

AI 工程学：面向智能任务可靠交付的学科纲领

——能力结构、可靠性关系与不确定性治理

作者：刘明

版本：中文正式稿 V1.0

日期：2026 年 8 月

摘要

大模型及其 Agent 系统已经能够生成代码、规划任务、调用工具并修改真实工程状态，但能力增长尚未同步转化为复杂任务的可靠交付。现有研究已分别从模型能力、上下文、检索增强、Agent 架构、软件工程、评测、可靠性与安全保证等方向处理相关问题，但这些成果之间仍存在一个需要进一步组织的问题：如何以智能任务的可靠交付为统一轴，将能力结构、任务—能力匹配、过程观测和工程干预连接起来。

本文在 KSC、SCR 和多轮审计收敛（Multi-Round Audit Convergence, MRAC）三项前置研究基础上，提出 AI 工程学的一套初始学科纲领。本文将 AI 工程学的核心研究对象定义为：**智能系统在开放、复杂且约束持续存在的任务中实现可靠交付的条件、机制、方法与基础设施**。在这套初始基础架构中，KSC 作为智能能力抽象框架，描述智能任务所需的能力结构；SCR 作为交付可靠性的关系理论，解释有效搜索空间、覆盖能力与交付可靠性之间的关系，并导出搜索空间治理这一主要工程方向；MRAC 则作为覆盖状态的过程观测机制，在缺少完备外部验证器时，通过多轮审计—修复轨迹提供二阶信息。

围绕智能任务的可靠交付，AI 工程学至少需要处理三个无法直接绕过的基础问题：**能力限制发生在哪里，什么条件下有限能力能够形成可靠交付，以及在覆盖边界无法被直接完整测量时，如何通过过程证据判断系统是否处于可靠处理范围**。KSC、SCR 与 MRAC 分别提供能力坐标、可靠性关系和过程观测机制。

AI 工程学通过持续积累证据识别能力边界和治理效果，并在任务接近或超出覆盖边界时，通过比较、回退和重新拆分恢复可靠性。一个重要的远期目标，是逐步构建能够在明确任务域内、以较低人工介入承担复杂项目交付的项目级闭环，并持续扩展其覆盖边界，即向工程 AGI 逼近 [5]。

关键词：AI 工程学；KSC；SCR；搜索空间治理；多轮审计收敛；MRAC；交付可靠性；工程 AGI；不确定性治理

1. 引言

生成式 AI 带来的根本变化，不只是软件系统新增了一类概率性组件，而是具有知识调用、结构生成、规划和行动能力的智能主体开始直接参与工程活动。AI 已经能够解释需求、选择方案、修改代码、调用工具和运行测试，部分工程认知过程由此开始由 AI 承担。

然而，局部正确、合理计划或通过当前测试，都不足以保证系统能够在复杂状态和持续约束下稳定完成任务。问题因此从“模型能否生成正确结果”进一步转向“系统能否在失败可发现、状态可恢复的条件下持续完成复杂任务”。近期 Agent 可靠性研究也开始把一致性、鲁棒

性、可预测性和安全性与单一准确率分开测量，说明能力或准确率提升并不能自动代表可靠性同步提升 [14]。

这提出了一个不能完全由模型能力研究或现有工程工具单独解决的问题：

在没有完备外部验证器、任务约束不能完全形式化、系统状态持续变化的条件下，如何观察、干预并逐步提高智能系统的交付可靠性？

本文将这一问题视为 AI 工程学的一条核心研究主线，并在 KSC、SCR 和 MRAC 三项前置研究基础上提出初始学科纲领。各项基础成果的完整论证由相应论文承担，这里着重说明它们在研究纲领中的位置，以及由此产生的工程范式、证据需求、研究议程和远期目标。

2. AI 工程学的研究对象与学科缺口

2.1 从生成能力到交付可靠性

单轮答案质量、Benchmark 得分和有限任务成功率可以证明系统拥有某种能力，却不足以说明它能够承担真实复杂项目。复杂任务中的约束通常分布在 Spec、代码、文档、依赖、历史决策和运行环境中；中间结果会改变后续状态，局部正确也不能自动推出整体正确。软件工程长期关注复杂性、维护和系统演化问题 [7]；机器学习系统研究进一步指出，模型之外的数据依赖、配置、耦合和技术债会形成持续的系统级工程风险 [8][9]。这些工作说明，局部模型能力或局部任务效果不足以替代系统层面的工程治理。

因此，需要区分生成能力、单个任务的完成能力和长期可靠交付能力。AI 工程学主要研究最后一种能力：如何通过工程治理，使 AI 已有的生成和任务执行能力转化为可持续的可靠交付。SWE-bench 已经使用真实 GitHub issues 及其对应 PR 构造仓库级软件工程任务 [10]；长任务能力研究进一步尝试用人类专家完成任务所需时间来描述 AI 可自主完成任务的尺度，并显示任务尺度本身是理解系统能力边界的重要维度 [15]。但从单个任务完成走向长期项目闭环，仍需要更完整的状态、约束、依赖、恢复和证据治理。

交付可靠性不同于单次正确性，也不是脱离条件的静态成功率。它关注系统在给定任务分布和工程条件下，能否持续产生可接受结果，并能否发现、隔离、修复和恢复失败。因此，AI 工程学不仅研究系统是否能够完成任务，还研究可靠性在什么条件下产生、为何变化，以及工程干预能否改善这种可靠性。

2.2 本文采用的学科化判断标准

本文不试图给出所有新学科成立的普遍最低条件，而采用一组适用于本研究纲领的判断标准：相对独立的研究对象，能够组织关键现象的基础关系，可实施的干预对象，可持续积累的外部证据，以及允许基础主张被检验、修正或替代的研究路线。

按照这一标准，智能任务的交付不确定性及其治理已经形成一个可以独立组织的研究对象。任务表达、上下文、工程状态、边界、验证和执行过程提供了干预对象；任务条件、执行轨迹、失败和真实反馈提供了证据来源。这一研究方向能否进一步成熟为稳定学科体系，仍取决于其基础框架能否解释关键问题，并在持续的经验研究中接受检验。

2.3 名称、边界与当前研究范围

“AI Engineering”并非本文首次提出。Carnegie Mellon University Software Engineering Institute (SEI) 在 2021 年已明确将 AI Engineering 描述为一项新兴学科，并以人本、鲁棒与安全、可扩展等方向组织相关研究和实践 [13]。本文既不主张名称首创，也不主张此前相关研究不存在，而是提出一个以复杂智能任务可靠交付为统一组织轴的更明确研究纲领：

AI 工程学研究智能系统可靠完成真实复杂任务的条件、机制、方法、证据与基础设施。

AI 工程学的边界不由应用领域决定，而由问题结构决定。凡是 AI 作为认知或行动主体参与复杂目标交付，任务具有多步骤、状态演化、持续约束、不完备验证或失败恢复需求，并需要通过外部工程结构提高可靠性，都可以成为 AI 工程学的研究对象。单纯的模型训练、一次性生成、普通功能集成或无法形成可观测交付过程的活动，并不因使用 AI 就自动属于这一研究范围。

现阶段，软件工程是 AI 工程学的主要研究入口。代码、Spec、测试、版本控制、差异比较和回退提供了相对完备的外部治理与证据基础。以软件工程作为初级阶段的主要研究对象，是由当前研究条件决定的，而不是对学科长期边界的限制。科研、产品设计、管理和其他复杂目标能否进入这一研究范围，需要根据其目标、状态、约束、过程和证据能否被外部化和治理而分别验证。

3. 知识来源、层级与开放吸收

3.1 既有工程基础与 AI 新增问题

AI 工程学并不是从零开始。设计科学、软件工程、系统与安全工程以及 AI 风险管理等领域已经提供了大量可继承知识 [6][7][11][12]，分别提供关于复杂设计、软件复杂性、系统安全和风险治理的成熟视角。

这些成熟知识在 AI 系统中继续有效，并不自动意味着出现了新的 AI 工程学成果。只有当 AI 作为具有生成性、概率性和部分自治能力的工程参与者，使原有研究对象、关系或适用条件发生实质变化时，才产生新的 AI 工程问题。例如，通过分解和层次结构管理复杂性属于既有工程思想 [6][7]；任务或模块边界如何改变 AI 面对的有效搜索空间，并使当前系统由无法可靠覆盖转向可以可靠覆盖，则是 AI 环境下需要重新研究的问题。

3.2 学科知识的不同层级

AI 工程学同时包含基础框架与理论、局部机制、工程方法、工具和经验性证据。这些知识都可能影响可靠交付，但承担不同作用，不能作为同等级的基础知识处理。

一项研究要成为学科的候选基础构件，至少应直接涉及智能任务的可靠交付，揭示 AI 环境带来的新抽象、关系或机制，具有跨现实意义，并允许外部证据修改其适用范围。局部机制、方法、工具和经验即使未达到这些条件，也仍然具有重要价值，只是处于不同的知识层级。

3.3 开放且可修正的初始基础架构

初始基础架构由三类构件组成：KSC 是智能能力的抽象框架，SCR 是交付可靠性的关系理论，MRAC 是覆盖状态的过程观测机制。搜索空间治理是 SCR 的工程推导；MRAC L1—L5 则进一

步讨论 MRAC 的论证层次、证据边界和研究路线 [4]。

这一架构保持开放。外部框架、理论或机制无需使用本文术语，只要能够更有效地解释智能任务的可靠交付，就应进入比较。未来研究可以补充或替代能力框架，修正 SCR 关系，提出其他基础关系、工程方向或观测机制。AI 工程学应始终以研究对象为中心，而不是以作者、术语或特定技术路线限定知识来源。

4. 初始基础架构的三个构件

本章只给出三个构件在学科纲领中所需的必要定义和位置。其详细推导、边界和经验命题由相应独立论文承担。

4.1 KSC：智能能力的抽象框架

KSC 以三种递进能力描述智能任务 [2]：

- **K (Knowledge, 知识)**：能够在任务中被系统有效调用的信息、经验、规则、事实和表示；
- **S (Structural Generation, 结构生成)**：基于已有知识形成推理、解释、计划、设计、代码或其他新结构的能力；
- **C (Constraint Maintenance, 约束保持)**：使结构生成持续服从目标、边界、状态和约束的能力。

为避免与 SCR 中的 Coverage Ability（覆盖能力）混淆，本文在需要使用字母缩写时，将约束保持写作 **KSC-C**，将 SCR 中的覆盖能力写作 **SCR-C**。

三者可以理解为信息存在、基于信息形成新结构，以及对结构生成实施控制。只有 K 而没有 S，信息不能参与生成；有 K 和 S 而缺少足够的约束保持（KSC-C），系统可以产生大量局部有效结构，却不能保证这些结构持续适合当前任务。约束保持不只指长上下文中的记忆，还包括目标选择、相关性判断、边界控制、约束协调、错误识别和停止条件。

KSC 是一种智能能力抽象框架和能力坐标系，而不是因果理论或操作指南，也不要求与模型内部模块一一对应。它把笼统的“AI 能力不足”分解到知识、结构生成和约束保持三个可以讨论的位置。按照 KSC 的分析，AI 在知识调用和结构生成方面已表现出较强能力，而复杂任务中的约束保持仍经常需要人类或外部工程结构补足 [2]，这构成了人机分工和 AI 工程治理的重要基础。KSC 回答智能任务依靠什么能力完成，但不能单独判断某项任务能否可靠交付；后者还取决于任务规模、工程状态和治理条件。

4.2 SCR：交付可靠性的关系理论

SCR 以三个变量描述 AI 任务中的核心工程关系 [3]：

- **有效搜索空间 (Search Space, S)**：在任务表达、工程状态和治理条件下，系统需要面对的相关解释、决策、路径、实现和失败可能性的集合；
- **覆盖能力 (Coverage Ability, C; 本文记为 SCR-C)**：模型、Agent、工具、上下文和验证机制对相关空间进行探索、排除、验证和修复的综合能力；
- **交付可靠性 (Delivery Reliability, R)**：系统在给定任务分布和工程条件下持续产生可接受交付并处理失败的能力。

有效搜索空间不是上下文窗口，也不要求能够被完整枚举；覆盖能力（SCR-C）也不是模型固定不变的全局属性，而是相对于具体任务存在的系统属性。SCR 首先表达一种方向性关系：在其他条件近似不变时，有效搜索空间扩大通常降低交付可靠性，覆盖能力增强通常提高交付可靠性。工程系统既可以增强覆盖能力，也可以通过缩小、分割、显化或稳定有效搜索空间，使既有能力更可靠地完成任务。

SCR 解释为什么系统拥有相关能力仍可能交付失败，以及为什么相同模型在不同工程状态下表现不同。它把能力研究连接到工程结果，并导出搜索空间治理这一主要工程方向。KSC 与 SCR 处于不同分析层次：前者提供能力抽象，后者解释任务条件、系统能力与交付结果之间的关系。

4.3 MRAC：覆盖状态的过程观测机制

许多复杂任务缺少完备验证器。一个结果通过测试，只能说明它通过了当前测试；一轮审计没有发现问题，也不能证明不存在共同盲区。因此，AI 工程学还需要利用多次检查、修复和重新执行中的变化，获得单次结果无法提供的二阶信息。

MRAC 据此提出一种过程观测机制：通过多轮审计—修复过程，观察新增有效问题是否趋于减少、已修复问题是否复发，以及问题发现是否逐渐趋于稳定 [1]。在任务对象、关键约束和审计协议保持稳定，审计意见有效且过程没有目标漂移的条件下，如果高影响问题持续出现，就强烈表明当前系统还不能在这一任务边界内实现可靠覆盖。收敛则提供强于单次成功的支持性证据，但不能直接等同于正确性或完整覆盖。

MRAC 不取代测试、Eval 或真实运行反馈，而是把多轮问题轨迹转化为继续、停止、回退或重新拆分的依据。其不收敛诊断、收敛证据、拆分后收敛和递归研究路线，已在 MRAC L1—L5 研究中展开 [4]。

5. 核心工程范式：搜索空间治理

5.1 从能力竞争到任务—能力匹配

生成式 AI 进入工程后，工程系统不仅需要治理确定性程序，还需要处理智能主体在语义理解、方案选择、状态修改和验证过程中面对的大量可能路径。

人类难以逐步规定 AI 的全部内部路径，也不应把 AI 退化为只能执行完全确定指令的工具。更可行的方向是治理外部条件：让有效路径更容易被发现，让无效路径更早被排除，使关键约束能够持续存在，并在任务超出当前覆盖能力时及时识别并重新组织任务。

本文把这类活动称为**搜索空间治理（Search Space Governance）**。它不是与三个基础构件并列的第四项基础成果，而是 SCR 从关系解释走向工程干预的直接推导。

5.2 治理对象与两类主要路径

搜索空间治理主要面向四类对象：任务治理负责目标、范围、约束和验收；状态治理负责代码、文档、数据、环境、依赖和历史决策；过程治理负责任务拆分、顺序、检查点、回退和人工介入；证据治理负责测试、审计、轨迹、差异和失败分类。

提高交付可靠性可以沿两类主要工程路径进行：增强覆盖能力，或者治理有效搜索空间。模型扩展、推理、工具调用和 Agent 改进主要扩大系统能够处理的范围；Spec、模块化、工程

状态治理、任务拆分和回退主要限制或重构系统需要处理的范围。具体方法可能同时作用于两侧，其实际效果应由任务条件和外部证据判断，而不宜固化为一一对应的变量映射。

这一分类用于组织可靠交付的工程干预。安全、责任、授权、成本等仍可作为独立治理目标；它们与搜索空间和覆盖能力之间的关系，需要进一步研究。

5.3 面向不确定性的工程治理

AI 工程学不要求消除一切不确定性，而是把不确定性限制在可治理的范围内。关键约束需要得到保持，失败需要能够被发现；超出覆盖能力的任务，则应能够被拆分、回退或转交。已经建立的可靠交付能力，还应随着工程状态和治理经验的积累得到保持和提升。

长期工程可靠性并不来自 AI 的每一步都变成确定性执行，而是来自另一种条件：随着项目复杂度增长，新任务面对的有效搜索空间仍然可控，失败始终处于可以发现和恢复的工程边界内。这构成 AI 工程学相对于单纯模型能力竞争的独立工程价值。

6. 证据驱动的 AI 工程过程

6.1 证据为什么具有基础地位

有效搜索空间和覆盖能力目前都难以被完整测量，因此单一 Benchmark、一次成功或一项静态能力结论不足以确定系统的可靠工作边界。

AI 工程学不能等待所有变量完全形式化后再开始研究。更现实的路径是先建立方向性关系，再通过跨任务、跨系统和长期项目证据，逐步识别这些关系的成立条件、作用强度和失效边界。证据不是基础主张提出后的附加验证，而是学科知识继续发展的基本方式。

6.2 从单路径执行到可比较过程

当任务接近或超过当前系统的覆盖边界，并且不存在完备验证器时，复杂任务未必能沿一条预先确定的路径稳定完成。系统需要比较候选路径，拒绝依据不足的方案，回退错误状态，或者重新拆分任务，使任务边界重新与当前覆盖能力匹配。

这些行为不仅用于处理失败，也会产生新的工程信息。某条路径被淘汰，说明它在当前条件下缺乏继续投入的依据；发生回退，说明当前状态不值得继续保留；任务拆分后重新出现收敛，则支持原任务边界与覆盖能力不匹配的判断。因此，面向可靠交付的 AI 工程系统不仅需要使 Agent 能够执行，还需要使候选路径能够被比较和否定、错误状态能够被回退，并保留已经验证有效的结果。

6.3 证据方向与当前边界

AI 工程证据需要回答四类基本问题：可靠性变化是否与某种条件或治理干预稳定相关；系统和方法在什么边界内有效；变化应归因于模型、Agent、任务还是治理；结论能否跨任务、项目和团队迁移。失败、不收敛、路径淘汰和真实缺陷，往往比单次成功更能揭示能力边界。

以证据持续检验工程判断，是 AI 工程学的基本研究方向，而不是一套已经完成的统一标准。本文不预设证据 Schema、综合分数或固定协议。具体指标和基础设施应由真实项目与长期证据逐步推动。

7. 基础问题的结构必要性

本章讨论问题位置的结构必要性：一旦 AI 工程学以智能任务的可靠交付为研究目标，至少会自然产生三个连续且无法直接绕过的基础问题。这里不讨论具体框架是否不可替代，也不证明基础问题集合已经完备。

7.1 从可靠交付目标推出至少三个必要问题

AI 工程学以智能系统在复杂任务中的可靠交付为中心。这个目标本身会约束基础研究必须回答什么，而不是由既有理论反向决定学科问题。

首先是**能力边界问题**：智能系统具有什么能力，限制发生在哪里，以及哪些限制需要由人类或外部系统补足。工程治理之所以必要，首先因为智能系统的能力有限，而且不同能力之间并不均衡。如果不能稳定描述能力结构，就难以解释 AI 为什么在某些任务中失败，也难以明确人机分工和工程治理需要补足什么。

但能力描述本身不能推出可靠交付，由此产生第二个问题：**在既定能力条件下，什么样的任务与工程条件能够形成可靠交付**。同一系统面对不同任务规模、有效搜索空间、工程状态和治理方式时，交付可靠性可能显著不同。因此，仅仅知道“系统拥有什么能力”还不够，还必须解释有限能力在什么条件下足以完成任务。

即使建立了这种可靠性关系，工程实践仍面临第三个问题：**当覆盖边界无法被直接完整测量时，如何判断系统当前是否处于可靠处理范围**。许多真实复杂任务缺少完备验证器，有效搜索空间和覆盖状态也难以直接测量，因此需要从执行、验证和修复过程中获得额外证据。

三个问题由此形成连续递进：**能力描述不能直接推出可靠性，因此需要研究可靠交付条件；可靠性关系又不能被直接完整观测，因此需要过程观测**。这一逻辑可以概括为：

有限能力 → 条件性可靠 → 不完全可观测

因此，能力边界、可靠交付条件和覆盖状态观测，不是为了容纳已有理论而人为划分的三个类别，而是从“可靠交付”这一学科目标逐层产生的至少三个必要问题。

7.2 当前基础构件的位置

基础问题	当前基础构件	当前回答
能力边界如何描述	KSC	提供能力坐标，定位知识、结构生成和约束保持上的能力限制及人机互补关系
有限能力何时能够可靠交付	SCR	解释有效搜索空间、覆盖能力与交付可靠性之间的关系
覆盖状态如何判断	MRAC	利用多轮审计—修复轨迹获得覆盖状态的二阶证据

在这套初始架构中，KSC、SCR 和 MRAC 分别对应能力边界、可靠性条件和工程观测，形成从能力描述到可靠交付再到过程判断的连续结构。

7.3 必要的是问题位置，而不是具体理论

学科目标约束的是必须回答的问题，而不是必须采用的答案。未来研究完全可能提出比 KSC 更有效的能力框架，修正或替代 SCR 的关系理论，也可能发展出比 MRAC 更有效的覆盖状态观测机制。只要 AI 工程学仍以有限智能系统的可靠交付为研究对象，能力限制、可靠交付条件和覆盖状态观测就仍需要以某种形式被回答。

KSC、SCR 和 MRAC 来自同一研究线，具有较强的结构一致性，也因此需要更多独立研究进行比较和检验。判断未来基础成果是否应进入 AI 工程学，不应取决于是否继承本文术语，而应取决于它能否更好地回答基础问题、解释可靠交付，并接受外部证据的持续检验。

8. 研究议程与工程 AGI

8.1 初始知识域

AI 工程学的初始知识域至少包括：能力结构与系统边界；任务、约束与有效搜索空间；知识、上下文和工程状态治理；模块边界、任务拆分与依赖编排；测试、Eval、审计和真实运行观测；回退、恢复与人机责任边界；交付可靠性及可复现证据；项目级闭环与工程 AGI。

这些知识域不是最终课程目录，而是一版围绕可靠交付组织研究的基础范围。它们使模型、Agent、软件工程、评测和组织治理研究能够围绕同一类问题展开协作。

8.2 核心研究方向

基础研究首先需要检验 KSC、SCR 和 MRAC 的适用范围，并与替代框架、理论和观测机制比较。工程研究需要识别 Spec、上下文、RAG、工程状态、模块化、任务拆分和回退等方法在什么条件下改善可靠性。基础设施研究则应支持任务条件、工程状态、执行轨迹、失败和交付结果的长期积累。

SCR 的实证研究尤其需要避免定义性循环：有效搜索空间、覆盖能力和交付可靠性应尽可能使用相互独立的代理指标、干预变量或实验设计进行操作化，再检验三者之间的关系。MRAC 则需要持续研究审计协议、模型能力、共同盲区、修复漂移和外部验证对收敛解释力的影响。

研究重点应从“系统是否曾经做对”扩展到“它在什么条件下能够可靠做对”，并逐步形成跨任务、跨系统和跨项目可比较的证据。

8.3 从可靠交付到工程 AGI

AI 工程学的近期目标，是提高单个智能任务在明确条件下的交付可靠性；进一步的目标，是把这种可靠性扩展到跨任务、跨阶段和持续变化的项目级闭环。本文采用前置研究提出的工程 AGI 工作定义 [5]：一个重要的远期目标，是使 AI 能够在低人工介入下，持续承担原本需要普通人类团队长期协作才能完成的复杂项目闭环，即向工程 AGI 逼近。

工程 AGI 更适合被理解为持续扩展的覆盖边界，而不是只能以“达到”或“未达到”描述的单一终点 [5]。系统可以先在边界清楚、规模有限的任务域中形成低人工介入的项目闭环，再逐步扩大任务类型、项目规模、持续时间和环境开放程度。在受限任务域内形成此类闭环，可以视为对工程 AGI 的局部逼近，而不等同于已经实现一般意义上的 AGI。

形成闭环后，工程研究仍需提高可靠性，减少无效路径、重复审计和回退，降低时间、算力与组织成本，并把闭环能力迁移到更广泛的问题。KSC 描述闭环所需的能力结构，SCR 描述可以可靠覆盖的任务边界，MRAC 则提供发现和扩展这一边界的过程信号。这里的低人工介入或无人化，指项目执行不再依赖人类持续参与，但不排除人类继续负责目标、授权、责任和高风险边界的治理。

9. 相邻学科、局限与发展方向

9.1 与相邻学科的关系

AI 工程学不取代机器学习、软件工程、系统工程、MLOps、Agent、AI 安全、人机协同或可靠性研究等领域，而是在这些成果之上组织一个共同问题：已有能力和工程方法如何在真实复杂任务中形成可靠交付。

因此，AI 工程学更适合作为连接这些领域的交叉研究层次，而不应被缩减为“如何使用当前 AI 编程工具”。工具会快速变化，任务搜索空间、覆盖边界、约束保持、失败发现和可靠交付则是更稳定的研究对象。Agent 可靠性和长任务能力等相邻研究也可以作为这一体系的证据来源 [14][15]。

9.2 当前局限

本文仍处于学科纲领提出而非成熟学科阶段，存在以下限制：

1. 有效搜索空间难以直接枚举或精确测量，覆盖能力也高度依赖任务、工程状态和协议，不适合压缩为单一全局指标；
2. SCR 中的有效搜索空间、覆盖能力与交付可靠性仍需发展相互独立的代理指标和干预设计，以避免循环解释；
3. KSC、SCR 和 MRAC 主要来自同一研究线，需要更多独立研究比较其构件、关系与适用边界；
4. 现有经验主要来自 AI 软件开发，其他领域未必能够直接复用相同治理方法；
5. MRAC 及其他过程观测机制依赖验证能力，验证器共同盲区可能制造虚假稳定。

9.3 跨领域扩展

跨领域扩展不取决于领域名称，而取决于目标、状态、约束、过程和证据能否形成可治理的外部结构。随着独立证据积累，相关研究可以逐步扩展到科研、管理和其他开放智能任务，并检验软件工程之外是否存在不同的验证、价值与责任边界。

10. 结论

AI 工程学把智能任务中的交付不确定性及其治理作为独立研究对象，研究有限智能能力如何在复杂状态、持续约束和不完备验证条件下转化为可靠结果。

围绕这一目标，至少会出现三个无法直接绕过的基础问题：能力边界如何描述，有限能力在什么条件下形成可靠交付，以及覆盖状态难以直接测量时如何获得过程证据。KSC、SCR 和

MRAC 分别提供能力框架、可靠性关系和过程观测机制；搜索空间治理进一步把可靠性关系转化为工程干预，使任务、状态和过程治理成为增强模型能力之外的另一类主要路径。

这是一套仍需跨模型、跨 Agent、跨项目和跨团队证据持续检验与修正的初始架构。其长期方向，是把可靠交付从单个任务扩展到持续的项目级闭环，并逐步扩大 AI 能够低人工介入承担复杂项目交付的覆盖范围，持续向工程 AGI 逼近。

参考文献

注：[1]—[5] 为本文作者的前置研究，其中 [5] 给出本文采用的工程 AGI 工作定义；其余文献用于说明相关工程传统、既有 AI Engineering 语境、Agent 可靠性与长任务能力研究背景。

- [1] Liu, M. (2026a). *Multi-Round Audit Convergence (MRAC)* [Preprint]. Zenodo. doi:10.5281/zenodo.21804751
- [2] Liu, M. (2026b). *KSC: A Minimal Framework for Understanding AI Capability and Human-AI Collaboration* [Preprint]. Zenodo. doi:10.5281/zenodo.21834579
- [3] Liu, M. (2026c). *Search Space Governance in AI Engineering* [Preprint]. Zenodo. doi:10.5281/zenodo.21902420
- [4] Liu, M. (2026d). *MRAC from L1 to L5: Evidence Boundaries and Research Roadmap* [Preprint]. Zenodo. doi:10.5281/zenodo.21917629
- [5] Liu, M. (2026e). *An Engineering Definition of AGI: Complex-Project Closed-Loop Capability as an Assessment Criterion* [Preprint]. Zenodo. doi:10.5281/zenodo.21927623
- [6] Simon, H. A. (1969). *The Sciences of the Artificial*. MIT Press.
- [7] Brooks, F. P., Jr. (1987). No Silver Bullet: Essence and Accidents of Software Engineering. *Computer*, 20(4), 10–19. doi:10.1109/MC.1987.1663532
- [8] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28.
- [9] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H. C., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 291–300). doi:10.1109/ICSE-SEIP.2019.00042
- [10] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.
- [11] Leveson, N. G. (2012). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.
- [12] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. doi:10.6028/NIST.AI.100-1

[13] Software Engineering Institute. (2021). *AI Engineering: An Emergent Discipline for Human-Centered, Robust and Secure, and Scalable AI* [Brochure]. Carnegie Mellon University.

[14] Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S., & Narayanan, A. (2026). *Towards a Science of AI Agent Reliability*. Proceedings of the 43rd International Conference on Machine Learning (ICML 2026). arXiv:2602.16666.

[15] Kwa, T., West, B., Becker, J., et al. (2025). *Measuring AI Ability to Complete Long Software Tasks*. Advances in Neural Information Processing Systems (NeurIPS 2025). arXiv:2503.14499.

附录 A：核心术语表

中文术语	英文术语	含义
AI 工程学	AI Engineering	研究智能系统可靠完成真实复杂任务的条件、机制、方法、证据与基础设施
知识	Knowledge (KSC-K)	系统能够在任务中有效调用的信息与经验
结构生成	Structural Generation (KSC-S)	基于知识形成推理、计划、设计、代码等新结构的能力
约束保持	Constraint Maintenance (KSC-C)	使生成持续服从目标、边界、状态和约束的能力
有效搜索空间	Effective Search Space (SCR-S)	当前任务与工程条件下需要面对的相关解释、路径、实现和失败可能性
覆盖能力	Coverage Ability (SCR-C)	模型、Agent、工具与治理系统探索、排除、验证和修复相关空间的综合能力
交付可靠性	Delivery Reliability (SCR-R)	在给定任务分布与工程条件下持续产生可接受交付并处理失败的能力
搜索空间治理	Search Space Governance	干预任务、状态、过程和证据，使搜索空间与覆盖能力重新匹配
多轮审计收敛	Multi-Round Audit Convergence (MRAC)	通过多轮审计—修复轨迹观察问题发现是否趋于稳定的机制
二阶观测	Second-order Observation	通过多次验证、修复或运行结果的变化获得单次结果不能提供的状态信息
工程 AGI	Engineering AGI	AI 在低人工介入下持续承担原本需要普通人类团队长期协作的复杂项目闭环；在受限任务域内形成此类闭环可视为对工程 AGI 的局部逼近