扩展 QuantumBench：面向量子问题求解的选择性符号增强与多视角评测框架

摘要

大语言模型在科学研究中已展现出加速知识获取与推理分析的潜力，并逐步被应用于物理等高技术门槛领域。然而，量子科学任务通常同时涉及形式化数学推导、精确数值计算以及高度专业化的符号表达，这对以自然语言建模为核心的大语言模型提出了更高要求，现有评测结果也表明其在该领域的推理可靠性与能力边界尚不清晰。针对量子科学的任务中的“形式化计算可验证不足”与“多项选择评测难以刻画真实推理能力”的结构性问题，本文在严格复现QuantumBench基准的基础上，提出一套面向量子推理的多视角扩展评测框架。具体而言，我们引入选择性符号增强的混合推理机制，在保持评测可比性的前提之下，将外部符号计算限定于高收益子分布，从而系统分析工具增强在量子任务中的适用边界。同时构建开放式量子推理任务集，并提出结合规则化自动评估，与 LLM-as-Judge 的双重评测协议，以刻画推理过程质量与正确性之间解耦关系。进一步地，我们设计研究生层次的多阶段量子推理基准，用于探索模型在更高阶理论推导场景的能力上限。实验结果表明，单一多选准确率难以充分反映模型在量子推理中的可靠性，而所提出的扩展评测框架能够系统揭示不同推理路径、工具使用与不同任务难度下的系统差异，为量子领域的大语言模型评测与应用提供更具诊断性的应用。

关键词： 量子推理评测；工具增强大语言模型；符号计算；开放式评估

---

1 引言

大语言模型（LLMs）已在通用知识问答与文本生成任务中展现出显著能力，并逐步被引入科学研究流程，包括科学写作、代码生成、实验规划与数据分析等（Brown et al., 2020; Achiam et al., 2023; Gemini Team, 2023; Dubey et al., 2024）。然而，越来越多研究表明，当任务涉及高度形式化的数学推导、可验证的数值计算以及严格的领域符号体系时，LLMs 往往表现出系统脆弱性，具体体现为计算性错误、推导步骤跳跃以及表面连贯，但实质不正确的幻觉式陈述。这类问题直接限制了模型在高可信科研场景中的适用性，也使其推理能力的真实边界变得难以评估（Wei et al., 2022; Wang et al., 2022;

Ji et al., 2023; Manakul et al., 2023）。这一问题在科学推理任务中尤为突出：即便模型能生成结构完整、语言连贯的推理文本，也可能在关键等价变形、近似条件或数值一致性环节发生偏差，导致“表面连贯但不正确”的结论（Kojima et al., 2022; Cobbe et al., 2021; Hendrycks et al., 2021; Stephan et al., 2024）。

量子科学构成了上述问题的极端且具有代表性的测试场景。一方面，量子力学及其相关子领域依赖高度形式化的数学结构，包括线性代数、算符方法等，要求模型在符号操作与物理语义之间保持严格一致。另一方面，量子概念本身具有显著的非直觉性，问题求解通常会依赖隐含的物理约定（如单位、常数、近似与边界条件）。在此背景下，模型生成的错误往往并非显性语法失误，而是潜藏于等价化简、近似条件滥用或推导链断裂之中，使得“语言上合理”的推理文本并不必然对应物理或数学上的正确结论。（Meurer et al., 2017; Yao et al., 2022; Pan et al., 2024）。为系统评估 LLM 在量子问题求解中的表现，作者 Minami 提出 QuantumBench ，该基准通过覆盖多个量子子领域的多项选择题，为量子领域 LLM 评测提供重要参考。（Minami et al., 2025）。然而，多项选择评测在方法论层面仍存在难以回避的结构性局限。首先选择题正确率不可避免地受到随机猜测与选择设计的影响，使得中等准确率区间内的性能差异难以映射为真实能力差异（Hendrycks et al., 2020; Wang et al., 2024）。其次，该评测范式仅关注最终选项，难以反映推理过程的正确性、论证的完整性以及中间步骤的可核验性，这与近年来针对“过程可见性≠正确性”的评估讨论相呼应，也促使开放式评测与更强的过程性指标被广泛关注（Zheng et al., 2023; Liu et al., 2023; Gu et al., 2024; Ye et al., 2024）。第三，以本科难度题目为主的设定亦难以刻画模型在更高阶理论推理场景下的能力上限（Rein et al., 2023; Zhu et al., 2025; Phan et al., 2025）。

基于上述问题，本文并非试图提出一种新的量子问题求解算法，而是从评测方法论的角度出发，对现有量子评测范式进行系统性的扩展。我们围绕 QuantumBench 的标准流程开展严格复现，并在此基础上引入三项互补机制，分别从可验证计算、推理过程刻画与难度外推三个维度拓展信号评测空间，以更真实地刻画 LLM 在量子科学任务中的能力边界与可靠性分析。本文的核心目标在于回答：在高度形式化的量子推理场景中，应如何构建更具诊断性的评测架构，而非仅依赖单一的多选准确率的指标。这一整体思路与近期“让模型在可验证计算与外部工具交互中获得更稳健推理”的趋势一致（Schick et al., 2023; Patil et al., 2023; Qin et al., 2023; Gou et al., 2024）。

本文主要贡献如下：

- **C1：提出一种面向量子推理评测的选择性符号增强混合框架。**不同于将工具增强视为争议模块的既有做法，本文基于题型与子领域标注构建确定性门控策略，仅在可验证计算收益显著子分布中启用符号执行，从而系统分析工具增强在量子任务中的正迁移与负迁移条件，并揭示其对评测结果稳定性的结构性影响。

- **C2：构建开放式量子推理任务集并提出双信号评测协议。**针对多项选择评测，难以刻画推理过程质量的问题，本文构建覆盖多类量子推理行为的开放式任务集，

并提出结合规则化自动评估与 LLM-as-Judge 的双重评分机制，以量化分析结构化推理能力与内容正确性之间解耦关系。

- **C3：提出研究生层次的量子推理外推机制QuantumBench-Grad。**在保持原有评测接口可复现性的前提下，引入多阶段推理约束与难度标注，用于系统考察模型在更高阶量子问题中的能力上限，并分析结构化推理行为是否能够可靠转化为最终正确性。

2 相关工作

在面向科学推理能力的系统评测中，通用基准已形成较成熟的谱系：除 GPQA 聚焦由领域专家撰写的研究生/博士难度生物、物理与化学选择题并以“Google-proof”作为核心设计目标之外（Rein et al., 2023），MMLU 也以覆盖多学科与职业领域的多选题形式，成为衡量模型通用知识与推理广度的经典基准之一（Hendrycks et al., 2021）；此外，BIG-bench 通过覆盖更广泛任务类型与能力维度补充了通用评测的生态位（Srivastava et al., 2022），而 BBH 则进一步以“更难子集”强化对复杂推理能力的区分度（Suzgun et al., 2022）。与之互补，SciBench 以大学层次的数学、化学与物理问题为主体，强调多步推理与计算求解过程，从而更细粒度地暴露模型在科学问题求解中的能力短板（Wang et al., 2023）；类似地，GSM8K 等算术推理数据集也常被用于检验模型在可计算任务上的稳定性与错误模式（Cobbe et al., 2021）。然而，上述通用科学基准通常并不针对量子科学特有的符号体系（如态矢、算符、张量积等）与子领域分布（如量子信息、量子计算、凝聚态与量子光学等）进行专门覆盖，因此其评测信号难以直接迁移为“量子领域能力”的可靠刻画。为弥补这一空白，QuantumBench 以量子科学为对象，汇集约 800 道、覆盖九个量子相关方向的八选一多项选择题，并进一步分析模型对题面格式变化的敏感性，为量子领域 LLM 评测提供了更贴近领域表征与任务形态的公开参照（Minami et al., 2025）。

与此同时，工具增强（tool augmentation）与混合推理框架为提升科学/量子推理的“可验证性”提供了重要方法论支撑。除了 Toolformer 证明语言模型可通过自监督方式学习“何时调用外部工具、如何将工具返回融合进生成”，从而在算术与检索等能力上获得显著增益之外（Schick et al., 2023），链式思维提示（CoT）也被广泛用于提升复杂推理任务的可分解性与可解释性（Wei et al., 2022），并进一步通过自一致性解码在多条推理路径之间进行边际化以改善稳健性（Wang et al., 2022）。ReAct 将推理轨迹中的“思考—行动”交织起来，使模型在与外部信息源或环境交互时同时保持可解释的中间推理过程，并在多类任务上提升成功率与可信度（Yao et al., 2022）。在“用工具做可执行推理”方面，PAL 通过让模型生成程序并交由解释器执行来降低算术与逻辑错误（Gao et al., 2022），而 ToRA 进一步将工具交互纳入数学问题求解代理的系统框架以提升可验证性与稳定性（Gou et al., 2024）。对量子推理任务而言，工具增强的关键价值往往不在于“额外知识注入”，而在于对代数化简、符号推导与数值一致

性等环节提供可复算、可审计的外部证据，从而把“看似合理的叙述”转化为“可核验的推导链”（Meurer et al., 2017）。与此同时，围绕真实 API/工具使用能力的评测与数据构建也在快速发展：例如 Gorilla 关注降低工具调用幻觉并提升 API 调用正确性（Patil et al., 2024），API-Bank 提供了更系统的工具增强评测与任务集合（Li et al., 2023），ToolLLM 则以大规模真实世界 API 指令数据推动模型工具使用能力的训练与评测（Qin et al., 2023），这些工作共同为本文“选择性符号计算增强”的门控与失败模式分析提供了更直接的研究背景。

最后，开放式生成任务的评价面临答案空间开放、表述多样且标准答案不唯一的结构性困难，促使 LLM-as-Judge 成为可扩展的近似人工评测方案。MT-Bench 系统讨论了 LLM 评判中的位置偏差、冗长偏差与自增强偏差等问题，并提出相应缓解策略，展示了强评判模型在与人类偏好的一致性方面可达到较高水平（Zheng et al., 2023）。G-Eval 进一步通过结构化评分表单与链式推理范式提升评判与人工评价的一致性，增强了开放式文本质量评估的可操作性（Liu et al., 2023）。与此同时，关于 LLM-as-Judge 的系统综述强调了其在可扩展评估中的机会与风险，并从一致性、偏差控制与可靠性验证等维度给出了更系统的研究框架与实践建议（Gu et al., 2024）；近年的实证研究也开始直接量化 Judge 在偏好、呈现方式与长度等因素上的系统性偏差（Ye et al., 2024），并指出在数学推导与可验证推理任务上 Judge 的不确定性与方差问题可能更突出（Stephan et al., 2024）。因而，在量子领域的开放式推理评测设计中，将“领域符号与推导可核验性”与“可扩展的判别式评价协议”相结合，能够在保证评测覆盖面的同时，最大限度降低开放式答案带来的主观性与不确定性。

3 QuantumBench 基准与复现设置

3.1 数据集与标注结构

QuantumBench 是一个面向量子科学推理能力评测的多项选择基准，数据集包含 769 道量子相关的 8 选 1 多选题，覆盖九大子领域，并为每道题提供题型标注（Conceptual Understanding、Algebraic Calculation、Numerical Calculation）（Minami et al., 2025）。在本研究的复现中，我们直接使用仓库内提供的数据文件 `quantumbench.csv` 与 `category.csv` 进行评测，以保证与原工作在数据来源与标注体系上的一致性。同时，为降低工程细节对结果的扰动并确保可复核性，我们严格遵循原评测脚本的选项打乱策略：在固定随机种子的条件下对每题选项进行置乱，使得选项顺序可由题号稳定复现，从而保证不同运行之间的评测输入等价。

为便于后续按题型与子领域展开对比分析，我们进一步基于本地 `category.csv` 对数据分布进行了统计，并汇总如下（统计结果由**我们实际 `category.csv` 计算得到**）。从整体结构看，题型分布高度偏向可计算类任务：代数计算题占比最高（74.8%），数值计算题次之（18.7%），而概念理解题占比较低（6.5%）；子领域分布则呈现“力学与

光学占主导、其余方向多样补充”的格局，这为后续讨论模型在不同量子子域的能力差异提供了必要前提。

表 1: QuantumBench（769题）数据分布

维度	类别	数量N	占比%
题型	Algebraic Calculation（代数计算）	575	74.8
题型	Numerical Calculation（数值计算）	144	18.7
题型	Conceptual Understanding（概念理解）	50	6.5
子领域	Quantum Mechanics	212	27.6
子领域	Optics	157	20.4
子领域	Quantum Field Theory	107	13.9
子领域	Quantum Chemistry	86	11.2
子领域	Quantum Computation	62	8.1
子领域	Photonics	57	7.4
子领域	Mathematics	37	4.8
子领域	String Theory	33	4.3
子领域	Nuclear Physics	18	2.3

3.2 模型、推理环境与复现协议

在模型与推理服务设置上，本研究选择 **Qwen2.5-7B** 作为被评测模型，并在本地环境中通过 Ollama 以 `qwen2.5:7b` 方式调用；模型相关技术细节以其官方技术报告为准（Qwen Team., 2024）。推理服务采用 Ollama 提供的本地 HTTP 接口，调用链路与仓库的本地推理方案保持一致，从而减少因框架差异引入的额外变量。实验在 Windows 11 平台完成，GPU 为 RTX 4060 8GB，使用 Python 3.12（uv 管理）执行全量 769 题评测及扩展实验记录。

在提示词（prompt）设计上，我们遵循仓库提供的 zero-shot 模板，要求模型以可解析的固定格式输出单一选项字母（如“The correct answer is (X).”），以降低输出格式噪声对计分与统计造成的影响。为保证结果可复现，本研究对所有涉及随机性的环节进行显式控制：多选题选项打乱固定 `seed=0`；同时在混合推理框架中，若存在回退策略的随机选择分支，则通过 `random.seed(seed + idx)` 将随机状态与题号绑定，使得相同数据与代码条件下可稳定复现实验轨迹（具体实现细节见第 4.2 节）。

3.3 评测指标

针对 QuantumBench 的多项选择题（MCQ）设置，本文以准确率（Accuracy）作为主评测指标，即模型输出选项与标准答案一致的比例。QuantumBench 将量子领域问

题组织为八选一选择题，并覆盖九个量子子领域及三类题型（代数计算、数值计算与概念理解），因此 Accuracy 能够在统一接口下对不同类别问题进行可比统计，并与原基准的评测口径保持一致。进一步地，为评估 MCQ 格式难以覆盖的“推导、解释与论证”能力，本文在创新点 2 中构建开放式任务并引入双重评分机制，输出 0–100 的综合得分；该得分由规则化自动评估与 LLM-as-Judge 按权重融合得到（第 4.3 节给出形式化定义）。此外，为刻画更高难度场景下的结构化推理能力，本文在创新点 3 中构建研究生层次基准 QuantumBench-Grad，在保持多选准确率统计的同时，额外报告回答是否显式覆盖 Part A/B/C/D 的“推理阶段覆盖度”，从而将“结论正确性”和“推理结构完整性”区分为两个可分析维度。

4 方法（Methods）

4.1 复现实验基线（Baseline Reproduction Protocol）

为保证与 QuantumBench 的原始评测流程严格对齐，本文复现其标准管线，即 8-way MCQ + zero-shot 指令提示 + 规则化答案抽取（Minami et al., 2025），从而建立可与原工作直接对比的统一参照。QuantumBench 的构建过程强调每题可独立求解、选项由人工策划为高质量干扰项，并将题目划分为九个子领域与三类题型以支持分组评测（Minami et al., 2025）。在该设定下，我们的基线复现包括三个相互衔接的实现环节：在题目与选项重排阶段，将 1 个正确答案与 7 个干扰项合并为 8 个选项，并以固定随机种子 `seed=0` 对选项顺序进行置乱，以降低模型利用选项位置先验（position bias）的可能性（Zheng et al., 2023; Wang et al., 2024）；在提示词设计上，采用 zero-shot 模板输入“问题 + 8 个选项 + 输出格式约束”，要求模型以 `The correct answer is (X).` 的固定格式输出选项字母，从而降低自由生成带来的解析不确定性并保持与仓库脚本一致；在答案抽取与失败处理上，使用规则化正则表达式从模型输出中抽取最后出现的合法选项字母 A–H 作为预测，若无法抽取则标记为解析失败并触发回退逻辑。需要强调的是，解析失败的回退策略会直接影响统计公平性与误差结构，因此本文在第 4.2.3 节进一步引入无偏化回退机制，以避免由固定回退选项导致的系统性偏差。

4.2 创新点1：选择性符号增强的混合推理（Selective Symbolic-LLM Hybrid Reasoning）

量子科学问题通常同时包含概念判断与代数/数值计算，且表达式化简、单位一致性与数值精度往往决定最终选项是否正确（Meurer et al., 2017; Gao et al., 2022; Gou et al., 2024）。即便在通用推理能力较强的模型上，代数化简中的符号歧义、符号/下标处理错误，以及数值近似引发的细微偏差，仍可能导致“局部推理看似合理但最终选错”的失误。外部符号计算工具（如 SymPy）能够提供可执行、可复算、可审计的计算路径（Meurer et al., 2017），从而在原则上提升计算可靠性；然而，若对所有题目无差别

启用工具，工具调用本身会引入额外失败模式（例如代码生成不稳定、表达式解析失败、超时与异常），并可能因链路复杂度上升而造成整体性能回落。基于上述张力，本文提出选择性符号增强混合推理框架，其目标并非在数值上最大化多选准确率，而是作为一种评测诊断工具，用于刻画外部符号计算在不同量子任务结果下的有效边界性。该思路与近年来“将语言推理与外部可验证计算结合”的研究趋势一致，例如 ToRA 在数学推理任务中通过工具交互提升可验证性与求解稳定性（Gou et al., 2024）。

4.2.1 选择性门控：何时启用 SymPy

记题目为 q，其题型标签为 $t(q)$，子领域标签为 $d(q)$。其中 $t(q) \in \{\text{Conceptual}, \text{Algebraic}, \text{Numer}$ 分别对应 Conceptual Understanding、Algebraic Calculation 与 Numerical Calculation。v4 的确定性门控函数定义为：

$$\mathbb{I}_{\text{sym}}(q) = \mathbf{1}[t(q) = \text{Numerical}] \ \vee \ \mathbf{1}[t(q) = \text{Algebraic} \wedge d(q) \in \mathcal{D}_+], \tag{1}$$

其中 $\mathbf{1}[\cdot]$ 为指示函数，且

$$\mathcal{D}_+ = \{\text{Mathematics}, \text{Quantum Computation}, \text{Quantum Chemistry}, \text{Quantum Mechanics}\}. \tag{2}$$

该门控策略的直观含义是：概念理解题不启用符号路径，因为其主要误差源来自物理概念与语义判断，符号执行难以直接提供增益；数值计算题始终启用符号路径，因为该类题目高度依赖数值一致性与可复算性，工具校验对减少低级算错更有效；代数计算题仅在 ($\mathcal{D}_+$) 中启用符号路径，因为该子集通常表达更规范、可解析性更强、化简与匹配收益更显著。为确保可复现性与可审计性，本文在实现中保持 ($\mathcal{D}_+$) 与 v4 脚本中的 `SYMPY_DOMAINS` 完全一致（`Mathematics / Quantum Computation / Quantum Chemistry / Quantum Mechanics`），使得门控决策可由题型与子领域标签确定性复现。

4.2.2 两阶段混合推理与无偏回退：语言求解优先，符号路径兜底

当且仅当 $\mathbb{I}_{\text{sym}}(q) = 1$ 时，本文采用“两阶段”策略以平衡成本与收益。第一阶段为 zero-shot 语言求解：沿用基线提示生成答案并解析选项字母；对代数计算题而言，若解析成功则直接返回以节省额外开销。第二阶段为符号增强路径：对数值计算题固定触发，对代数计算题仅在第一阶段解析失败时触发，通过“生成可执行的 SymPy 代码—受限执行—将执行结果与选项匹配”的方式获得可验证答案。符号执行后端采用 SymPy（Meurer et al., 2017），并在受限执行环境中控制可复现性与鲁棒性，包括预置常用符号与函数（如 `x,y,z,t,hbar,omega` 等）、设置执行超时（例如 30s）、捕获异常并记录日志，以及将执行得到的 `result` 与选项进行数值/字符串匹配以输出最终字母预测。通过将工具使用限定为“校验/兜底”而非“全程主导”，该框架旨在降低工具链路引入的新失败模式，并将工具优势集中在最易产生可验证收益的计算环节。

在工程复现中，答案解析失败是影响 MCQ 统计稳定性的主要噪声源之一。若采用“解析失败固定回退为 A”等简单策略，会引入强位置偏差并污染统计结论。为此，本文在 v3/v4 中引入分层的鲁棒抽取与无偏回退：首先扩展正则覆盖多种常见表述（如 `The answer is (X)`、`The correct answer is X` 以及括号/加粗等变体）；若仍失败，则在输出文本尾部窗口中搜索候选字母；最终仍失败时，在 A–H 中按固定随机种子进行随机选择，以避免固定偏置并保持跨运行可复现。该处理将“解析失败”从系统性偏差源转化为可控的随机噪声，从而使不同方法之间的对比更接近真实能力差异。

4.3 创新点2：开放式任务构建与双重评测框架（Open-ended + Dual Evaluation）

QuantumBench 的 8-way MCQ 形式便于规模化统计，但真实科研与高阶教学情境中的量子问题往往要求完整推导、概念解释与论证，而非仅输出选项字母。QuantumBench 也指出，未来更贴近实践的评测应纳入开放式描述性问题与结构化任务分解等形式（Minami et al., 2025）。因此，本文构建开放式任务集，并提出“自动评估 + LLM-as-Judge”的双重评测框架：自动评估提供结构与形式层面的可解释约束，LLMJudge 尽可能近似对内容正确性与论证质量的判别信号。需要注意的是，近两年研究已系统指出 LLM-as-Judge 在冗长偏好、呈现方式敏感性与一致性方面可能存在结构性偏差（Ye et al., 2024），且在数学/推导类内容上这种偏差可能被放大并表现为更高的不确定性（Stephan et al., 2024）。基于这一现实，本文采用双信号融合而非单一 Judge，以降低对单一评估源的过拟合风险，并在局限性部分进一步讨论该设置的可解释边界。

4.3.1 任务集构建：5 类开放式题型，共 165 题

开放式任务以 JSON 形式存储于 `data/open_ended_tasks.json`，并保留与原题的关联字段 `original_question_id` 以保证可追溯性与可复核性。任务设计覆盖从“生成推导”到“诊断错误”的多类能力切面，具体包括 `free_derivation`（50 题）、`concept_explanation`（50 题）、`analytical_qa`（15 题）、`error_diagnosis`（30 题）与 `multi_step_reasoning`（20 题），合计 165 题。该划分使后续分析能够区分模型在推导正确性、概念阐释质量、错误定位能力与多步组织能力等维度上的差异，而不局限于 MCQ 的离散答案空间。

4.3.2 评分器与融合：AutoEvaluator、LLMJudge 与最终得分

自动评估器输出 0–100 分，并由五个可解释子维度加权求和：关键词覆盖率、推理深度（基于步骤标记与连接词统计）、响应长度适配性、公式/符号使用（如 LaTeX 与狄拉克记号模式）以及结构完整性（requirements 覆盖检查）。加权定义为：

$$S_{\text{auto}} = \sum_{k=1}^{5} w_k s_k, \quad w = \{0.20, 0.25, 0.15, 0.20, 0.20\}. \tag{3}$$

该设计的目标是在不依赖外部参考答案的前提下，为开放式回答提供稳定、可复核的“结构与形式”约束信号，从而降低仅凭主观评价导致的不可重复性。

LLMJudge 按五项维度输出结构化 JSON：物理准确性、数学正确性、逻辑连贯性、表达清晰度与完整性，并将五项得分取均值作为 Judge 总分：

$$S_{\text{judge}} = \frac{1}{5}\sum_{i=1}^{5} s_i. \tag{4}$$

考虑到可复现性与成本控制，本文在评估中使用同一模型（Qwen2.5-7B）通过不同 system prompt 扮演被测模型与 Judge。该设置具备工程可复现性，但可能引入自评偏置，并受到 LLM-as-Judge 已知偏差模式的影响（Ye et al., 2024；Stephan et al., 2024）。因此，本文将其作为可复现的基线评估方案，并在第 7 节对潜在风险与改进方向作出明确限定，而不将其等同于人工标注的无偏真值。所以本文将自动评估与 Judge 评分进行加权融合，定义最终得分为：

$$S_{\text{final}} = 0.4\, S_{\text{auto}} + 0.6\, S_{\text{judge}}. \tag{5}$$

其中更高权重赋予 S_{judge} 的动机在于，开放式解答的物理正确性与论证质量难以被纯规则指标充分捕捉；与此同时，S_{auto} 提供结构与形式的“下限约束”，可在一定程度上缓解 Judge 对冗长与表述风格的偏好所带来的偏差风险（Ye et al., 2024）。

4.4 创新点3：研究生层次 QuantumBench-Grad 构建与评测（Graduate-level Benchmark）

QuantumBench 的题目主要来源于公开课程与教学材料，并且从其统计特征可见：整体更接近本科标准难度，不包含最高难度与最高专业度等级。然而研究生层次量子问题通常需要更强的理论框架与多阶段推理组织（从建模到推导再到物理解释），仅靠 MCQ 的一次性作答难以刻画“推理过程是否完整”。因此，本文构建 QuantumBench-Grad：在不破坏原评测接口（仍为 8-way MCQ）的前提下，引入多阶段结构化推理要求，使评测既能保持与原管线兼容的可复现统计，又能额外观察模型在高阶推理组织上的行为特征。

4.4.1 数据规模、来源与标注文件

QuantumBench-Grad 共 71 题（`data/grad_benchmark/quantumbench_grad.csv`），由两部分构成：一部分为 21 道人工设计的研究生风格多阶段模板题（Question id 0–20），用于覆盖典型“分段推导—物理解释—结论归纳”的回答结构；另一部分为从原 QuantumBench 抽样并升级的 50 题（Question id 21–70），对题干统一添加 A–D 多阶段推理前缀以提高推理要求，同时保留 8 个选项与标准答案，从而确保可直接复用既有的“提示—解析字母—统计准确率”流程。为支持领域与难度维度的可分析性，本文提

供 category_grad.csv，包含 Subdomain_lecture（用于领域统计）、DifficultyLevel 与 ReasoningSteps 等字段。该设计使 QuantumBench-Grad 在数据层面与原基准保持相同的“可分组统计”能力，同时允许在实验章节对“难度分层—准确率变化—推理结构覆盖”进行更细粒度讨论。

4.4.2 研究生评测协议：准确率 + 推理阶段覆盖度

在评测协议上，模型需对 Part A–D 分段作答并最终输出 The answer is X。除准确率外，本文通过规则抽取回答中的阶段标记（如“Part A”、“Step 1” 等）估计推理结构覆盖度，作为推理完整性的辅助量化指标。该指标并不直接等价于正确性，但能够揭示模型是否被提示稳定触发结构化推理行为，从而为后续分析“长推理文本是否转化为更高正确率”提供证据基础。

5 实验与结果

5.1 实验设置

本节系统报告基线复现与三项扩展（选择性混合推理、开放式评估、研究生基准）的实验配置与结果。为保证可比性，除非特别说明，所有实验均采用同一模型与同一本地推理环境，并保持输出格式与 QuantumBench 原始 CSV 结构兼容，使不同方法的结果可直接对齐统计。被测模型为 Qwen2.5-7B，本地通过 Ollama 部署调用，其模型细节以官方技术报告为准（Qwen Team, 2024）。硬件环境为 RTX 4060 8GB，软件环境为 Windows 11 与 Python 3.12（uv 管理）。评测数据包括三部分：其一为 QuantumBench-MCQ 全量 769 题（创新点 1 的主评测对象）；其二为开放式任务 165 题（本节在 free_derivation 子集上报告 10 题试运行与 30 题扩展实验）；其三为 QuantumBench-Grad 全量 71 题（创新点 3）。指标方面，MCQ 与 Grad 采用准确率，开放式任务采用第 4.3 节定义的融合得分 $S_{\text{final}} = 0.4S_{\text{auto}} + 0.6S_{\text{judge}}$。

需要说明的是，数据标注中存在轻微噪声：如第 3.1 节所述，category.csv 中 Question id=659、660 的 Subdomain_question 出现非规范值。为避免该噪声影响领域对比的可解释性，本文在总体准确率统计中仍使用全量 769 题；而在按子领域统计时默认仅汇报 9 个规范子领域（合计 767 题），并在相应表格与文字中显式标注统计口径差异来源。

5.2 基线复现结果

我们首先在 QuantumBench-MCQ 全量 769 题上复现 zero-shot 基线，以建立后续所有改进的统一参照。基线总体准确率为 38.49%（296/769）。为刻画不同认知负载与能力成分下的差异，我们进一步按题型统计基线性能，结果如表 5-1 所示：代数计算题占比最大且准确率最低（36.70%），因此成为限制总体性能的主要瓶颈；数值计算题准

确率相对更高（44.44%）但仍存在可观提升空间；概念理解题并未显著优于计算类题目（42.00%），提示模型在量子概念体系掌握上同样存在不稳定性。

表 2: 基线按题型准确率（QuantumBench-MCQ）

题型（QuestionType）	题量	Baseline Acc.
Algebraic Calculation	575	36.70%
Numerical Calculation	144	44.44%
Conceptual Understanding	50	42.00%

5.3 创新点1：Symbolic-LLM Hybrid 的 MCQ 结果

本节比较 Baseline 与 Hybrid v1–v4 的全量评测结果，并进一步分析选择性门控（v4）为何能够在抑制工具副作用的同时获得更稳定的净增益。总体结果如表 5-2 所示：v1 与 v2 出现不同程度退化，表明无差别或弱约束的符号增强并不必然带来收益；v3 基本回到基线附近；v4 在全量准确率上提升幅度有限（+0.78pp），该结果并不只在证明一种新的性能最优解，而是作为一种评测诊断信号，表明在严格控制工具副作用后，符号增强在特定量子分布中能够产生稳定的正迁移。进一步的领域集分析显示，该争议呈现显著的结构性集中，而非均匀分布，这一现象恰恰揭示了多选准确率难以反映的能力差异，并验证了选择性门控在评测分析中的必要性。

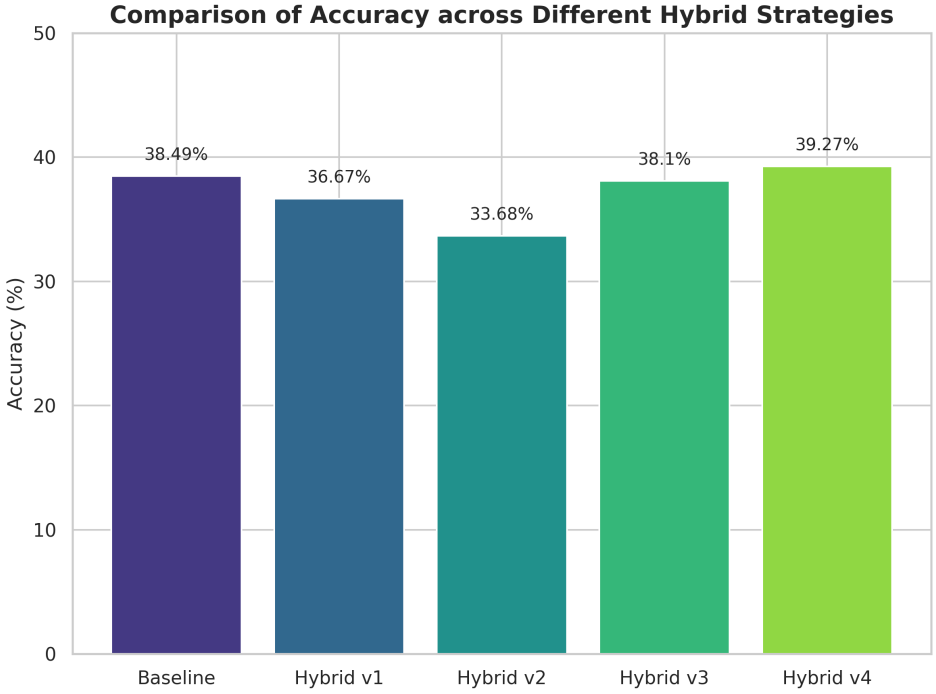

图 1: Baseline 与 Hybrid v1–v4 性能对比

表 3: Baseline 与 Hybrid v1–v4 总体准确率（769题）

方法	正确数/总数	Acc.	相对 Baseline（百分点）
Baseline (Zero-shot)	296/769	38.49%	+0.00
Hybrid v1	282/769	36.67%	−1.82
Hybrid v2	259/769	33.68%	−4.81
Hybrid v3	293/769	38.10%	−0.39
Hybrid v4 (Selective)	**302/769**	**39.27%**	**+0.78**

为检验 Baseline 与 v4 的配对差异是否具有统计显著性，我们采用双侧精确 McNemar 检验。在 Baseline 与 v4 的差异中，由对变错为 $b = 103$，由错变对为 $c = 109$，得到 $p = 0.731$。该结果表明，全量层面的 +0.78 个百分点提升尚不足以判定为统计显著，更适合作为“门控策略抑制副作用”的方向性证据。

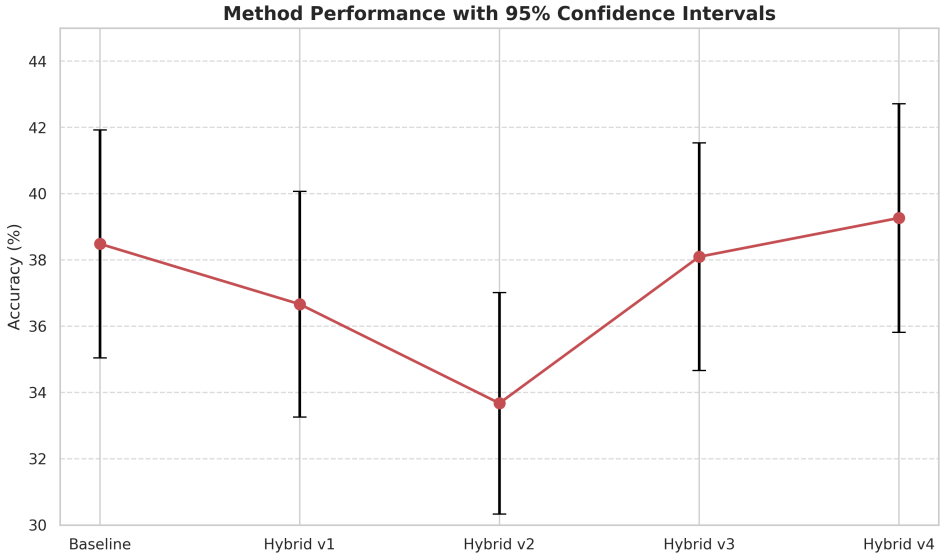

图 2: 准确率误差条分析

为了理解增益来源与潜在副作用，我们进一步按题型与子领域分解结果。按题型统计（表 5-3）显示，v3 虽在数值计算题上获得较大收益（47.92%），但概念题出现明显负迁移（28.00%），说明工具链路在不需要工具的题型上可能放大解析与匹配失败；v4 通过对概念题显式禁用 SymPy 路径，使概念题准确率恢复至基线水平，同时在代数与数值类题目上维持小幅提升，从而体现选择性启用的净效应来源。

表 4: 按题型准确率（Baseline vs v1–v4）

题型	Baseline	v1	v2	v3	v4
Algebraic Calculation	36.70%	36.87%	32.35%	36.52%	**37.57%**
Conceptual Understanding	42.00%	36.00%	38.00%	28.00%	**42.00%**
Numerical Calculation	44.44%	36.11%	37.50%	**47.92%**	**45.14%**

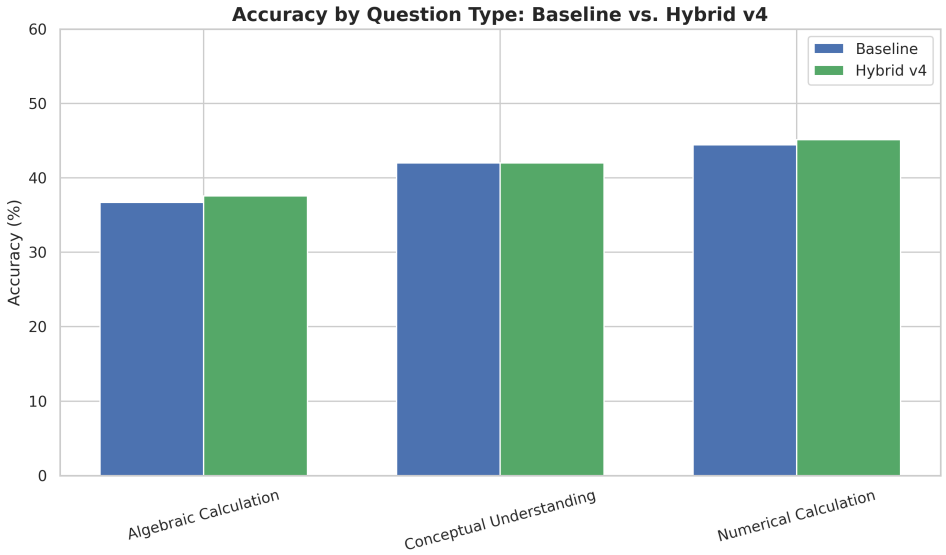

图 3: 按题型分类的性能对比

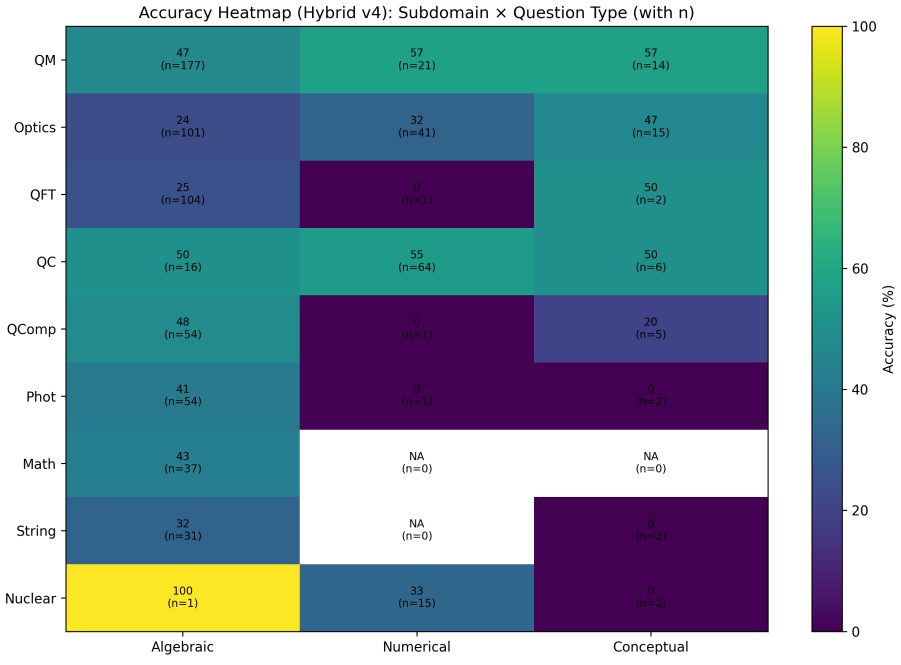

图 4: 不同方法在各题型上的表现热力图

进一步地，在仅包含 9 个规范子领域的 767 题统计口径下（表 5-4），v4 的增益呈现明显的领域集中性，其中 Quantum Computation 子领域提升最大（+18.33 个百分点）。在该子领域内进行双侧精确 McNemar 检验得到 $p = 0.0127$，但由于未进行多重比较校正，该结果应视为具有提示性的统计信号而非最终结论。这一现象意味着 v4 的门控子集 $\mathcal{D}_+$ 确实捕捉到了一部分“表达更规范、计算更可验证”的高收益问题，同时也提示在其他子领域（如 Optics、Photonics）中，符号路径仍可能受限于表达式解析与选项匹配难度，从而出现收益不足以抵消工具引入失败率的情况。

表 5: 子领域准确率（Baseline vs v4；9个规范子领域，767题）

子领域	题量	Baseline	v4	Δ（pp）
Quantum Mechanics	212	44.81%	48.58%	+3.77
Optics	157	31.21%	28.03%	−3.18
Quantum Field Theory	107	28.04%	25.23%	−2.80
Quantum Chemistry	86	53.49%	53.49%	+0.00
Quantum Computation	**60**	**26.67%**	**45.00%**	**+18.33**
Photonics	57	43.86%	38.60%	−5.26
Mathematics	37	45.95%	43.24%	−2.70
String Theory	33	33.33%	30.30%	−3.03
Nuclear Physics	18	27.78%	33.33%	+5.56

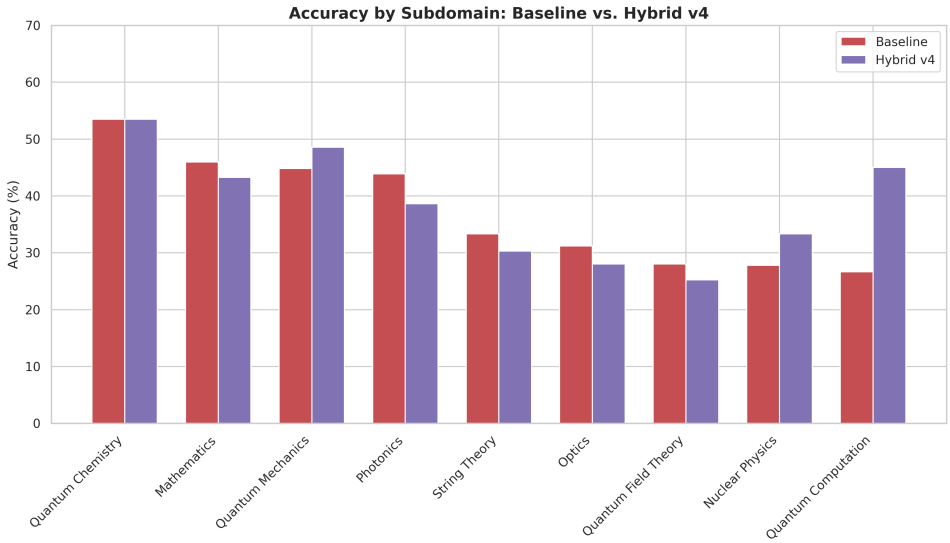

图 5: 各量子子领域准确率差异分析

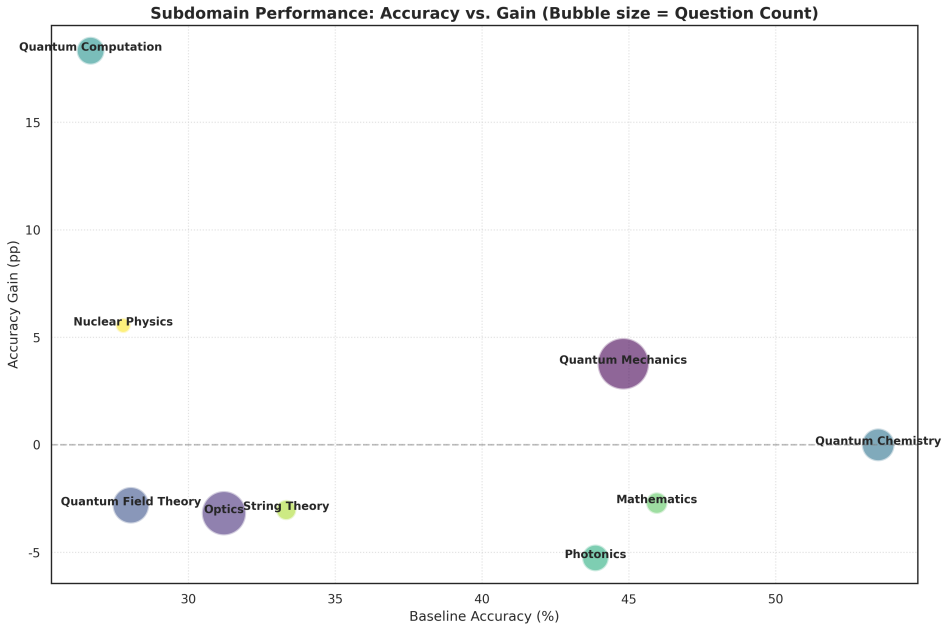

图 6: 子领域样本量与增益的气泡图分析

为直接观察门控是否把工具用于更适合工具的样本，我们记录了 v4 每题路由结果（`Strategy` $\in$ {`sympy`, `zeroshot`}）。如表 5-5 所示，SymPy Hybrid 覆盖 55.7% 的样本，并在该覆盖子集上达到 46.26% 的准确率，高于 zero-shot 子集的 30.50%。需要强调的是，该对比反映不同策略覆盖子集的经验表现，并非同分布条件下的严格公平对照，但它说明门控确实在经验上将更可能受益的样本路由至工具路径，从而在总体上形成净增益。与此同时，v4 也带来显著成本上升：Baseline 平均 token（Prompt + Completion）约 936.6，而 v4 约 1492.5；在 RTX 4060 8GB 的全量运行记录中，v4 完成 769 题总耗时约 4 h 34 m 32 s，平均 21.42 s/题。该结果表明 v4 更像是一种以额外推理与执行开销换取计算一致性的工程策略，因此在资源受限或高吞吐场景中仍需探索更细粒度门控或更轻量校验机制以改善成本—收益比。

表 6: v4 路由策略分布与准确率（769题）

策略	题量	占比	Acc.
SymPy Hybrid	428	55.7%	**46.26%**
Zero-shot	341	44.3%	30.50%

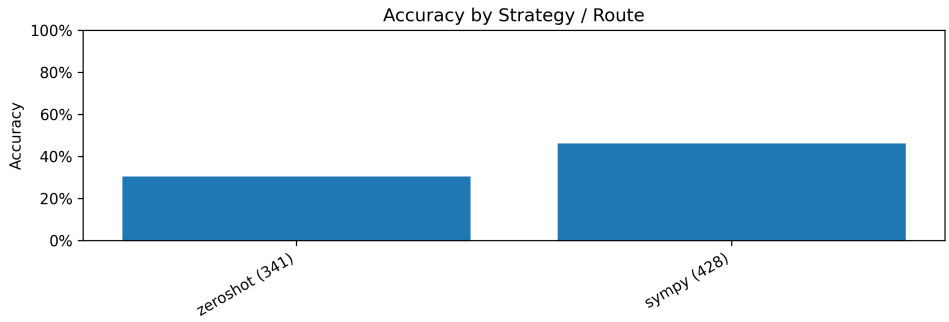

图 7: v4 版本的路由策略选择分布

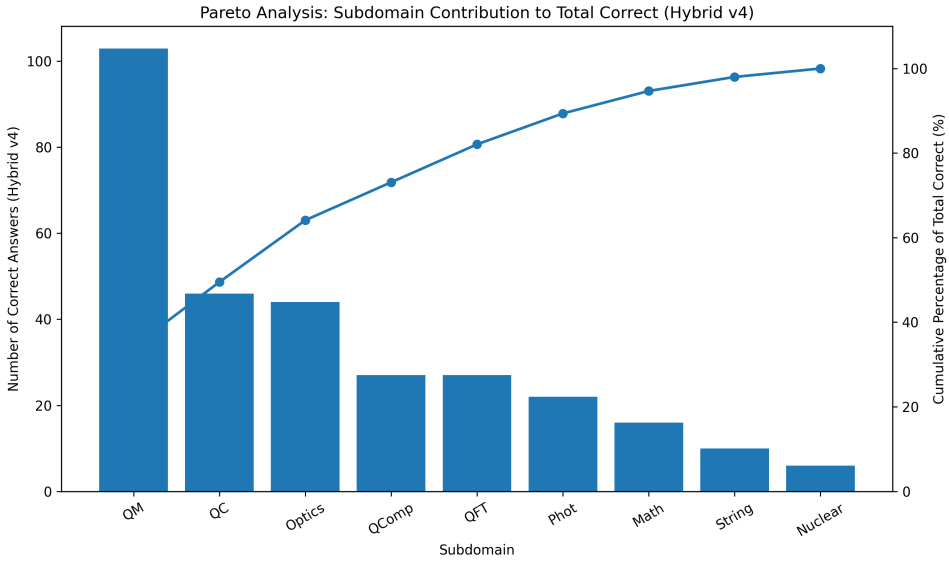

图 8: 准确率与推理成本的帕累托分析

5.4 创新点2：开放式推理任务的实验结果

本节在 `free_derivation` 子集上报告开放式评测框架的运行结果，并分析 AutoEvaluator 与 LLMJudge 的互补性。开放式评测的核心难点在于答案空间开放且标准答案不唯一，因此评估信号必须在可扩展性与可靠性之间权衡。近两年的研究表明，LLM-as-Judge 可能存在系统偏差，尤其可能偏好冗长回答或受到风格与呈现方式影响（Ye et al., 2024），并在数学推导类任务中表现为更高的不确定性与更大的方差（Stephan et al., 2024）。基于这一现实，本文采用 AutoEvaluator 提供结构/形式的稳定约束，并以 LLMJudge 提供面向正确性的主信号，通过加权融合降低单一评估源的偏差风险。

表 5-6 给出 10 题试运行与 30 题扩展实验的汇总结果。可以看到，综合得分 S_{final} 的均值从 10 题到 30 题变化较小（73.7→71.7），提示该框架在该子集上具有一定稳定性；同时，LLMJudge 的方差显著高于 AutoEvaluator，反映开放式回答质量波动较大，且 Judge 对细微正确性差异更敏感，这与近期关于 Judge 不确定性行为的结论一致（Stephan et al., 2024）。

表 7: Open-ended（free_derivation）评估汇总

规模	N	Auto (均值±std)	LLM (均值±std)	Final (均值±std)	推理步数 (均值)	端到端用时 (实验记录)
试运行	10	58.6 ± 6.4	83.7 ± 14.3	73.7 ± 8.6	12.4	约 7 min
扩展	**30**	**60.3 ± 5.4**	**79.4 ± 14.5**	**71.7 ± 9.5**	**12.5**	约 21 min

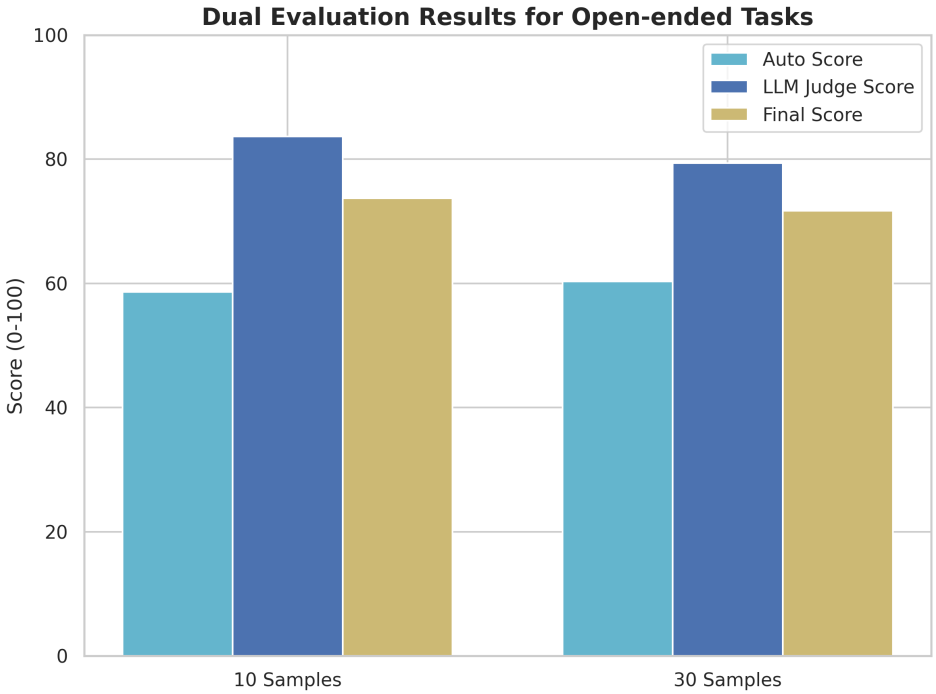

图 9: 开放式任务得分分布

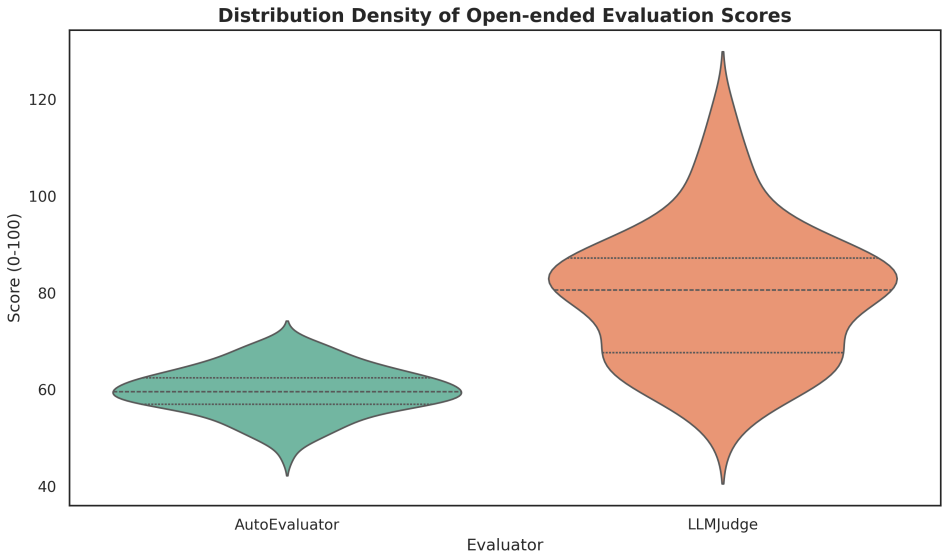

图 10: Auto 与 Judge 评分的小提琴图分布

在 30 题扩展实验中，本文进一步按领域统计 Final 均值（表 5-7）。总体上，不同领域之间的均值差异可能由领域对严格推导与背景假设的依赖程度不同所致；同时，由于部分领域样本量较小（例如 Mathematics 仅 2 题），该表更适合用于定性观察而非强结论。为检验双评测的互补性，本文计算 Auto 与 LLMJudge 的相关系数并得到 $r \approx 0.26$，该低相关性提示二者确实关注不同侧面：Auto 更强调形式与结构一致性，Judge 更强调内容正确性与论证质量，因此融合机制能够在一定程度上降低“只看结构”或“只看主观正确性”的单侧偏差。

表 8: 30题扩展实验：按领域 Final 得分

领域	题数	**Final（均值）**
String Theory	3	76.5
Quantum Computation	3	74.9
Quantum Mechanics	8	72.4
Quantum Field Theory	4	70.9
Optics	6	70.8
Photonics	4	70.5
Mathematics	2	64.0

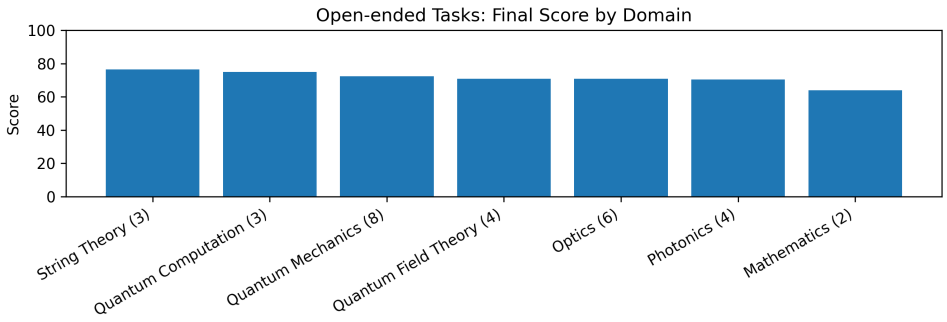

图 11: 按领域划分的开放式任务 Final 得分

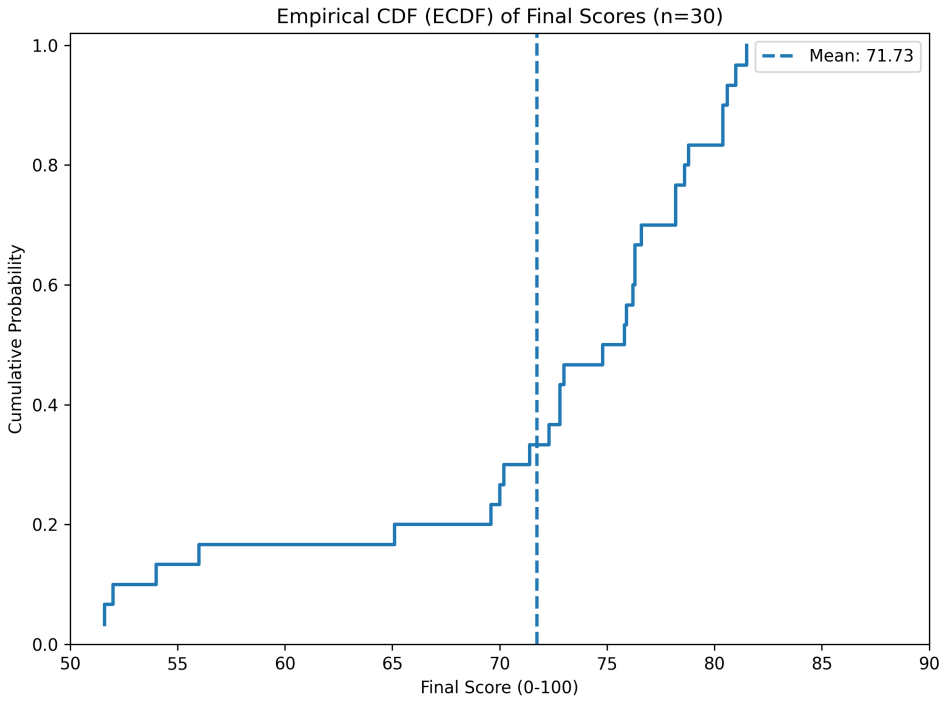

图 12: 得分累积分布函数（CDF）曲线

5.5 创新点3：QuantumBench-Grad（研究生基准）实验结果

本节报告 QuantumBench-Grad 全量 71 题评测结果，重点考察难度提升是否在统计上体现为整体准确率下降，以及多阶段提示是否能稳定触发结构化推理行为并与正确性产生关联。总体准确率为 35.21%（25/71），相较原 QuantumBench 基线（38.49%）下降 3.28 个百分点，表明在保持 8-way MCQ 接口不变的前提下，多阶段推理要求确实提高了任务难度并压缩了模型的正确率空间。按难度分层结果如表 5-8 所示，其中 Graduate-3 的样本量仅 2 题，因此不具统计意义，本文仅将其作为现象记录。

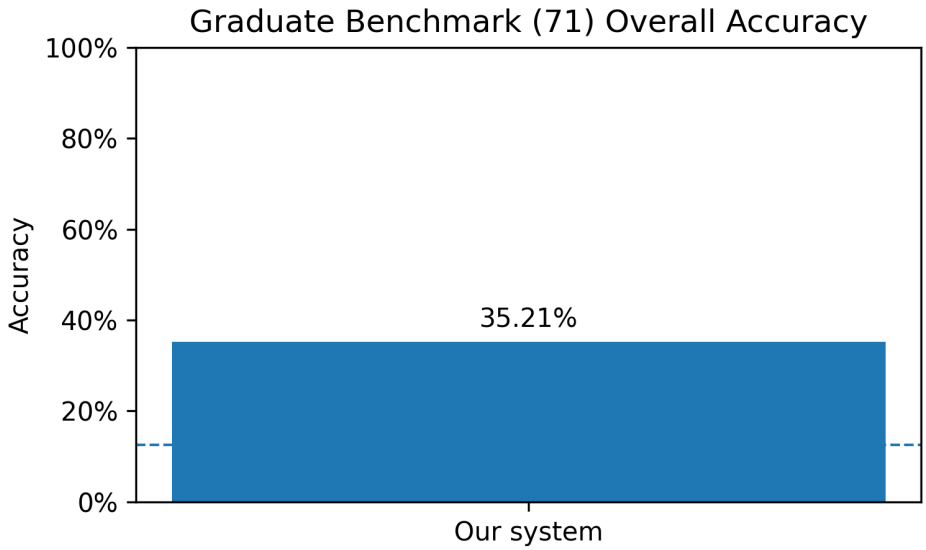

图 13: QuantumBench-Grad 总体准确率表现

表 9: QuantumBench-Grad 按难度分层准确率

难度级别	正确/总数	Acc.
Graduate-1	18/56	32.1%
Graduate-2	5/13	38.5%
Graduate-3	2/2	100.0%*

注：Graduate-3 样本量过小，不具统计意义。

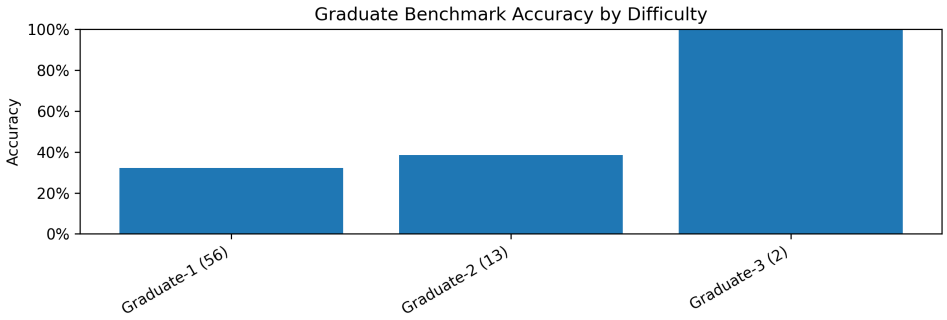

图 14: 按难度等级划分的 Grad 准确率

从领域维度看（表 5-9），不同 lecture-level 子领域之间差异较大，且存在多个样本量极小的领域（单题领域），因此该结果更适合作为薄弱领域定位的诊断性信号，而非用于断言模型在某子领域具有确定性优势或劣势。

表 10: QuantumBench-Grad 按领域准确率（71题）

领域	正确/总数	**Acc.**
Quantum Optics	1/1	100.0%*
Quantum Mechanics	7/13	53.8%
Condensed Matter	1/2	50.0%
Quantum Field Theory	4/8	50.0%
Photonics	6/15	40.0%
Quantum Computation	4/10	40.0%
Optics	2/11	18.2%
String Theory	0/5	0.0%
Quantum Chemistry	0/3	0.0%
Quantum Information	0/1	0.0%
Atomic Physics	0/1	0.0%
Particle Physics	0/1	0.0%

注：单题领域仅作现象记录。

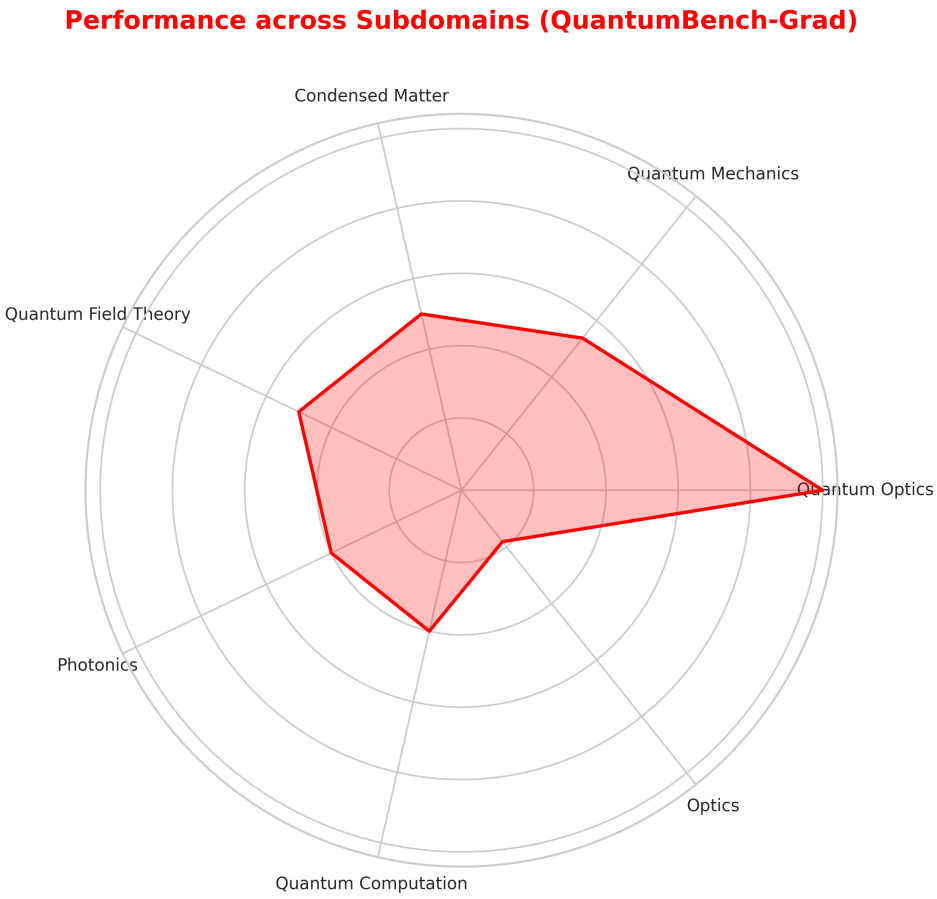

图 15: 各高阶子领域能力的雷达图

关于“结构化推理是否被触发”，实验显示在 Part A–D 的提示约束下模型几乎总能生成分段结构（Complete = 71/71，100%），阶段标记数量均值为 9.06（预期为 4），且最小 4、最大 35。这一结果表明，多阶段提示能够在行为层面稳定诱导更长、更结构化的推理文本；然而阶段数与正确性相关性接近 0，意味着生成更长推理链并不必然带来更高的最终正确率。该现象与近期关于“可见推理过程与真实正确性可能脱钩”的观察相一致，也进一步支持本文在开放式评测中采用“结构/形式 + 正确性”双信号联合评估的动机：仅凭过程可见性难以保证内容可靠，而仅凭最终对错又难以定位失败机制。

6 讨论

尽管选择性符号增强在全量 QuantumBench 上带来的总体提升仅为 +0.78pp，但其增益呈现出显著的领域集中性：收益主要聚焦于 Quantum Computation，而在 Optics、Photonics 与 Quantum Field Theory 等子领域中则可能出现不同程度的负迁移。这一现象表明，符号工具对大语言模型的有效增益并非“普适成立”，而是强烈依赖于题目表

达的规范性、可执行性以及“执行结果—选项空间”的可匹配程度。当题面表达式具有较高结构化程度、变量与常量约定清晰且目标计算可直接落在选项可比对的形式上时，符号执行能够显著提升代数化简与数值一致性的可靠性；相反，在更依赖近似、隐含单位/常数约定或表达形式多样的题目中，符号链路容易在表达式解析、数值化策略与选项匹配环节累积误差，从而抵消工具带来的潜在收益并最终表现为负迁移。因而，本文结果更支持一种“工具增强需被任务结构约束与选择性触发”的观点，即工具不应被视为通用增益模块，而应被纳入可解释的门控策略之中，使其仅在收益显著的子分布上发挥作用。

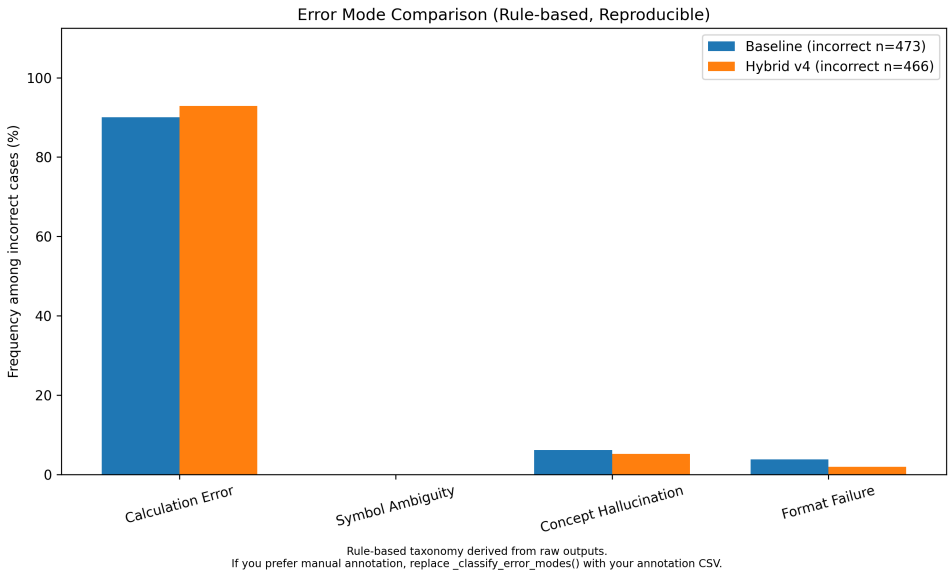

图 16: 得分累积分布函数（CDF）曲线

开放式评测进一步揭示了 MCQ 指标难以捕捉的关键现象：模型在生成层面具备较强的“结构化表达能力”，但这种能力并不等价于推理正确性。具体而言，在开放式实验中模型普遍能够生成较长的推理文本（平均约 12.5 步），然而 LLMJudge 的方差显著高于 AutoEvaluator，且两者相关系数较低 $r \approx 0.26$，说明“形式结构与内容正确性”在该任务中呈现明显的解耦关系——结构指标能够稳定刻画回答的组织性与形式完整度，却无法替代对物理与数学正确性的评估；反过来，基于 Judge 的正确性评分又更敏感于推导关键环节的细微错误与论证跳步。这一观察也为研究生基准中“阶段覆盖度几乎满分但准确率仍下降”的现象提供了统一解释：多阶段提示能够有效诱导模型产出更完整的分段叙述，但模型仍可能在关键推导点发生算符代数错误、边界条件遗漏或单位一致性破坏，从而导致最终结论不正确。换言之，过程可见性确实提高了可诊断性，但并不能自动转化为更高的结果正确率，因此在量子推理评测中将“结构质量信号”与“正确性信号”联合建模与评估是必要的。

从难度外推角度看，QuantumBench-Grad 相比原基准的整体准确率下降表明该构造在宏观上实现了“难度提升”的目标，但其统计稳健性仍受样本量与标注粒度限制。未来工作可沿着三个方向进一步增强该基准的诊断性与可解释性：其一，扩大更高难度

样本规模，并引入更细粒度的知识点与推理步标注，以支持更可靠的分层比较；其二，将评测从“仅评最终答案”推进到“分段可验证”，例如对中间结论设置可核验检查点，从而更直接地区分“推理链组织能力”与“关键结论正确性”；其三，构建面向量子推理的错误类型学，对常见失误（如算符代数、近似条件、单位与常数约定、表达式等价性判定失败等）进行系统归类，以形成更具可操作性的改进指引与更精细的能力画像。整体而言，这些结果共同指向一个结论：量子领域的 LLM 评测需要从单一 MCQ 正确率扩展到“可验证计算、结构化推理与开放式正确性”的多视角框架，才能更真实地刻画模型在高形式化科学任务中的能力边界与可靠性风险。

7 局限性与可复现性说明

尽管本文在 QuantumBench 复现、选择性符号增强与开放式评测方面给出了较为系统的实验结果，但仍存在若干重要局限。首先，为兼顾可复现性与成本控制，开放式评测中采用同一模型（Qwen2.5-7B）分别充当被测模型与 Judge，这一设置虽便于复核与快速迭代，却不可避免地引入潜在的“自评偏置”。已有研究表明，LLM-as-Judge 可能在表达风格、冗长偏好与一致性上存在结构性偏差，因此本文的开放式得分应主要被理解为可扩展、可对比的相对信号，而非等同于人工标注的绝对质量指标。未来工作可通过引入异源或更强模型作为外部 Judge，并在小规模样本上结合人工标注进行校准，以量化并缓解该类系统性误差。

其次，本文采用的“符号执行结果—选项匹配”策略具有较强工程假设，对一般形式的表达式等价、单位与量纲一致性及近似条件下的等价判定仍不够完备，可能在部分子领域产生误判或负迁移；因此相关结果更支持“工具增强需要选择性门控与误差控制”的方法论结论，而非宣称已彻底解决量子符号等价性问题。同时，子领域层面的显著性检验尚未进行严格的多重比较校正，相关发现应视为探索性结果。加之实验主要基于单一模型与单一硬件配置，其外推性仍有限。尽管本文通过固定随机过程与统一统计口径最大化了可复现性，但仍有必要在跨模型、跨硬件与跨推理框架的设置下进行进一步验证，以明确所提出框架的稳定性与适用边界。

8 结论

本文的核心发现并非在于提出一种性能更优的量子问题求解方法，而在于系统性表明：在高度形式化的量子科学任务中，单一多项选择准确率不足以刻画大语言模型的真实推理能力与可靠性边界。通过对工具增强、开放式推理评测与难度外推的联合分析，本文揭示了结构化推理行为，可验证计算与最终正确性之间的复杂关系，为量子领域 LLM 评测提供了一种从“结果导向”走向“过程与可靠性并重”的研究范式。我们在与原基准一致的 8-way 多选题管线下建立可比的基线参照，并引入三项互补扩展：其一，提出选择性符号增强的混合推理机制，在全量 769 题上将准确率从 38.49% 提升至

39.27%，并观察到增益在 Quantum Computation 子领域呈现更强信号，表明工具增强在量子任务中的效用具有显著的子分布依赖；其二，构建包含 165 题的开放式量子推理任务集，并提出结合自动指标与 LLM-as-Judge 的双重评测框架，在 30 题自由推导子集上获得 71.73/100 的综合得分，从而将推理过程质量纳入可规模化的量化评估；其三，构建 71 题研究生层次的 QuantumBench-Grad，通过多阶段结构化推理要求对模型进行难度外推测试，揭示模型在更高阶知识整合与关键推导可靠性方面仍存在明显不足。总体而言，本文结果表明，仅依赖多选正确率难以充分刻画量子领域 LLM 的真实能力边界；更具代表性的评测应同时覆盖最终答案正确性、计算可验证性、推理过程质量以及在更高难度任务上的外推表现，以避免结构化叙述能力与真实正确性之间的混淆。

面向未来，更完善的量子领域评测体系仍需在三个方向上继续推进：其一，在开放式评测中引入更强且异源的 Judge 或少量人工校准，以进一步提升正确性评分的可信度与可迁移性；其二，将评估从“结果对齐”推进到“分段可验证”，在关键中间结论处设置可核验检查点，使推理链的可靠性可以被更直接地量化；其三，扩展更大规模、更高难度且更贴近真实科研流程的任务形态（例如目标设定、程序化分解、验证与解释以及更系统的实验规划），从而使基准不仅衡量答题能力，也能反映模型在量子科学研究与工程开发场景中的实际适用性与可靠性边界。整体而言，本文提出的扩展框架为量子领域 LLM 的系统评估提供了更贴近任务本质的多视角工具，并为后续构建更接近真实科研工作流的量子智能体评测标准奠定了实验基础。

参考文献

[1] Achiam, J., Adler, S., Agarwal, S., et al. (2023). **GPT-4 Technical Report**. *arXiv preprint* arXiv:2303.08774.

[2] Brown, T. B., Mann, B., Ryder, N., et al. (2020). **Language Models are Few-Shot Learners**. *NeurIPS 2020*.

[3] Chen, W., Ma, X., Wang, X., et al. (2022). **Program-of-Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks**. *arXiv preprint* arXiv:2211.12588.

[4] Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). **Training Verifiers to Solve Math Word Problems**. *arXiv preprint* arXiv:2110.14168. (GSM8K)

[5] Dubey, A., Jauhri, A., Pandey, A., et al. (2024). **The Llama 3 Herd of Models**. *arXiv preprint* arXiv:2407.21783.

[6] Dupuis, C., Bhatia, A., Harkins, F., et al. (2024). **Qiskit Code Assistant: Using LLMs for Generating Quantum Computing Code**. *arXiv preprint* arXiv:2405.19495.

[7] Gao, L., Madaan, A., Zhou, S., et al. (2022). **PAL: Program-Aided Language Models**. *NeurIPS 2022 Workshop / arXiv* arXiv:2211.10435.

[8] Gemini Team. (2023). **Gemini: A Family of Highly Capable Multimodal Models**. *arXiv preprint* arXiv:2312.11805.

[9] Gou, Z., Shao, Z., Gong, Y., et al. (2024). **ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving**. *ICLR 2024 / arXiv* arXiv:2309.17452.

[10] Gu, J., Jiang, X., Shi, Z., et al. (2024). **A Survey on LLM-as-a-Judge**. *arXiv preprint* arXiv:2411.15594.

[11] Hendrycks, D., Burns, C., Basart, S., et al. (2021). **Measuring Massive Multitask Language Understanding**. *ICLR 2021 / arXiv* arXiv:2009.03300. (MMLU)

[12] Hendrycks, D., Burns, C., Kadavath, S., et al. (2021). **Measuring Mathematical Problem Solving With the MATH Dataset**. *NeurIPS 2021 / arXiv* arXiv:2103.03874.

[13] Ji, Z., Lee, N., Frieske, R., et al. (2023). **Survey of Hallucination in Natural Language Generation**. *ACM Computing Surveys / arXiv* arXiv:2202.03629.

[14] Kashani, H., Farimani, A. B., & others. (2024). **QuantumLLMInstruct: Instruction-Tuning LLMs for Quantum Tasks**. *arXiv preprint* arXiv:2407.10150.

[15] Kojima, T., Gu, S. S., Reid, M., et al. (2022). **Large Language Models are Zero-Shot Reasoners**. *NeurIPS 2022*.

[16] Li, M., Li, J., Tang, J., et al. (2023). **API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs**. *EMNLP 2023 / arXiv* arXiv:2304.08244.

[17] Liu, Y., Iter, D., Xu, Y., et al. (2023). **G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment**. *EMNLP 2023 / arXiv* arXiv:2303.16634.

[18] Manakul, P., Liusie, A., & Gales, M. (2023). **SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models**. *EMNLP 2023 / arXiv* arXiv:2303.08896.

[19] McNemar, Q. (1947). **Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages**. *Psychometrika*, 12, 153–157.

[20] Meurer, A., Smith, C. P., Paprocki, M., et al. (2017). **SymPy: Symbolic Computing in Python**. *PeerJ Computer Science*, 3, e103.

[21] Mikuriya, T., Hamamura, I., Ishigaki, T., et al. (2025). **QCoder: A Benchmark for Evaluating Quantum Code Generation with LLMs**. *arXiv preprint* arXiv:2510.26101.

[22] Minami, S., Ishigaki, T., Hamamura, I., et al. (2025). **QuantumBench: A Benchmark for Quantum Problem Solving**. *arXiv preprint* arXiv:2511.00092.

[23] Pan, F., Zhang, Y., Ebert, U., et al. (2024). **Large Language Models for Quantum Many-Body Physics Calculations**. *arXiv preprint* arXiv:2403.03154.

[24] Patil, S. G., Zhang, T., Wang, X., & Gonzalez, J. E. (2023). **Gorilla: Large Language Model Connected with Massive APIs**. *arXiv preprint* arXiv:2305.15334.

[25] Phan, L., Nadkarni, A., Xu, J., et al. (2025). **Humanity' s Last Exam**. *arXiv preprint* arXiv:2501.14249.

[26] Qin, Y., Liang, S., Ye, Y., et al. (2023). **ToolLLM: Facilitating Large Language Models to Master 16,000+ Real-world APIs**. *arXiv preprint* arXiv:2307.16789.

[27] Qwen Team. (2024). **Qwen2.5 Technical Report**. *arXiv preprint* arXiv:2412.15115.

[28] Rein, D., Hou, B. L., Stickland, A. C., et al. (2023). **GPQA: A Graduate-Level Google-Proof Q&A Benchmark**. *arXiv preprint* arXiv:2311.12022.

[29] Schick, T., Dwivedi-Yu, J., Dessì, R., et al. (2023). **Toolformer: Language Models Can Teach Themselves to Use Tools**. *arXiv preprint* arXiv:2302.04761.

[30] Srivastava, A., et al. (2022). **BIG-bench: Beyond the Imitation Game Benchmark**. *arXiv preprint* arXiv:2206.04615.

[31] Stephan, A., Zhu, D., Aßenmacher, M., et al. (2024). **From Calculation to Adjudication: Examining LLM Judges on Mathematical Reasoning Tasks**. *arXiv preprint* arXiv:2409.04168.

[32] Suzgun, M., et al. (2022). **Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them**. *arXiv preprint* arXiv:2210.09261. (BBH)

[33] Touvron, H., Lavril, T., Izacard, G., et al. (2023). **LLaMA: Open and Efficient Foundation Language Models**. *arXiv preprint* arXiv:2302.13971.

[34] Wang, X., Hu, Z., Lu, P., et al. (2023). **SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models**. *arXiv preprint* arXiv:2307.10635.

[35] Wang, X., et al. (2022). **Self-Consistency Improves Chain of Thought Reasoning in Language Models**. *arXiv preprint* arXiv:2203.11171.

[36] Wang, Y., et al. (2024). **MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark**. *NeurIPS 2024 / arXiv* arXiv:2406.01574.

[37] Wei, J., Wang, X., Schuurmans, D., et al. (2022). **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. *NeurIPS 2022 / arXiv* arXiv:2201.11903.

[38] Yao, S., Zhao, J., Yu, D., et al. (2022). **ReAct: Synergizing Reasoning and Acting in Language Models**. *arXiv preprint* arXiv:2210.03629.

[39] Ye, J., Wang, Y., Huang, Y., et al. (2024). **Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge**. *arXiv preprint* arXiv:2410.02736.

[40] Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena**. *arXiv preprint* arXiv:2306.05685.

[41] Zhu, Z., Goddard, C., Wang, Z., et al. (2025). **CritPt: A Physics Critique Benchmark for Evaluating LLM Physics Reasoning**. *arXiv preprint* arXiv:2502.01639.