-误译	100	32.5	4.81			
Qwen2.5-Max-误译	100	32.5	4.516	42.397	0.000***	0.023
GLM4-Plus-误译	100	33	4.657			
			增译			
DeepSeek-R1-增译	100	19	3.023			
Qwen2.5-Max-增译	100	20	2.494	17.78	0.000***	0.24
GLM4-Plus-增译	100	17.5	3.573			
			省译			
DeepSeek-R1-省译	100	26	3.202			
Qwen2.5-Max-省译	100	25	3.112	0.54	0.763	0.052
GLM4-Plus-省译	100	26	3.211			
			过译			
DeepSeek-R1-过译	100	8	4.075			
Qwen2.5-Max-过译	100	5	2.608	56.86	0.000***	0.457
GLM4-Plus-过译	100	9	3.37			
			欠译			
DeepSeek-R1-欠译	100	13	6.704			
Qwen2.5-Max-欠译	100	9	5.021	49.031	0.000***	0.472
GLM4-Plus-欠译	100	15	7.809			

综合来看，三款模型在准确性维度的表现呈现 “两极分化与共性短板并存” 的特征：Qwen2.5-Max 在影响最大的过译、欠译错误上表现最优，是其准确性维度整体领先的核心原因；GLM4-Plus 则在过译、欠译上表现最差，且错误波动性强，信息传递完整性与准确性不足；DeepSeek-R1整体表现居中，无显著优势或劣势；而省译错误是三款模型的共同薄弱环节，误译错误则因农业科技文本核心概念的固定性而表现趋同。[4]

（四）风格维度差异分析

风格维度通过 “语域错误”“不自然表达”“不地道表达”三类错误展开评估 —— “语域错误” 指译文学术正式度与农业科技

文本适配偏差，“不自然表达” 为句式生硬、不符合英文学术写作逻辑，“不地道表达” 则是非母语化的词汇搭配或语序问题。[5]

根据表4显示，从 “语域错误” 来看，三款模型表现差异极具显著性且影响程度极强。Friedman 检验统计量达 198.061（p<0.001），Cohen's f=2.5（极强效应），是风格维度中差异最突出的错误类型。样本量均为 100 的情况下，三款模型的语域错误罚分中位数呈现明显层级：DeepSeek-R1 最低（35），表明其对农业科技文本 “学术正式度” 的把握最贴合实际需求；GLM4-Plus 次之（41）；Qwen2.5-Max 最高（46.5），语域判断标准显著严苛。从标准差来看，三者均处于较低水平（Qwen2.5-Max 1.795、GLM4-Plus 1.676、DeepSeek-R1 2.146），说明各模型的语域判断逻辑相对稳定——Qwen2.5-Max 的高罚分并非源于随机误判，而是对 “学术正式度” 的阈值设定过高，而 DeepSeek-R1 的判断标准更契合农业领域读者对学术语域的普遍期待，较少出现此类过度判定。

在 “不自然表达” 错误上，三款模型差异显著且 Qwen2.5-Max 优势突出。Friedman 检验统计量为 188.18（p<0.001），Cohen's f=1.921（强效应），反映出模型间的自然度控制能力差异对译文可读性影响重大。中位数数据显示，Qwen2.5-Max 的不自然表达罚分最低（16），其译文在句式逻辑与学术表达流畅性上表现最优；GLM4-Plus 次之（21）；DeepSeek-R1 最高（27），表达生硬问题最为明显。标准差方面，Qwen2.5-Max（2.462）与 GLM4-Plus（2.385）接近，DeepSeek-R1（2.002）略低，说明 DeepSeek-R1 的不自然表达虽多但类型相对固定，而 Qwen2.5-Max 则能通过优化语序提升表达自然度，更符合英文学术写作的逻辑习惯。

“不地道表达” 错误同样呈现显著差异，且 Qwen2.5-Max 的母语化表达优势明确。Friedman 检验统计量为 140.323（p<0.001），Cohen’s f=0.573（中等偏上效应），表明模型间的地道性差异已对译文的 “母语化质感” 产生实质影响。从位数对比来看，Qwen2.5-Max 的不地道表达罚分最低（21），在词汇搭配与表达习惯上最贴近母语者撰写的农业科技文本；DeepSeek-R1 次之（23）；GLM4-Plus 最高（24），非母语化表达问题最突出。三款模型的标准差均处于 2.0-2.3 区间（Qwen2.5-Max 2.066、GLM4-Plus 2.059、DeepSeek-R1 2.249），错误分布相对均匀。GLM4-Plus 的不地道表达多体现在学术常用搭配偏差，而 Qwen2.5-Max 则更能精准使用符合母语习惯的搭配，提升译文的地道性与专业性。

综合来看，三款模型在风格维度的表现呈现 “优势分化与特性鲜明” 的特征：Qwen2.5-Max 虽因语域判断标准严苛导致语域错误罚分最高，但在不自然表达（强效应）与不地道表达（中等偏上效应）上表现最优，整体风格适配性最符合农业科技文本的 “自然化、母语化” 需求；DeepSeek-R1 以语域错误最少为核心优势，但其不自然表达问题显著，需优化句式逻辑以提升可读性；GLM4-Plus 在三类错误中均处于中间或劣势水平，尤其不地道表达问题突出，需强化对英文学术表达习惯的习得。[6]

表4 风格维度各错误类型罚分值的多配对样本 Friedman检验

变量名	样本量	中位数	标准差	统计量	P	Cohen’s f
语域错误						
DeepSeek-R1-语域错误	100	35	2.146	198.061	0.000***	2.5
Qwen2.5-Max-语域错误	100	46.5	1.795			
GLM4-Plus-语域错误	100	41	1.676			
不自然表达						
DeepSeek-R1-不自然表达	100	27	2.002	188.18	0.000***	1.921
Qwen2.5-Max-不自然表达	100	16	2.462			
GLM4-Plus-不自然表达	100	21	2.385			
不地道表达						
DeepSeek-R1-不地道表达	100	23	2.249	140.323	0.000***	0.573
Qwen2.5-Max-不地道表达	100	21	2.066			
GLM4-Plus-不地道表达	100	24	2.059			

三、结论

（一）核心研究发现

本研究基于 MQM 2.0框架，对 DeepSeek-R1、Qwen2.5-Max、GLM4-Plus 三款国内领先大语言模型（LLM）在农业科技文本英译任务中的质量展开系统性评估，结果表明，三款模型均能满足农业科技文本 “专业信息传递” 的基础需求，但在四大评估维度（术语、准确性、语言惯例、风格）中表现出显著分化特征。其中，术语、准确性、风格三大维度存在统计学意义上的显著差异，仅语言惯例维度表现趋同，这种趋同源于模型普遍集成的通用语法纠错模块，且农业科技摘要的简洁性特征减少了复杂句式导致的表达不一致问题。[7]

从各模型核心特征来看，DeepSeek-R1 以术语规范性最优为鲜明优势，但在风格维度存在不自然表达问题。Qwen2.5-Max 则在准确性与风格自然度、地道性上表现突出：准确性维度中，其过译与欠译罚分均为三款模型最低；风格维度中，其母语化表达能力最优，但存在语域判断标准严苛的缺陷，易将农业科技文本常规表达误判为非正式。GLM4-Plus则在多维度表现相对薄弱，整体存在术语偏差、信息遗漏及母语化表达不足的共性问题。

（二）实践启示

基于上述研究发现，本研究为国内 LLM 在农业科技翻译领域

的应用与优化提供三方面实践启示。在模型优化方向上，需针对各模型核心短板精准施策：DeepSeek-R1 需补充农业科技文本风格专项训练，重点优化句式逻辑以减少不自然表达（如降低冗长嵌套句使用频率，适配英文学术写作习惯）；Qwen2.5-Max 需调整语域判断阈值，通过引入更多农业科技期刊摘要语料校准“学术正式度”判定标准；GLM4-Plus 则需双管齐下，一方面强化农业科技专属术语库建设，另一方面优化语义完整性控制机制，减少欠译等问题。[7]

在翻译工具选择上，需结合具体应用场景匹配模型优势：对于学术论文发表等对 “语义准确性” 与 “表达自然度” 要求较高的场景，优先选择 Qwen2.5-Max，其在准确性维度的过译、欠译控制最优，风格维度的自然度与地道性表现突出，能更好满足学术文本的可读性与严谨性需求；对于技术标准、术语手册等“术语密集型” 文本，优先选择 DeepSeek-R1， 其术语规范性优势显著，可最大程度降低术语误译导致的概念混淆风险。

在人工校对重点上，需针对各模型薄弱环节制定差异化策略：GLM4-Plus 的译文需重点核查两方面内容，一是农业专属术语的准确性，二是核心信息的完整性；Qwen2.5-Max 的译文需聚焦语域判断合理性，排查是否存在 “常规学术表达被误判为非正式” 的过度修正情况；DeepSeek-R1 的译文则需关注表达自然度，优化生硬句式与非母语化搭配，提升文本可读性。

（三）研究局限与未来方向

本研究虽构建 MQM 2.0 框架下国内 LLM 农业科技翻译质量的量化评估体系，但仍存在三方面局限。其一，语料来源与类型相对单一，仅选取《中国农业科学》的中文摘要作为研究对象，未涵盖农业科技专著、技术手册、田间试验报告等其他文本类型。 这类文本在术语密度、逻辑复杂度、表达风格上与摘要存在差异，可能导致研究结论的普适性受限。其二，未纳入国际主流 LLM 进行对比分析，无法清晰界定国内 LLM 在农业科技翻译领域与国际模型的差距，也难以揭示国内外模型在设计机制上的本质差异。其三，评估体系以客观量化指标为主，缺乏目标读者的主观反馈，无法验证 “模型错误是否实际影响学术阅读体验与信息理解效率”，评估结果的实践指导性仍有提升空间。

针对上述局限，未来研究可从三方面推进：一是拓展语料类型与来源，纳入农业科技专著章节、技术标准文本、试验报告等，覆盖 全谱系，同时增加不同语种农业科技文本，提升研究结论的普适性；二是引入国际主流 LLM 作为参照组，对比分析国内外模型在术语准确性、语义完整性、风格适配性上的差异，挖掘国内模型的优势与短板，为模型优化提供更精准的对标方向；三是构建 “客观指标 + 主观反馈” 的综合评估体系，邀请农业科研人员、学术翻译从业者对译文质量进行评分与反馈，结合眼动追踪等行为数据，量化模型错误对阅读体验的影响，使评估结果更贴合实际应用需求，为农业科技领域智能翻译工具的迭代与应用提供更全面的实证支撑。[6,7]

参考文献

[1]COUNCIL T M. Introduction to TQE[EB/OL]. The MQM Council, 2025[2025-03-01].

[2] 杨博超，冷冰冰 . 基于 MQM 质量评估模型的专业文本机器翻译错误类型实证分析 [J/OL]. 上海理工大学学报 (社会科学版), 2024: 1-7.

[3]COUNCIL T M. The MQM Error Typology[EB/OL]. The MQM Council, 2023[2025-03-01].

[4] 赵衍，张慧，杨祎辰 . 大语言模型在文本翻译中的质量比较研究——以《繁花》翻译为例 [J/OL]. 外语电化教学，2024(04): 60-66+109.

[5]COUNCIL T M. The MQM Scoring Models[EB/OL]. The MQM Council, 2023[2025-03-01].

[6] 成爽，张玉双 . 翻译教学质量评估新视角：行业评估模型 [J/OL]. 当代外语研究，2024(05): 45-56+79.

[7]LOMMEL A, GLADKOFF S, MELBY A, et al. The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control[EB/OL]. (2024-05-27). DOI:10.48550/arxiv.2405.16969.

