班 级 2203011
学 号 22009200186

西安电子科技大学

本科毕业设计论文

题 目 基于推理式大语言模型的化学推理

学 院 计算机科学与技术学院

专 业 计算机科学与技术

学 生 姓 名 钱之浩

导 师 姓 名 李宇楠

西安电子科技大学
毕业设计（论文）诚信声明书

本人声明：本人所提交的毕业论文《 基于推理式大语言模型的化学推理 》是本人在指导教师指导下独立研究、写作的成果，论文中所引用他人的无论以何种方式发布的文字、研究成果，均在论文中加以说明；有关教师、同学和其他人员对本文的写作、修订提出过并为我在论文中加以采纳的意见、建议，均已在我的致谢辞中加以说明并深致谢意。

本论文和资料若有不实之处，本人承担一切相关责任。

论文作者：钱之浩（签字） 时间：2026 年 5 月 26 日

指导教师已阅：李宇楠（签字） 时间： 2026 年 5 月 26 日

摘要

随着大语言模型在科学发现中的深入应用，引导模型在特定自然科学数据流形上推理成为焦点。针对化学反应预测，现有的显式思维链（Explicit CoT）受限于高质量化学反应机理数据的稀缺，而隐式思维链（Latent CoT）仍处于起步阶段，业内对其在不同情景下的效能边界缺乏系统性研究 。

为此，本文系统探究了大语言模型在化学反应机理中的推理表现。基于化学反应规则，本文自动化构建了首个面向自然语言的高质量化学反应机理数据集 Flux-min，并以轻量级模型为基座，分别训练了显式语言推理 Flux-R 与隐空间推理模型 Flux-Z。多维度对比实验表明：在 Flux-min 数据集规模下，显式思维链凭借强语言约束，在准确率与可解释性上优势显著；而引入更大规模数据时，隐式思维链成功突破标注瓶颈，不仅在宏观产物预测上反超基准模型，更能以小于 5‰的模型参数量与小于 5%的词元（token）消耗量，达到万亿参数大模型约 80%的预测性能。本研究为化学领域提供了高效的轻量化模型特化方案，也为科学推理模型的架构选择提供了有价值的探索。

关键词：大语言模型 隐式思维链 显式思维链 化学反应预测 科学推理

Abstract

With the deepening application of large language models (LLMs) in scientific discovery, guiding these models to reason on specific natural science data manifolds has become a focus. For chemical reaction prediction, existing explicit chain-of-thought (Explicit CoT) is constrained by the scarcity of high-quality mechanism data, while latent chain-of-thought (Latent CoT) remains in its infancy, lacking systematic research regarding its performance boundaries across different scenarios.

To address this, this thesis systematically investigates the reasoning performance of LLMs in chemical reaction mechanisms. Based on chemical reaction rules, we automatically constructed Flux-min, the first high-quality, natural language-oriented mechanism dataset, and trained an explicit language reasoning model (Flux-R) and a latent space reasoning model (Flux-Z) using a lightweight model as the backbone. Multi-dimensional comparative experiments show that at the scale of the Flux-min dataset, explicit CoT exhibits significant advantages in accuracy and interpretability due to its strong linguistic constraints; however, when introduced to larger-scale data, latent CoT successfully breaks through the annotation bottleneck, not only surpassing the baseline model in macroscopic product prediction but also achieving approximately 80% of the predictive performance of trillion-parameter models with less than 5‰ of their model parameters and less than 5% of their token consumption. This study provides an efficient, lightweight model specialization solution for the chemical domain, and offers a valuable exploration for the architectural choice of scientific reasoning models.

Keywords: Large language model Latent chain-of-thought

Explicit chain-of-thought Chemical reaction prediction Scientific reasoning

目 录

第一章 绪论

1.1 研究背景和意义

近年来，大语言模型（Large Language Models, LLMs）在自然语言理解、代码生成与多步推理等任务上取得了显著进展，并逐步渗透至自然科学研究的多个领域[1]。在化学领域，模型被尝试用于反应产物预测、逆合成分析、分子性质评估以及实验路径规划等任务，呈现出“AI for Science”范式下连接通用智能与专业知识的潜在价值[2]。然而，与一般语言任务不同，化学推理要求模型在严格的物理化学规则约束下，沿着电子转移、键合形成与断裂等反应机理展开严密推断，这对模型推理过程的规则一致性、可解释性与可控性提出了远高于普通问答任务的要求。

思维链（Chain-of-Thought, CoT）[3]被认为是引导模型完成复杂推理的关键技术。其中，显式思维链以自然语言形式描述中间推理步骤，使模型决策在符号层面可读、可审计，但其有效性高度依赖于高质量的步骤化标注数据。在化学反应预测中，能够刻画电子推移机理的逐步标注极度稀缺：已有的大规模反应数据多以“反应物—产物”端到端形式呈现，缺少机理层面的中间过程；而专业反应机理库的人工标注成本高昂，规模十分有限。与此同时，隐空间思维链（Latent CoT）尝试将推理过程压缩进模型的内部连续表示，绕过自然语言标注瓶颈，但目前相关研究仍处于起步阶段，针对其在不同数据规模、不同隐空间扰动与不同推理深度下能力边界的系统性实证仍不充分。

在上述背景下，系统比较两类思维链在化学推理任务上的能力差异，对推动科学推理模型的架构设计具有重要意义。本研究面向化学反应预测任务，自动化构建了具有电子推移机理的自然语言数据集，并以轻量级模型为基座对显式与隐式思维链进行多维度对照实验。研究一方面为化学领域提供了高效的轻量化模型特化方案，另一方面也为科学推理模型在“语言可解释性”与“隐式泛化能力”之间的架构权衡提供了关键的实证依据。

1.2 国内外研究现状

1.2.1 显式思维链研究现状

显式思维链由 Wei 等人[3]首次系统提出。他们发现，在少样本提示中显式给出中间推理步骤，能显著提升模型在算术、常识与符号推理任务上的表现。Kojima 等人[4]进一步表明，即便不提供示例，仅以“Let's think step by step”作为零样本提示，足够规模的模型也能自发生成多步推理链。在此基础上，Wang 等人[5]提出自一致性解码，通过对多条推理路径采样并多数投票，显著提升答案稳健性；Yao 等人[6]提出的思维树（Tree of Thoughts）将链式结构扩展为可回溯的树状搜索。这些工作奠定了显式思维链作为通用推理增强机制的基础。

针对显式思维链对标注数据的强依赖，研究者尝试通过自生成与蒸馏的方式扩充训练资源。Zelikman 等人[7]提出的 STaR 方法利用模型自身生成的正确推理链回流再训练，迭代增强推理能力；Magister 等人[8]则将大模型生成的思维链作为监督信号蒸馏至小模型，使小规模学生模型也具备多步推理能力。近期，以 OpenAI o1 与 DeepSeek-R1 为代表的“长思维链”（Long CoT）模型进一步将强化学习引入显式推理训练，DeepSeek-R1 [25]仅依靠可验证的最终答案信号便驱动模型自发出现自我反思、回溯与验证等长链行为，在数学与代码推理上接近闭源前沿模型水平；Chen 等人[26]对该范式进行了系统综述，将长思维链与传统短思维链的差异归纳为深度推理、广泛探索与可行反思三个核心特征。上述方法在通用领域行之有效，但一旦迁移到强专业约束的科学场景，模型自生成的中间步骤往往难以满足化学规则的严密性，错误推理链反而会污染训练分布。总体而言，显式思维链的优势在于过程可读、规则可校验，但其训练成本受到高质量步骤化标注数据稀缺的根本制约——在化学反应机理这一对正确性要求极高的细分领域，这一矛盾尤为突出。

1.2.2 隐空间思维链研究现状

为缓解显式思维链对自然语言标注的依赖，研究者尝试将推理过程内嵌到模型的连续表示空间。Goyal 等人[9]提出在输入序列中插入可学习的“暂停符”（pause tokens），使模型在生成最终答案前获得额外的隐式计算步数，从而提升推理质量。Pfau 等人[10]系统性地证明，仅依赖与具体任务无关的填充符号，

Transformer 同样可以利用其残差通路完成隐式中间计算，提示推理能力可以在一定程度上脱离自然语言显式表达而存在。

在更直接的隐式建模路线上，Deng 等人[11]通过知识蒸馏将显式思维链中的中间步骤压缩进模型内部状态，使学生模型无需输出推理过程即可给出答案，并正式提出了“隐式思维链”这一概念。Hao 等人[12]提出的 Coconut 框架进一步将上一步隐藏表示循环回送作为下一步输入，使模型在连续隐空间中执行多步推理而无需经过文本解码，在数学与逻辑任务中取得了与显式思维链相当甚至更优的性能。沿此方向，Geiping 等人[27]提出基于循环深度的隐空间推理架构，通过反复迭代同一组循环模块来动态延展推理路径，在测试时仅靠加深而非扩宽模型即可提升复杂推理能力，为“测试时计算扩展”（test-time scaling）提供了一条与显式长思维链互补的路径。Zhu 等人[28]的最新综述将上述工作统一刻画为“在隐空间中以连续表示进行多步推理”的研究范式，并系统讨论了其在表达带宽、计算效率与对齐风险等维度的权衡。

尽管隐式思维链（Latent CoT）在计算效率和规避文本标注成本上展现出巨大潜力，但其在严密科学推理中的可靠性仍面临严峻挑战。由于缺乏自然语言离散符号的强约束，模型在连续隐空间中进行多步无监督优化时，倾向于寻找仅服务于最终预测的“表征捷径（Representation Shortcut）”。这导致其生成的中间隐式特征逐渐脱离了预训练时的自然语言分布，产生严重的“隐式流形偏移”（Latent Manifold Shift）现象。这种偏移不仅使得通过 VAE 等解码器还原的文本呈现无意义的退化状态，彻底丧失了反应机理的可解释性，也成为了制约隐空间推理进一步对齐人类科学逻辑的核心瓶颈。

围绕隐空间推理，已有工作初步讨论了循环深度对推理能力的影响、表示路径上的扰动与正则化策略，以及与显式监督的混合训练方式。然而，相关结论多在通用推理基准上获得，针对自然科学数据流形上的系统性证据仍较为有限：在数据规模变化、隐空间路径噪声以及推理步数等维度上，隐式思维链的能力边界与适用场景尚缺乏一致的实证刻画。换言之，隐空间思维链的方法工具已初步成型，但其理论与实证研究仍处于早期阶段。

1.2.3 化学推理思维链的研究现状

在化学领域，端到端反应预测的代表性工作可追溯至 Schwaller 等人[13]提出的 Molecular Transformer，该模型将反应预测建模为 SMILES 字符串之间的序列翻译任务，并在多个公开数据集上取得了优异的产物预测精度。Irwin 等人[14]提出的 Chemformer 通过大规模化学语料上的自监督预训练，进一步提升了反应预测与逆合成等下游任务的表现。这类工作虽然在性能上颇为出色，但其推理过程被压缩为黑箱式的端到端映射，缺少对反应机理的显式刻画，难以满足化学研究对推理可解释性的要求。

在通用大模型与化学结合的方向上，M. Bran 等人[15]提出的 ChemCrow 借助工具调用与思维链提示，使大模型能够规划合成路径并辅助实验设计；Taylor 等人[16]发布的 Galactica 则在大规模科学语料上进行预训练，展示了通用模型处理化学符号系统的潜力。在面向机理的可解释推理层面，Tavakoli 等人[17]构建的 RMechDB 较早提供了基元反应步骤的标注资源，但其规模相对有限，且推理步骤主要以模板化的箭头推移与原子映射形式表示，并非自然语言描述。近期出现的 FlowER[18] 与 ChemCoTBench[19] 等工作在反应流与化学思维链评测方向上做出了进一步尝试，但其中间步骤同样以结构化或半结构化形式呈现，缺少以自然语言完整表达电子推移机理的大规模训练资源。

1.3 本文主要研究内容

综合来看，现有大语言模型在端到端的化学反应预测层面已取得一定进展，但在面向机理的逻辑推理方向仍存在明显空白：一方面，自然语言形式的电子推移机理标注极度稀缺，严重制约了显式思维链（Explicit CoT）的学习上限；另一方面，基于连续向量的隐空间思维链（Latent CoT）在化学等具有强规则约束的科学场景中的表现尚未得到充分验证，其与显式推理的性能边界及内在权衡仍缺乏系统的实证比较。本研究正是在这一空白处展开。

为解决上述问题，本文系统性地探究了大语言模型在化学反应数据流形上的推理范式。首先，利用化学信息学方法自动化构建了覆盖约 27 万条反应轨迹，160 万条反应单步的化学反应机理数据集。在此基础上，以 0.6B 轻量级模型为

基座，分别设计并实现了显式语言推理模型（Flux-R）与隐空间固定长度推理模型（Flux-Z）。为确保实验结论的严谨性，本文开展了广泛而深入的实证研究，不仅在小模型内部进行了超过 25 组不同配置与权重的消融实验，还引入了千亿参数级别的先进通用大模型（如 DeepSeek-V4-Pro、Qwen3.6-35B）进行跨量级的对比评测。

本研究的意义主要体现在以下几个方面：

（1）构建了稀缺的反应机理语料库： 本研究精心构建并清洗的自然语言电子推移数据集，是首个以自然语言形式构建的化学反应中间体推理路径的数据集，并且是以白盒形式构筑，有效填补了面向化学反应机理推理的可用、可信语料空白，为后续科学大模型学习化学底层逻辑提供了宝贵的数据基础。

（2）系统揭示了不同推理范式的效能边界： 本研究在科学推理场景下，对“显式自然语言约束”与“隐空间连续计算”两种思维链范式进行了深度的对比分析。客观指出了隐式思维链在突破标注瓶颈后的高性能，以及其面临的流形偏移等问题，为科学大模型底层架构的选择提供了关键的理论与实证支撑。

（3）提供了高性价比的轻量化模型特化实践： 本文通过严谨的基准测试证明，经过领域数据对齐与特定架构（如隐空间机制）加持的小参数模型（0.6B），能够在垂直推理任务上逼近甚至媲美千亿参数通用大模型的性能。这为受限算力条件下的“AI for Science”研究提供了一种兼具实用价值与应用前景的落地思路。

1.4 本文章节安排

本文围绕大语言模型在化学反应机理推导中的显式思维链与隐式思维链推理范式展开研究。全文共分为五个章节，具体结构安排如下：

第一章，绪论。 该章主要介绍本课题的研究背景与现实意义，系统梳理了近年来国内外在显式思维链、隐空间推理以及大模型驱动的化学反应预测等领域的研究现状。最后，概括了本文的主要研究内容、创新点及全文的组织结构。

第二章，相关理论简介。 该章对本文研究所依托的基础理论进行了简要概述，主要包括数据流形假设、大语言模型的基本推理机制（显式与隐式思维链的原理）、以及化学反应机理基础（如化学分子表示法、反应拓扑与电子推移机理等），为后续的数据集构建与模型设计提供理论支撑。

第三章，反应机理数据集构建与推理模型设计。 本章是本文方法实现的核心。首先，详细阐述了如何基于化学物理规则与拓扑学分析，自动化构建涵盖电子转移步骤的高质量化学推理数据集；随后，基于该数据集，详细介绍了本文所提出的显式语言推理模型（Flux-R）与隐空间固定长度推理模型（Flux-Z）的具体网络架构、计算流程与训练策略。

第四章，实验设计与结果分析。 本章对所提出的模型进行了充分的评测与论证。首先介绍了实验所采用的基准测试集、评价指标以及作为对比框架的先进大模型（如 DeepSeek-V4-Pro 等）；随后，通过多维度消融实验，深入剖析了两种推理模型在不同数据规模下的宏观预测准确率与机理解释性；最后，着重探讨了隐空间推理在实验中展现出的计算效率优势以及伴随的“隐式流形偏移”现象。

第五章，总结与展望。 该章对全文的核心工作和实验结论进行了概括性总结，客观分析了本研究中存在的局限性，并对未来科学推理大模型的架构发展和优化方向进行了展望。

第二章 相关理论简介

为便于后文本研究进行介绍，本章将对大语言模型执行推理任务的基本推理机制（包括数据流形假设、显式思维链、隐式思维链），以及本研究关注的自然科学领域的化学反应机理中的相关理论进行介绍。

2.1 数据流形假设

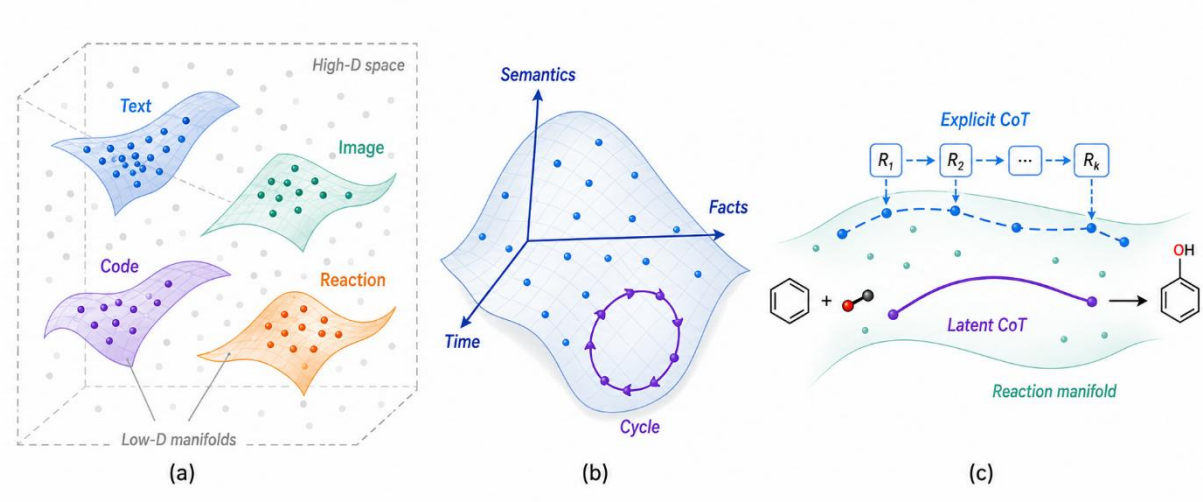


图 2.1 数据流形假设

机器学习中的“数据流形假设”认为：自然语言(text)、图像(image)、化学反应(reaction)、代码(code)等真实世界数据虽然通常以高维向量或离散序列形式表示，但其有效样本并非均匀填充整个高维空间，而是集中在低内在维度、具有局部连续结构的子空间或多个子流形上，如图 1(a)所示。近年来，这一观点仍是深度学习和生成模型解释的重要基础。Loaiza-Ganem 等[21]在 TMLR 综述中指出，现代深度生成模型的成功与其学习受未知低维流形支撑的数据分布密切相关；Brown 等[20]也在 ICLR 2023 中对常用图像数据集进行了实证验证，发现真实数据更符合“多个子流形联合”的结构，而非单一均匀流形。

在大语言模型中，数据流形假设可进一步理解为：预训练模型并非仅记忆表层 token 共现，而是在隐藏状态空间中形成了对语义、事实、时空关系和任务结构的连续表示。已有实证研究表明，LLM 的激活空间中存在可解释的几何结构：例如 Gurnee 与 Tegmark [22]发现语言模型能够在不同尺度上以近似线性方式表示空间和时间信息；Marks 与 Tegmark [23]发现真/假事实陈述在大模型内部表示中呈现可迁移且具有因果作用的线性结构；Engels 等[24]进一步指出，部分语言模型

特征并非一维线性方向，而可能表现为星期、月份等具有周期性的低维多维结构，如图 1(b)所示。这些结果共同说明，大模型的内部推理并不完全依赖显式文本步骤，而可以在连续隐空间中沿着具有语义和任务含义的几何结构进行计算。

因此，本文将“数据流形假设”作为理解化学推理大模型的理论起点：化学反应样本不仅具有自然语言描述层面的分布规律，还受到反应物结构、官能团转换、电子推移方向和反应条件等物理化学约束，其有效推理路径应落在相对结构化的化学反应数据流形附近。如图 1(c)所示，显式思维链可以看作使用自然语言符号对该流形上的推理轨迹进行约束和展开；隐空间思维链则尝试在连续表示中直接压缩和推进这一轨迹。后文对 Flux-R 与 Flux-Z 的比较，正是围绕“语言可解释性约束”与“隐空间流形泛化能力”之间的差异展开。

2.2 推理式大语言模型

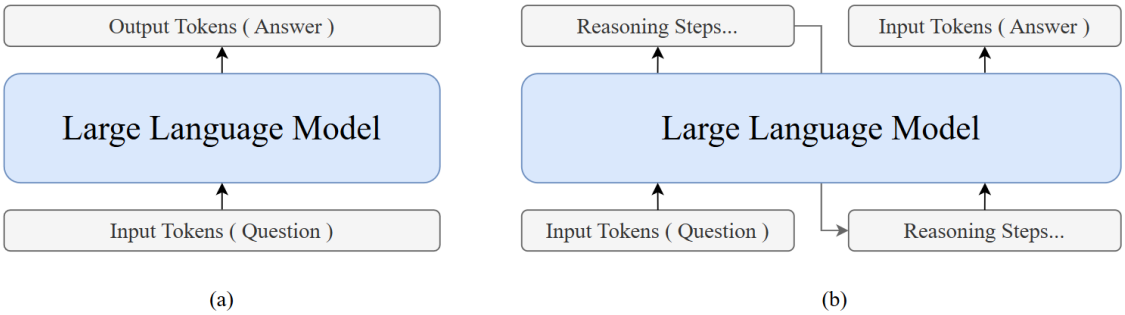


图 2.2 非推理式大语言模型与推理式大语言模型对比

如图 2.2(a) 所示，早期的大语言模型遵循“输入即问题、输出即答案”的端到端范式：模型直接对条件概率 $p(y \mid x)$ 进行建模，在输入文本 x 的最后位置一次性回归出答案文本 y，整个推理过程没有可观测的中间过程。这种结构在通用问答与文本生成上简洁高效，却难以应付需要分阶段计算或长程依赖的复杂任务。

如图 2.2(b) 所示，本文将输入与最终输出之间存在若干显性或隐性中间状态的大语言模型称为推理式大语言模型：其建模目标变为由一系列中间步骤 z 所诱导的联合分布 $p(z, y \mid x)$，使整体推理过程可被分解为若干较易学习的子计算。Feng 等人[36]在 NeurIPS 2023 的工作从电路复杂度视角证明，固定深度的 Transformer 在不引入中间步骤时无法以多项式规模直接求解一类基本算术与等式约束问题，但在允许其生成中间推理链后，常数规模的 Transformer 即可对该

类问题以及更一般的动态规划任务给出正确解；Merrill 与 Sabharwal[37]进一步在 ICLR 2024 中刻画思维链对 Transformer 表达能力的扩张：线性步数的中间生成足以使解码器识别全部正则语言，多项式步数则使其精确对应于 P 类可解问题。这些结果共同表明，引入中间推理步骤本质上等价于对原问题进行分治式分解，使其在计算复杂度等级上从“浅层不可解”迁移到“可解”这一更宽阔的复杂度类。

中间推理步骤的具体载体存在多种选择：可以是离散的自然语言符号，也可以是模型内部的连续隐藏表示，甚至可以是结构化的中间结果（如程序、表达式树）。本文聚焦其中最具代表性的两类——显式思维链（Explicit CoT）与隐式思维链（Latent CoT），并将在 2.3、2.4 节分别给出形式化定义与原理解析。

2.3 显式思维链

显式思维链（Explicit Chain-of-Thought, Explicit CoT）是指模型在生成最终答案之前，以离散的自然语言符号显式地输出一系列中间推理步骤的范式。形式化地，给定输入 x 与目标答案 y，标准的“输入—答案”映射可以建模为条件概率 $p(y \mid x)$；而显式思维链则将该过程分解为可观测的中间序列 $z=(z_1, z_2, \ldots, z_t)$ 使联合分布写为：

$$p(y \mid x)=\sum_{z} p(z \mid x) p(y \mid x, z) \approx p(y \mid x, \hat{z}), \hat{z} \sim p_{\theta}(z \mid x) \tag{2-1}$$

其中 $\hat{z}$ 为模型自回归采样得到的语言推理链。该建模将最终答案与若干可观测的中间步骤显式耦合，使决策依据在符号层面对人类可读、可校验，构成了显式思维链最核心的归纳偏置。

为提升推理结果的稳健性，研究者在解码层面进一步引入多路径推理与聚合机制：对同一输入采样若干条思维链，再通过多数投票或启发式搜索从多分支推理空间中筛选最优解。这些方法的共同思想可形式化为对答案后验的近似：

$$\hat{y}=\arg\max_{y} \sum_{i=1}^{N} 1\{y=f(z^{(i)})\}, z^{(i)} \sim p_{\theta}(z \mid x) \tag{2-2}$$

即通过 N 条独立采样的推理路径估计答案的边缘分布。

在训练侧，显式思维链的有效性高度依赖含完整中间步骤的标注数据：无论是教师—学生蒸馏还是基于自生成正确链的回流再训练，都无法摆脱对正确推理链作为监督信号的需求。

总体而言，显式思维链的核心优势在于其推理过程以自然语言表达，过程可读、规则可校验，便于与领域知识进行对齐与审计；其根本局限则在于性能强烈依赖高质量步骤化标注数据，且推理时需要完整解码长串符号，时间与显存开销较大。在化学反应机理这类对正确性要求极高、机理标注极其稀缺的细分领域，这一矛盾尤为突出，也直接构成了本文进一步研究隐式思维链的动机。

2.4 隐式思维链

隐式思维链（Latent / Implicit Chain-of-Thought）是相对于显式思维链而提出的另一种推理范式，其核心思想是将原本以自然语言离散符号承载的中间推理过程，压缩进模型的连续隐藏表示空间内执行，从而绕过对中间文本标注的依赖。形式化地，记模型在第 t 步的隐空间向量为 $z_t \in \mathbb{R}^d$，则隐空间推理可统一抽象为如下迭代更新：

$$z_{t+1} = f_\theta(z_t, x), t = 0,1,\ldots,T-1; \hat{y} = g_\varphi(z_T) \tag{2-3}$$

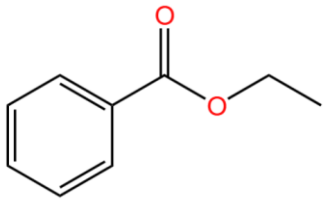
其中 f_θ 为共享参数的隐空间变换算子，T 为推理深度，g_φ 为最终的解码头。与显式思维链相比，整个推理过程不再经由词表 V 的离散采样 $w_t \in V$，而是直接在低维稠密流形上传播。

在最直接的连续隐空间建模中，整个推理过程不再经由词表 V 的离散采样 $w_t \in V$，而是由共享参数的隐空间变换算子 f_θ 在低维稠密表示上递进展开。隐空间向量在每一步既承载上文累计的信息，也作为下一步推理的“虚拟词元”输入到模型，使得多步思考在不进行任何文本解码的前提下完成。形式化地，第 t 步的隐空间状态 z_t 通过共享算子作用于 (z_{t-1}, x) 得到，并最终由解码头 g_φ 投影回离散答案空间，从而绕过对中间文本标注的依赖。

可以看出，隐式思维链以连续表示替代离散符号，在突破自然语言标注瓶颈和提升计算效率上展现出明显潜力；但其代价是推理过程不再具备直接可读的自然语言形式，带来流形偏移与可解释性下降等新问题。围绕这一权衡，本文将在第三章中分别构建显式语言推理模型 Flux-R 与隐空间固定长度推理模型 Flux-Z，并在第四章对两类范式在不同数据规模下的能力边界进行系统的实证比较。

2.5 化学反应机理

化学反应机理推理是本研究面向的目标任务。要让大语言模型在该任务上展开严密推理，须先回答两个基础问题：分子如何在计算机中被符号化表示，以及反应过程如何由自然语言所刻画。本节先简要介绍当前主流的分子表示方法，再阐述化学反应、基元反应与反应机理之间的层次关系，并说明本研究为何坚持以自然语言形式描述反应机理。



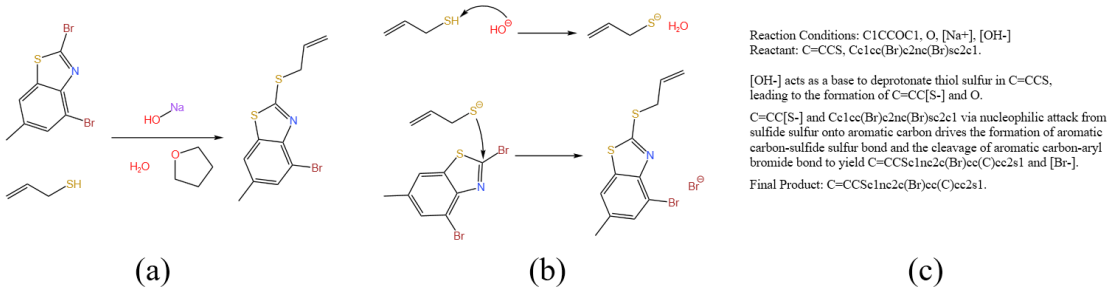
(a)

CCOC(=O)C1=CC=CC=C1

(b)

图 2.3 化学分子键线式与计算机 SMILES 表示

如图 2.3 所示，化学家通过键线式（skeletal formula）以二维线条直观刻画分子的拓扑骨架，其语义清晰却不便被计算机直接读取。为此，文献中常用三类机器可读的分子表示方法：（1）SMILES（Simplified Molecular Input Line Entry System）以一维 ASCII 字符串编码原子、键级与立体信息，凭借高压缩率与文本接口的天然兼容性，已成为大语言模型处理分子的主流输入格式；（2）SMARTS 在 SMILES 基础上引入通配符与子结构匹配语义，被广泛用于反应模板、子结构检索与机理位点描述；（3）分子图（Molecular Graph）将原子视作顶点、键视作边，便于图神经网络等模型直接对拓扑结构进行学习。三种表示各有侧重，但本质上都是同一分子流形在不同维度上的投影，并可通过 RDKit 等工具相互转换。



(a) (b) (c)

图 2.4 化学反应、化学反应机理和反应机理描述

一个完整的化学反应通常由反应物（reactant）、反应条件（reaction condition）与产物（product）三部分构成，如图 2.4(a) 所示。

然而，单步反应只是宏观层面的概括。从微观尺度看，多数典型反应可沿电子推移路径分解为若干基元反应（elementary reaction）的链式组合，如图 2.4(b) 所示，这一链式组合即称为反应机理（reaction mechanism）。相比单步反应，反应机理具有更高的逻辑严密性与因果可解释性：化学专家可在实验之前依据已知机理规律推断中间体走向并预测产物，这也为大语言模型沿机理思维链进行产物预测提供了客观的事实依据。若仅以 SMILES 等结构化字符串表示反应机理，模型容易陷入对符号的字面记忆而难以建立电子推移层面的语义理解。要让大语言模型真正利用机理知识进行推理，需将每一基元步骤所涉及的亲核位点、亲电位点、电子流方向与键变化以自然语言显式表达。然而，目前学界尚缺乏以自然语言完整刻画电子推移机理的大规模可用语料，这恰是本研究第三章着力填补的关键缺口。如图 2.4(c) 所示，本研究将每一基元步骤渲染为一句以电子流为主语的英文叙述，使反应物—产物之间的因果关系在文本层面显式可见。

2.6 本章小结

本章对支撑后续工作的理论基础进行了简要梳理。首先，从数据流形假设出发，说明了真实数据集中分布于低维子流形附近的客观规律。其次，给出了显式思维链与隐式思维链的形式化定义与计算特征，明确了二者在可解释性与泛化能力之间的权衡关系。最后，介绍了化学反应机理的层次结构与常用分子表示方法，指出现有结构化语料无法直接支持机理级思维链训练的根本不足。上述理论与背景共同构成了第三章数据集构建与模型设计的出发点。

第三章 反应机理数据集构建与推理模型设计

3.1 反应机理数据集构建

3.1.1 高质量反应机理训练数据集 Flux-min

要让大语言模型沿反应机理进行推理，训练语料须以自然语言形式描述每一基元步骤。然而已有公开资源大多止步于机器可读的化学符号：FlowER[18] 提供了原子映射完备、电子守恒的多步基元反应轨迹，但每一步均以 SMILES/SMARTS 字符串表示；ORDerly[29] 则汇聚了大规模反应物—产物端到端记录，缺少机理中间体。直接将上述语料用于训练易使模型陷入对符号的字面记忆，难以建立对电子转移过程的语义理解。本节针对这一缺口，设计了一条将带原子映射的基元反应轨迹自动转写为自然语言电子推移描述的处理管线，整体流程如图 3.1 所示，包括上下文构建、拓扑差异提取、机理决策与模板渲染四个阶段。

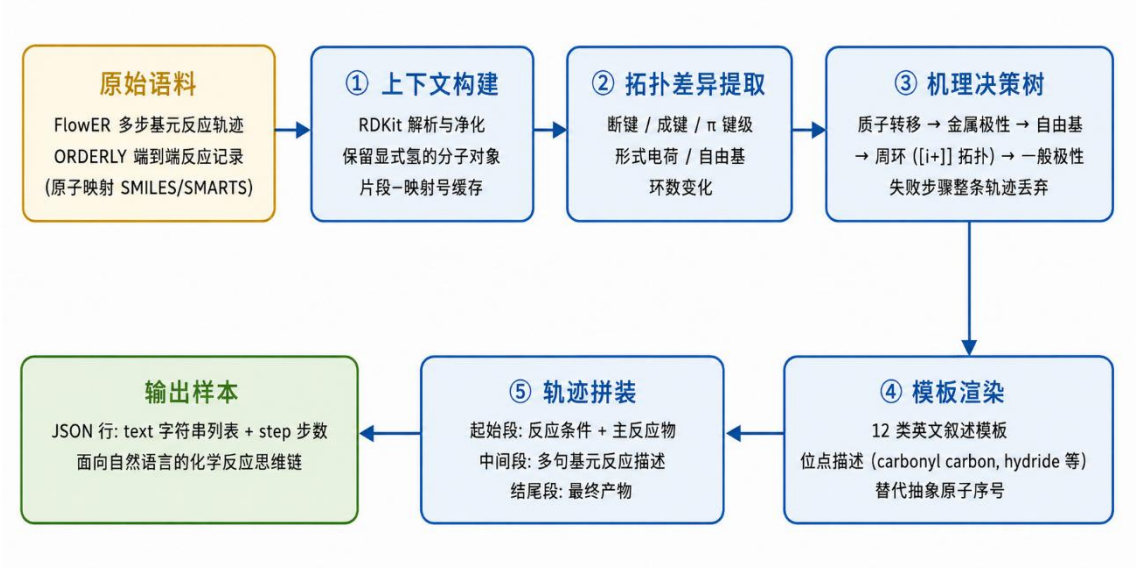


图 3.1 反应机理自然语言数据集构建管线

上下文构建以 RDKit[30] 完成分子加载与净化，对反应物与产物分别建立保留显式氢的分子对象，并按原子映射号缓存原子、片段与映射查询表；任何在解析或净化中报错的步骤即时丢弃，从源头排除原子映射错误或化学计量异常的样本。

拓扑差异提取面向同一原子映射号在反应前后的状态变化进行细粒度比对，统一刻画为断键集合、成键集合、π 键级变化、形式电荷变化、自由基电子数变化与环数变化六类信号，为后续机理判定提供可枚举的判别依据。机理决策按照先严格、后兜底的优先级在差异信号上展开决策树，依次判定严格质子转移、金属中心参与的极性反应、含未配对电子的自由基反应（涵盖偶联、加成、夺取、β-断裂、均裂与双自由基分解六种模式）、不涉及电荷变化的周环反应（涵盖环加成、电环化、σ-迁移、螯合及其逆过程，并在最短路径上计算 [i+j] 拓扑标记）以及一般含电荷变化的极性反应；对一般极性反应进一步依据形式电荷增量识别亲核位点与亲电位点，结合断键与成键事件生成结构化电子流描述。任何无法被任一分支接受的步骤被标记为未分类，整条轨迹直接丢弃。

表 3.1 反应机理自然语言数据样例

字段	内容
step	3
text[0]	Reaction Conditions: CCOC(C)=O, CI, CN(C)C=O, [H-], [Na+] Reactant: Cn1c(-c2ccccc2Cl)nnc1C1(C(=O)Nc2ccc(Cl)cc2)CCC1.
text[1]	Cn1c(...)CCC1 and [H-] via nucleophilic attack from hydride onto amine proton drives the formation of amine proton-hydride bond and the cleavage of amide nitrogen-amine proton bond to yield Cn1c(...)([N-]...)CCC1 and [H][H].
text[2]	CI and Cn1c(...)([N-]...)CCC1 via nucleophilic attack from anionic nitrogen onto primary carbon drives the formation of anionic nitrogen-primary carbon bond and the cleavage of alkyl iodide-primary carbon bond to yield CN(C(=O)C1(...))c1ccc(Cl)cc1 and [I-].
text[3]	Final Product: CN(C(=O)C1(c2nnc(-c3ccccc3Cl)n2C)CCC1)c1ccc(Cl)cc1.

模板渲染将上一阶段得到的机理类别与机理变量代入预先编制的英文叙述模板，统一采用单句、电子流主导的叙事方式，并以基于原子环境的位点描述（如 carbonyl carbon、hydride、Mg metal center 等）替代抽象原子序号，使输出文本既保留反应物—产物的 SMILES 锚点，又显式呈现电子推移方向与键变化的因果关系。一条完整轨迹的输出由三段拼装而成：起始段以集合形式列出反应条件与主反应物（不含碳的物种归入条件，含碳物种归入反应物）；中间段为按时间顺序排列的多句基元反应描述；结尾段在剔除条件与无机片段后给出最终产物。最终样本以 JSON 行格式存储，字段 text 为字符串列表，字段 step 为基元反应步数，便于按机理深度分桶训练。一条三步样本的形态如表 3.1 所示。

至此，本节构建的反应机理自然语言数据集 Flux-min，是据本文调研所知首个以自然语言完整刻画电子推移机理的大规模化学反应思维链语料，为第 3.2、3.3 节中显式与隐式思维链模型的训练提供了统一的监督信号基础；其规模、长度分布与模板分布的具体统计将在第 4.3 节中给出。

3.1.2 高质量反应机理基准测试

与传统机器学习模型不同，大语言模型的推理过程为完整的自回归解码，单次评测往往涉及上千乃至上万 token 的连续生成，整体计算与显存开销随采样次数线性放大；因此其基准测试不再像传统监督学习那样动辄使用上万条样本，而需以“小样本、多采样、高保真”的方式精心组织。近年面向大模型科学推理能力的若干权威基准也清晰地体现了这一规律：Rein 等人[31]提出的 GPQA 在生物、物理、化学三大学科中由 PhD 级专家命题，全集仅 448 道研究生水平选择题，却已成为评估前沿大模型科学推理能力的事实标准；Wang 等人[32]提出的 SciBench 收集自大学物理、化学与数学的标准教材，涵盖 695 道开放式数值推理题，在 ICML 2024 上被广泛用作大模型科学解题能力的对比基线；在化学专业领域，Mirza 等人[33]构建的 ChemBench 由化学家设计，其面向多种 LLM 的人机对照子集亦控制在数千题量级，使评测可在专业人员可承受的工作量内完成。基于此原则，本节面向化学反应机理推理任务专门构建了 Flux-bench 基准测试集，覆盖端到端产物预测、机理思维链生成与对抗背诵三类核心能力。

Flux-bench 共包含三个相互独立的子集：ORDerly-300、ORDerly-MC 与 Fukuyama-A。其中 ORDerly-300 与 ORDerly-MC 由 ORDerly[29] 数据库经清洗后抽取，并通过严格的产物去重确保所有测试反应的目标产物均不在训练集 Flux-min 与 Flux-max 中出现，从源头杜绝数据泄露并避免对本研究模型性能的高估；对照地，第 4.7 节使用的开源大模型由于其训练语料不可控，无法保证未接触过原始数据，因此对其能力反而是不被低估的。Fukuyama-A 子集则由福山机理题练习册 A 组中精选 10 条多步反应轨迹组成，其难度不低于本科有机化学专业的反应机理课程要求；本研究还引入具备化学专业知识的人工标注以校核机理推理的逐步正确性，从而显著提升了机理评测的可靠性。三个子集的具体规模与评测目的将在第 4.3 节统一给出。

3.2 显式思维链推理大语言模型

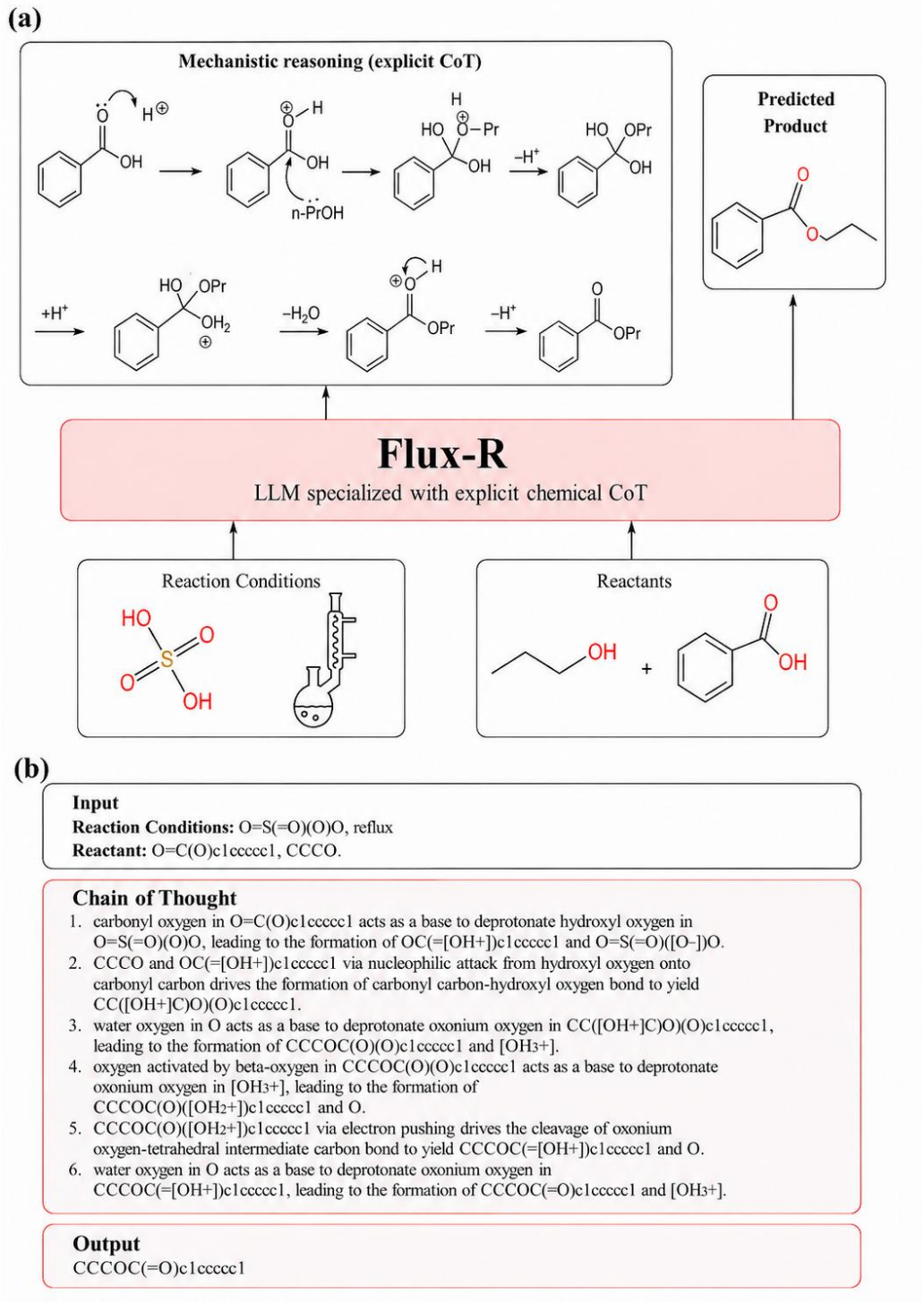


图 3.2 显式思维链推理范式

如图 3.2(a) 所示，本文设计的 Flux-R（针对显式化学思维链优化的 LLM）的整体工作流如下：模型首先接收特定的反应条件（如图中的硫酸催化剂与回流装置）以及反应物结构（如苯甲酸和正丙醇）作为输入。在输出预测产物之前，

Flux-R 会执行化学反应机理推理的中间过程。图示中表现为一系列按时间顺序发生的基元反应机理，展示了电子推移、质子化、亲核进攻以及脱水等中间化学态的转化细节。这种直观的物理化学图像在模型中具体的文本交互例子示意如图 3.2(b) 所示。

表 3.2 显式思维链推理模型 Flux-R 的具体实现步骤

模型输入：反应条件文本 x_c 与反应物文本 x_r，分词器 T，参数化模型 p_θ

模型输出：预测产物序列 $\hat{y}$ 及完整的显式机理推理链 $\hat{w}$

过程：

1. 将输入 $x=[x_c;x_r]$ 经分词器 T 切分为词元序列 $t=(t_1,\ldots,t_L)$；
2. 经嵌入层得到嵌入向量序列 $E=(e_1,\ldots,e_L)$，其中 $e_i=\mathrm{Embed}(t_i)$；
3. 在 t 之后自回归地预测下一个词元，对每一步 k 计算条件分布 $p_\theta(\cdot\mid t_{<L+k})$，并按 softmax 采样得到 t_{L+k}；
4. 重复步骤 3，逐步生成中间机理叙述 $\hat{w}=(\hat{w}_1,\ldots,\hat{w}_T)$，每条 $\hat{w}_i$ 描述一基元反应的电子推移过程；
5. 当解码至特殊段 “Final Product:” 时，模型继续解码出最终产物的 SMILES 序列 $\hat{y}$；
6. 训练阶段，对全序列施加交叉熵损失 $L=-\sum_i \log p_\theta(t_i\mid t_{<i})$ 进行参数更新；
7. 返回产物预测 $\hat{y}$ 与完整推理链 $\hat{w}$。

结合上述图示，更进一步的精细化推理过程可描述如下：如表 3.1 中 text[0] 所示，反应条件与反应物均以自然语言形式给出，经分词器（tokenizer）切分为词元（token）后，由嵌入层映射为稠密向量并送入模型；模型在自回归解码过程中，依据已生成的上下文逐步预测下一个词元，并在每一步以词表分布上的 softmax 采样选择具体输出，由此自然地涌现出按时间顺序排列的中间机理叙述，构成完整的显式思维链；当模型生成至 “Final Product:” 段时切换为产物 SMILES 的解码，最终以自然语言形式输出预测产物。为便于后续讨论，本节将该流程整理为算法 1，详见表 3.2。

3.3 隐式思维链推理大语言模型

3.3.1 概述

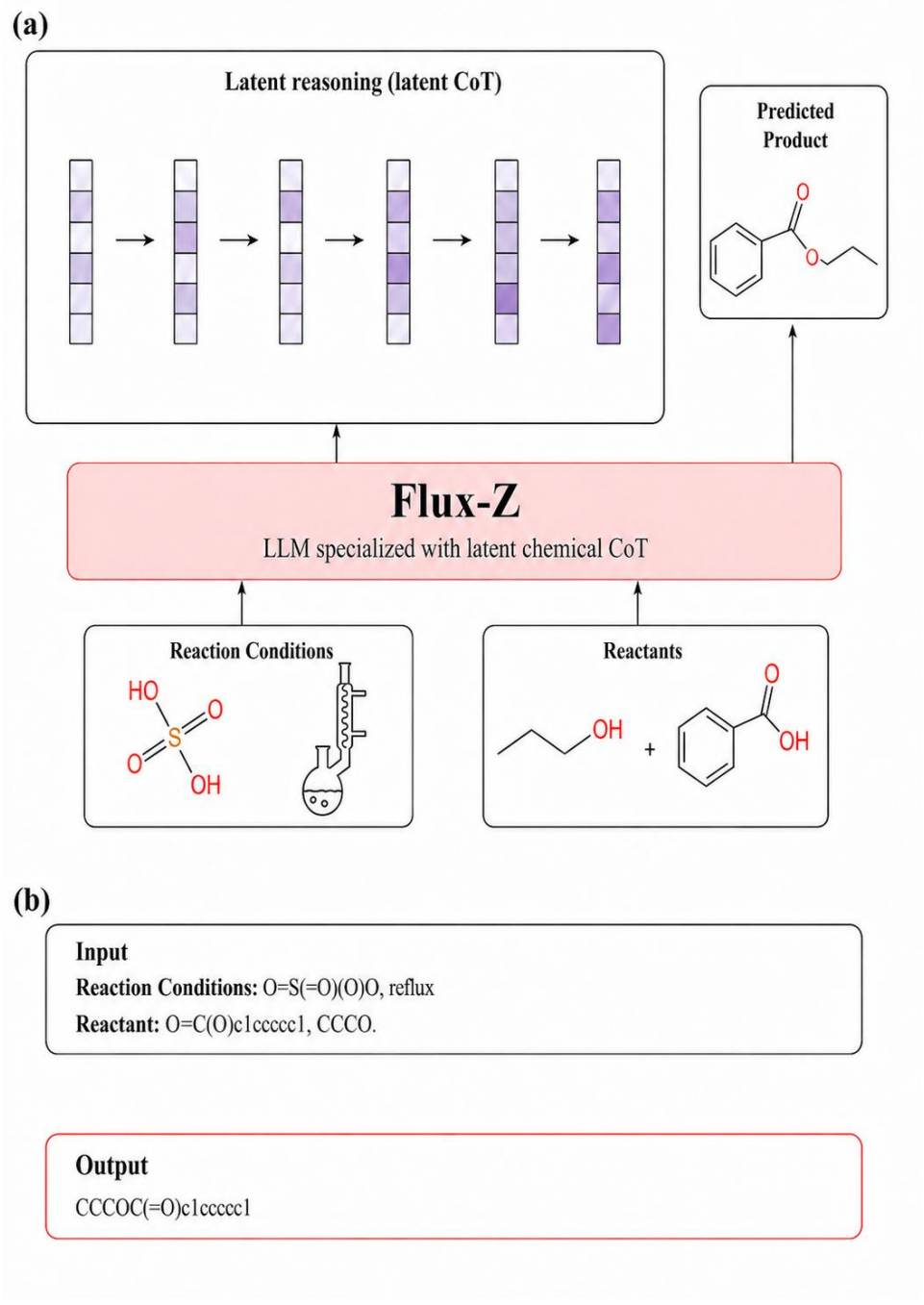


图 3.3 隐式思维链推理范式

如图 3.2(a) 所示，本文设计的 Flux-Z（针对隐式化学思维链优化的 LLM）的整体工作流如下：模型同样首先接收特定的反应条件（如图中的硫酸催化剂与回流装置）以及反应物结构（如正丙醇与苯甲酸）作为输入。在输出预测产物之

前，与显式生成中间机理文本的范式不同，Flux-Z 会在内部执行隐式化学推理过程（Latent reasoning 或 latent CoT）。图示中表现为一系列按顺序排布的隐状态向量块（以深浅不同的紫色方块表示），展示了模型在连续隐空间内进行的隐式思考过程，而无需将其解码为具体的物理化学文本。这种跳过中间自然语言输出的直接交互范式，在模型中具体的文本交互例子示意如图 3.2(b) 所示，可见模型在接收输入提示后，并未输出中间步骤，而是直接给出了最终的预测产物。

更进一步，显式与隐式思维链两者的差别在解码端体现：在隐式思维链中，模型不再对每一中间位置执行 “softmax + 采样”，而是直接将该位置在最后一层的隐空间向量作为下一位置原本的嵌入向量送回模型。如此，自然语言层面不再产生任何中间词元，模型的多步思考完全发生在连续隐空间内；在累计达到设定步数后，模型再回到普通自回归解码模式，输出最终产物的自然语言描述。为便于后续讨论，本节将该流程整理为算法 2，详见表 3.3。

表 3.3 隐式思维链推理模型 Flux-Z 的具体实现步骤

模型输入：反应条件与反应物文本 x_c 与 x_r，隐空间块数 U，块内向量数 L
模型输出：预测产物序列 $\hat{y}$
过程：
1. 将输入 $x=[x_c;x_r]$ 经分词器 T 切分为词元序列 $t=(t_1,\ldots,t_L)$，并嵌入为 $E=(e_1,\ldots,e_L)$；
2. 在序列末尾追加特殊词元 <latent>，标识隐式思维链开始；
3. 依次进入第 $u=1,\ldots,U$ 个隐空间块，在每块内一次生成 L 个隐空间向量 $z_{u,1},\ldots,z_{u,L}$；
4. 每个隐空间向量直接作为下一位置的输入嵌入；
5. U 个隐空间块结束后追加特殊词元 </latent> 以触发自然语言解码恢复；
6. 模型恢复普通自回归解码，依据隐空间块累积的上下文生成 “Final Product:” 段，输出最终产物 SMILES 序列 $\hat{y}$；
7. 在产物 token 上施加交叉熵损失，优化目标完全由最终预测驱动；
8. 返回产物预测 $\hat{y}$。

3.3.2 4D 因果注意力掩码

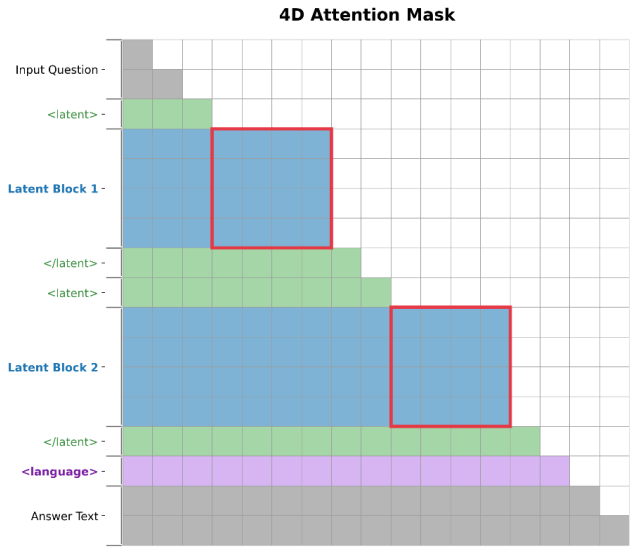


图 3.4 4D 因果注意力掩码示意

为提升模型对隐空间推理过程的可控性与可适应性，本研究在基座模型 Qwen3 上进行了两点关键改造：其一，对词表（vocabulary）进行扩展，引入 <latent>、</latent>、<language> 三个特殊词元（special token）以显式标识隐空间块的开始、结束以及自然语言解码的恢复（其中 <language> 仅在后续 Flux-I 模型中使用）；其二，参考 Kang 等人[38]的注意力掩码修改方法，对底层注意力（attention）机制进行自定义的方法，使其在普通词元与隐空间向量之间施加差异化的可见性约束。

如图 3.4 所示，本研究将这一改造称为“4D 因果注意力掩码”。其名称中的“4D”实质上指代底层实现中注意力掩码张量（Tensor）的四维数学形状，即批次大小（Batch Size）、注意力头数（Num Heads）、查询序列长度（Query Sequence Length）以及键值序列长度（Key Sequence Length）。在该四维标准数据结构的基础上，本研究对传统的下三角掩码进行了核心逻辑的重新定义，将其重构为“全局因果与局部双向”相结合的混合可见性矩阵。具体而言，对于普通的自然语言序列以及不同的隐空间块（Latent Block）之间，掩码严格遵循因果约束（即当前位置仅能关注自身及历史上下文），保证了块与块之间、以及文本生成时的严格因果顺序，防止未来信息泄露；而对于处于同一个隐空间块内部的连续隐向量，掩码限制被解除，使其相互完全可见（构成块内全注意力）。这种定

制化的四维掩码设计，既维持了自回归模型的全局生成范式，又极大促进了同一推理步内隐变量之间细粒度的信息交互与融合。

在后文中，本研究的纯粹隐式思维链 Flux-Z，语义引导预训练 Flux-I 和由 Flux-I 进一步后训练得到的 Flux-ZI 等系列隐式推理模型，都使用了上述 4D 因果注意力掩码。Flux-Z 和 Flux-ZI 固定了隐式推理块的个数，故块间不再设置<lantent>、</latent>标记，仅在连续块两端设置。

3.3.3 隐空间编解码模型 Flux-vae

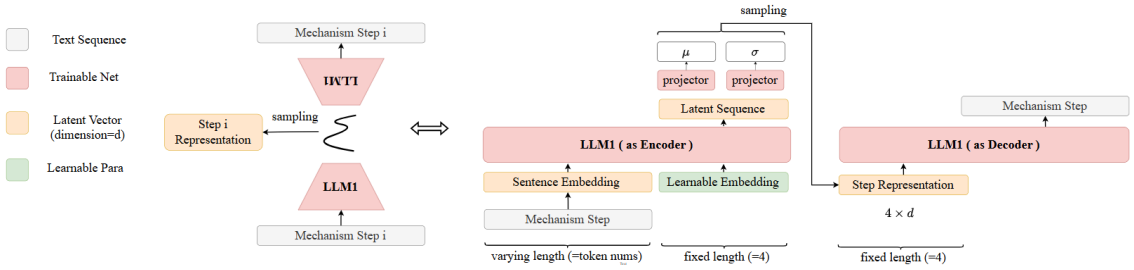


图 3.5 隐空间编解码模型 Flux-vae 的整体结构

为对隐式思维链推理过程中产生的隐空间向量给予可解释的还原渠道，本研究参考 Kang 等人[38]的工作，设计了一个基于变分自编码器（Variational AutoEncoder, VAE）的隐空间编解码模型 Flux-vae，整体结构如图 3.5 所示。Flux-vae 以一句完整的基元反应自然语言描述为输入，先由编码器将其压缩为形状为 4×1024 的隐空间张量，再由解码器在该张量条件下重构原始文本。在 Flux-min 数据集上完成预训练后，Flux-vae 既可作为隐空间向量的“翻译器”，将隐式思维链产物逐步还原为人类可读的英文机理描述，也可在反向上将文本机理标签编码为隐空间监督信号，为 3.3.4 节的语义引导预训练提供训练目标。

至此，本研究的系统在隐空间维度上同时具备两个互补能力：将一段自然语言机理描述编码进入 4×1024 隐空间张量的能力，以及对推理模型生成的 4×1024 隐空间张量进行解读的能力。这一双向通道为后续显式监督与隐式推理的混合训练提供了基础设施支持。

3.3.4 语义引导预训练模型 Flux-I

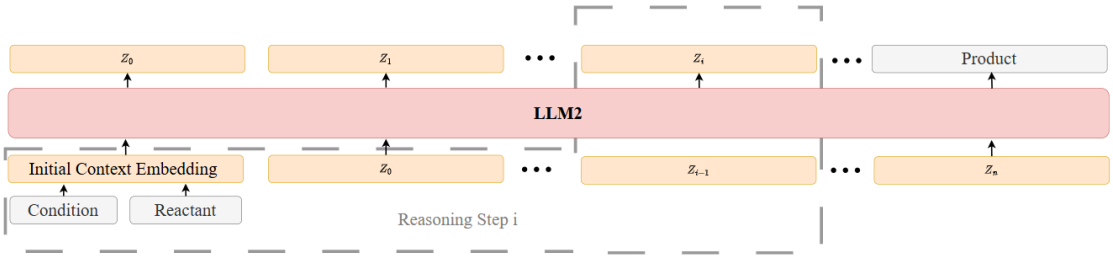


图 3.6 语义引导预训练模型 Flux-I 的训练流程

本研究参考 Kang 等人[38]的工作，在 Flux-vae 训练完成的基础上，进一步设计了语义引导的隐空间预训练模型 Flux-I，整体流程如图 3.6 所示。借助 Flux-vae 的编码能力，可将 Flux-min 中每条多步反应轨迹的每一基元步骤逐句压缩为对应的隐空间张量，从而为隐空间推理提供逐步对齐的真实标签——表 3.1 中 text[1] 所示的一句机理描述即可被映射为一个形状为 4×1024 的隐空间向量。

在此监督信号下，Flux-I 学习“上一步隐空间向量到下一步隐空间向量”的转移规律：当模型执行第 i 步隐空间推理时，输入端拼接反应物与反应条件的文本词元，以及第 0 步至第 $i-1$ 步全部由 Flux-vae 编码得到的隐空间向量；模型需在输出端预测第 i 步对应的隐空间张量，损失以与真实隐空间标签的均方误差形式给出，其计算如式 (3-1) 所示。

$$\mathcal{L}_{Flux-I}(\theta)=\begin{cases}-\sum_{t}\log p_\theta(y_t\mid x,y_{<t},z_{<i}), & is\_text=True\\ \frac{1}{L}\|\hat{z}-z\|^2, & is\_text=False\end{cases} \tag{3-1}$$

完成上述语义引导预训练后，本研究将预训练的权重作为初始化，再在固定长度的隐空间推理任务上微调，得到正式投入对比的模型 Flux-ZI。第 4.5 节的实验将定量说明：相比直接以最终答案为唯一监督的端到端隐式训练 Flux-Z，这一两阶段训练方式具有显著的性能与稳定性增益。

3.3.5 隐空间流形扰动正则机制

如 3.3.4 节所述，Flux-Z 在隐空间中以连续向量 z_t 替代离散词元执行多步推理。然而，z_t 是高维稠密表示，其有效区域可视为分布在反应机理流形附近的低维子集；若推理过程仅以最终答案为目标进行无监督优化，模型容易锁定流形上的某些极窄通路：一旦上一步输出偏离这些通路，下一步便落入流形外的“未定义区域”，进而产生显著的级联误差。这一现象在第二章中已被概括为“隐式流形偏移”（Latent Manifold Shift）。

为缓解上述问题，本研究在隐空间推理过程中引入流形扰动正则机制。其核心思想是：在每一步隐空间向量 z_t 进入下一步推理之前，于该向量的局部邻域内施加一个由其自身分布决定的随机扰动 ε_t，从而强制模型在反应机理流形的局部邻域上保持一致预测。形式化地，记 z_{t} 在推理时的方差估计为 σ_{t}2，则扰动后的隐空间向量定义为：

$$z' = z_t + \varepsilon_t, \varepsilon_t \sim N(0, \lambda\sigma_t^2 I) \tag{3-2}$$

其中 λ 控制扰动强度。该方案在数学上等价于以扰动期望损失 $E_\varepsilon[L(z_t+\varepsilon_t)]$ 替代精确点损失 $L(z_t)$：对原损失进行二阶 Taylor 展开可知，期望损失等于原损失叠加一项与隐空间二阶曲率（Hessian 迹）相关的隐式正则项，从而抑制模型在高曲率区域过拟合，使学习到的特征沿流形局部光滑、其反向传播梯度对小扰动具备稳健性。这一等价关系最早由 Bishop[34] 在 1995 年的经典工作 “Training with Noise is Equivalent to Tikhonov Regularization” 中给出严格证明；近年来，Camuto 等人[35]在 NeurIPS 2020 上进一步从频域视角揭示，高斯噪声注入对深层神经网络存在一项显式正则项，可惩罚函数在高频方向上的剧烈变化，并改善分类边界的校准性，为本节流形扰动机制在训练阶段的稳定性与泛化性提供了理论支撑。

从流形几何的视角看，加入扰动等价于将每一步的隐空间表征由“单点”扩展为以 z_t 为中心、以 σ_t 为半径的“邻域云团”，迫使解码器与下一步推理模块对该邻域内的任意采样点都给出一致输出，从而填补隐空间中潜在的“死区”，

并保证特征间的局部利普希茨连续性。本研究将在第 4.6 节通过四档不同强度的扰动水平 λ 与不同隐空间步数 U 的网格化实验，定量展示流形扰动正则机制对隐式思维链推理鲁棒性的影响。

3.4 本章小结

本章面向化学反应机理推理这一目标任务，依次完成了从数据到模型的方法构建。在数据侧，本章设计了“上下文构建—拓扑差异提取—机理决策—模板渲染”四阶段的自动化处理管线，将带原子映射的基元反应轨迹转写为以电子流为主语的自然语言机理描述，得到首个大规模电子推移机理思维链语料库 Flux-min；在此基础上，本章进一步组织了规模适中、覆盖全面、严控数据泄露的 Flux-bench 基准测试集，用于支撑后续多维度评估。在模型侧，本章基于 Qwen3 提出了显式思维链推理模型 Flux-R 与隐式思维链推理模型 Flux-Z 两条路线，并通过 4D 因果注意力掩码、隐空间编解码器 Flux-vae、语义引导预训练 Flux-ZI 与隐空间流形扰动正则机制等关键组件，使 Flux-Z 在突破自然语言标注瓶颈的同时仍尽可能保持机理可解释性与推理鲁棒性。第四章将基于本章构建的数据集与模型，对显式与隐式两条思维链路线在不同数据规模与扰动条件下进行系统的实证比较。

第四章 实验设计与结果分析

为确保评测结果的统计显著性与严谨性，本研究针对构建的评估基准执行了高强度的多采样推断（Multi-sample Inference）。在评测阶段，包含了从 0.6B 本地轻量化模型到 35B 以及 1.6T 级云端超大模型（DeepSeek-V4-Pro）的广泛参数区间。

由于化学反应机理推理涉及冗长的自回归生成过程，单次评估的计算开销极大。本实验累计生成并验证了超过一亿个 Tokens，单个本地模型基准测试耗时高达数百个 GPU 卡时。详尽解码与采样策略（Pass@1, 3, 10），一定程度缓解了大模型生成随机性带来的偶然误差，确保了本文对隐空间与显式语言空间性能的探索建立在坚实的数据基础之上。

4.1 实验环境配置

本研究全部训练与推理实验在同一台多卡服务器上完成。该机器以四张 NVIDIA RTX PRO 6000 显卡作为算力主体；软件层面统一使用 PyTorch 2.8.0 与 CUDA 12.8 作为底层框架，借助 Accelerate 完成多卡并行调度，并在训练与推理阶段统一采用 bfloat16 精度。具体软硬件配置如表 4.1 所示。

表 4.1 实验软硬件环境配置

类别	项目	配置参数
硬件环境	GPU	NVIDIA RTX PRO 6000 (96 GB) × 4
	CPU	Intel Xeon Platinum 8470Q (88 vCPU)
	内存	440 GB
	存储	系统盘 30 GB / 数据盘 600 GB
软件环境	操作系统	Ubuntu 22.04
	编程语言	Python 3.12
	深度学习框架	PyTorch 2.8.0 / CUDA 12.8
	分布式加速	HuggingFace Accelerate
训练设置	训练精度 / 推理精度	bfloat16 / bfloat16
	训练方式	全参数微调（Full Fine-tuning）

4.2 评测使用指标说明

本研究的评测涵盖产物预测、机理推理与对抗背诵三类核心能力，因而需要在不同子集上配套不同维度的指标。其中：Product Hit Rate 衡量预测产物是否与真实产物的 SMILES 完全等价；SMILES Validity 用 RDKit 解析模型输出的字符串能否还原为合法分子，以反映生成的化学合规性；Bert-Score 与 BLEU 分别从语义与字面两个层面度量推理链与人工标注机理的相似程度；FG Selectivity 与 Product Selectivity 则面向同源反应物的多条件实验，分别考察模型是否能依据条件差异输出正确的官能团变化与最终产物选择。各指标的具体含义与对应基准子集如表 4.2 所示。

表 4.2 评测指标含义

评估指标	具体含义	使用指标的基准测试子集
Product Hit Rate	产物命中率	ORDerly-100
SMILES Validity	化学标记有效性	ORDerly-100
Bert-Score	推理链在 Bert-large 隐空间中真实标签语义相似度	Fukuyama-A
BLEU	推理链与真实标签直接字符串相似度	Fukuyama-A
FG Selectivity	官能团选择性	ORDerly-mc
Product Selectivity	产物选择性	ORDerly-mc

4.3 反应机理数据集 Flux-min、Flux-max 与基准测试 Flux-bench 统计数据

本节给出第 3.1.1 节构建的反应机理训练数据集 Flux-min、Flux-max 与第 3.1.2 节构建的 Flux-bench 基准测试集在数据规模、长度分布、模板分布以及子集划分上的统计结果，为后续训练超参与评测口径的设定提供量化依据。

首先，训练数据集的整体规模与丢弃情况如表 4.3 所示。在 FlowER 全量轨迹上运行第 3.1.1 节构建的处理管线后，共保留 271 597 条多步反应轨迹，丢弃 17 427 条（占 6.0%），合计含 2 014 238 个基元步骤的自然语言描述。在此之外，本文进一步基于 ORDerly[29] 中清洗后的反应物—产物对补充了 471 683 条端到端反应数据，用于补充训练语料得到扩展数据集 Flux-max。

表 4.3 反应机理数据集规模统计

项目	数值
Flux-min 保留多步反应轨迹数	271 597
Flux-min 丢弃轨迹数（净化失败 / 未分类）	17 427
Flux-min 总基元反应步骤数	2 014 238
Flux-max 端到端反应补充量（来自 ORDerly）	471 683

其次，机理决策树触发的模板分布如表 4.4 所示。整体上极性反应占绝对多数，周环与自由基反应数量稀少，与有机化学的常识规律一致；该分布也直接决定了后续训练中按机理类别加权采样的策略。

表 4.4 机理模板使用频次统计

模板类别	出现次数
1_Polar_Complex_Mechanism	868 904
1_Polar_Inter_ProtonTransfer	463 042
1_Polar_Metal_General	192 812
1_Polar_Intra_ProtonTransfer	60 089
2_Pericyclic_Sigmatropic	12 353
2_Pericyclic_Cycloaddition	672
2_Pericyclic_(Retro_)Ene	699
2_Pericyclic_Retro_Cycloaddition	630
3_Radical_Homolysis	519
3_Radical_Coupling	458
4_General_Metathesis_Exchange	176

在长度分布上，单条基元反应描述与整条轨迹上下文的统计如表 4.5、表 4.6 所示。绝大多数基元步骤可在 256 token 内完整表达；整条轨迹上下文有约 96% 落在 2 048 token 以内。这两项统计为后文显式与隐式思维链模型确定单步生成上限与上下文窗口提供了直接依据。

表 4.5 单条基元反应描述长度分布

统计项	数值	占比
≤ 256 token	2 011 304	99.85%
≤ 512 token	2 014 238	100.00%
平均长度	95.8 token	—
最大长度	482 token	—

表 4.6 完整轨迹上下文长度分布

统计项	数值	占比
≤ 256 token	26 039	9.59%
≤ 512 token	146 729	54.02%
≤ 1024 token	210 938	77.67%
≤ 2048 token	261 389	96.24%
平均长度	710.5 token	—
最大长度	8 584 token	—

最后，第 3.1.2 节构建的 Flux-bench 基准测试集由 ORDerly-300、Fukuyama-A 与 ORDerly-MC 三个子集组成，规模与定位如表 4.7 所示。ORDerly-300 用于评估端到端产物预测的普适能力；Fukuyama-A 用于评估机理思维链的逐句准确性；ORDerly-MC 在共反应物、不同条件的多组反应上联合考察，旨在抑制模型对训练样本的字面背诵，严格地考察其对反应机理规律的学习。

表 4.7 Flux-bench 基准测试集统计

子集	样本量	评测目的
ORDerly-300	300 条反应物—产物对	测试普适意义上的产物预测能力
Fukuyama-A	10 条多步轨迹（共 29 句）	测试机理思维链生成的准确性
ORDerly-MC	30 组多条件反应	对抗背诵，考察机理规律的真实学习

4.4 隐空间编解码器 Flux-vae

隐空间编解码器 Flux-vae 是连接显式机理文本与隐式推理向量的关键基础设施，其重构能力直接决定了后续语义引导预训练（Flux-I）所提供的监督信号的可信度，也决定了对 Flux-Z 输出隐空间向量进行人类可读还原的上限。为此，本节以 Flux-bench 的 Fukuyama-A 子集为评测来源，将其 10 条多步轨迹拆解为 29 句相互独立的基元反应描述，依次输入 Flux-vae 的编码器以得到形状为 4×1024 的隐空间张量，再经其解码器自回归还原为自然语言句子。重构质量同时从字面与语义两个层面进行衡量：精确匹配率（Exact Match Rate）统计还原句子与原句逐字符完全一致的比例，反映编解码过程是否存在信息无损保留；BLEU 则容许同义改写与词序细微变化，从字符串相似度上给出更宽松的对齐度量。具体结果如表 4.8 所示。

表 4.8 Flux-vae 在 Fukuyama-A 单句上的重构评测

评测项目	数值
评测样本数	29
精确匹配数（Exact Match Count）	11
精确匹配率（Exact Match Rate, %）	37.93
BLEU 分数	79.28

从结果上看，Flux-vae 在 29 句样本上取得了 79.28 的 BLEU 与 37.93% 的精确匹配率：约四成句子能在亲核位点、亲电位点、键变化等关键化学要素上被一字不差地还原，其余样本虽未达到字面完全一致，仍能在词序与同义改写层面保持较高的字符串重合度。这表明 4×1024 的低维隐空间已足以承载一句完整电子推移描述的核心语义信息，能够为 Flux-I 的语义引导预训练提供质量可信的监督信号；但从仍未达到 100% 的精确匹配率来看，编解码过程不可避免地会在一些细节用词与立体描述上出现损失，也提示后续若引入更长上下文或更复杂的多步机理，仍需进一步扩大隐空间维度或在解码端增强词级约束。

4.5 不同数据集规模下的模型性能

本节系统比较显式思维链与隐式思维链在不同数据规模下的表现差异，旨在明确两个核心问题：在机理标注稀缺时，显式与隐式两类范式各自能达到的性能上限；以及当训练数据由稀缺的机理语料 Flux-min 扩展到更大规模的端到端语料 Flux-max 时，隐式思维链是否能够突破自然语言标注瓶颈并获得显著增益。为此，本节在两组训练规模下分别对 Flux-Base、Flux-R、Flux-Z 与 Flux-ZI 等模型进行对照评测，并以 ORDerly-300 上的 Hit Rate 与 Validity、Fukuyama-A 上的 BERTScore 与 BLEU、ORDerly-MC 上的 FG Selectivity 与 Product Selectivity 共三类六项指标作为衡量维度。所有模型均在相同的 0.6B 基座（Qwen3）上训练，并在 pass@1、pass@3、pass@10 三档采样次数下报告结果，以兼顾单次稳定性与多采样上限。为便于阅读，先约定本节出现的模型记号：Flux-Base 表示在反应物—产物对上端到端训练的对照模型；Flux-R 为显式思维链推理模型，输出包含完整中间机理；Flux-Z-LAUB 为隐式思维链推理模型，其中 L 与 U 分别记隐空间每块向量数与块数，二者共同决定隐式思考的总步长；Flux-ZI 在隐式推理之前增加了由 Flux-vae 提供的语义引导预训练阶段。

4.5.1 Flux-min 反应机理数据集上的模型对比

本节将四个模型限制在 Flux-min（约 27 万条机理轨迹）单一来源上进行训练，重点考察机理标注稀缺时不同思维链范式的相对优势。结果如表 4.9 所示。

从表 4.9 可以观察到三点关键现象。其一，在所有 Hit Rate 与 Validity 指标上，Flux-R 与 Flux-Base 取得了最优或次优的成绩；与之相对，仅依靠最终答案监督训练的 Flux-Z 在 ORDerly-300 与 ORDerly-MC 上表现明显落后。这表明在相对较小的数据规模条件下，显式思维链借助自然语言符号的强约束，能够在端到端产物预测上保持稳定优势，而完全无机理监督的隐式推理则缺乏足够的归纳偏置。其二，在度量机理逐句正确性的 Fukuyama-A 子集上，Flux-R 的 BERTScore 与 BLEU 远高于 Flux-Z 与 Flux-ZI——后者甚至出现了负的 BERTScore，说明 Flux-Z 输出的隐空间序列经 Flux-vae 解码后已严重偏离自然语言流形，呈现出第二章所述的“隐式流形偏移”。其三，引入 Flux-vae 监督的 Flux-ZI 相对原始 Flux-Z 在 Hit Rate、FG Selectivity 等多数指标上均有提升，且

表 4.9 Flux-min 训练下不同模型在 Flux-bench 上的性能对比（pass@k，%）

子集	指标	pass@k	Flux-Base	Flux-R	Flux-Z-L4U4	Flux-ZI-L4U4
ORDerly-300	Hit Rate	pass@1	19.67	**22.33**	8.00	15.33
		pass@3	21.67	**25.67**	8.67	21.33
		pass@10	28.67	**31.33**	13.00	26.33
	Validity	pass@1	**98.33**	84.67	96.67	95.67
		pass@3	99.33	99.33	**100.00**	99.00
		pass@10	**100.00**	99.67	**100.00**	**100.00**
Fukuyama	BERTScore	pass@1	-	**46.52**	-26.88	10.13
		pass@3	-	**53.54**	-23.16	9.64
		pass@10	-	**53.70**	-24.80	10.77
	BLEU	pass@1	-	**44.17**	9.19	9.84
		pass@3	-	**56.56**	11.09	9.19
		pass@10	-	**58.15**	9.24	10.42
ORDerly-MC	FG	pass@1	**40.92**	37.33	44.00	48.60
		pass@3	**63.44**	52.30	58.50	60.75
		pass@10	**73.92**	62.70	73.07	69.30
	Prod	pass@1	**6.67**	3.33	0.00	5.00
		pass@3	**13.33**	10.00	3.33	5.00
		pass@10	**15.00**	10.00	6.66	10.00
模型参数量				0.6 B		

注：表中数值均为百分比；同一指标-pass@k 维度下的跨模型最优值以加粗标出，次优值以下划线标出。表中的字母 M、B、T 分别代表百万（Million）、十亿（Billion）和万亿(Trillion)。

Fukuyama-A 上的 BERTScore 由负值回到正区间，验证了语义引导预训练对约束隐空间表示走向自然语言流形的有效性。综合而言，在小规模标注的极端场景下，显式思维链仍是化学机理推理的更稳健选择，但隐式范式可以通过引入额外的语义对齐信号显著缩小与之的差距。

4.5.2 数据规模扩展对隐式思维链性能的影响

表 4.10 Flux-max 训练下隐式思维链模型的性能对比（pass@k，%）

子集	指标	pass@k	Flux-Base	Flux-Z-L1U1	Flux-Z-L1U4
ORDerly-300	Hit Rate	pass@1	39.33	**41.67**	40.33
		pass@3	45.00	**47.67**	45.00
		pass@10	53.00	52.67	**53.33**
	Validity	pass@1	**95.67**	95.33	94.67
		pass@3	**99.67**	99.00	**99.67**
		pass@10	100.00	100.00	100.00
ORDerly-MC	FG	pass@1	66.09	**73.70**	72.73
		pass@3	75.77	79.89	**82.93**
		pass@10	83.26	**87.11**	85.25
	Prod	pass@1	36.67	**45.00**	41.67
		pass@3	50.00	**56.67**	46.67
		pass@10	53.33	**56.67**	55.00
模型参数量				0.6 B	

注：表中数值均为百分比；同一指标-pass@k 维度下的跨模型最优值以加粗标出，次优值以下划线标出。表中的字母 M、B、T 分别代表百万（Million）、十亿（Billion）和万亿(Trillion)。

为进一步检验隐式思维链能否突破自然语言标注瓶颈，本节将训练数据由 Flux-min 扩展至 Flux-max（合计含约 47 万条端到端反应物—产物对的补充语料），在隐式范式下保留 Flux-Base 与 Flux-Z 系列模型，结果如表 4.10 所示。

对照表 4.9 与表 4.10 可以看出，当训练规模由 27 万机理轨迹扩大到包含 47 万端到端反应物—产物对的混合语料后，隐式思维链模型的整体表现迎来质变。在 ORDerly-300 上，Flux-Z-L1U1 的 pass@1 Hit Rate 由 8.00% 跃升至 41.67%，已反超同等条件下 Flux-Base 的 39.33%；在更严苛的 ORDerly-MC 上，Flux-Z 系列在 FG Selectivity 与 Product Selectivity 两项指标上几乎全面占优，pass@1 Product Selectivity 较 Flux-Base 提升约 8 个百分点。这说明，当数据规

模充分扩大、监督信号变得密集后，隐式思维链不再受限于显式机理标注的稀缺，能够从大量端到端样本中自发组织出有效的隐空间推理路径。

另外两点细节同样值得关注：其一，Flux-Z-L1U1 与 Flux-Z-L1U4 在大多数指标上的差距并不显著，提示在当前数据规模与基座参数量下，仅靠扩展隐空间块内向量数（U）所带来的性能增益已趋于饱和；其二，在 SMILES Validity 上，所有模型均接近 100%，表明扩大数据后隐式范式生成的 SMILES 仍能保持高度的化学合规性，未因隐空间推理而牺牲分子表示的合法性。综上，本节的实验明确回答了 4.5 节开头的第二个问题：在更大规模数据上，隐式思维链能够突破自然语言机理标注的瓶颈，并在端到端产物预测与对抗背诵任务上同时反超无思考的基线模型。

4.6 隐空间流形扰动正则机制的鲁棒性探索

在 4.5.2 节的数据规模变化实验中本文得出了这样的结论：随着训练数据规模的扩大，隐式思维链模型的整体性能持续提升，并最终反超作为基线对照的无思考模型 Flux-Base。

而在 4.4 节的 Flux-vae 重构实验，以及 4.5.1 节的小数据机理评测中，本文同时发现了一个伴随现象：隐式思维链容易发生显著的“语义漂移”，模型在多步无监督优化下倾向于滑向数据流形中的领域外区域（Out-of-Domain），表现为隐空间向量难以经 Flux-vae 还原为合规自然语言；引入语义引导预训练（Flux-I 及随后训练得到的权重 Flux-ZI）可在一定程度上缓解该问题。

一方面，当数据集规模继续扩大时，语义引导预训练所需的高质量机理标注成本将随之线性上升，其代价已不小于直接构建可信的显式思维链；另一方面，隐式思维链的核心优势恰恰在于绕开显式机理标注的瓶颈。因此，能否在训练阶段以低成本方式提升隐式思维链对流形偏移的鲁棒性，成为隐式范式能否真正规模化应用的关键问题之一。本节即针对该问题展开实证研究，以第 3.3.5 节提出的隐空间流形扰动正则机制为主线，系统考察训练时引入不同强度高斯扰动对模型性能的影响。

如第 3.3.5 节所述，在隐空间向量传递过程中按其自身分布注入高斯扰动，等价于在原损失上叠加一项与隐空间二阶曲率相关的隐式正则项，其作用是抑制模型在高曲率区域的过拟合，并将每一步的隐空间表征由“单点”扩展为“邻域云团”，从而填补流形上的潜在死区。本节将这一机制实际落地为训练时扰动：以扩大数据集训练得到的 Flux-Z 权重为底，在训练阶段按照不同噪声水平 λ 与不同隐空间块数 U 对每一步隐空间向量执行采样扰动，并按 ORDerly-300 的 Hit Rate 指标在 pass@1、pass@3 与 pass@10 三档下汇总，得到图 4.1 至图 4.3 所示的网格化热力图。

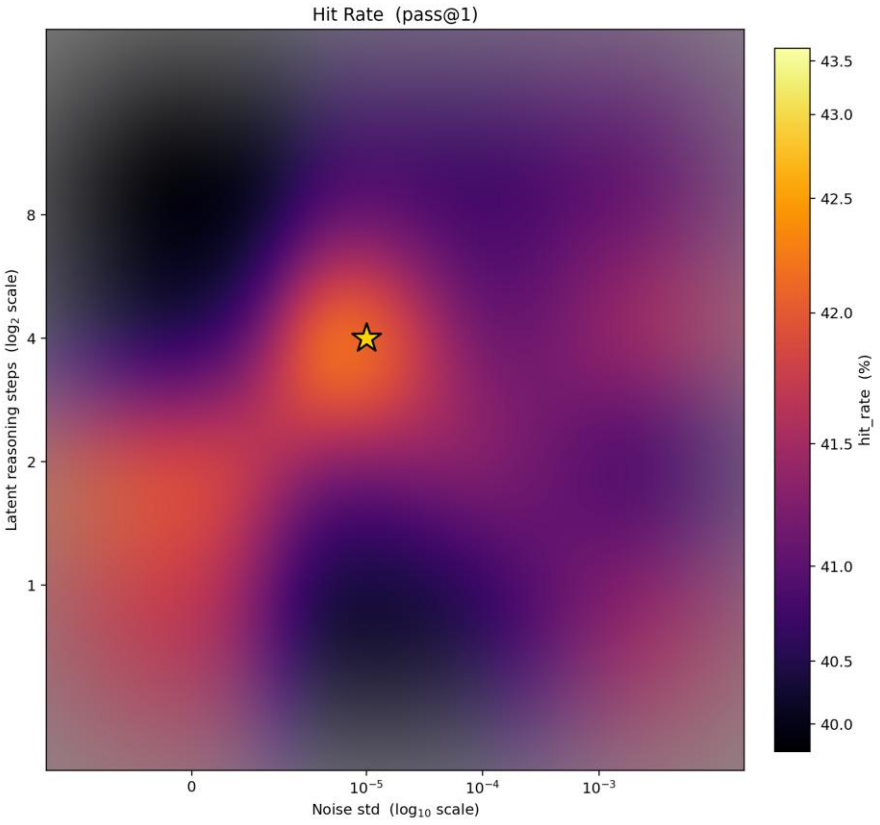


图 4.1 不同加噪水平和不同隐式思维链长度下的模型性能热力图（pass@1）

如图 4.1 所示，在 pass@1 的严格单次采样设定下，模型性能随噪声强度与隐式思维链长度的变化呈现出明显的非单调特征。图中横轴表示训练阶段注入的高斯扰动标准差，纵轴表示隐空间推理步数，颜色越亮代表 ORDerly-300 上的 Hit Rate 越高。可以观察到，性能最优点并不出现在无扰动区域，也不出现在最长推理链或最大噪声强度处，而是集中于弱扰动与中等隐式推理长度的组合附近。

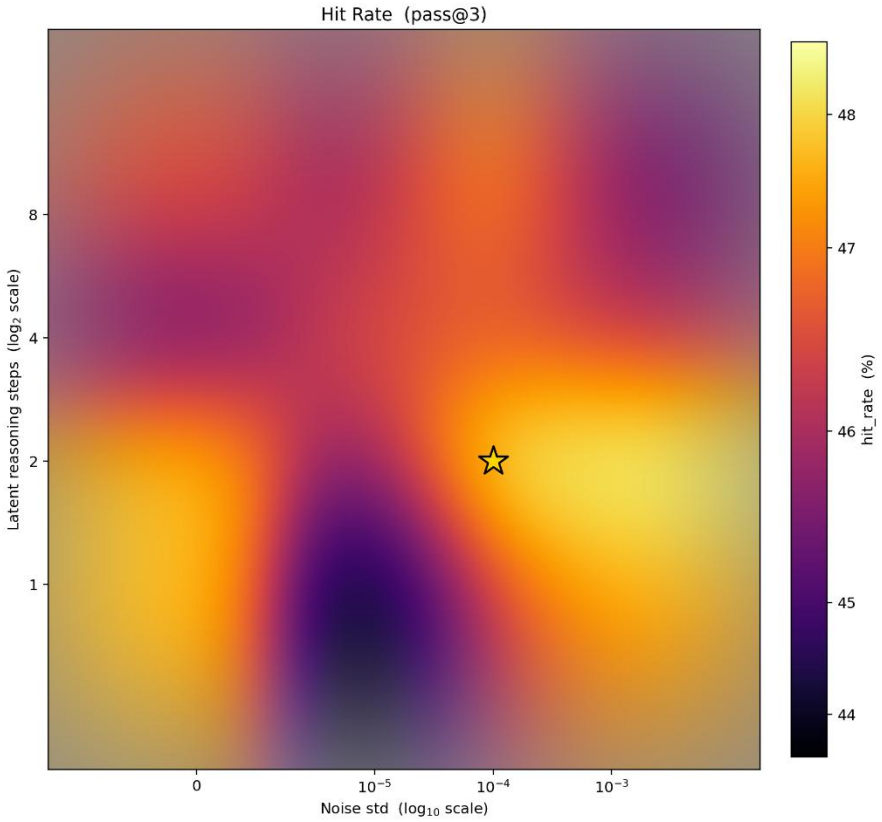


图 4.2 不同加噪水平和不同隐式思维链长度下的模型性能热力图（pass@3）

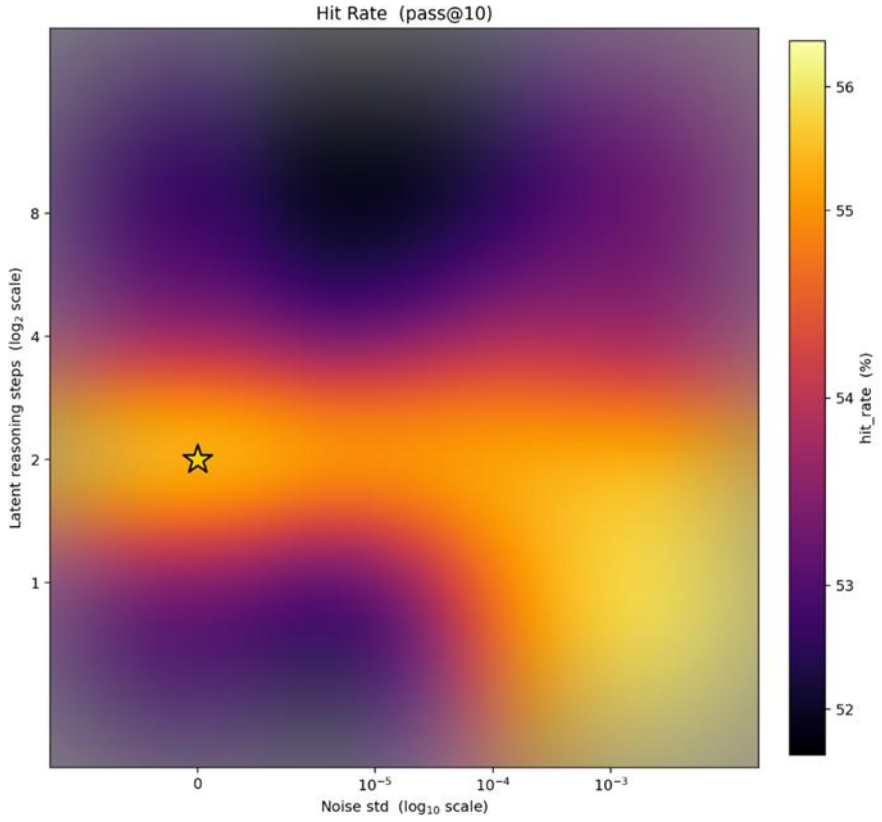


图 4.3 不同加噪水平和不同隐式思维链长度下的模型性能热力图（pass@10）

表 4.11 隐空间高斯扰动正则化的噪声水平设置

档位	噪声系数 λ	含义
L0	0.0	无扰动，等价于原始 Flux-Z 推理
L1	10^{-5}	弱扰动
L2	10^{-4}	中等扰动
L3	10^{-3}	强扰动

本文设置了四档不同强度的高斯噪声水平，扰动方差以隐空间向量的运行时方差 σ^2 为基准动态缩放，按 $\varepsilon \sim N(0, \lambda\sigma^2 I)$ 进行采样，其中 λ 取 0、10^{-5}、10^{-4}、10^{-3} 四个档位，分别对应“无扰动”、“弱扰动”、“中等扰动”与“强扰动”四种推理时正则强度，具体如表 4.11 所示。

对照图 4.1 至图 4.3 中的热力结果可以观察到三点规律。

（1）在隐空间块数 U 固定的前提下，性能最优区域往往并非位于“无扰动”一端，而是普遍出现在 $\lambda = 10^{-5}$—10^{-4} 的弱至中等扰动区间，说明适度的隐空间扰动能够稳定地改善 Hit Rate；这一观察与第 3.3.5 节中将扰动等价为隐式 Tikhonov 正则项的理论分析一致。值得注意的是，由于 Flux-Z 的训练目标为交叉熵损失，该扰动机制在本文任务中实际上化归为对隐空间二阶 Fisher 信息曲率的惩罚，从而有效地平滑了概率流形并改善了 Hit Rate。

（2）随着扰动水平进一步增大到 $\lambda = 10^{-3}$，性能在不同 U 值下均出现回落，提示扰动强度过大时模型会被推离原本所处的流形邻域，损害推理的学习。

（3）从纵向比较来看，最佳扰动水平随 U 增大而略向更大的 λ 偏移，可以解释为：在推理阶段，当推理步数变长时，每一步的“语义偏移”会沿隐空间链式累积，因此在训练时需要提高单步扰动方差以对抗推理阶段时误差被反复放大。综合三组热力图，本节验证了流形扰动正则机制的有效性——在不引入额外标注的前提下，仅通过推理时的轻量化随机扰动，即可稳健提升隐式思维链对流形偏移的鲁棒性，为下一节面向通用大模型的对比实验提供了本研究性能最稳定的隐式模型。

4.7 与前沿开源通用大模型的性能对比

为进一步刻画本研究所提模型在更宽广能力坐标系中的相对位置，本节选取两类具有代表性的开源通用大模型作为外部参照：DeepSeek-V4-Pro（约 1.6T 参数）作为万亿级超大规模通用推理模型的代表；Qwen3.6-35B-A3B 作为采用 Mixture-of-Experts 结构的中规模通用推理模型代表，其在推理时的激活参数量约为 3B。两者均在通用语料上完成预训练并具备较强的科学推理能力，对照之下可同时反映本研究 0.6B 量级特化模型的“垂直专精”与“算力性价比”两方面表现。本节分别在 Fukuyama-A 与 ORDerly-300 / ORDerly-MC 两组任务上展开对比：前者考察机理思维链的逐句生成质量，对应 4.5.1 节中训练得到的显式模型 Flux-R；后者考察端到端产物预测与多条件选择性，对应 4.6 节中得到的最终隐式模型 Flux-ZN-L1U4。Flux-ZN-L1U4 与 4.5.2 节 Flux-Z-L1U4 具有相同的隐空间推理块长和块数，但在 Flux-ZN-L1U4 训练时加入弱噪声扰动。结果分别如表 4.12 与表 4.13 所示，其中“Token 使用量”统一为整轮评测累计消耗的 token 数，隐式思维链已将隐空间向量等效计入。

表 4.12 不同模型在 Fukuyama-A 基准上的性能对比（pass@k，%）

子集	指标	pass@k	Flux-R	DeepSeek-V4-Pro	Qwen3.6-35B-A3B
Fukuyama	BERTScore	pass@1	**46.52**	19.90	10.24
		pass@3	**53.54**	23.23	14.71
		pass@10	**53.70**	26.49	23.00
	BLEU	pass@1	**44.17**	25.06	15.88
		pass@3	**56.56**	29.33	19.17
		pass@10	**58.15**	32.21	28.19
模型参数量			**0.6 B**	1.6 T	35 B
Token 使用量			**2.33M**	77 M	25 M

注：表中数值均为百分比；同一指标-pass@k 维度下的跨模型最优值以加粗标出，次优值以下划线标出。“参数量”一列的 A3B 变体表示推理时的激活参数量；“Token 使用量”指评测过程中累计的 token 消耗量。表中的字母 M、B、T 分别代表百万（Million）、十亿（Billion）和万亿(Trillion)。

由表 4.12 可见，在以机理逐句生成质量为衡量标准的 Fukuyama-A 上，仅 0.6B 量级的 Flux-R 在 BERTScore 与 BLEU 两项指标上均显著领先于参数量数量级更大的 DeepSeek-V4-Pro 与 Qwen3.6-35B-A3B：以 pass@10 为例，Flux-R 的 BERTScore 达 53.70%、BLEU 达 58.15%，而 DeepSeek-V4-Pro 的对应数值分别为 26.49% 与 32.21%，Qwen3.6-35B-A3B 仅为 23.00% 与 28.19%；与此同时，Flux-R 的整轮 token 消耗仅约 2.33M，远低于通用大模型在自由长链推理下产生的数千万 token 开销。

表 4.13 不同模型在 ORDerly-300 与 ORDerly-MC 基准上的性能对比（pass@k，%）

子集	指标	pass@k	Flux-ZN-L1U4	DeepSeek-V4-Pro	Qwen3.6-35B-A3B
ORDerly-300	Hit Rate	pass@1	43.66	**55.00**	24.33
		pass@3	46.66	**65.67**	40.67
		pass@10	53.00	**72.67**	51.33
	Validity	pass@1	94.00	**97.33**	87.33
		pass@3	**99.67**	**99.67**	98.33
		pass@10	**100.00**	**100.00**	**100.00**
ORDerly-MC	FG	pass@1	**64.71**	63.47	64.68
		pass@3	79.16	**82.12**	76.15
		pass@10	82.36	**91.35**	84.23
	Prod	pass@1	28.33	**38.33**	23.33
		pass@3	45.00	**58.33**	35.00
		pass@10	48.33	**68.33**	50.00
模型参数量			**0.6 B**	1.6 T	35 B
Token 使用量			**0.16M**	77 M	25 M

注：表中数值均为百分比；同一指标-pass@k 维度下的跨模型最优值以加粗标出，次优值以下划线标出。“参数量”一列的 A3B 变体表示推理时的激活参数量；“Token 使用量”指评测过程中累计的 token 消耗量。表中的字母 M、B、T 分别代表百万（Million）、十亿（Billion）和万亿(Trillion)。

表 4.12 表明，面向高度专业化的电子推移机理任务，“小参数、强对齐”的特化方案在生成质量与算力性价比上均明显优于通用大模型基于通用提示工程的解题方式。我们对此指标结果的进一步分析如下：

（1） Flux-R 原生思维链即为反应机理，换句话说，我们在训练阶段让模型本身学会了使用化学反应机理进行思考。DeepSeek、Qwen 等模型的显式思维链冗长，可能高达上万 Token，且内容随机性强，本身与化学反应机理关联不紧密。DeepSeek 和 Qwen 在结果中输出化学机理时，两者更多的是从世界知识中作化学领域知识的检索和召回。这一重要差异既影响了生成质量，又更加显著地影响了词元消耗。

（2） 电子推移描述的反应机理具有相对专业、明确、简洁的格式。虽然我们明确基准测试 Flux-bench 与训练数据 Flux-min 不存在数据泄露的情况，但是两者同样作为电子推移描述的反应机理，在语言风格和结构上必然相似。Flux-R 在 Flux-min 上训练并学会使用专业语言去推理，将在指标结果上天然具有一定的优势，但并不意味着 DeepSeek 和 Qwen 用更灵活的自然语言输出了错误的化学机理。

（3） 同样作为通用大模型，DeepSeek-V4-Pro 明显优于 Qwen3.6-35B-A3B，这与两者的模型参数量有直接关系，前者具有 1.6T 的超大规模参数，而后者是 35B 的中小规模参数：更大规模的参数意味着更多的世界知识。在如（1）中执行化学反应机理领域知识的检索和召回时，DeepSeek-V4-Pro 也将具有更高的准确率。

表 4.13 进一步给出使用 Flux-ZN-L1U4 推理的端到端产物预测与多条件选择性的对比。在 ORDerly-300 上 Flux-ZN-L1U4 作为 0.6B 模型，无论宽泛意义上的产物预测准确率还是化学表示的有效性，均领先参数规模 50 倍以上的 Qwen3.6-35B-A3B，同时 Hit Rate 推进到与 1.6T 通用大模型相邻的水平，在多项指标上取得次优的结果。在更严格的 ORDerly-MC 上，Flux-ZN-L1U4 的 pass@1 FG Selectivity 已反超两类通用大模型，而 Product Selectivity 显示均取得次优的结果，领先于 Qwen3.6-35B-A3B 而弱于 DeepSeek-V4-Pro。

综合两表可以得出一致结论：本研究通过机理语料对齐与隐空间结构改造所得到的 0.6B 特化模型，在化学反应机理推理任务上能够稳定逼近、并在多项指标上反超量级远大于自身的通用大模型，验证了“领域语料 + 结构改造”路线

相对纯粹堆叠参数路线在科学推理任务上的实用价值；同时也应客观指出，在 ORDerly-300 这一相对开放的任务上，1.6T 量级的 DeepSeek-V4-Pro 仍保持着相对领先的 Hit Rate，说明在极宽广反应分布上的兜底能力依然受益于规模红利。

4.8 模型能力边界分析

本节对预测失败样本作归因分析，回答两类思维链模型在哪些反应类型上更易发生“幻觉”或推理断裂。为同时分离“训练分布”与“思维链范式”两层效应，本节引入三个评测对象：4.7 节中的 Flux-R（显式思维链，Flux-min 训练）与 Flux-ZN-L1U4（隐式思维链，Flux-max 训练），以及 4.5.1 节中的 Flux-ZI-L4U4（隐式思维链，Flux-min 训练）。其中 Flux-R 与 Flux-ZI-L4U4 共享 Flux-min 训练集而范式不同，Flux-ZI-L4U4 与 Flux-ZN-L1U4 共享隐式思维链范式而训练数据规模不同，二者共同构成一组对训练分布与范式效应解耦的对照实验。评测在 Flux-bench 中规模最大、底物分布最宽泛的 ORDerly-300 子集（300 条反应）上展开，结果汇总于表 4.14。特别说明，ORDerly 在数据清洗阶段统一剥除立体描述符，本子集所有 SMILES 均不携带 @/@@、/、\ 等立体标记，故立体化学一类无法在本子集上直接量化。

表 4.14 ORDerly-300 子集上不同反应类别的三模型失败率对比

反应类别	样本数	Flux-R	Flux-ZI-L4U4	Flux-ZN-L1U4
整体（全集）	300	77.33%	85.00%	54.67%
含金属催化	45	93.33%	91.11%	46.67%
卤代芳烃为底物	23	91.30%	73.91%	52.17%
含 B/Si 偶联试剂	16	87.50%	93.75%	56.25%
含杂芳环底物	155	74.19%	83.23%	50.32%
多组分（≥2 个独立反应物片段）	214	71.96%	83.18%	54.67%
非金属催化（对照）	255	74.51%	83.92%	56.08%

注：表中失败率 = 1 − pass@1 Hit Rate；Flux-R、Flux-ZI-L4U4 训练于 Flux-min（约 27 万条机理轨迹），Flux-ZN-L1U4 训练于 Flux-max（额外引入约 47 万条 ORDerly 端到端反应物—产物对，与本评测使用的 ORDerly-300 已完全去泄露）；同一样本可同时属于多个类别（如 Pd 催化的卤代芳烃既计入"含金属催化"亦计入"卤代芳烃为底物"）。

由表 4.14 可作如下两组解耦分析。

（1）同训练集对照（Flux-R vs Flux-ZI-L4U4，均使用 Flux-min）：两者整体失败率为 77.33% 与 85.00%，与 4.5.1 节“在机理标注稀缺时显式思维链借助自然语言符号的强约束更具优势”的结论一致；进一步按类别细分，Flux-R 与 Flux-ZI-L4U4 在含金属催化类别上的失败率分别为 93.33% 与 91.11%，均显著高于各自整体失败率，且二者差距仅 2.22 个百分点。这说明在同为 Flux-min 训练时，金属催化反应对两类思维链范式都构成同一量级的困难：不能将金属配位与氧化态转换处的分支选择错误简单归因于显式范式本身。

（2）同范式对照（Flux-ZI-L4U4 vs Flux-ZN-L1U4，均为隐式思维链）：两者整体失败率为 85.00% 与 54.67%，差距 30.33 个百分点；按类别细分，差距在含金属催化（91.11% vs 46.67%，44.44 个百分点）、含 B/Si 偶联试剂（93.75% vs 56.25%，37.50 个百分点）、含杂芳环（83.23% vs 50.32%，32.91 个百分点）等类别上最为明显。

表 4.15 ORDerly-300 子集上各金属催化反应在三模型下的失败率

金属	样本数	Flux-R	Flux-ZI-L4U4	Flux-ZN-L1U4
Pd	22	90.91%	86.36%	31.82%
Al	7	100.00%	100.00%	71.43%
Sn	3	100.00%	100.00%	66.67%
Pt	2	100.00%	100.00%	50.00%
Zn	2	100.00%	100.00%	50.00%
Ti	2	100.00%	100.00%	50.00%
Mg	2	100.00%	100.00%	50.00%
Cu	2	50.00%	50.00%	50.00%
Fe	1	100.00%	100.00%	0.00%
Ag	1	100.00%	100.00%	100.00%
Os	1	100.00%	100.00%	100.00%

注：表中样本数按反应条件 SMILES 中显式出现的金属元素逐条计入，同一样本若条件中含两种金属则均计；失败率统计口径与表 4.14 完全一致。

表 4.15 按金属物种进一步拆分。Pd 是最高频的金属催化剂，Flux-R 与 Flux-ZI-L4U4 在 Pd 催化下失败率分别为 90.91% 与 86.36%，均处于高失败区间，而 Flux-ZN-L1U4 仅为 31.82%；Al、Sn 等强 Lewis 酸或活泼金属虽样本量较小，两个 Flux-min 训练模型仍普遍接近全错。

显式思维链的可解析性为进一步揭示失败时的微观推理形态提供独有的观察窗口。为此，本节对 Flux-R 在 ORDerly-300 上的全部 232 条失败样本逐条解析其推理过程文本，按字符级启发式规则识别五类推理断裂模式，如表 4.16 所示。

表 4.16 Flux-R 显式思维链推理断裂模式统计（基于 232 条失败样本）

推理断裂模式	触发样本数	占失败样本比例
推理过程引入反应物 / 条件之外的新元素原子	168	72.41%
模型最终给出的产物 SMILES 未作为推理链路中的生成目标出现	39	16.81%
模型输出的预测产物中混入未收敛的英文推理片段	38	16.38%
推理过程从未提及反应条件中显式给出的金属催化剂	26	11.21%
推理片段出现长段连续重复（卡入文本循环）	4	1.72%

注：5 类规则非互斥，同一失败样本可同时归类于多条规则。

表 4.16 揭示了 Flux-R 失败样本的五类典型微观形态。元素幻觉（72.41%）覆盖了绝大多数失败样本，模型在推理过程中引入了反应物 / 条件之外的新元素原子。答推分离（16.81%）刻画模型最终给出的产物 SMILES 从未作为任一推理步骤生成目标的情形，是显式范式下“答案”与“推理”之间最直接的不一致表现。催化剂缺失（11.21%）的绝对数量虽不大，但约占 Flux-R 金属催化失败样本（42 条）的 61.90%，说明 Flux-R 在面对金属催化时倾向于直接绕过金属配位环节、按“无催化剂”的纯有机电子推移规则生成机理的微观情形，也是 Flux-R 在金属催化失败中的可解释证据。输出未收敛（16.38%）则源于解码已抵达 token 上限但机理仍未走完。

下面摘取两条具代表性的 Flux-R 失败样本作直观对照：

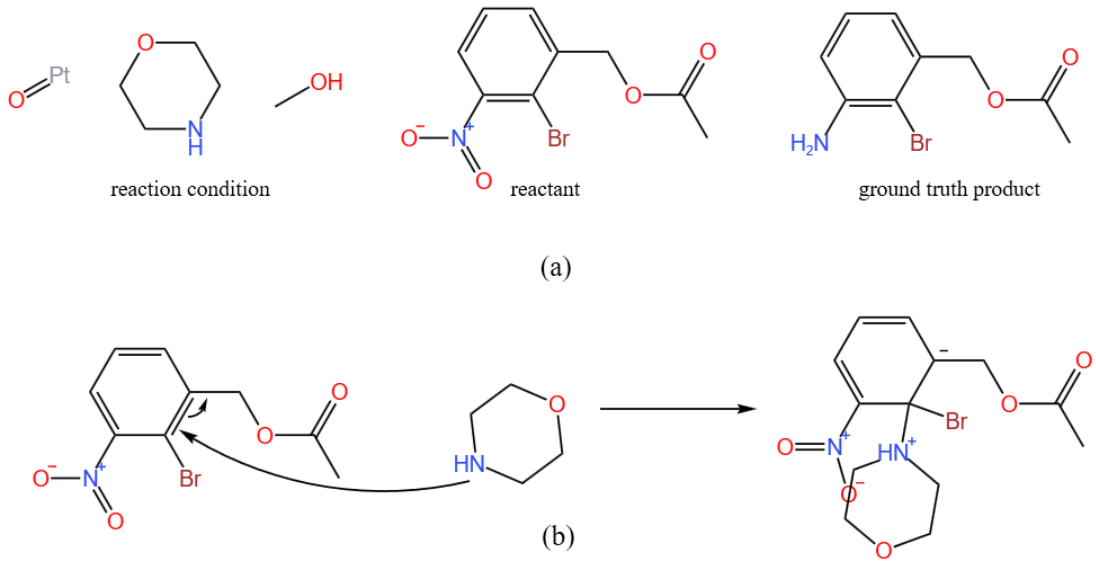


图 4.4 Pt 催化硝基还原进攻位点错误

（1）案例 A（Pt 催化硝基还原）：如图 4.4（a）所示，反应底物为 CC(=O)OCc1cccc([N+](=O)[O-])c1Br，反应条件含 [O]=[Pt]、吗啉与甲醇溶剂，真实情况为将硝基还原为氨基的产物 CC(=O)OCc1cccc(N)c1Br。Flux-R 与 Flux-ZI-L4U4 在该样本上均失败，前者推理过程完全忽略铂催化剂，如图 4.4（b）所示，转而以吗啉的二级胺氮亲核进攻芳环碳，生成取代型的错误产物，对应“催化剂缺失 + 反应类型误判”双重断裂；后者则未识别 Pt 催化场景，回吐了输入分子本身（硝基未被还原）；Flux-ZN-L1U4 在同样本上正确给出预测结果。

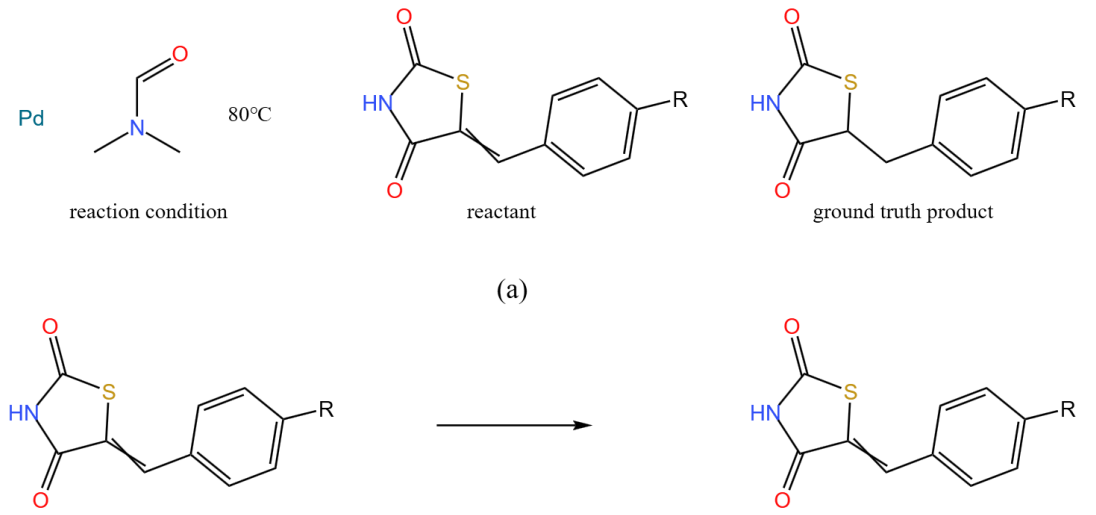


图 4.5 Pd 催化烯烃氢化未反应

（2）案例 B（Pd 催化烯烃氢化）：如图 4.5（a）所示，反应底物为含 α，β-不饱和噻唑酮的 CCCCCCCCCNC(=O)c1cccc(-c2ccc(C=C3SC(=O)NC3=O)cc2)c1，反应条件为 [Pd]、DMF、80°C，真实情况为外环 C=C 加氢后的产物。Flux-R 推理输出大量 “the metal center Pd ...” 模板化片段，但始终未将 Pd 与外环烯烃的协同氢化建模成功，最终如图 4.5（b）所示回吐了输入分子本身；Flux-ZI-L4U4 同样未给出加氢产物；Flux-ZN-L1U4 同样本正确给出加氢后产物。

综合表 4.14—4.16 及上述案例，可将本研究两类思维链范式当前的能力边界归纳如下：（1）数据集对金属催化族反应的覆盖不足会同时限制显式与隐式两类范式。过渡金属催化反应是三模型中两个 Flux-min 训练模型共同高失败的反应类别， Flux-min 机理语料中此类反应因多中心金属轨迹难以单链化为电子推移基元式而被天然稀释。（2）显式范式面对难以突破（1）中所述限制，而隐式范式则可以利用具有复杂反应的数据集突破限制。例如利用 Flux-max 补充训练，Flux-ZN-L1U4 对比同为隐式范式的 Flux-ZI-L4U4 ，在金属催化类反应上的失败率从 91.11% 降至 46.67%。（3）对 Flux-R 显式推理链路的微观断裂分析显示，元素幻觉与答推分离是两类最具代表性的“幻觉”形态。 （4）由于在清洗阶段统一剥除立体描述符，本节未能对立体专一反应（如不对称催化）作直接评测，本文所训练模型也不具备部分立体专一反应相关推理能力。

4.9 本章小结

本章基于第三章构建的数据集与模型，从评测设施、单模型评测和跨模型对比三个层面展开实证研究。首先统一软硬件环境与评价指标，并给出 Flux-min、Flux-max 训练语料及 Flux-bench 基准测试集的规模与分布。其次评测 Flux-vae 的隐空间重构能力，以及不同范式模型在两种数据规模下的表现：Flux-min 数据条件下，显式思维链 Flux-R 凭借自然语言符号约束，在准确率与可解释性上占优；数据规模扩大后，隐式思维链突破自然语言机理标注瓶颈，各项性能反超无思考基线模型。最后引入推理时高斯扰动正则机制，刻画隐式范式鲁棒性，并将 Flux-R、Flux-ZN-L1U4 与 DeepSeek-V4-Pro、Qwen3.6-35B-A3B 对照。结果表明，本文 0.6B 量级特化模型能以更低算力代价，在化学反应机理推理上取得相当甚至更优性能，实证回应了显式与隐式思维链能力边界问题。

第五章 总结与展望

5.1 工作总结

本文围绕大语言模型在化学反应机理推导中的显式思维链与隐式思维链推理范式，系统开展了从数据构建到模型设计、再到多维度实证评估的研究工作，主要成果可归纳为以下三方面。

第一，构建了首个面向电子推移机理的自然语言化学推理语料体系。本文设计了“上下文构建—拓扑差异提取—机理决策—模板渲染”四阶段自动化处理管线，将带原子映射的基元反应轨迹转写为以电子流为主语的英文叙述，得到包含约 27 万条多步轨迹与 200 余万个基元步骤的反应机理数据集 Flux-min；并以 ORDerly 为补充来源构建了端到端反应数据集 Flux-max，以及由 ORDerly-300、ORDerly-MC 与 Fukuyama-A 三个子集组成、严控数据泄露的 Flux-bench 基准测试集，为后续模型训练与评测提供了规模适中、覆盖全面的数据基础。

第二，提出了显式与隐式两类思维链推理模型并完成了系统改造。在 Qwen3 0.6B 基座上，本文实现了显式语言推理模型 Flux-R 与隐空间固定长度推理模型 Flux-Z；在隐式范式上，进一步引入 4D 因果注意力机制以支持“块状思考”，设计并预训练了隐空间编解码器 Flux-vae 以打通文本机理与隐空间向量的双向通道，提出了基于 Flux-vae 监督的语义引导预训练 Flux-I，并在第 3.3.5 节给出了与 Tikhonov 正则等价的隐空间流形扰动机制，使隐式推理在突破标注瓶颈的同时提升鲁棒性。

第三，从多维度实证刻画了两类思维链范式的能力边界。本文在 Flux-min 与 Flux-max 两种规模下完成 25 余组实验，引入 DeepSeek-V4-Pro、Qwen3.6-35B-A3B 等通用大模型作为外部参照，得到的主要结论是：在机理标注稀缺的条件下，显式思维链以可读、可校验的中间步骤在准确率与机理一致性上明显占优；当数据规模扩大、监督信号变得密集后，隐式思维链能够突破自然语言标注瓶颈，在端到端产物预测与对抗背诵任务上反超无思考基线，并在叠加推理时流形扰动后保持稳定；在与千亿参数级通用大模型的对比中，本研究所提 0.6B 特化模型以远低于通用大模型的 token 开销，在机理思维链质量与端到端产物预测上达到

甚至局部反超千亿参数模型的水平，验证了“领域语料 + 结构改造”在科学推理任务上的实用价值。

5.2 工作不足

受限于研究周期与算力条件，本文工作仍存在若干不足，需在后续研究中进一步完善。其一，反应机理数据来源主要集中在以 FlowER 为代表的极性、自由基与周环反应类型上，数据分布在金属催化、光化学等机理类别上的覆盖仍不完整，在长链多步、强空间立体效应反应上的标注密度也较有限，可能导致模型在对应分布尾部上的泛化能力被高估。其二，本文所有模型均建立在单一 0.6B 基座之上，跨基座、跨参数量级的对照实验尚未展开，对于显式与隐式范式在更大基座上是否仍保持本文揭示的相对关系，目前缺少直接证据。其三，隐式思维链虽然通过 Flux-vae 与流形扰动机制在一定程度上缓解了“隐式流形偏移”，但其推理过程仍难以像显式思维链那样被逐步精读，这也是本文在实验中观察到的客观局限。

5.3 未来展望

围绕上述不足，本文从数据、模型与应用三个层面对未来工作做如下展望。在数据层面，可在现有处理管线之上引入金属催化、光化学等机理类别的专门解析规则，并结合人工与半自动审校建立面向化学专业的细粒度评测集，以更全面地刻画模型在不同机理子流形上的能力差异。在模型层面，一方面可在更大参数量级的基座上重复显式与隐式范式的对照实验，验证本文结论在跨规模条件下的稳定性；另一方面可探索显式与隐式两类范式的混合训练策略，例如在隐空间推理过程中以一定概率回到自然语言空间进行可读化输出，使模型在“高效隐式计算”与“可读显式审计”之间获得更细粒度的折中。在应用层面，可将本文构建的反应机理推理能力进一步与逆合成规划、反应条件推荐、实验路径搜索等下游任务相结合，为“AI for Science”场景下的轻量化垂直模型提供更具系统性的工程范式。综合而言，本文工作对推理式大语言模型在化学领域的应用提供了一个面向机理的实证起点，相关发现对在其他自然科学数据流形上探索显式与隐式思维链的边界问题亦具备一定的参考价值。

结束语

最初拿到这个课题时，我产生了一个很自然的想法：化学反应机理是链式的，而大语言模型的思维链也是链式的，两者在逻辑上高度契合。结合我曾在全国中学生化学奥林匹克竞赛（CChO）中获得省一等奖的学科背景，以及本科期间作为计算机专业学生发表过共同第一作者论文的科研经历，直觉告诉我，这个研究方向是切实可行的。

“价值、取舍、工程”是我在开展本研究时不可避免需要直面的三大考问。

关于“价值”：当下的研究往往都在追逐“最优”或“SOTA”，而本文探讨的隐空间推理模型，不可避免地需要与主流、成熟的大语言模型进行对比。在 AI 技术日新月异、各大模型榜单几乎每天都在刷新的今天，盲目与大厂比拼“刷榜”显然毫无意义，如何挖掘出具有实质性学术价值的研究成果，是本文面临的一大难点。为此，我阅读了大量文献，系统梳理了大语言模型的推理机制，最终确立了以比较显式与隐式思维链作为突破口的切入角度。

关于“取舍”：在实际研究过程中，我所运行的实验远不止正文中展示的这些，但许多探索性的尝试最终都留待了未来。这个过程让我深刻体会到一个朴素的道理：科研注定是在曲折与试错中不断前进的。

关于“工程”：大语言模型的训练与传统机器学习大不相同。我从在 AutoDL 上租赁算力开始，逐步掌握了数据收集与清洗、多卡环境部署、自定义模型结构与训练逻辑，直至跑通训练、推理、评测及结果汇总的全流程。其间，我遭遇了无数工程壁垒，如多卡进程死锁、混合精度配置、FlashAttention 的调用等，最终都将其逐一克服。这让我初步具备了驾驭大模型全栈研发的能力，也是面向未来科研所必备的基石。

受限于客观的算力与时间条件，本研究仍有许多未尽的篇章。在未来条件允许时，我将继续深化拓展，使之更加完善与成熟。

致谢

本文是在我的导师李宇楠老师的悉心指导下完成的。李老师前沿的专业视野和宽容的待人风范，是我顺利完成毕业设计的重要保障。对一些人而言，毕业设计可能是一场煎熬；但对我来说，它更像是一次深刻的激励。很庆幸，在导师给予的充分信任与支持下，我成为了后者。在此，谨向我的导师致以最诚挚的感谢。

此外，特别感谢我在科研实习期间的导师——香港中文大学的刘圣超老师对本文提供的诸多指导。刘老师严谨的治学态度和高瞻远瞩的科研视野，令我在学术道路上获益匪浅。

感谢曾与我并肩奋斗、共同发表一作论文的大师兄冯冠文，在我的本科科研生涯中，他给予了我关怀与实质性的帮助。

感谢我的父母钱仁道先生与黄凌云女士。生育之恩，养育之情，没齿难忘；是你们的默默付出，让我在充满爱的环境中长成。

感谢我的挚友们，我心本无住，遇君生烟火。

最后，感谢我自己。你四年前就已知道，你始终知道，来到的每一站终会离开，你言：应无所住，而生其心。

参考文献

[1] Bubeck, S., Chandrasekaran, V., Eldan, R., et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4[EB/OL]. arXiv:2303.12712, 2023.

[2] Wang, H., Fu, T., Du, Y., et al. Scientific discovery in the age of artificial intelligence[J]. Nature. 2023, 8, 620（7972）. 47–60.

[3] Wei, J., Wang, X., Schuurmans, D., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[C]//Advances in Neural Information Processing Systems 35 (NeurIPS), 2022: 24824–24837.

[4] Kojima, T., Gu, S. S., Reid, M., et al. Large Language Models are Zero-Shot Reasoners[C]//Advances in Neural Information Processing Systems 35 (NeurIPS), 2022: 22199–22213.

[5] Wang, X., Wei, J., Schuurmans, D., et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models[C]//International Conference on Learning Representations (ICLR), 2023.

[6] Yao, S., Yu, D., Zhao, J., et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models[C]//Advances in Neural Information Processing Systems 36 (NeurIPS), 2023: 11809–11822.

[7] Zelikman, E., Wu, Y., Mu, J., Goodman, N. D. STaR: Bootstrapping Reasoning with Reasoning[C]//Advances in Neural Information Processing Systems 35 (NeurIPS), 2022: 15476–15488.

[8] Magister, L. C., Mallinson, J., Adamek, J., et al. Teaching Small Language Models to Reason[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers, 2023: 1773–1781.

[9] Goyal, S., Ji, Z., Rawat, A. S., et al. Think before you speak: Training Language Models With Pause Tokens[C]//International Conference on Learning Representations (ICLR), 2024.

[10] Pfau, J., Merrill, W., Bowman, S. R. Let's Think Dot by Dot: Hidden Computation in Transformer Language Models[C]//Conference on Language Modeling (COLM), 2024.

[11] Deng, Y., Prasad, K., Fernandez, R., et al. Implicit Chain of Thought Reasoning via Knowledge Distillation[EB/OL]. arXiv:2311.01460, 2023.

[12] Hao, S., Sukhbaatar, S., Su, D., et al. Training Large Language Models to Reason in a Continuous Latent Space (Coconut)[EB/OL]. arXiv:2412.06769, 2024.

[13] Schwaller, P., Laino, T., Gaudin, T., et al. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction[J]. ACS Central Science. 2019, 9, 5（9）. 1572–1583.

[14] Irwin, R., Dimitriadis, S., He, J., Bjerrum, E. J. Chemformer: A Pre-Trained Transformer for Computational Chemistry[J]. Machine Learning: Science and Technology. 2022, 1, 3（1）. 015022.

[15] M. Bran, A., Cox, S., Schilter, O., et al. Augmenting large language models with chemistry tools[J]. Nature Machine Intelligence. 2024, 5, 6（5）. 525–535.

[16] Taylor, R., Kardas, M., Cucurull, G., et al. Galactica: A Large Language Model for Science[EB/OL]. arXiv:2211.09085, 2022.

[17] Tavakoli, M., Mood, A., Van Vranken, D., Baldi, P. RMechDB: A Public Database of Elementary Radical Reaction Steps[J]. Journal of Chemical Information and Modeling. 2023, 2, 63（4）. 1114–1123.

[18] Joung, J. F., Fong, M. H., Roh, J., et al. Reproducing reaction mechanisms with machine-learning models trained on a large-scale mechanistic dataset (FlowER)[EB/OL]. arXiv:2507.07775, 2025.

[19] Li, Y., Liu, S., Wang, J., et al. ChemCoTBench: Redefining Molecular Reasoning with Chemical Chain-of-Thought Benchmarking[EB/OL]. arXiv:2505.07551, 2025.

[20] Brown, B. C. A., Caterini, A. L., Ross, B. L., Cresswell, J. C., Loaiza-Ganem, G. Verifying the Union of Manifolds Hypothesis for Image Data[C]//International Conference on Learning Representations (ICLR), 2023.

[21] Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., Cresswell, J. C. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections[J/OL]. Transactions on Machine Learning Research. 2024, 9.

[22] Gurnee, W., Tegmark, M. Language Models Represent Space and Time[C]//International Conference on Learning Representations (ICLR), 2024.

[23] Marks, S., Tegmark, M. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets[C]//Conference on Language Modeling (COLM), 2024.

[24] Engels, J., Michaud, E. J., Liao, I., Gurnee, W., Tegmark, M. Not All Language Model Features Are One-Dimensionally Linear[C]//International Conference on Learning Representations (ICLR), 2025.

[25] Guo, D., Yang, D., Zhang, H., et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[EB/OL]. arXiv:2501.12948, 2025.

[26] Chen, Q., Qin, L., Liu, J., et al. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models[EB/OL]. arXiv:2503.09567, 2025.

[27] Geiping, J., McLeish, S., Jain, N., et al. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach[C]//Advances in Neural Information Processing Systems 38 (NeurIPS), 2025.

[28] Zhu, R.-J., Peng, T., Cheng, T., et al. A Survey on Latent Reasoning[EB/OL]. arXiv:2507.06203, 2025.

[29] Wigh, D. S., Arrowsmith, J., Pomberger, A., et al. ORDerly: Data Sets and Benchmarks for Chemical Reaction Data[J]. Journal of Chemical Information and Modeling. 2024, 5, 64（9）. 3790–3798.

[30] Landrum, G. RDKit: Open-Source Cheminformatics Software[EB/OL]. https://www.rdkit.org, 2024.

[31] Rein, D., Hou, B. L., Stickland, A. C., et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark[C]//Conference on Language Modeling (COLM), 2024.

[32] Wang, X., Hu, Z., Lu, P., et al. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models[C]//Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235, 2024: 50622–50649.

[33] Mirza, A., Alampara, N., Kunchapu, S., et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists[J]. Nature Chemistry. 2025, 7, 17（7）. 1027–1034.

[34] Bishop, C. M. Training with Noise is Equivalent to Tikhonov Regularization[J]. Neural Computation. 1995, 1, 7（1）. 108–116.

[35] Camuto, A., Willetts, M., Şimşekli, U., Roberts, S. J., Holmes, C. C. Explicit Regularisation in Gaussian Noise Injections[C]//Advances in Neural Information Processing Systems 33 (NeurIPS), 2020.

[36] Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., Wang, L. Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective[C]//Advances in Neural Information Processing Systems 36 (NeurIPS), 2023.

[37] Merrill, W., Sabharwal, A. The Expressive Power of Transformers with Chain of Thought[C]//International Conference on Learning Representations (ICLR), 2024.

[38] Kang, H., Zhang, Y., Kuang, N. L., et al. LaDiR: Latent Diffusion Enhances LLMs for Text Reasoning[EB/OL]. arXiv:2510.04573, 2025.