研究论文

DOI：10.11949/0438-1157.

一种可再生能源波动下绿氨生产的大语言模型建模多智能体协作框架

任育奎，贺革，吉旭

（四川大学化学工程学院，四川 成都 610065）

摘要： 绿氨生产是可再生能源储存与零碳航运燃料的关键路径，涉及风光波动供电下电解、哈伯-博世合成及多单元集成等复杂过程。精确过程建模对系统柔性操作和优化至关重要，但传统方法高度依赖经验、开发周期长。大语言模型（LLM）有望通过知识检索与推理加速建模，却面临事实性幻觉、工程经验缺失及多维度合规性不足等挑战。为此，本文提出一种多智能体协作增强框架，由四类专业智能体（工艺建模、设备工程、安全合规、数据分析）与一个总控收敛智能体组成，通过独立生成、交叉辩论与共识收敛机制，迫使LLM在迭代中自主修正错误、补全工程约束、符合安全规范，从而输出更高质量、更可靠的建模方案。基于该框架，设计了5个复杂度递进的绿氨生产案例，通过单模型与多智能体对照实验，量化揭示了协作辩论带来的质量提升随任务复杂度增加而增强的规律：多智能体协作使平均综合得分提升9.0%，智能体间评分极差下降78.2%。四款模型的整体问题检出率为81.8%；以问题检出率最高的Gemini 3.1 Pro为例，其整体问题检出率为87.1%，其中设备问题、安全问题和机理问题的检出率分别为100.0%、80.0%和86.7%。实验同时明确了当前机制的能力边界——设备类问题识别较充分，而安全问题与深层隐性机理问题仍存在漏检，需通过专家监督进一步把关。本研究为风光波动供电下的绿氨工艺智能化建模提供了可复现的人机协同方法，以多智能体辩论促进LLM输出质量提升。

关键词： 多智能体；绿氨生产；大语言模型；协作框架

中图分类号： TK91；TQ113.26　　**文献标志码：** A

文章编号： 0438-1157（XXXX）XX-0001-22

A multi-agent collaboration framework for large language model-based green ammonia production modeling under renewable energy fluctuations

REN Yukui, HE Ge, JI Xu

(*School of Chemical Engineering*, *Sichuan University*, *Chengdu* 610065, *Sichuan*, *China*)

Abstract: Green ammonia offers a pathway to storing renewable energy and supplying zero-carbon shipping fuel. Its production involves electrolysis under fluctuating wind and solar power supply, Haber-Bosch synthesis, and integration of multiple process units. Accurate process modeling is essential for flexible system operation and optimization, yet conventional methods require extensive expertise and long development cycles. Large language models (LLMs) can potentially accelerate modeling via knowledge retrieval and reasoning, but they are plagued by challenges such as factual hallucinations, limited engineering knowledge, and difficulty satisfying multiple engineering constraints. To address these issues, this paper proposes a multi-agent collaborative enhancement framework consisting of four types of professional agents (process modeling, equipment engineering, safety compliance, and data analysis) and one coordinating agent. Through independent generation, cross-agent debate, and consensus-building mechanisms, the framework enables LLMs to autonomously correct errors, supplement

收稿日期： 2026-08-07　　**修回日期：** 2026-09-10

通信作者： 吉旭(1968—)，男，博士，教授，jixu@scu.edu.cn；贺革(1987—)，男，博士，副研究员，hege@scu.edu.cn

第一作者： 任育奎(2002—)，男，硕士研究生，renyukui@stu.scu.edu.cn

引用本文： 任育奎, 贺革, 吉旭. 一种可再生能源波动下绿氨生产的大语言模型建模多智能体协作框架[J]. 化工学报, XXXX, XX(XX): 1-22

Citation: REN Yukui, HE Ge, JI Xu. A multi-agent collaboration framework for large language model-based green ammonia production modeling under renewable energy fluctuations[J]. CIESC Journal, XXXX, XX(XX): 1-22

engineering constraints, and align with safety specifications during successive iterations, thereby producing higher-quality and more reliable modeling schemes. Based on the proposed framework, five green ammonia production cases with progressively increasing complexity are designed. Comparative experiments between a single LLM and the multi-agent system quantitatively reveal that the quality improvement brought by collaborative debate increases with task complexity: the multi-agent collaboration raises the average comprehensive score by 9.0% and reduces the range of agent evaluation scores by 78.2%. Across the four LLMs, the overall problem detection rate is 81.8%. For Gemini 3.1 Pro, which achieves the highest overall detection rate (87.1%), the detection rates for equipment, safety, and mechanism-related problems are 100.0%, 80.0%, and 86.7%, respectively. These results clarify the limits of the current framework: equipment-related problems are identified more completely, whereas safety problems and subtle mechanistic issues can still be missed and therefore require expert supervision. This study presents a reproducible human – AI collaborative approach that uses multi-agent debate to improve LLM-generated models of green ammonia processes under fluctuating wind and solar power.

Keywords: multi-agent; green ammonia production; large language model; collaborative framework

引 言

全球能源体系正加速向低碳化转型，风能、太阳能等可再生能源装机容量持续增长，但其固有的间歇性与波动性对能源储存与运输提出了新的技术要求[1-2]。绿氨作为一种零碳能源载体，具备高能量密度、便于液化运输、可直接用作燃料或进一步裂解制氢等优势，被认为是连接波动性可再生能源与稳定终端用能领域的关键纽带[3-5]。

绿氨生产通常依托风光发电，通过电解水制氢，再与空气分离得到的氮气进行哈伯-博世合成，这一过程涉及电解槽、压缩机、反应器、分离单元等多个设备，涉及电化学、热力学、反应动力学及多相流传质过程的强耦合[6-8]。在风光供电的波动条件下，电解槽无法持续供应恒定流量的氢气，导致合成氨原料供应不稳定。传统的哈伯-博世合成氨工艺通常以稳态模式运行，调节速度较慢，而可再生能源的电力供应变化可达每分钟3%以上，两者之间存在显著的时间尺度错配[9-12]。因此，绿氨生产系统需要在宽负荷范围内（通常为额定负荷的30%–100%）协同运行，负荷快速变化会引发反应器和管道内的压力波动，对设备造成机械疲劳，长期运行将显著增加运维成本[13-15]。

精确的数学建模是实现工艺优化、柔性调控和故障诊断的基础。绿氨流程涉及电解槽极化特性、合成反应器非理想流动、多组分分离等多个子系统，且各单元之间存在复杂的物料和能量耦合关系[16-18]。在变负荷工况下，H_2/N_2比、惰性气体含量等操作参数的调整直接影响反应平衡与整体效率，模型不仅要描述稳态行为，更需要反映工况切换时的动态响应边界[19-21]。然而，传统建模方法高度依赖领域专家的经验知识，研究人员需要从大量文献中筛选适用于特定工艺的模型结构，理解不同简化假设的适用边界，推导控制方程并标定关键参数。Du和Yang明确指出，手动参数调优和对专家主观经验的依赖是化工过程模拟中长期存在的挑战[22]；Van de Vijver等亦将专家用户高度参与和耗时计算列为21世纪该领域的主要障碍[23]。上述问题在绿氨这一多学科强耦合场景中尤为突出：风光波动条件下，传统哈伯-博世反应器受热惯性制约，与波动性低碳能源输入匹配较差。传统建模过程通常耗时数周甚至数月，且学习门槛较高，已成为制约绿氨技术快速迭代与工程人才培养的现实瓶颈[24-26]。

近年来，大语言模型（LLM）在知识检索、逻辑推理和文本生成方面展现出强大能力，为突破这一瓶颈提供了新思路。基于LLM的多智能体系统正在成为化学工程领域的前沿方向。已有研究将多智能体系统用于化学工程，探索其在复杂化工工作流拆解、场景化监测架构定制及多源过程数据融合等方面的应用。在具体应用方面，多个代表性的多智能体LLM框架已被提出。CeProAgents构建了一个层次化多智能体系统，通过知识、概念、参数三层智能体的协同分工，实现化学过程开发的端到端自动化[27]。ChemGraph针对计算化学工作流，利用多智能体将复杂任务分解为子任务，在13个基准任务上验证了多智能体协作执行专业任务的可行性[28]。此外，还有工作提出了从文本描述到可执行化学过程模拟的多智能体工作流，将设计任务分解为任务理

解、拓扑生成、参数配置和评估分析等专用智能体。

上述研究在方法论上主要关注利用多智能体系统的分工协作能力，实现对复杂化工设计或模拟任务的“自动化执行”或“端到端完成”。相比之下，如何针对智能体协作输出的结果质量（如模型选择合理性、控制方程正确性、参数工程可行性及多领域约束一致性）构建系统化的质量增强与校验机制，目前仍缺乏深入探索。具体而言，LLM固有的“事实性幻觉”（生成看似合理但实质错误的方程或参数）、“工程经验缺失”（输出参数可能偏离设备可行区间）以及“安全合规意识薄弱”（易忽略高温高压、易燃易爆介质等安全约束）等缺陷，在多智能体协同执行任务的过程中并不会自动消除[29-31]。如何有效抑制模型错误、补齐工程实践约束、实现多专业视角的一致性对齐，仍是该领域亟待解决的关键难题。

针对上述问题，本文提出一种辩论驱动的多智能体协作增强框架。该框架首先由LLM根据结构化提示词独立完成初始建模方案，涵盖模型推荐、控制方程、参数范围及关键假设等要素。随后，四个专业智能体分别负责工艺建模、设备工程、安全合规和数据分析，并从各自专业视角对该初始方案进行独立评审，指出其中的错误、缺失、不一致或不合规之处，并给出量化评分与专业依据。总控收敛智能体收集各专业智能体的评审意见，识别分歧点，组织定向辩论。在辩论过程中，各专业智能体需基于内置领域知识库相互挑战对方的判断，并提出具体的修正建议；原LLM在此过程中被迫重新审视其输出，吸收有效批评，对方案进行修正。上述“评审—辩论—修正”环节可多次迭代，直至各专业智能体的评分极差不大于预设阈值（如1分）或达到最大迭代次数，最终输出一份经过多维度交叉验证、符合工程规范的高质量建模方案。该框架的核心价值并非简单评估输出好坏，而是通过多轮交叉辩论，迫使LLM在迭代中自主完成质量提升，类似于人类科研团队通过内部同行评议完善研究成果的过程。

为验证上述辩论驱动的多智能体协作增强框架的有效性，本研究以风光波动供电下的绿氨生产工艺为应用场景，设计了五个复杂度递进的典型建模任务，涵盖电解水制氢单元稳态建模、哈伯-博世合成反应器动力学建模、变负荷条件下动态响应分析、全流程系统综合建模以及合成工段异常工况诊断。选取四种主流LLM（GPT-5.4、DeepSeek V3.2、Qwen 3.5 Max、Gemini 3.1 Pro），以单智能体迭代输出为基线，以多智能体协作增强输出为实验组，邀请三位绿氨工艺领域专家对两类输出进行双盲评估，从合规性、鲁棒性、准确性、完整性四个维度量化评分。本研究的主要贡献体现在以下方面：（1）提出了一种辩论驱动的多智能体协作增强框架，通过先建模、后多专业交叉辩论与共识收敛的机制，迫使LLM在迭代中自主修正错误、补全工程约束、对齐多专业视角要求，从而显著提升建模方案的质量与工程可靠性；（2）以绿氨生产为场景，通过五个梯度任务系统验证了该框架在不同复杂度下的增强效果，揭示了辩论增益随任务复杂度增加而增强的规律；（3）明确了当前框架的问题识别能力边界——四款模型整体问题检出率为81.8%；以问题检出率最高的Gemini 3.1 Pro为例，其设备问题、安全问题和机理问题检出率分别为100.0%、80.0%和86.7%，表明设备类问题识别较充分，而安全与深层机理问题仍存在漏检，需要人工复核，从而为“人机协作、以人为主”的智能化工建模范式提供了可复现的操作框架与量化依据。

1 方法论

本节详细阐述了辩论驱动的多智能体协作增强框架的设计与实现方法。首先介绍总体工作流程，然后分别介绍各专业智能体的角色与知识库配置、提示工程、辩论收敛机制以及质量评估方法。

1.1 总体工作流程

本框架的核心思想是通过多个专业智能体的多轮评审、辩论与方案修正，迫使模型逐步提升输出质量，直至达成共识收敛，具体框架如图1所示。系统首先接收一个绿氨生产工艺建模任务，例如电解槽稳态建模或合成反应器动力学建模，再根据结构化提示词调用LLM生成初始建模方案。该方案包括推荐的数学模型类型、核心控制方程、关键本构关系、符号定义、参数取值范围以及所采用的简化假设。初始方案生成后进入多专业独立评审阶段，四个专业智能体，即工艺建模智能体、设备工程智能体、安全合规智能体和数据分析智能体，分别从各自专业视角对当前方案进行独立评审。每个智能体按照预定义的评估维度，包括合规性、鲁棒性、准确性和完整性，给出1到10分的量化评分，并

输出具体的评审意见，包括优点、缺陷、核心问题以及修改建议，且所有评审均需依据内置的领域知识库。随后进入辩论与方案修正阶段，总控收敛智能体收集四个专业智能体的评审结果，计算各维度评分的极差。若某一维度的极差大于1分，则判定为存在分歧，总控智能体针对分歧点组织定向辩论，要求相关智能体基于知识库补充专业依据并相互挑战对方的判断。辩论结束后，总控智能体将各专业智能体达成的共识意见汇总形成修正指导，这些共识意见包括错误修正、缺失补全和约束对齐等内容。然后，建模专家，即原LLM或专门负责建模的智能体，依据该修正指导对当前的建模方案进行针对性修订，生成修正后的方案。修正后的方案不再直接输出，而是重新进入多专业独立评审阶段，由四个专业智能体再次评审。上述评审、辩论、修正与再评审的闭环迭代可重复多次，直至所有评估维度的评分极差均小于等于1分，或达到预设的最大迭代次数，本研究中最大迭代次数设为3次。此时，总控智能体终止迭代，输出最终经过多维度交叉验证的建模方案，同时生成一份完整的评估报告，记录各轮评分变化、分歧焦点、辩论摘要及收敛结论。

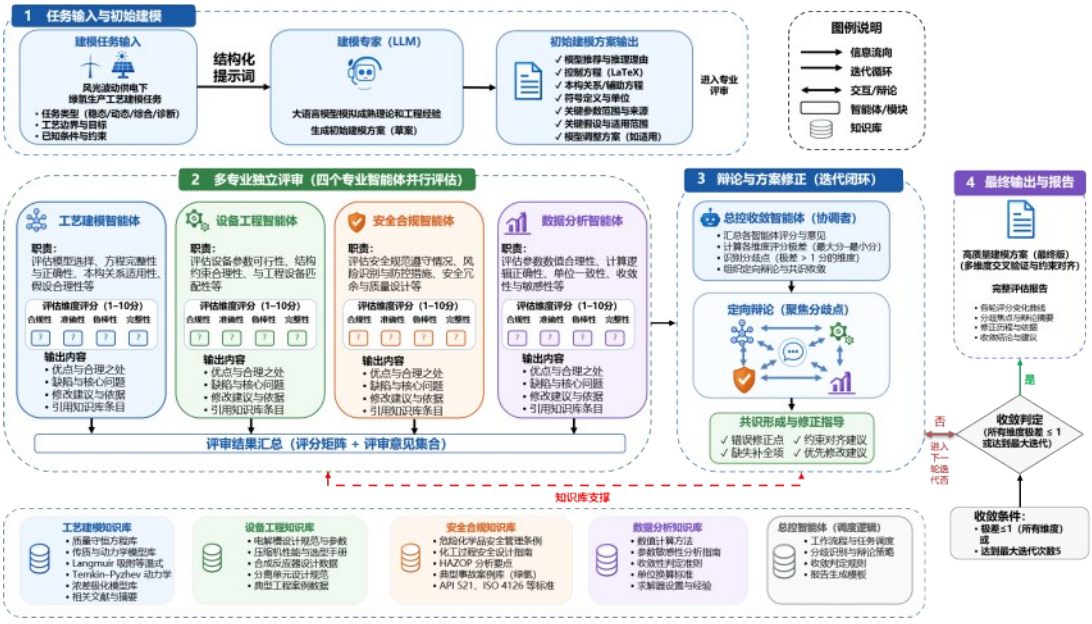

图1　用于绿色氨工艺建模的多智能体协作增强框架

Fig.1　Multi-agent collaborative enhancement framework for green ammonia process modeling

1.2　面向绿氨生产的多智能体系统实现

为使上述多智能体协作框架能够针对绿氨生产建模任务具体执行，本文基于Coze低代码平台构建多智能体工作流。系统由建模智能体、四个专业评估智能体和总控收敛智能体组成，各智能体作为相互独立的功能节点进行配置，并通过结构化文本字段完成信息传递。建模智能体负责根据任务信息生成初始建模方案，四个专业智能体分别完成独立评审，总控智能体负责评分解析、分歧识别、辩论调度和修正指导生成。整个执行过程不依赖本地模型训练，各LLM均通过云端应用程序编程接口（API）调用，本地计算资源主要承担工作流调度、结果解析及实验记录。

为保证各智能体围绕同一绿氨生产问题开展推理，每次任务均按照统一的数据结构输入四类信息，即工艺描述、模拟目标、已知假设与工程约束以及迭代历史信息。其中，工艺描述包含当前任务涉及的主要单元、设备规格及运行状态；模拟目标规定需输出的模型类型、控制方程及分析目标；已知假设与约束用于限定模型边界和允许的简化条件；在第二轮及后续迭代中，系统进一步加入上一轮专业评审意见及总控生成的修正指导。由此形成“工艺背景输入—建模输出—专业反馈—修正输入”的闭环信息传递过程，避免后续智能体在迭代过程中脱离原始绿氨工艺条件重新生成答案。

针对绿氨生产涉及电化学、反应动力学、设备运行及安全规范等多学科知识的特点，本文采用专业知识分区的方式配置智能体知识库。工艺建模智能体挂载质子交换膜（PEM）电解槽极化模型、Temkin–Pyzhev合成氨反应动力学、质量与能量守

恒以及反应器建模等知识，用于判断模型选择、控制方程及简化假设的合理性；设备工程智能体挂载PEM电解槽、压缩机和合成反应器等设备标准与典型参数范围，用于识别工作压力、电流密度、负荷范围及设备匹配性问题；安全合规智能体挂载氢气使用、压力容器、可燃及有毒气体检测以及超压保护等相关标准，用于识别高压氢气、氨介质及压力系统中的安全约束遗漏；数据分析智能体配置数值计算方法、收敛判据和单位换算规则，用于检查参数量纲、数值计算逻辑和求解条件。通过这种知识分区方式，不同智能体对同一建模方案形成互补而相对独立的专业判断。

在运行过程中，建模智能体首先依据结构化提示词生成包含模型推荐、控制方程、关键参数和简化假设的初始方案。随后四个专业智能体独立读取相同的方案及绿氨工艺背景，并分别给出合规性、鲁棒性、准确性和完整性评分及对应的专业依据。对于评价维度 j，总控智能体计算四个专业评分的极差：

$$D_j = \max_i(S_{ij}) - \min_i(S_{ij}) \tag{1}$$

式中，D_j 为第 j 个评价维度上四个专业智能体评分的极差，S_{ij} 为第 i 个专业智能体对第 j 个评价维度的评分；i=1,2,3,4 对应四个专业智能体，j=1,2,3,4 对应合规性、鲁棒性、准确性和完整性四个评价维度。当任一 $D_j > 1$ 时，系统认为该维度存在实质性专业分歧，并提取该维度评分最高与最低智能体的评审依据，形成针对具体争议点的辩论任务。当所有 $D_j \leqslant 1$ 时，则认为各评价维度的评分分歧已满足收敛要求。例如，在变负荷合成氨案例中，若工艺建模智能体认为降低回路压力有利于柔性运行，而设备工程智能体认为该方案可能进入压缩机低效率或设备受限区域，总控智能体不会直接平均两者意见，而是生成针对“降低回路压力的工艺收益与设备可行边界”的定向辩论议题，要求双方分别基于反应动力学、设备性能及知识库依据重新论证。

需要说明的是，D_j 的功能定位是表征同一评价维度下不同专业智能体之间的横向意见分歧，用于触发定向辩论并判定专业共识，其本身不反映单个智能体在不同评价维度之间的评分差异。为补充描述评价结果的维度均衡性，本文采用维度标准差作为纵向差异指标，其计算式为：

$$\sigma_{\text{dim}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2}{4}} \tag{2}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4} \tag{3}$$

式中，σ_{dim} 为四个评价维度得分的总体标准差，$\bar{x}$ 为四个维度得分的算术平均值，x_j 为第 j 个维度的得分(j=1,2,3,4)。当针对单个智能体计算时，$x_j=S_{ij}$；当针对最终建模方案或不同协作范式进行比较时，x_j 为四个评价维度的聚合得分。维度标准差越小，表明评价结果在四个维度上的分布越均衡。需要强调的是，维度标准差仅作为结果诊断指标，不纳入辩论触发或收敛判据，因为合规性、鲁棒性、准确性和完整性分别对应不同质量属性，维度间存在真实得分差异并不等同于专业意见冲突。因此，D_j 与维度标准差分别刻画“同维度跨智能体一致性”和“同评价对象跨维度均衡性”，二者共同构成对评价结果的互补描述。

辩论完成后，总控智能体不直接修改建模方案，而是将各智能体形成的共识内容整理为结构化修正指导，包括待纠正错误、需要补充的方程或参数以及需增加的工程和安全约束。建模智能体依据该指导生成新的建模方案，并重新进入专业评审。上述过程重复进行，直到四个评价维度的评分极差均不大于1分，或达到预设的最大迭代次数。由此，本文将一次性的LLM绿氨建模输出转化为由领域知识约束、专业分歧触发和反馈修正共同驱动的闭环迭代过程。

1.3 专业智能体角色与知识库配置

为了实现多视角、专业化的交叉校验，本框架设置了四个专业智能体，分别对应建模过程中的核心专业维度：工艺建模、设备工程、安全合规与数据分析。每个智能体均具备明确的角色定位和职责边界，并依据其专业领域挂载相应的知识库，以确保评估意见的准确性和可追溯性。工艺建模智能体负责评估模型选择的合理性、控制方程的完整性与正确性、本构关系的适用性以及简化假设的工程可接受度。其知识库主要包含质量守恒方程库、典型传质与动力学模型(如Langmuir吸附等温式、Temkin-Pyzhev反应速率方程)、浓差极化模型库以及相关领域文献摘要。设备工程智能体负责评估建模方案与工程设备之间的匹配性，重点检查操作压力、温度、流量等关键参数是否处于设备允许的

可行区间，以及设备结构约束（如塔径高比、膜面积、催化剂床层高度）是否得到合理体现。其知识库主要包括电解槽、压缩机、合成反应器、分离单元等设备的设计规范、标准参数范围以及典型工程案例数据。安全合规智能体负责评估建模方案是否符合高温高压及易燃易爆介质（特别是氢气和氨）的安全规范，审查设计冗余、泄漏防控、紧急切断等安全措施的充分性，并识别可能引发连锁事故的潜在风险点。其知识库涵盖危险化学品安全管理条例、化工过程安全设计指南、危险与可操作性分析（HAZOP）要点、绿氨工厂典型事故案例库以及相关安全标准（如美国石油学会（API）521标准、国际标准化组织（ISO）4126标准等）。数据分析智能体负责评估建模方案中量化计算部分的准确性，包括参数取值的合理性、单位一致性、计算逻辑的正确性以及收敛性判断的可靠性。其知识库包括常用数值计算方法（如有限差分、有限元、求解器设置）、参数敏感性分析指南、模型收敛性判定准则以及单位换算标准。上述四个智能体均通过结构化提示词实现角色绑定，并在Coze低代码平台中通过上传本地文档或链接方式挂载对应的知识库，具体的知识库清单如表1所示。表中GB、GB/T、TSG和API STD分别为强制性国家标准、推荐性国家标准、特种设备安全技术规范和美国石油学会标准的代号；Nm^3表示标准状态下的立方米。此外，框架还设置了一个总控收敛智能体，其职责是协调工作流程、汇总各专业智能体的评分与评审意见、识别分歧维度、组织定向辩论并判定收敛条件，总控收敛智能体不参与具体专业评估，因此不需要挂载领域知识库，其核心能力来源于任务调度逻辑和收敛判定规则。

表1　各智能体知识库清单

Table 1　List of knowledge bases for each agent

智能体	知识类型	具体内容
工艺建模	核心文献	Springer T.E. et al. *J. Electrochem. Soc.*, 1991, 138(8): 2334 – 2342（PEM电解槽极化模型）；Nielsen A. *Catalytic Ammonia Synthesis*. Plenum Press, 1983（Temkin–Pyzhev动力学）；Froment G.F. et al. *Chemical Reactor Analysis and Design*. Wiley, 2011（多床层反应器建模）
	方程库	质量守恒/能量守恒方程模板；Temkin–Pyzhev反应速率方程；Langmuir吸附等温式；浓差极化过电位模型（以LaTeX格式存储）
设备工程	设计规范	GB/T 45539–2025《PEM电解槽技术要求》；GB/T 45541–2025《PEM电解槽性能测试方法》；API STD 617:2022 *Axial and Centrifugal Compressors*
	工程案例	典型绿氨装置设备规格：PEM电解槽产氢量500 Nm^3/h、额定电流密度1.5 A/cm^2；合成氨回路年产18万吨，三床层间冷式反应器
安全合规	行业标准	GB 4962–2008《氢气使用安全技术规程》；TSG 21–2016《固定式压力容器安全技术监察规程》（含2021年第1号修改单）；GB/T 50493–2019《可燃气体和有毒气体检测报警设计标准》；API STD 521:2020 *Pressure–relieving Systems*；ISO 4126–1:2013 *Safety devices for protection against excessive pressure*
	事故案例	合成氨领域典型事故案例：日本昭和电工（Showa Denko）川崎工厂合成氨设备爆炸；内蒙古鄂尔多斯九鼎化工公司净化工段换热器泄漏引发爆炸；四川眉山金象化工合成氨车间阀门故障致管道氢气泄漏引发爆炸；宁夏捷美丰友化工公司安全阀起跳后液氨直接进入火炬管线，因未设气液分离罐致液氨从火炬口喷出气化扩散，造成200米范围内41人急性氨中毒；高温高压下氢气/氮气对钢材的氢蚀及渗氮加速设备疲劳腐蚀，引发物理爆炸；合成工段高压设备密封失效导致粗合成气瞬间泄漏喷出。
数据分析	计算方法	电流效率计算公式（$\eta_{current}$ = 实际产氢量/理论产氢量 × 100%，法拉第效率基于电流积分）；氨转化率定义（X_{NH3} = 出口NH_3摩尔流量/理论最大NH_3摩尔流量 × 100%）；能量效率计算（基于氢气低位热值（LHV）为120 MJ/kg）；有限差分法（前向/中心差分格式）与Newton–Raphson迭代（收敛容差1×10^{-6}）
	单位换算	化工过程常用单位换算表（压力、温度、流量、电流密度）

为系统抑制LLM在化工建模输出中的事实性幻觉，各专业智能体的评审规则中嵌入了面向不同错误类型的幻觉识别逻辑。根据本文实验统计的幻觉表现形式，将幻觉分为五类：机理幻觉（编造不存在的反应路径或传递机理）、方程幻觉（自行推导错误的控制方程形式或本构关系）、参数幻觉（给出偏离工业可行区间的参数值或虚构参数）、安全幻觉（忽视或淡化高温高压、易燃易爆介质的固有风险）、符号幻觉（使用非标准符号、单位混用或量纲不一致）。四类专业智能体的评审规则针对上述幻觉类型设置了差异化的识别机制：工艺建模智能体重点拦截机理幻觉与方程幻觉，要求所有非经验方

程必须注明出处，经验关联式必须给出适用条件与文献依据；设备工程智能体重点拦截参数幻觉，要求关键参数（压力、温度、电流密度等）必须落在工业可行区间内，并标注数值来源（设计规范或工程案例）；安全合规智能体重点拦截安全幻觉，要求建模方案中必须体现防爆分区、超压保护、泄漏检测等安全约束，禁止以“简化假设”为由省略法定安全措施；数据分析智能体重点拦截符号幻觉，要求所有符号按照化工标准定义物理意义与单位，并进行量纲一致性校验。若任意智能体在评审中发现输出方案存在上述幻觉迹象，需在“核心问题”部分明确标注幻觉类型、具体位置及专业依据，将其作为强制修正项纳入修正指导，确保幻觉问题在后续迭代中被定向消除。

上述幻觉识别逻辑与附录A中各智能体的提示词条款形成直接映射关系。具体而言：机理幻觉与方程幻觉的防控对应工艺建模专家提示词中“禁止无依据假设与事实性幻觉”“若存在知识边界、机理矛盾或事实错误，需明确指出并说明专业依据”等条款，以及建模提示词中“严禁编造不存在的方程、参数或数据”的硬性约束；参数幻觉的防控对应设备工程专家提示词中“若输入参数超出工业可行区间或设备适配矛盾，需明确指出并给出工程化修正范围”的要求；安全幻觉的防控对应安全合规专家提示词中“若输出存在安全风险或合规漏洞，需明确指出风险等级与合规依据”的评审职责；符号幻觉的防控对应数据分析专家提示词中要求“所有量化判断必须依托化工标准计算方法，禁止数据捏造与计算逻辑错误”“若量化结果存在逻辑矛盾或单位错误，需明确指出并给出正确计算方法”等条款；各类幻觉的标注、追溯与强制修正机制则由总控收敛专家提示词中“禁止无依据补充或修改专家观点”“若专家间存在分歧需明确列出分歧点及各方依据”等原则提供流程性保障。上述条款以负面清单与正面约束相结合的方式，在各智能体的结构化提示词中落地执行，并通过预实验对各智能体的提示词进行了校准，确保其在不同案例复杂度下均能有效识别幻觉。

1.4 结构化提示工程

结构化提示工程是本框架实现高质量输出的关键环节，通过精心设计的提示词模板，可以引导LLM生成格式统一、内容完整、便于后续专业智能体评估的建模方案，同时也为各专业智能体的评审提供明确的输出规范。本框架针对初始建模、四个专业评估和总控收敛三类核心任务分别设计了提示词模板，三者在设计原则上遵循共同的核心思想，即明确角色定位、分解任务模块、强制结构化输出。

初始建模提示词的核心目标是使LLM在无人工干预的情况下生成一份可直接供专业智能体评审的建模草案。为此，提示词首先为模型设定了“绿氨工艺建模专家”的角色身份，并明确要求其回答以成熟理论和工程实践为依据，禁止编造不存在的方程或参数。接着，提示词将建模任务强制分解为三个递进的子任务：模型推荐、控制方程构建和模型调整。在模型推荐部分，要求模型说明推荐理由、关键假设、适用范围以及主要优缺点；在方程构建部分，要求以LaTeX格式呈现核心控制方程，同时给出关键本构关系或辅助方程，并对所有符号进行明确的物理意义和单位定义；在模型调整部分，要求针对特定工艺现象（如结垢、催化剂失活、膜污染等）说明模型的修改方案和调整依据。此外，提示词还要求模型提供关键参数的典型取值范围，并标注引用的参考文献或工程数据来源。这种“任务分解-格式约束-依据追溯”的设计，确保了初始方案具备足够的专业深度和可评估性。

专业评估提示词为四个专业智能体分别定制，其核心思想是限定每个智能体的评估边界，避免跨领域评论或发散性输出。以工艺建模智能体为例，提示词明确限定其只能从“工艺机理适配性”和“方程构建合理性”两个维度进行评价，禁止讨论设备工程、安全合规或数据分析的内容。评估任务同样被分解为固定模块：首先要求智能体基于案例背景和用户输入内容提炼评估依据摘要，然后按照合规性、鲁棒性、准确性、完整性四个维度分别给出1-10分的量化评分，接着依次评价模型推荐和方程构建的优点与缺陷，汇总核心问题，最后提出优化建议并注明专业依据。其他专业智能体的提示词结构类似，但评估侧重点根据其知识领域进行调整，设备工程智能体聚焦设备匹配性与参数可行性，安全合规智能体聚焦安全约束与排放标准，数据分析智能体聚焦量化准确性与单位一致性。所有评估提示词均要求智能体在输出中标明依据来源，从而保证评审结果可追溯、可复验。需要说明的是，此处“工艺机理适配性”和“方程构建合理性”仅用于限定工艺建模智能体的专业内容边界，并非替代后

续四个统一质量评价维度；四维评分均在各智能体自身的专业职责范围内完成。

总控收敛提示词的设计侧重于流程调度与观点整合。该提示词首先要求总控智能体扮演“中立的总控专家”角色，不参与具体专业评估。其工作流程被固定为：收集四个专业智能体的评分与评审意见，计算各维度的平均分、最高分、最低分及极差，判断是否收敛（极差≤1分）。若未收敛，则要求总控智能体按专业维度分类梳理分歧点，明确列出各方观点及其依据，然后生成定向辩论议题（例如“请工艺建模智能体与设备工程智能体就参数工程可行性分歧补充依据”），并分发至相关智能体。在辩论反馈后，总控智能体需更新共识结论和优化建议，重新判定收敛状态，最终输出一份包含评分汇总、共识结论、分歧说明和统一优化建议的综合评估报告。该提示词还强调，所有总结和判定必须基于各专业智能体的原文输出，不得添加总控智能体自身的专业判断，从而保证报告的中立性和客观性。

上述三类提示词模板的完整内容见附录A。在实际运行中，所有提示词均包含固定的占位符（如[工艺描述]、[模拟目标]、[已知假设]），由系统在调用时自动填入具体案例信息。此外，为了平衡生成内容的确定性与多样性，所有模型推理时的温度参数均设为0.1（经预实验验证为最佳值），以降低随机采样带来的输出波动，确保评审结果的可重复性。

1.5　辩论收敛机制

辩论收敛机制是本框架实现输出质量增强的核心环节，该机制旨在通过多轮交叉挑战与共识引导，迫使建模方案在迭代中逐步消除专业视角之间的分歧，最终达到各维度要求的一致性收敛。收敛过程由总控收敛智能体全程协调，具体包括分歧判定、定向辩论、方案修正与收敛条件四个关键子环节。

分歧判定是启动辩论的前提。总控收敛智能体收集四个专业智能体对当前建模方案在各评估维度上的评分后，分别计算合规性、鲁棒性、准确性和完整性四个维度的评分极差（即最高分与最低分之差）。若任一维度的极差大于1分，则判定该维度存在实质性分歧，需要启动辩论；若所有维度的极差均小于等于1分，则认为当前方案已达成共识，可直接进入输出阶段。这一阈值（1分）基于预实验确定。

为系统论证上述参数的合理性，本文选取案例体系中复杂度最高的案例4（全流程系统综合建模）与案例5（合成工段故障诊断）作为校准基准。预实验阶段采用单一基础模型进行参数压力测试，其目的在于比较不同收敛阈值与迭代上限下的收敛行为、输出质量和计算成本，而非进行跨模型性能比较；正式实验则在预实验完成参数校准后，将同一收敛规则统一应用于GPT-5.4、DeepSeek V3.2、Qwen 3.5 Max和Gemini 3.1 Pro四款模型。预实验分别测试三个收敛阈值（0.5分、1.0分、1.5分）与三个迭代上限（2轮、3轮、4轮），以综合得分与单案例平均处理耗时为评价指标。各配置均按实际执行轮次统计耗时，即若提前满足所有评价维度$D_j \leqslant 1$，则立即终止；否则运行至设定的最大迭代轮次。结果如表2、3所示。

表2　案例4预实验对比

Table 2　Comparative pre-experiments of case 4

收敛阈值	迭代上限	实际迭代轮次	收敛状态	综合得分	单案例耗时(min)
0.5分	2轮	2轮	未收敛(达上限)	7.1	24
0.5分	3轮	3轮	未收敛(达上限)	7.3	35
0.5分	4轮	4轮	未收敛(达上限)	7.9	61
1.0分	2轮	2轮	未收敛(达上限)	7.0	24
1.0分	3轮	3轮	已收敛	7.9	39
1.0分	4轮	3轮	已收敛	7.8	37
1.5分	2轮	2轮	已收敛	6.9	21
1.5分	3轮	2轮	已收敛	7.1	23
1.5分	4轮	2轮	已收敛	6.5	19

表3　案例5预实验对比

Table 3　Comparative pre-experiments of case 5

收敛阈值	迭代上限	实际迭代轮次	收敛状态	综合得分	单案例耗时(min)
0.5分	2轮	2轮	未收敛(达上限)	6.4	25
0.5分	3轮	3轮	未收敛(达上限)	6.6	37
0.5分	4轮	4轮	未收敛(达上限)	7.3	58
1.0分	2轮	2轮	未收敛(达上限)	6.7	25
1.0分	3轮	3轮	未收敛(达上限)	7.7	35
1.0分	4轮	4轮	已收敛	7.8	61
1.5分	2轮	2轮	已收敛	6.3	26
1.5分	3轮	2轮	已收敛	6.4	24
1.5分	4轮	2轮	已收敛	6.7	22

预实验结果表明，收敛阈值的选取对收敛行为与输出质量具有决定性影响。当阈值为0.5分时，即使迭代至4轮，案例4与案例5的评分极差仍未收

敛，所有实验均在达到轮次上限后终止，表明该标准过于严苛，超出了当前多智能体辩论机制在复杂工程问题上的收敛能力边界。当阈值为1.5分时，案例4与案例5均于第2轮即满足该宽松标准，但两案例的综合得分均显著低于1.0分阈值下经充分辩论后的得分——案例4在1.5分/4轮配置下得分仅6.5，较1.0分/3轮配置低1.4分；案例5在1.5分/4轮配置下得分仅6.7，较1.0分/3轮配置低1.0分——说明过于宽松的收敛条件会导致辩论在达成真正共识前被过早截断，输出质量显著下降。

不同任务类型在收敛速度上亦呈现差异：案例4（全流程综合建模）在1.0分阈值下于第3轮达成收敛；而案例5（故障诊断，溯因推理任务）在1.0分阈值下于第3轮尚未收敛（得分7.7），至第4轮方可达标（得分7.8）。这一差异源于故障诊断问题固有的多解性——不同专业视角对“最可能的故障原因”的判定存在更持久的分歧，需要更充分的辩论方能收敛。

从质量—效率权衡看，3轮更适合作为统一的最大计算预算，而非保证所有任务均完全收敛的充分条件。在1.0分阈值下，案例4于第3轮达到完全收敛，继续增加轮次未带来稳定的质量增益；案例5则在第3轮后仍保留局部专业分歧，至第4轮方满足所有评价维度$D_j \leq 1$。与此同时，案例5由3轮增加至4轮时，综合得分仅由7.7提高至7.8，增益0.1分，而处理时间由35 min增加至61 min，增加74.3%。若上限仅设为2轮，案例4与案例5均未达到完全收敛，且综合得分分别仅为7.0和6.7。由此可见，继续增加至4轮虽然能够进一步消除残余分歧，但边际质量收益较小而计算成本显著增加。综合考虑收敛程度、输出质量与处理耗时，本文最终将1.0分设为收敛阈值，将3轮设为正式实验的统一最大迭代预算；若3轮内提前满足全部$D_j \leq 1$则提前终止，否则在第3轮截断输出并记录残余分歧，供后续专家复核。

需要进一步区分预实验、正式实验单次运行以及第3.2节汇总统计的三个层级。预实验采用单一基础模型，用于校准收敛阈值和最大迭代预算；其“完全收敛”严格按单次运行的四个评价维度判断，只有当所有$D_j \leq 1$，即$max(D_j) \leq 1$时才记为完全收敛。因此，案例5在1.0分阈值、3轮上限的预实验配置下仍有至少一个评价维度$D_j > 1$，故记为“未收敛（达上限）”，至第4轮全部维度满足条件后才记为完全收敛。正式实验则对4款基础模型×5个案例统一采用“全部$D_j \leq 1$或达到3轮上限”的双终止机制，以保持不同模型之间一致的计算预算；达到3轮上限时若仍存在$D_j > 1$，则保留当前输出及残余分歧记录，不将其等同于完全收敛。第3.2节表8进一步对正式实验结果进行跨维度、跨模型汇总，因此其“平均极差”仅用于描述总体分歧缩减程度，不能替代单次运行的收敛判据。

定向辩论是分歧消除的主要手段。总控收敛智能体首先识别出存在分歧的维度以及在该维度上评分最高和最低的专业智能体，随后召集相关智能体进行定向讨论。辩论过程中，总控智能体要求每个参与智能体基于自身知识库提供具体的专业依据（如引用标准参数范围、安全规范条款或经典方程形式），并针对对方的质疑逐条回应。辩论采用回合制：每轮辩论中，评分较低的智能体首先说明其低分理由，评分较高的智能体则对其理由进行反驳或补充；之后双方互换角色。辩论持续进行，直至双方不再提出新的分歧点，或总控智能体判定当前辩论已覆盖所有主要争议。辩论全程记录留存，供后续修正参考。

方案修正是辩论成果向质量提升转化的关键步骤。辩论结束后，总控收敛智能体将各专业智能体在辩论中达成的共识性结论（包括必须修正的错误、需要补全的缺失信息、应当对齐的工程约束等）汇总形成修正指导。修正指导采用结构化自然语言描述，明确列出每条修改建议的依据来源和预期修改方向。随后，建模专家（即初始进行建模的LLM）依据修正指导对当前建模方案进行针对性修改。修改完成后，生成的新方案不再直接输出，而是重新进入评审阶段，由四个专业智能体再次独立评分。这一“评审—辩论—修正—再评审”的闭环确保了每一次迭代都基于前一轮的共识结论，避免漫无目的的反复。

本框架设置双重收敛条件：第一，当所有评估维度的评分极差均小于等于1分时，判定为共识收敛，迭代终止；第二，为避免无限循环，设置最大迭代次数为3次。当任一收敛条件满足时，总控收敛智能体终止迭代，将当前版本的建模方案作为最终输出，并生成一份完整的收敛报告。该报告记录了每轮迭代的评分变化、各轮辩论摘要、修正指导的具体内容以及最终的收敛判定依据，为后续人工审查或方法复现提供完整的追溯信息。

通过上述机制，本框架将一次性的、不可控的LLM输出，转化为多轮可控的、逐步逼近多专业共识的迭代过程。辩论与修正交替进行，迫使模型在专业视角的交叉挑战中持续提升输出质量，最终输出经过系统化验证的、工程上可用的建模方案。

1.6 质量评估维度与量化标准

为系统衡量建模方案的质量并支持辩论收敛过程中的分歧判定，本框架定义了四个正交的评估维度：合规性、鲁棒性、准确性和完整性。每个维度采用1 - 10分制进行量化评分，分数越高表示该维度的表现越优。四个专业智能体（工艺建模、设备工程、安全合规、数据分析）均需独立对当前建模方案在所有四个维度上分别给出评分，从而形成4×4的评分矩阵。这种设计使得同一建模方案在不同专业视角下获得多维度的交叉评估，为后续分歧识别与辩论收敛提供了充分的比对依据。各维度的具体定义与评分锚定标准详见表4。

表4 评估维度定义

Table 4 Definition of evaluation dimensions

得分	合规性	鲁棒性	准确性	完整性
9–10	完全符合适用规范及专业约束	对扰动、工况变化或数值条件具有良好稳健性	专业判断、模型、参数或计算准确，无明显错误	专业核心要素完整，含必要的深层分析
7–8	基本符合，有次要疏漏	总体稳健，极端工况或边界条件说明不足	主体正确，存在次要错误或遗漏	核心要素齐全，非关键遗漏
5–6	存在明显不符合项，接近专业底线	对部分扰动或工况变化较敏感	核心内容存在明显错误	遗漏多个子问题，支撑不足
1–4	严重违反适用规范或专业底线	稳健性差，关键条件下失效或不稳定	关键判断错误或违背基本原理	核心要素严重缺失或无关

需要说明的是，四类专业智能体虽然采用相同的四个质量维度和1 - 10分量表，但并非以完全相同的专业对象重复评分。本文采用“通用质量维度+专业职责映射”的两层评分规则：表4规定合规性、鲁棒性、准确性和完整性的统一含义及分值锚点，用于保证评分尺度一致；各专业智能体则仅在自身职责边界内解释四个维度并给出分数。若发现明显属于其他专业领域的问题，仅在评审意见中标注并交由对应智能体处理，不作为本专业重复扣分项。由此，同一评价维度下不同智能体的评分反映的是该质量属性在不同专业约束下的一致性，而非对同一检查项的重复评价。

具体评分侧重点如下：(1)工艺建模智能体：合规性重点判断模型是否遵循质量/能量守恒、成熟机理及适用条件；鲁棒性关注模型与假设在变负荷和扰动下是否仍成立；准确性关注模型选型、控制方程、本构关系和机理描述；完整性关注方程组、边界条件、关键假设及必要耦合关系是否齐全。(2)设备工程智能体：合规性关注设备选型和操作参数是否满足设计规范与允许区间；鲁棒性关注宽负荷运行、设备裕度及喘振/超限等边界；准确性关注设备参数、性能曲线和工艺—设备匹配关系；完整性关注设备结构约束、操作边界及关键限制条件是否覆盖。(3)安全合规智能体：合规性关注法规、标准和安全底线；鲁棒性关注扰动或异常工况下保护措施、冗余和失效防护；准确性关注危险源识别、风险判断及保护要求是否正确；完整性关注泄漏检测、超压保护、联锁与紧急切断等关键安全要素是否齐全。(4)数据分析智能体：合规性关注符号、单位和数值计算规范；鲁棒性关注求解收敛性、数值稳定性和敏感性；准确性关注计算逻辑、单位换算与参数处理；完整性关注变量定义、求解条件、收敛判据及必要的数据处理步骤。上述专业化映射同时作为附录A四类专业评估提示词的评分口径。

在实际评分过程中，各专业智能体首先依据知识库中的相关资料对当前建模方案依次进行比对，然后参照表4的锚定标准给出分数。若方案表现介于两个分数段之间，智能体应优先选择较低的分数，以体现质量评估的保守性原则。每个维度的评分均需附带简要的文字说明，明确指出扣分原因或加分依据。这些评分不仅用于最终的质量判定，更是辩论收敛机制中分歧识别（极差计算）和修正指导生成的核心输入。通过将抽象的“质量”概念分解为四个可量化、可操作的专业维度，本框架为多智能体协作提供了统一的判据，确保了评估过程的客观性和可复现性。

为进一步明确综合得分的形成方式，本文对四个评价维度采用等权线性加权。设第 e 位评价者对

某一评价对象在第 j 个评价维度上的评分为 s_{ej}，维度权重为 w_j，则该评价者给出的综合得分 C_e 计算为：

$$C_e = \sum_{j=1}^{4} w_j s_{ej}, \qquad w_1 = w_2 = w_3 = w_4 = 0.25, \quad \sum_{j=1}^{4} w_j = 1 \tag{4}$$

式中，e=1,2,3 为三位评价专家的序号，j=1,2,3,4 为四个评价维度的序号；s_{ej}、w_j 和 C_e 分别为第 e 位专家在第 j 个维度的评分、该维度权重及该专家的综合得分。本文将合规性、鲁棒性、准确性和完整性视为并列的质量属性，现阶段不存在充分的理论或工程证据支持对其中某一维度赋予更高优先级；采用等权设置可减少人为赋权带来的主观偏差，并保证不同模型、案例及协作模式之间的比较具有一致尺度。对于实验结果中的人工双盲评价，3位专家之间同样采用等权处理，因此某一模型在某一案例下的最终综合得分 C 为：

$$C = \frac{1}{3}\sum_{e=1}^{3} C_e = \frac{1}{12}\sum_{e=1}^{3}\sum_{j=1}^{4} s_{ej} \tag{5}$$

其中，表中报告的单项维度得分为3位专家在该维度上的算术平均值；案例综合得分按照上述公式计算。进一步计算模型在5个案例上的“平均得分”时，对5个案例的综合得分等权取算术平均；计算“四种模型平均”时，则对4款模型的对应得分等权取平均。

2 绿氨生产案例体系设计

为系统验证辩论驱动的多智能体协作增强框架在不同复杂度建模任务中的有效性，本章设计了五个梯度化测试案例，覆盖绿氨生产的核心工艺流程与典型建模任务。案例设计遵循三项原则：工艺类型全覆盖（电解水制氢、哈伯-博世合成、多单元集成）、任务类型多元化（稳态建模、动态响应、系统综合、故障诊断）以及难度层级梯度化（从单单元稳态建模逐步递进至全流程耦合与溯因推理）。各案例的复杂度量化结果见表5，核心建模挑战及与工艺流程的对应关系在下文进一步说明。

为减少复杂度分级的主观性，本文进一步从变量/单元规模（V）、耦合程度（C）、工况与时间尺度复杂性（T）以及任务推理难度（R）四个维度进行半定量评分。由于本文测试对象是“建模任务”而非预先固定的数学模型，不同LLM可能生成不同数量的状态变量，因此V不直接采用某一具体模型的变量个数，而以任务涉及的主要工艺单元数量、状态变量类型和求解对象规模作为可复现的代理指标。四项指标均按1–4分计：V=1表示单单元、少量核心变量，V=4表示全流程多单元或多源高维诊断变量；C=1表示弱耦合或单一机理，C=4表示跨单元物料/能量或多专业强耦合；T=1表示稳态单工况，T=4表示宽负荷、多模式或多时间尺度动态；R=1表示正向模型/方程构建，R=4表示信息不完整条件下的反演与故障溯因。

案例综合复杂度得分定义为 K=V+C+T+R。本文将K=4–6定义为低复杂度，K=7–9定义为中等复杂度，K=10–12定义为中高复杂度，K=13–16定义为高复杂度。该分级仅用于描述测试任务本身的结构与推理负担，不参与模型质量评分或综合得分计算。按照上述规则，五个案例的量化结果见表5。

表5 测试案例复杂度量化判定

Table 5 Quantitative complexity assessment of test cases

案例	工艺单元/任务	V（规模）	C（耦合）	T（工况）	R（推理）	K/复杂度
1	PEM电解槽稳态建模	1	1	1	1	4/低
2	哈伯-博世合成氨反应器建模	2	3	1	2	8/中
3	变负荷动态响应	3	3	3	3	12/中高
4	全流程系统综合	4	4	4	3	15/高
5	合成工段故障诊断	4	4	4	4	16/高

案例1聚焦电解水制氢单元的稳态建模，要求LLM推荐适用的极化曲线模型并给出电流密度-电压关系方程，属于低复杂度的单物理场建模任务。案例2转向哈伯-博世合成氨反应器，需处理强放热反应的热平衡与反应速率的耦合关系，建立多床层间冷式反应器的轴向温度分布模型，复杂度提升至中等。案例3引入风光供电的波动性，要求模型在40%–100%的宽负荷范围内描述合成氨回路的动态响应，并讨论H_2/N_2比、惰气含量等参数的柔性调整策略，复杂度升至中高。案例4整合电解、空分、压缩、合成、分离等单元，构建全流程的物料衡算与能量衡算框架，重点考察模型在多单元耦合与不同

时间尺度动态(电解槽秒级、合成塔小时级)下的综合描述能力,复杂度为高。案例5设定合成塔出口氨浓度下降、床温周期性波动等异常工况,要求LLM推荐跨模式故障诊断框架、指出关键缺失参数并设计多负荷下的诊断策略,属于信息不完整条件下的溯因推理任务。

上述五个案例由简入繁、从单单元到全流程、从稳态到动态、从正向建模到故障反演,形成了完整的复杂度梯度,能够系统考察LLM在不同类型建模任务中的输出质量,并为多智能体辩论与收敛机制的有效性验证提供了多样化的测试场景。各案例的详细工程参数(设备规格、操作条件、物料组成及输出要求)见附录B。按照上述量化规则,案例1–5的复杂度得分分别为4、8、12、15和16分,对应低、中、中高、高和高四个等级,从而为后续任务复杂度与协作增益的比较提供一致的判定依据。

3 实验结果与分析

本章所有实验均基于统一的LLM调用参数配置,各模型的具体参数见表6。在此基础上,基于第1.1节所述总体工作流程,并结合第2章设计的五个梯度化测试案例,系统开展对照实验并呈现结果。实验采用对照设计:以单智能体迭代输出作为对照组,以多智能体协作输出作为实验组。两种模式运行相同的迭代轮次,唯一区别在于修正意见的来源——单智能体模式下由同一模型对其自身输出进行自我评估并修正,多智能体模式下由四个专业智能体交叉辩论后形成共识性修正指导。由3名化工领域专家对两组输出进行双盲评估,从合规性、鲁棒性、准确性、完整性四个维度量化评分。通过对比两组得分差异,并结合分歧度、问题检出率等指标,系统验证多智能体协作机制相较于单智能体自我迭代的优越性。

其中,各模型温度参数统一设为0.1,所有调用均采用同步请求模式,单次请求超时时间120秒,失败时最多重试3次。批量大小设为1,即每次仅处理单个案例的单个迭代轮次,以避免多案例并行调用带来的上下文相互干扰及API速率限制冲突。

表6 各模型API版本与调用参数

Table 6 API versions and invocation parameters of each model

模型	API版本/标识符	上下文窗口	最大生成长度	批量大小
GPT-5.4	gpt-5.4	128k tokens	4096 tokens	1
DeepSeek V3.2	deepseek-chat	128k tokens	4096 tokens	1
Qwen 3.5 Max	qwen-max	128k tokens	4096 tokens	1
Gemini 3.1 Pro	gemini-3.1-pro-preview	128k tokens	4096 tokens	1

3.1 综合得分对比分析

为评估专家评分的客观性与可重复性,在呈现实验数据之前,首先对3位专家的评分一致性进行量化检验。本文以全部专家评分为样本(3位专家×5个案例×4个评估维度×2种输出模式=120个评分),计算各维度的Fleiss′ Kappa系数。结果表明:合规性Kappa=0.82,鲁棒性Kappa=0.79,准确性Kappa=0.85,完整性Kappa=0.76,融合四个评价维度得到总体Kappa=0.81。依据Landis和Koch提出的Kappa系数分级标准,0.61–0.80对应高度一致(substantial agreement),0.81–1.00对应近乎完美一致(almost perfect agreement)。本研究各维度Kappa均达到高度一致及以上区间,其中合规性、准确性与综合总体Kappa落入近乎完美一致范围,说明三位专家的标注评分一致性良好,人工标注信度可靠。这一结果与LLM人工评估领域的已有研究一致:同类专家人工评测实验中,Kappa=0.78即被视为具有较高的标注一致性。本文总体Kappa=0.81,高于该参照值。基于可靠的专家标注结果,表7进一步对比四款模型在五个测试案例下的综合得分表现。

表7中的各案例综合得分均按照第1.6节所述等权公式计算,即先对四个评价维度按0.25等权聚合,再对3位专家的综合评分等权平均;表中的模型平均值和四种模型平均值分别进一步按案例和模型取算术平均。

为检验多智能体协作对综合得分的总体提升是否具有统计显著性,本文以表7中的"模型—案例"组合为配对统计单元:4款模型×5个案例共形成20组配对观测(n=20),每一对分别对应同一模型在同一案例下的单智能体综合得分与多智能体综合得分。采用双侧配对t检验,原假设为两种模式的配对差值均值为0,显著性水平设为α=0.05。需要说明的是,该检验以经4个评价维度等权聚合并由3位专家等权平均后的模型—案例综合得分作为统计样本,而非以单个专家的原始维度评分作为独立样本;配对关系严格限定在相同模型和相同案例之间,不跨模型或案例进行配对。

表7 各模型单/多智能体各案例综合得分

Table 7 Comprehensive scores of each case for single-agent and multi-agent models

模型	输出模式	案例1	案例2	案例3	案例4	案例5	平均得分	绝对提升/分
GPT-5.4	单智能体	9.0	8.2	7.8	7.6	7.3	8.0	—
GPT-5.4	多智能体	9.5	8.9	8.5	8.3	8.1	8.7	0.7
DeepSeek V3.2	单智能体	8.7	7.9	7.5	7.3	7.0	7.7	—
DeepSeek V3.2	多智能体	9.2	8.6	8.2	8.0	7.8	8.4	0.7
Qwen 3.5 Max	单智能体	8.5	7.6	7.2	7.0	6.7	7.4	—
Qwen 3.5 Max	多智能体	9.1	8.4	8.0	7.8	7.6	8.2	0.8
Gemini 3.1 Pro	单智能体	9.2	8.4	8.0	7.8	7.5	8.2	—
Gemini 3.1 Pro	多智能体	9.6	9.0	8.7	8.5	8.3	8.8	0.6
四种模型平均	单智能体	8.9	8.0	7.6	7.4	7.1	7.8	—
四种模型平均	多智能体	9.4	8.7	8.4	8.2	8.0	8.5	0.7

从表7可以看出，四款模型在多智能体协作模式下均取得了更高的综合得分，四种模型平均综合得分由7.8分提升至8.5分，绝对提升0.7分。相对提升率以单智能体平均得分7.8分为基准，按未舍入的原始得分计算，为9.0%。在上述20组一一配对的模型—案例综合得分上，配对差值的平均值为0.7分，双侧配对t检验结果为$t(19)=25.18$，$p<0.001$，表明多智能体协作带来的综合得分提升具有统计学显著性。从模型维度看，Qwen 3.5 Max提升最为显著（+0.8分），Gemini 3.1 Pro提升相对较小（+0.6分），这与各模型基线表现相关——基线越高的模型，提升空间相对有限。从案例维度看，多智能体协作在高复杂度案例中的提升效果更为突出：案例4（全流程综合）四款模型平均提升0.7分，案例5（故障诊断）平均提升0.8分，而低复杂度案例1平均提升0.5分。结合第2章的四维复杂度量化结果，案例复杂度由4分增加至15–16分时，协作增益由案例1的0.5分提高至案例4的0.7分和案例5的0.8分，整体呈现高复杂度任务受益更明显的趋势。

图2以折线图形式展示了四款模型综合得分随案例复杂度变化的趋势。单智能体输出与多智能体输出均呈现随复杂度增加而下降的趋势，但多智能体输出的下降幅度更为平缓，同一模型的两条折线在低复杂度案例中较为接近，随着复杂度增加逐渐分离，直观印证了多智能体协作机制在高复杂度任务中的优势。

图3以柱状图形式呈现了四款模型在单智能体与多智能体模式下的平均综合得分对比。四组比较中，多智能体模式的平均综合得分均高于单智能体模式，且提升幅度相近，说明多智能体协作机制具有良好的跨模型通用性。

从各维度来看，多智能体协作在鲁棒性维度提升最为显著，平均提升0.8分，反映多智能体通过交叉验证与补充论证，有效增强了模型输出对参数波动和工况变化的稳健性。准确性维度提升0.8分，完整性维度提升0.7分，表明模型在机理方程准确性和输出要素完整覆盖方面的改进效果同样显著。合规性维度提升相对较小（0.5分），说明在规范遵守方面仍有提升空间，部分模型在高复杂度任务中对压力容器安全、防爆分区、泄漏检测等安全合规要求的覆盖仍有不足。

图4~7分别以热力图形式展示了四款模型在单智能体与多智能体模式下各案例各维度的得分分布。从图中可以清晰看到，多智能体输出的热力图在合规性、鲁棒性、准确性、完整性及综合得分等所有维度上，单元格颜色普遍更深（分数更高），且随着案例复杂度增加，颜色优势更为明显，直观反映了多智能体协作对输出质量的全面提升。值得关注的是，四款模型在单智能体模式下的基线表现存在明显差异（如Qwen 3.5 Max在案例5的合规性得分仅6.5，而Gemini 3.1 Pro达到7.4），但在多智能体协作模式下，各模型的热力图颜色分布趋于一致，差异缩小。这一现象表明，多智能体协作机制在提升整体输出的同时，具有一定的“能力均衡”效应——能够在保持模型原有优势的同时，弥补其短板，实现“扬长补短”的效果。

3.2 分歧度收敛分析

分歧度反映多智能体协作过程中四类专业智能体（工艺建模、设备工程、安全合规、数据分析）之间意见的一致性程度，是衡量协作机制有效性的关键指标。对于某一款基础模型在某一案例上的一次具体运行，本文仍以同一评价维度下四类专业智

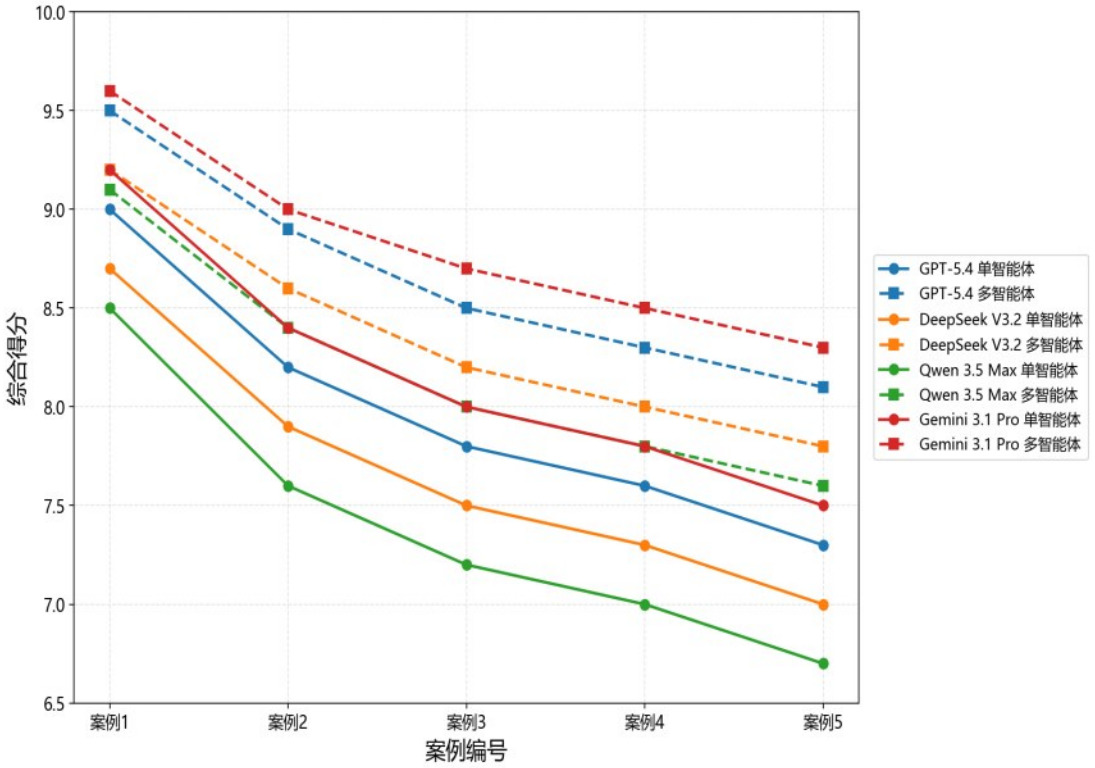

图2　四款模型综合得分随复杂度变化趋势折线图

Fig.2　Line chart of comprehensive scores of four models varying with task complexity

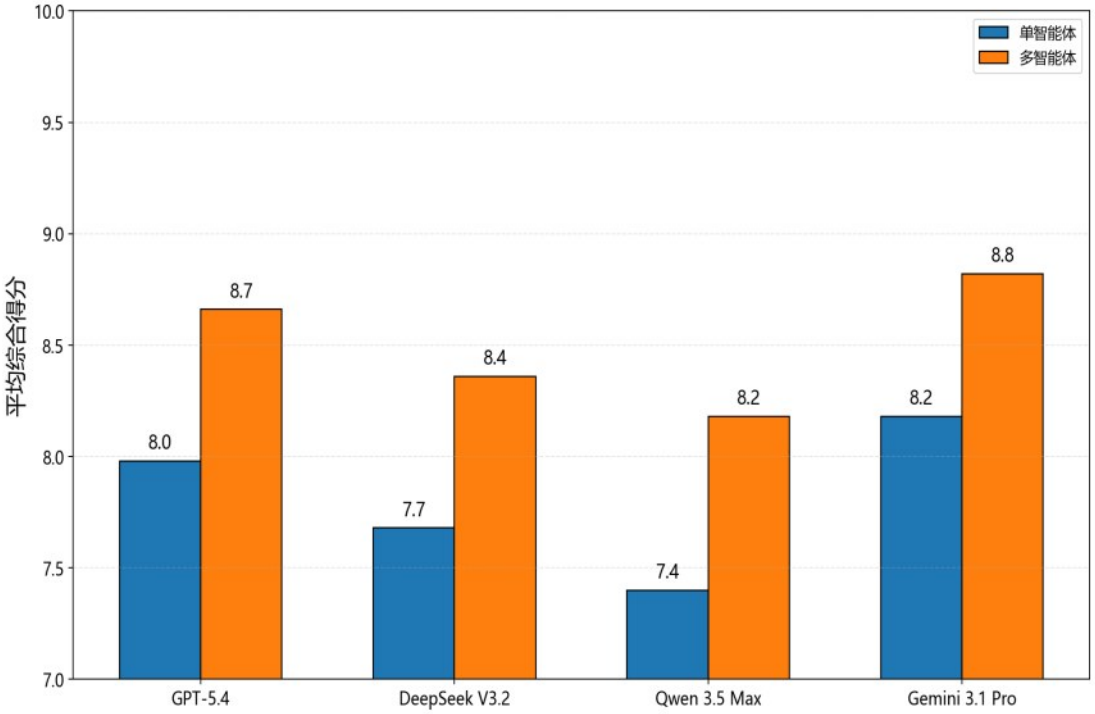

图3　四款模型单智能体与多智能体平均综合得分对比柱状图

Fig.3　Bar chart of average comprehensive scores of four models: single-agent vs. multi-agent

能体评分的极差D_j作为分歧度，并以“所有评价维度$D_j \leq 1$”作为完全收敛标准。本节进一步在正式实验的4款基础模型上对该指标进行汇总，用于比较讨论前后的总体分歧变化。

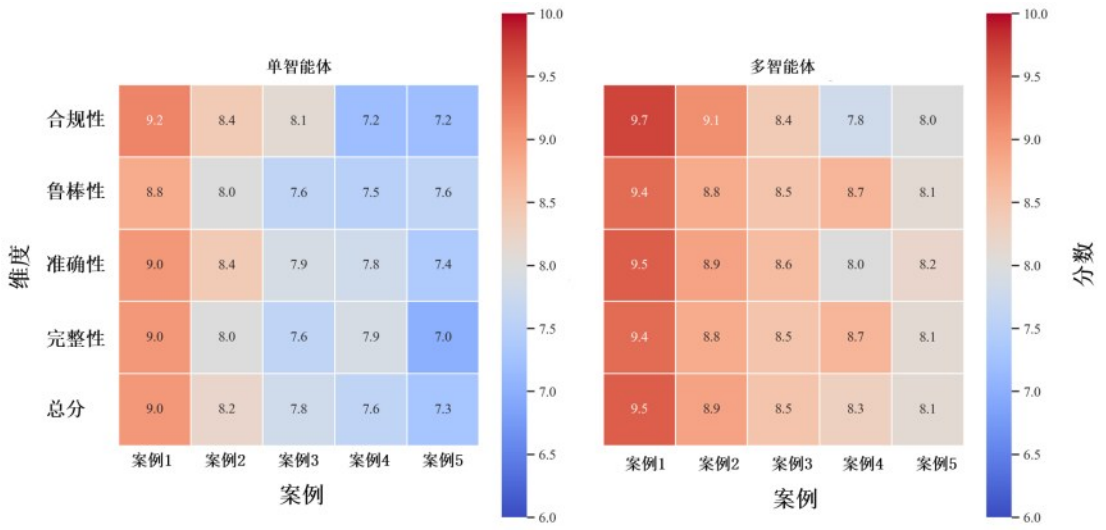

图4 GPT-5.4单智能体与多智能体各案例各维度得分热力图

Fig.4 Heatmaps of scores across cases and evaluation dimensions for GPT-5.4 in single-agent and multi-agent modes

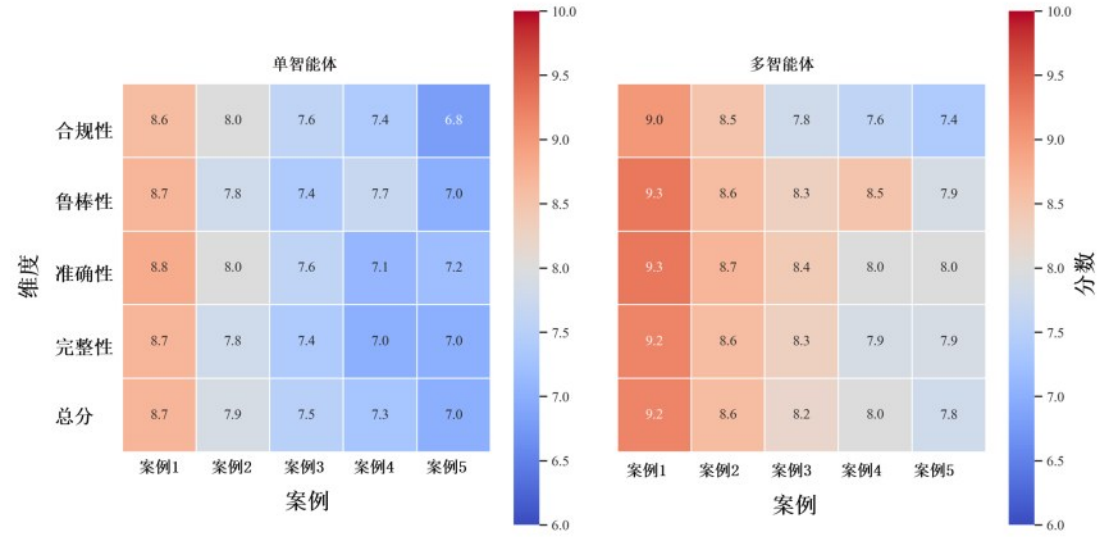

图5 DeepSeek V3.2单智能体与多智能体各案例各维度得分热力图

Fig.5 Heatmaps of scores across cases and evaluation dimensions for DeepSeek V3.2 in single-agent and multi-agent modes

具体而言，设D_{mkj}为第m款基础模型在第k个案例、第j个评价维度上的评分极差，其中m=1,2,3,4，k=1,2,3,4,5，j=1,2,3,4。首先，对同一“模型—案例”下4个评价维度取算术平均，得到该模型在该案例的维度平均极差$\bar{D}_{mk}=\frac{1}{4}\sum_{j=1}^{4}D_{mkj}$；随后对4款基础模型进一步取平均，得到表8所报告的案例层面四模型平均极差$\bar{D}_{k}=\frac{1}{4}\sum_{m=1}^{4}\bar{D}_{mk}=\frac{1}{16}\sum_{m=1}^{4}\sum_{j=1}^{4}D_{mkj}$。由此，表8中的每个数值均综合了4款基础模型和4个评价维度，是正式实验的描述性汇总指标，而非某一具体模型—案例运行的最终最大极差。单次运行是否完全收敛仍严格依据$\max_{j}\left(D_{mkj}\right)\leqslant 1$判断，因此$\bar{D}_{k}\leqslant 1$并不能推出所有模型、所有评价维度均已满足$\max_{j}\left(D_{mkj}\right)\leqslant 1$。

表8展示了4款模型在五个案例中讨论前后平均极差的汇总结果，图8给出相应可视化。讨论前，各案例的四模型平均极差分别为2.1、2.3、2.3、2.4和2.8分，五个案例的均值为2.4分；讨论后分别降至0.3、0.6、0.5、0.3和0.9分，五个案例的均值降至0.5分，整体下降78.2%（各下降幅度均按未舍入数据计算）。所有案例的四模型平均极差下降幅度均超过65%，说明多智能体辩论能够在不同基础模型上显

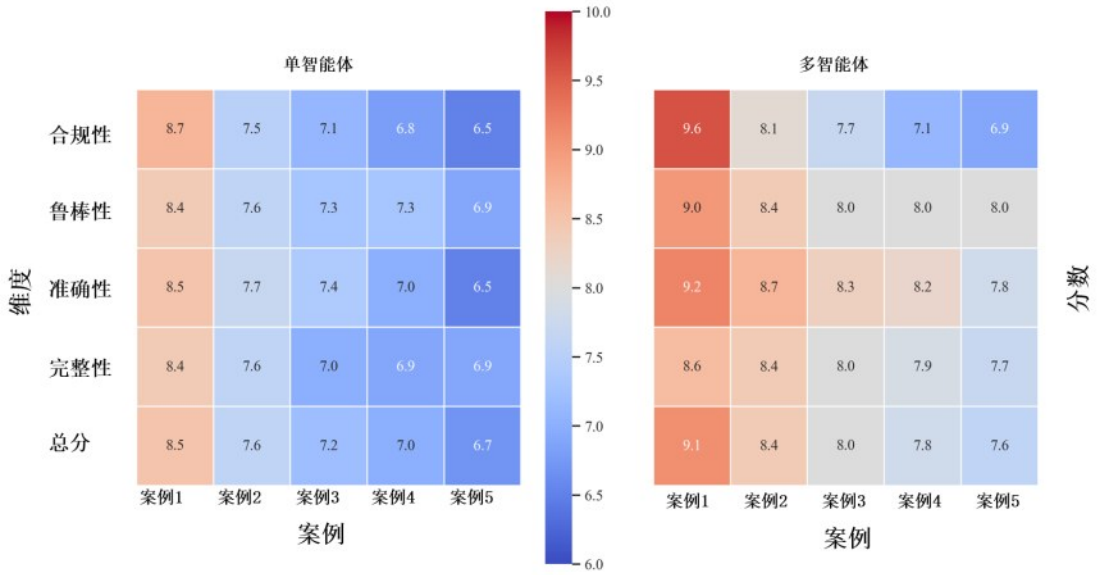

图6 Qwen 3.5 Max单智能体与多智能体各案例各维度得分热力图

Fig.6 Heatmaps of scores across cases and evaluation dimensions for Qwen 3.5 Max in single-agent and multi-agent modes

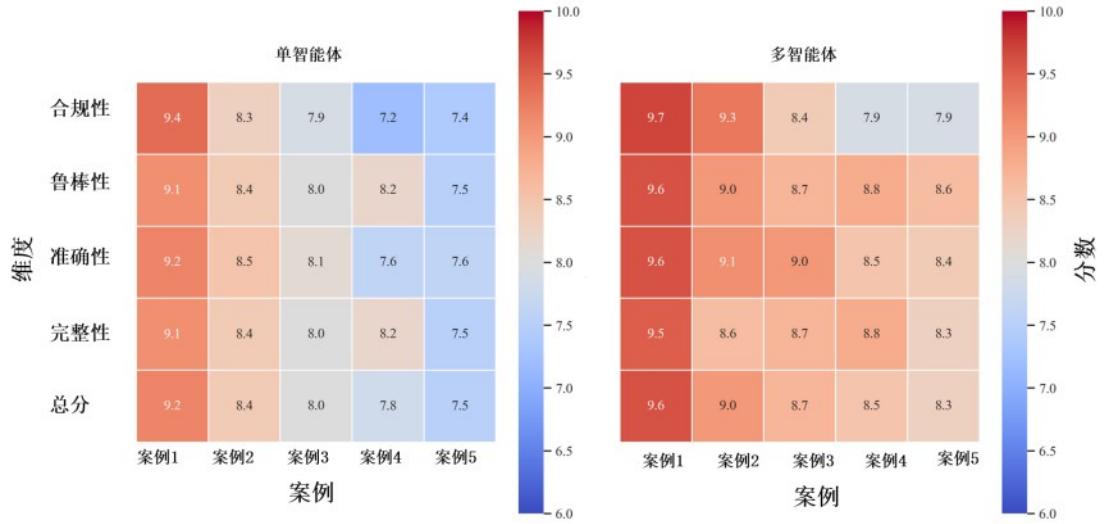

图7 Gemini 3.1 Pro单智能体与多智能体各案例各维度得分热力图

Fig.7 Heatmaps of scores across cases and evaluation dimensions for Gemini 3.1 Pro in single-agent and multi-agent modes

著压缩专业视角之间的总体分歧。需要强调的是，这些数值用于衡量4款模型正式实验结果的总体分歧水平及其变化，不用于直接判断某一次具体运行是否达到完全收敛。

值得关注的是，案例5（故障诊断）讨论后的四模型平均极差为0.9分，仍为五个案例中最高。该0.9分是4款基础模型在案例5上的跨维度、跨模型汇总平均值，而不是某一款模型单次运行的收敛结果，因此不能据此认定4款模型在案例5中均已满足全部$D_j \leq 1$。案例5的故障诊断任务具有较强多解性，因此即使总体平均分歧已明显下降，个别模型或个别评价维度仍可能保留$D_j > 1$的残余分歧。正式实验统一采用3轮最大预算，达到上限后若仍未完全收敛，则按截断输出处理并保留分歧记录，而不宣称所有模型均已完全收敛。

3.3 问题检出能力分析

问题检出能力反映多智能体协作过程中各专业智能体识别模型输出中潜在问题的能力，是验证专业分工价值的关键指标。本研究将模型输出中的问题分为四类：机理问题（如电化学极化模型选

表8 四款模型在各案例中讨论前后平均极差对比(描述性统计)

Table 8 Comparison of average ranges across four models before and after discussion for each case (descriptive statistics)

案例编号	讨论前四模型平均极差	讨论后四模型平均极差	极差下降幅度
案例1	2.1	0.3	85.7%
案例2	2.3	0.6	73.9%
案例3	2.3	0.5	78.3%
案例4	2.4	0.3	87.5%
案例5	2.8	0.9	67.9%
平均值	2.4	0.5	78.2%

择错误、反应动力学方程缺失、多单元耦合传质逻辑错误)、设备问题(如电解槽工作压力超出设计范围、压缩机喘振约束未考虑、换热器面积与热负荷不匹配)、安全问题(如氢气/氨气防爆分区缺失、压力容器安全阀设置遗漏、有毒气体泄漏检测要求未体现)和数据问题(如电流密度单位错误、能耗计算逻辑不正确、参数取值范围偏离工程实际)。其中,幻觉指事实性错误类问题,包括机理幻觉、方程幻觉、参数幻觉、安全幻觉和符号幻觉。为避免与问题修正效果混淆,本文将问题检出率定义为 $R_{\mathrm{det}}=\frac{N_{\mathrm{det}}}{N_{\mathrm{total}}}\times 100\%$,其中 R_{det} 为问题检出率,N_{det} 为多智能体检出的问题数,N_{total} 为专家双盲评审确认的总问题数。

为深入考察多智能体协作机制对不同LLM的适应性,本研究按模型统计了多智能体协作模式下的问题检出情况。各模型在多智能体协作过程中均采用相同的专业智能体配置与知识库,通过对比不同模型的问题检出率,以验证协作机制的跨模型普适性。表9展示了四款模型在多智能体协作模式下的问题检出统计。

表9 各模型总问题检出统计

Table 9 Statistics of total detected problems for each model

模型	总问题数	多智能体检出数	检出率
GPT-5.4	37	30	81.1%
DeepSeek V3.2	34	26	76.5%
Qwen 3.5 Max	35	29	82.9%
Gemini 3.1 Pro	31	27	87.1%
合计	137	112	81.8%

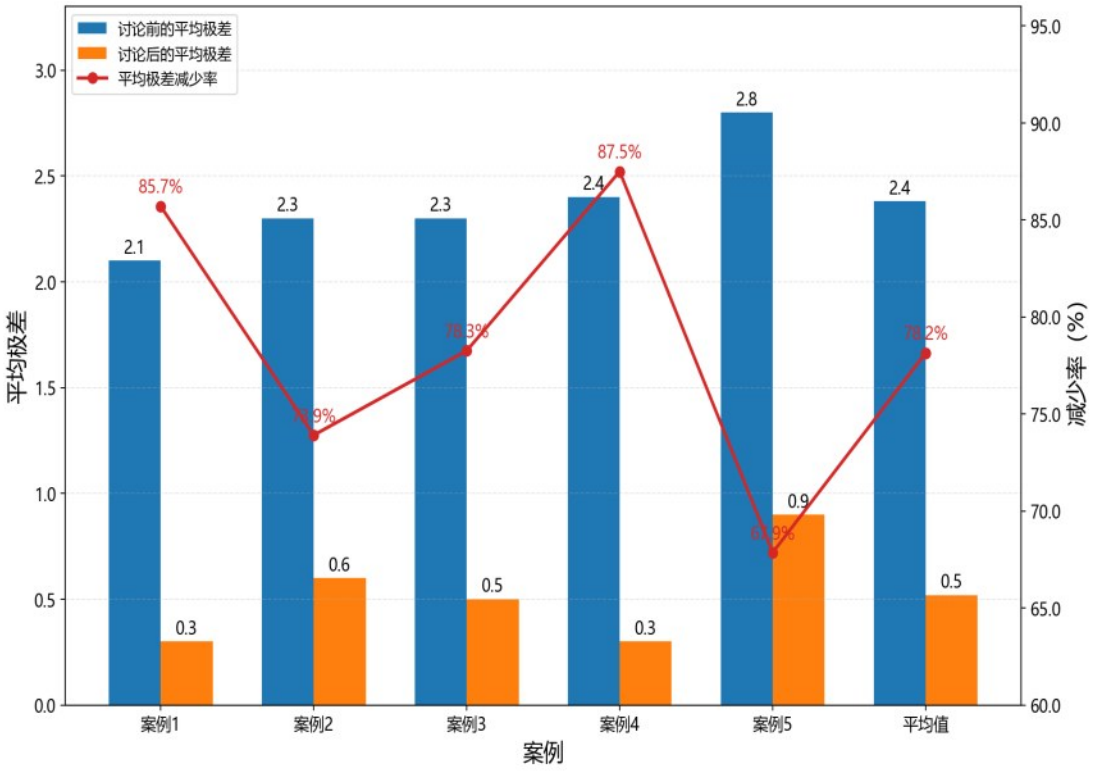

图8 四款模型在各案例中讨论前后平均极差及下降幅度

Fig.8 Average ranges across four models and reduction rates before and after discussion

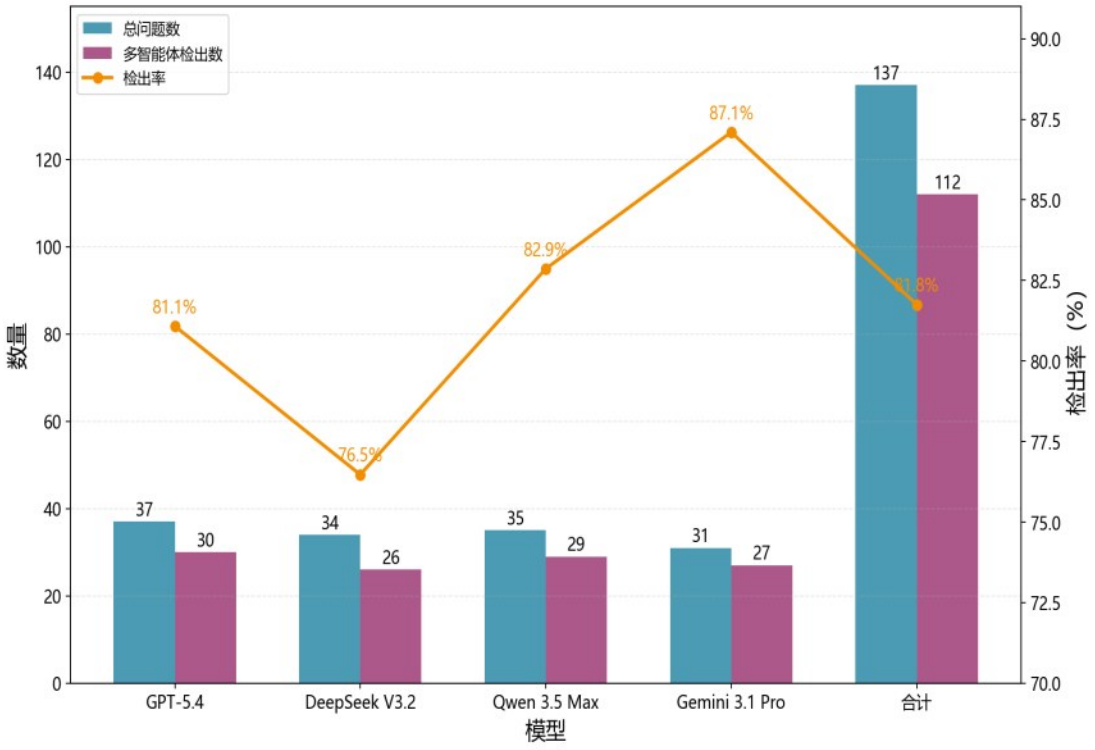

图9 四款模型的问题检出率

Fig.9 Problem detection rates of four large language models

图9展示了不同模型的总问题检出率，四款模型在多智能体协作模式下的问题检出率均处于较高水平，整体检出率达81.8%。其中Gemini 3.1 Pro的检出率最高（87.1%），Qwen 3.5 Max次之（82.9%），GPT-5.4为81.1%，DeepSeek V3.2为76.5%。各模型的原始问题数量存在差异，但检出率均超过75%，表明多智能体协作机制对不同基础模型的输出均具有一定的问题识别能力。

根据上述对问题检出率的分析，Gemini 3.1 Pro表现最佳，因此本文以其为例进一步分析各案例中问题的分布情况。表10统计了Gemini 3.1 Pro模型在五个案例中存在的总问题数，以及多智能体协作过程中各专业智能体检出的问题数。总问题数以专家盲评识别的并集为准。从整体来看，五个案例共发现31个问题，多智能体协作共检出27个，整体检出率达87.1%。未检出的4个问题主要集中在案例5（故障诊断）和案例4（全流程综合）的高复杂度任务中，且多为需要深度领域知识才能识别的隐性机理问题（如多工况耦合传递特性、催化剂动态失活对反应动力学的影响等）。

各专业智能体在各自擅长领域的问题检出能力如表11所示。设备工程智能体在设备问题检出方面表现最佳（100.0%），数据分析智能体在量化问题检出方面达到83.3%，安全合规智能体在安全问题检出方面为80.0%，工艺建模智能体在机理问题检出方面为86.7%。

表10 各案例问题检出统计

Table 10 Statistics of detected problems for each case

案例编号	总问题数	多智能体检出数	检出率
案例1	5	5	100.0%
案例2	5	5	100.0%
案例3	5	4	80.0%
案例4	7	6	85.7%
案例5	9	7	77.8%
合计	31	27	87.1%

表11 各专业智能体问题检出能力

Table 11 Problem detection capabilities of each professional agent

智能体类型	擅长问题类型	该类型总问题数	检出数	检出率	贡献占比
工艺建模智能体	机理问题	15	13	86.7%	48.2%
设备工程智能体	设备问题	5	5	100.0%	18.5%
安全合规智能体	安全问题	5	4	80.0%	14.8%
数据分析智能体	量化问题	6	5	83.3%	18.5%
总计	—	31	27	87.1%	100.0%

从表11可以看出，工艺建模智能体对应的机理问题数量最多（15个，占总问题数的48.4%），且贡献占比最高（48.2%），其检出率也达到86.7%，体现了较好的识别能力。机理问题涉及电化学极化机理、反应动力学、多相传质及热质耦合等核心物理化学过程，本身具有更强的隐蔽性和复杂性。部分深层次问题（如变负荷条件下催化剂热循环应力的动态建模、电解槽膜渗透损失与压力耦合的非线性关系等）需要结合大量工程经验与专业知识才能准确识别。设备工程智能体在设备问题上实现了100.0%检出，安全合规智能体和数据分析智能体也分别达到80.0%和83.3%的较高水平。整体而言，多智能体协作机制在各专业领域整体表现较好，验证了专业分工与交叉校验的有效性。

3.4 不同复杂度案例的处理耗时与质量—效率权衡分析

为评估框架的计算效率及其质量收益，本节结合表7中的综合得分提升与各案例处理耗时，对五个不同复杂度案例进行质量—效率权衡分析。耗时统计范围为从输入案例描述到输出最终建模方案的全过程，包含API调用时间、推理时间及轮次间调度开销。两种模式的迭代轮次设置保持一致——案例1与案例2为2轮，案例3、案例4与案例5最多执行3轮。需说明的是，案例5在预实验单一模型、1.0分阈值条件下需第4轮方能满足所有维度$D_j \leqslant 1$的完全收敛条件；但从第3轮增加至第4轮仅带来0.1分综合得分增益，而耗时由35 min增至61 min。基于该质量—效率权衡，正式实验统一采用3轮最大迭代预算：若提前满足全部$D_j \leqslant 1$则提前终止，若达到3轮仍存在局部分歧则截断输出并记录。表8中案例5讨论后的0.9分是对4款模型、4个评价维度汇总得到的平均极差，仅用于描述总体分歧水平，不表示该案例的所有正式运行均在3轮内完全收敛。单智能体模式下每轮由同一模型依据自我评估结果修正方案，多智能体模式下则由四类专业智能体交叉辩论后形成修正指导。表12进一步将各案例的综合得分提升与处理时间增量进行对应比较。

从时间成本看，多智能体模式的总耗时为18–40 min，约为单智能体模式的2.22–3.13倍，额外耗时为12–24 min。但该时间增量同时对应0.5–0.8分的综合得分提升。为进一步刻画额外时间投入转化为质量收益的效率，本文定义描述性指标$\eta_k = \frac{10\Delta S_k}{\Delta t_k}$，其中k=1,2,3,4,5对应五个测试案例，$\Delta S_k$为第k个案例多智能体相对于单智能体的综合得分提升，Δt_k为对应的额外处理时间（min），η_k表示每增加10 min处理时间获得的综合得分提升。按表7与表12计算，案例1–5的η_k分别为0.42、0.41、0.36、0.30和0.37分/10 min。可见，随着任务复杂度提高，绝对质量收益总体由0.5分增加至0.8分，但单位额外时间对应的边际质量收益并未同步增加，说明复杂任务需要更多专业评审与辩论开销来获得更高的绝对质量增益。

表12 各案例综合得分提升与平均处理耗时对比

Table 12 Comparison of comprehensive score improvement and average processing time for each case

案例	复杂度	得分提升/分	单智能体耗时(min)	多智能体耗时(min)	增量(min)
案例1	低	0.5	6	18	+12
案例2	中	0.7	8	25	+17
案例3	中高	0.7	12	32	+20
案例4	高	0.7	14	38	+24
案例5	高	0.8	18	40	+22

这一结果表明，本框架的实际应用价值不宜仅以“得分越高越好”或“耗时越短越好”单独判断，而应依据任务复杂度和风险等级进行自适应选择。对于案例1等低复杂度任务，多智能体协作提升0.5分但需额外12 min，可优先采用单智能体或较少辩论轮次以提高效率；对于案例4–5等高复杂度任务，尤其是多单元耦合、异常诊断及安全约束较强的场景，0.7–0.8分的质量提升能够更充分体现多专业交叉校验的价值。其中案例5以额外22 min获得0.8分提升，为五个案例中绝对增益最高。由此可见，该框架更适合作为离线建模审核、工艺方案论证和复杂故障分析中的质量增强工具，而对于时间约束严格的在线控制任务，则应通过复杂度分级、减少迭代轮次或仅在高风险任务中触发多智能体辩论，以实现质量与效率之间的动态平衡。

3.5 与层级分工范式的横向对比

为凸显本框架辩论共识机制的独特优势，本节选取引言中提及的CeProAgents作为代表性基线进行横向对比。CeProAgents是中国科学院大连化物所与中科大联合提出的层级多智能体系统，包含知识层、概念层和参数层三个专业智能体群组，通过纵向分层协作实现化工过程开发的端到端自动化，代表了当前多智能体LLM在化工领域的主流协作

范式——层级分工执行。两者构成清晰的范式对照：CeProAgents强调“纵向分层、逐级推进”，本框架强调“同一方案、多视角交叉辩论与共识收敛”。

由于CeProAgents的完整代码尚未公开，本文采用方法论复现策略：根据其文献中描述的智能体架构与协作逻辑，在相同的Coze平台上构建对应的多智能体系统，在相同的5个案例上运行，并由同一组3位专家按四个维度盲评。复现过程中保留原文的智能体分工方式。两种范式均基于Gemini 3.1 Pro模型，并采用相同的绿氨工艺知识库配置，以保证对比条件一致。两种范式在5个案例上的综合得分对比如表13所示。

表13 层级分工与辩论共识范式的综合得分对比

Table 13 Comparison of comprehensive scores between hierarchical division and debate consensus paradigm

协作范式	代表性框架	案例1	案例2	案例3	案例4	案例5	平均
层级分工执行	CeProAgents	9.0	8.3	7.9	7.5	7.3	8.0
辩论共识收敛	本框架	9.6	9.0	8.7	8.5	8.3	8.8

本框架的平均综合得分为8.8分，高于CeProAgents的8.0分，提升幅度约10.3%。从案例维度看，本框架的优势随任务复杂度增加而扩大：在低复杂度案例1中，两者差距为0.6分；而在高复杂度案例5中，差距扩大至1.0分（13.7%）。这一趋势表明，辩论共识范式在复杂推理任务中的优势更为突出——任务越复杂，单一视角的局限性越明显，多专业交叉验证的价值越大。

为进一步揭示两种范式性能差异的内在原因，本文以案例4（全流程系统综合建模）为例，从专家评分的维度均衡性角度进行对比分析。案例4涵盖电解、空分、压缩、合成、分离五个单元的全流程耦合，是正向建模任务中复杂度最高的场景，最能体现各维度质量均衡发展对综合性能的影响，对比结果如表14所示。本节所采用的“维度标准差”按照第1.2节上述定义，对同一评价对象四个评价维度的聚合得分计算总体标准差，其值越小表示维度表现越均衡。

在案例4中，本框架的维度标准差（0.37）小于CeProAgents（0.45），降幅约17.8%，表明本框架在各维度上的表现更为均衡。CeProAgents在鲁棒性和完整性维度上得分偏低（7.0和7.1），与合规性和准确性（8.0和7.9）形成明显落差，反映出层级分工范式下的维度失衡现象——模型在机理层面（准确性）看似合理，但在设备参数可行性（鲁棒性）和工程约束覆盖（完整性）方面存在短板，导致综合得分受限（7.5）。相比之下，本框架各维度得分分布更为均衡（7.9 – 8.8），尤其鲁棒性与完整性较CeProAgents分别提升1.8分和1.7分，综合得分达8.5。值得注意的是，本框架的合规性得分（7.9）略低于CeProAgents（8.0），这表明辩论机制并非在所有维度上都带来得分提升；各专业智能体通过相互质询寻求专业共识。总体而言，辩论共识机制通过多专业视角的交叉校验，使建模方案在各维度上均获得充分审视与修正，从而实现了更均衡、更可靠的整体质量。

表14 案例4维度得分与均衡性对比

Table 14 Comparison of dimension scores and balance for case 4

协作范式	合规性	鲁棒性	准确性	完整性	综合得分	维度标准差
层级分工执行（CeProAgents）	8.0	7.0	7.9	7.1	7.5	0.45
辩论共识收敛（本框架）	7.9	8.8	8.5	8.8	8.5	0.37

上述对比表明，辩论共识收敛范式在建模质量与维度均衡性上均优于层级分工执行范式，尤其在高复杂度任务中优势更为显著。这验证了本框架的核心设计理念——通过辩论驱动的交叉校验而非单纯的任务分工，是提升LLM化工建模输出质量与工程可靠性的更优路径。

4 结 论

本研究通过对五个绿氨生产工艺梯度案例的实验发现，LLM在该类专业门槛高、耦合性强的化工过程中展现出“高效率、低可靠”的显著特征。虽然模型能够快速构建结构化的建模框架，但在单模型模式下受限于“事实性幻觉”，平均综合得分仅为7.8分，且常在反应动力学选型、压力容器安全约束及变负荷动态逻辑上出现难以工程实施的错误。针对这一困境，本研究构建的“4+1”多智能体协作体系通过专业分工、观点交叉检验与共识收敛机制，实现了建模方案质量的系统性提升。实验数据显示，协作模式使综合得分平均提升0.7分（相对提升9.0%），其中鲁棒性维度提升最为显著（+0.8分），且四款模型在五个案例上的汇总平均极差由2.4降至0.5，下降幅度达78.2%。这种模拟人类科研小组的内部质控流程，证明了多智能体协作是抑制LLM幻觉、提升绿氨工艺建模工程适用性的有效路径。

基于第2章的四维复杂度量化结果，实验进一步显示协作增益整体随任务复杂度增加而增强：在低复杂度的电解槽稳态建模案例中，协作提升相对有限（+0.5分），而高复杂度案例5的提升达到+0.8分。从问题识别能力看，四款模型的整体问题检出率为81.8%；以问题检出率最高的Gemini 3.1 Pro为例，其五个案例整体问题检出率为87.1%，其中设备问题、安全问题和机理问题的检出率分别为100.0%、80.0%和86.7%。这表明多智能体对设备类问题的识别较充分，但安全问题和机理问题仍存在漏检，其中未检出的多为需要深度领域知识才能识别的隐性机理问题。在溯因推理等具有多解性的任务中，专家智能体间仍存在一定分歧（案例5讨论后的四模型平均极差为0.9，为各案例最高）。进一步分析机理问题检出率未达到100.0%的原因，可从三个层面解析：（1）知识覆盖方面，当前各专业智能体挂载的知识库以显性知识（标准规范、经典方程、典型参数）为主，缺乏对前沿机理（如催化剂动态失活的介观尺度模型、多工况耦合传递特性的第一性原理描述）的系统覆盖，LLM亦无法“理解”尚未被充分文献化的隐性工程经验；（2）辩论机制方面，多智能体辩论能有效识别逻辑矛盾与参数越界，但对于需要深层领域直觉才能发现的机理错误（如多物理场耦合中的非线性失稳），仅靠文本层面的辩论难以触及——近期研究亦指出，多智能体辩论中的“共识”可能掩盖真正的推理缺陷，智能体可能因顺从其他智能体的观点而非真正推理达成一致；（3）方程验证方面，LLM对数学方程的处理本质上是符号模式匹配，而非对物理规律的真正理解，当方程涉及复杂偏微分方程组或多尺度耦合时，模型难以验证其数学闭合性与物理一致性。因此，该范式目前应定位为高效的“人机协同”工具，而非完全去人工化的决策系统。

在上述能力边界的基础上，该框架在以下两类更复杂场景中面临进一步挑战：（1）多模态数据融合建模方面，当前框架仅处理文本形态的工艺描述与建模方案，无法直接融合工艺时序数据（如集散控制系统（DCS）趋势图）或设备图像数据（如管道与仪表流程图（P&ID）、红外热像），多模态数据的引入将带来跨模态语义对齐与时序动态建模等挑战；（2）极端工况建模鲁棒性方面，当前案例设计以单一故障或单一负荷变化为主，未覆盖设备故障叠加负荷骤变的复合极端场景，此类场景下各专业智能体的知识库可能缺乏对复合故障模式的先验描述，辩论机制亦因缺乏依据而难以有效收敛。

未来研究可从以下方向深化：一是进一步挂载绿氨工艺的专用知识库与热力学数据库，增强智能体对新型催化剂、低压合成等前沿技术的理解；二是探索多轮迭代辩论机制对分歧的深度收敛效果，平衡时间成本与输出质量；三是将本框架推广至其他可再生能源化工过程（如绿甲醇、合成燃料等），验证其通用性；四是引入物理信息神经网络（PINN）作为框架的“机理验证层”，将质量守恒、能量守恒等物理定律作为损失函数的正则化项，自动验证LLM输出方程是否满足物理一致性，并利用其微分方程求解能力补全缺失的本构关系或边界条件；五是构建“LLM多智能体辩论 + PINN代理模型 + 数字孪生”的三层架构——辩论层生成初始建模方案，验证层由PINN对方案进行快速数值验证与参数标定，孪生层将验证通过的模型部署为数字孪生的机理核心，接收实时传感器数据进行在线校正与状态估计；六是探索自适应辩论机制，根据任务复杂度动态调整辩论轮次与收敛阈值，在质量与效率之间实现更精细的平衡。最终目标是构建更贴近真实工业场景的智能化决策支持体系，推动人工智能（AI）辅助工具在安全关键工程中的可靠应用。

参考文献

[1] Ergun S, Dik A, Boukhanouf R, et al. Large-scale renewable energy integration: tackling technical obstacles and exploring energy storage innovations[J]. Sustainability, 2025, **17**(3): 1311.

[2] Oliver M E, Chapman O, Grijalva S, et al. Managing zero-marginal-cost, intermittent renewable energy: a survey of the engineering, economic, and policy challenges[J]. Renewable and Sustainable Energy Reviews, 2026, **226**: 116334.

[3] Sun M W, Ling Z Q, Mao J N, et al. Ammonia-based clean energy systems: a review of recent progress and key challenges[J]. Energies, 2025, **18**(11): 2845.

[4] AlZohbi G. Ammonia from hydrogen: a viable pathway to sustainable transportation?[J]. Sustainability, 2025, **17**(18): 8172.

[5] Zhao H, Fan W, Deng X, et al. Coupled hybrid neural network models with alarm threshold optimization for early warning of abnormal load variations in green ammonia production[J]. Journal of the Taiwan Institute of Chemical Engineers, 2025, **174**: 106224.

[6] Ishaq H, Shehzad M F, Crawford C. Transient modeling of a green ammonia production system to support sustainable development [J]. International Journal of Hydrogen Energy, 2023, **48**(99): 39254-39270.

[7] Xu Z C, Gao S D, Feng K K, et al. Process integration and optimization of green ammonia synthesis under fluctuating renewable energy supply[J]. International Journal of Hydrogen Energy, 2026, **212**: 153736.

[8] Li G D, Fan W, Zhao J, et al. Structural design and numerical simulation of a combined gas-liquid separator for renewable energy-driven water electrolysis systems[J]. International Journal of Hydrogen Energy, 2026, **209**: 153612.

[9] Armijo J, Philibert C. Flexible production of green hydrogen and ammonia from variable solar and wind energy: Case study of Chile and Argentina[J]. International Journal of Hydrogen Energy, 2020, **45**(3): 1541-1558.

[10] Wang C L, Walsh S D C, Longden T, et al. Optimising renewable generation configurations of off-grid green ammonia production systems considering Haber-Bosch flexibility[J]. Energy Conversion and Management, 2023, **280**: 116790.

[11] Xie Z Y, Fan Y F, Hou J J, et al. Multi-time-scale coordinated optimization scheduling strategy for wind - solar - hydrogen - ammonia systems[J]. Electronics, 2026, **15**(4): 795.

[12] He G, Dang Y G, Zhou L, et al. Architecture model proposal of innovative intelligent manufacturing in the chemical industry based on multi-scale integration and key technologies[J]. Computers & Chemical Engineering, 2020, **141**: 106967.

[13] Gottheil L, Bremer J. Enabling load-flexible ammonia synthesis via polytropic fixed-bed reactors[M]//Manenti F, Reklaitis G V. 34th European Symposium on Computer Aided Process Engineering / 15th International Symposium on Process Systems Engineering. Computer Aided Chemical Engineering, Vol. **53**. Amsterdam: Elsevier, 2024: 1891-1896.

[14] Zhao J, Hu Z B, Yang J, et al. Adaptive feature selection and random forest modeling for green ammonia production process[J]. Chemical Engineering & Technology, 2025, **48**(7): e70076.

[15] 贺革, 吉旭, 苏鹏, 等. 规模化绿电驱动下绿氨多稳态柔性调控技术与工程实践[J/OL]. 化工进展, 2026. DOI: 10.16085/j.issn.1000-6613.2026-1055.
He G, Ji X, Su P, et al. Multi steady state flexible regulation technology and engineering practice for green ammonia driven by large scale renewable power[J/OL]. Chemical Industry and Engineering Progress, 2026. DOI: 10.16085/j. issn. 1000-6613.2026-1055.

[16] Zhou J H, Zhang Z Y, Zhang R Z, et al. Optimal capacity and multi-stable flexible operation strategy of green ammonia systems: Adapting to fluctuations in renewable energy[J]. Energy Conversion and Management, 2024, **314**: 118720.

[17] Kalamaris N N, Maravelias C T. A general optimization framework for designing chemical & energy systems subject to multi-scale temporal variability[J]. Computers & Chemical Engineering, 2025, **203**: 109315.

[18] Yousri D, AlKuwaiti R H, Farag H E Z, et al. Advanced modeling of PEM electrolyzers for microgrid systems: Incorporating electrochemical and thermal models[J]. International Journal of Hydrogen Energy, 2024, **83**: 755-773.

[19] Sun Z X, Zhang Y C, Huang H J, et al. Modeling and simulation of dynamic characteristics of a green ammonia synthesis system[J]. Energy Conversion and Management, 2024, **300**: 117893.

[20] Zhang X M, Li G D, Zhou Z T, et al. How to achieve flexible green ammonia production: insights via three-dimensional computational fluid dynamics simulation[J]. Industrial & Engineering Chemistry Research, 2024, **63**(28): 12547-12560.

[21] 熊敏, 刘冬妹, 王智超, 等. 变负荷条件下绿氨生产操作参数的调控与优化[J]. 化工学报, 2025, **76**(6): 2791-2801.
Xiong M, Liu D M, Wang Z C, et al. Optimization and adjustment of operating parameters for green ammonia production under variable load conditions[J]. CIESC Journal, 2025, **76**(6): 2791-2801.

[22] Du W L, Yang S Y. The potential and challenges of large language model agent systems in chemical process simulation: from automated modeling to intelligent design[J]. Frontiers of Chemical Science and Engineering, 2025, **19**(10): 99.

[23] Van de Vijver R, Devocht B R, Van Geem K M, et al. Challenges and opportunities for molecule-based management of chemical processes[J]. Current Opinion in Chemical Engineering, 2016, **13**: 142-149.

[24] Yan Z B, Zhang W T, Liu Y T, et al. A physics-informed hybrid-fidelity surrogate model for chemical process modeling and optimization[J]. Journal of the Taiwan Institute of Chemical Engineers, 2026, **185**: 106679.

[25] Kitchin J R, Alves V, Laird C D. Beyond the fourth paradigm of modeling in chemical engineering[J]. Nature Chemical Engineering, 2025, **2**(1): 11-13.

[26] 吉旭, 林今, 聂李红, 等. 适用可再生能源不确定特性的合成氨多稳态柔性工艺技术[J]. 洁净煤技术, 2024, **30**(2): 23-35.
Ji X, Lin J, Nie L H, et al. Multistable-flexible ammonia process adapted to renewable energy[J]. Clean Coal Technology, 2024, **30**(2): 23-35.

[27] Yang Y H, Li R K, Ma J F, et al. CeProAgents: a hierarchical agents system for automated chemical process development[PP/OL]. V1. arXiv (2026-03-02) [2026-10-09]. https://arxiv.org/abs/2603.01654. DOI: 10.48550/arXiv.2603.01654.

[28] Pham T D, Tanikanti A, Keçeli M. ChemGraph as an agentic framework for computational chemistry workflows[J]. Communications Chemistry, 2026, **9**: 33.

[29] Chakraborty A, Taskiran N P, Kottooru R, et al. Building hybrid AI models in chemical engineering: a tutorial review[J]. Computers & Chemical Engineering, 2025, **201**: 109236.

[30] Galitsky B, Rybalov A. Neuro-symbolic verification for preventing LLM hallucinations in process control[J]. Processes, 2026, **14**(2): 322.

[31] Huang J, Han T, Yu N, et al. Socratic Elenchus-inspired multi-agent debate for mitigating hallucinations in large language models [J]. Expert Systems with Applications, 2026, **320**: 132208.