从 AI Agent 到 Agentic Workflow，25 篇论文全面了解智能体工作流

王吉伟

摘要：

• 25 篇关于智能体工作流的论文，关注 Agentic workflow 不要错过

• 想要了解智能体工作流，一定不要错过这 25 篇 LLM 及 workflow 相关论文

• 从 LLM 到 AI Agent 再到 workflow，25 篇论文#深度好文计划#全面了解智能体工作流

• 从架构到系统，从基准到方法论，6 大类 25 篇论文助你吃透智能体工作流

• 想要系统了解智能体工作流，看这 25 篇论文就够了

• 什么是智能体工作流？Agentic workflow 有哪些系统和工具？一篇文章看明白

著名AI学者、斯坦福大学教授吴恩达提出了AI Agent的四种设计方式后，Agentic Workflow（智能体工作流）立即火爆全球，多个行业都在实践智能体工作流的应用，并推动了新的Agentic AI探索热潮。

技术的发展与应用已经进入新的拐点，从大语言模型（Large Language Models，LLM）到AI Agent再到Agentic workflow，这些新的技术一经出现便得到快速应用。而AI Agent和Agentic workflow作为LLM的落地应用方式，鉴于它在各种场景的普适性和灵活性，其普及速度比我们想的快很多。

比如在4月份文心智能体平台就已汇聚超5万开发者，创建智能体超过3万，还有30万创作者在文心一言APP创建了智能体，上线了40万个功能丰富的智能体，智能体调用量达8亿。在Coze平台，单是构建智能体能调用的插件就已超过100个。

王吉伟频道盘点过的80多个AI Agent构建平台中，有很多平台已经有不少用户和数量可观的智能体。其中，OpenAI的GPTs数量在今年1月份就已经超过300万个。

扩展阅读：AI智能体构建智能未来，全球80+AI Agent构建平台大盘点

智能体来势汹汹，已经引起很多人的担心。比如《互联网的未来》一书的作者哈佛大学法学院教授 Jonathan Zittrain 就已在《The Atlantic》 杂志上发文，他认为当智能体形成数百万量级的庞大生态时，其行为可能不受控制，进而对人类社会产生重大危害，所以应该立即对智能体的行为进行规范，并改进现有互联网标准，从而更好地控制智能体，防止它们失控。

文章链接：
https://www.theatlantic.com/technology/archive/2024/07/ai-agents-safety-risks/678864/

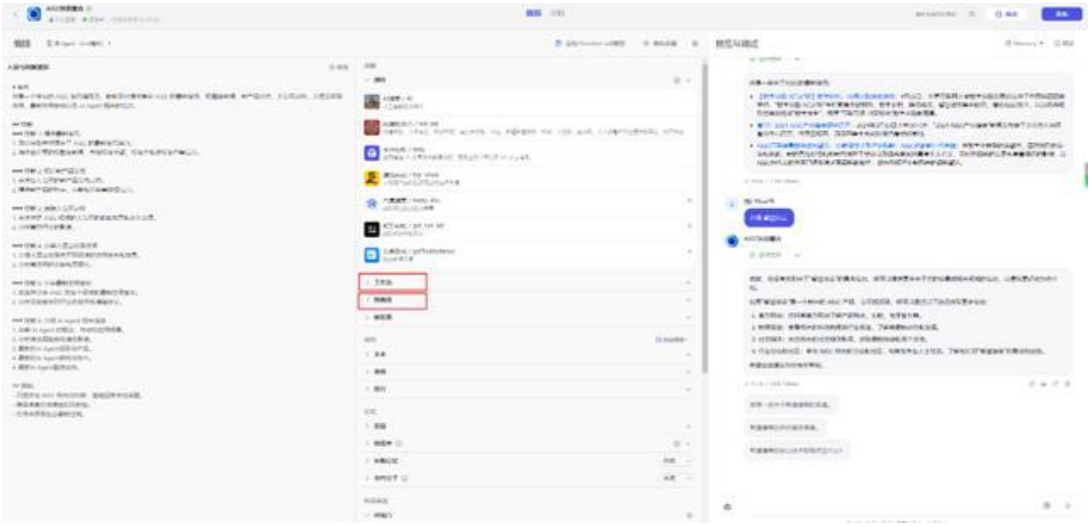
这篇文章，也从侧面见证了 Agentic workflow 的野蛮生长。

虽然 Agentic Workflow 已获得惊人的进展，但业内外对其认知还存在一定的偏差。

吴恩达教授在介绍 Agentic Workflow 时，认为它是与 LLM 交互和完成任务的一种方法，可以将任务分解成多个步骤，在不同环节进

行迭代，指导最终生成期望的结果。并将 Agentic Workflow 的设计模式总结为反思、工具使用、规划和多智能体协作四种。

扩展阅读：Agentic Workflow 加速 Agentic AI 到来，AI Agent 成为重要实现方式

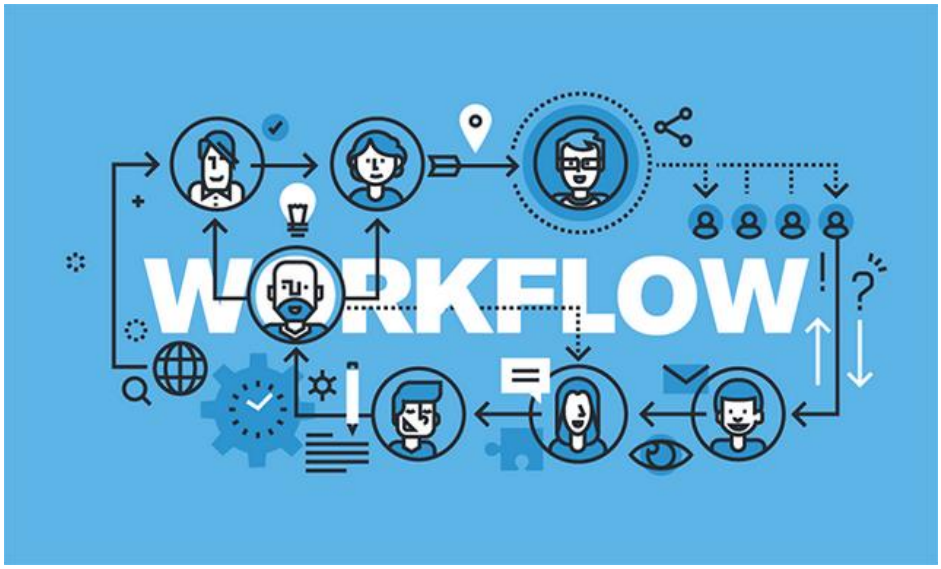


Coze 的 Bot 构建页面

但在实际应用中，我们经常会看到智能体工作流的应用模式远不止这四种模式。比如 Coze 不只推出了多智能体和工作流功能，还衍生出了图像流。

而最终通过插件、大模型、代码、知识库、工作流、图像流、选择器、文本处理、消息、变量、数据库等构建的工作流，又会被置入「技能」模块而最终构建成为一个智能体（Coze 平台称之为 Bot）。更多的智能体，可以执行更多的任务，参与相对复杂的业务流程。

还有，如果仔细观察你会发现，在 LLM 应用越发普及化的前提下，很多工作流都是混合了传统业务流程与智能体工作流。其中既有“四种模式”的工作流，也有传统应用嵌入 GenAI 的工作流，还有简单的直接应用大语言模型的工作流。



一个典型的案例就是，目前通过 AI Agent 构建平台构架的智能体工作流尚无法完成操作企业管理软件等复杂业务流程（受 API 及连结能力限制），而通过 RPA 等超自动化工具连结更多的简单智能体工作流就是不错的方式。

与此同时，RPA 等超自动化工具现在也已经进化成 RPA Agent，使用 RPA 本身也是对智能体工作流的一种应用。并且这种方式，正在被越来越多地应用于企业级业务场景。

在王吉伟频道看来，Agentic Workflow 并非简单的智能体工作流，而是包含传统软件（工具、解决方案）、大语言模型、AI Agent 等在内的新型业务流程的集合。当传统业务流程包含了 LLM 工作流或者 Agent 工作流，都可以视作 Agentic Workflow。

尤其是在大语言模型 Agent 化以及智能助手（Copilot 也具备反思、规划、工具使用能力并能调用 Agent）Agent 化的趋势下，显然它们更符合 Agentic Workflow 的定义。

所以要研究 Agentic Workflow，不只要看 AI Agent 以及 Agentic Workflow 本身，更要关注大语言模型及 RPA 等传统业务流程在 LLM 及 Workflow 方面的进展。

为了让大家更好地学习与理解 Agentic Workflow，本文精选了 25 篇智能体工作流相关的论文，并将其分为技术框架、系统（套件与工具）、评估测试基准、编程语言、模型与工作流及方法论六大类，希望对大家有所帮助。

注：为了方便不能科学上网的朋友，已将本文提到的所有论文打包。后台发消息 **Workflow**，获取论文资源。

一、技术框架

1、Sibyl：用于复杂现实世界推理的简单而有效的智能体框架

Sibyl: Simple yet Effective Agent Framework for Complex Real-world Reasoning

论文地址：https://arxiv.org/abs/2407.10718

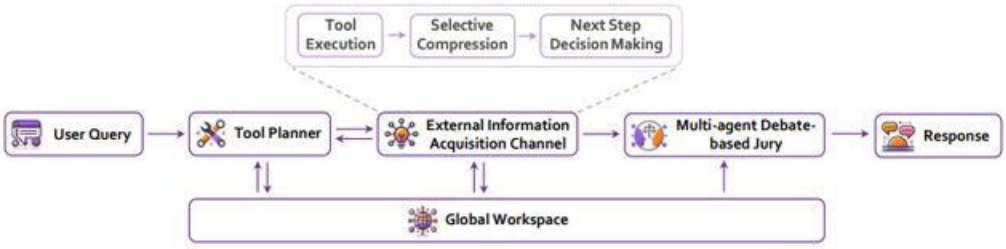


Figure 1: The overall pipeline of *Sibyl* framework.

大型语言模型（LLM）集成了固有知识、上下文学习和零样本能力，展现出强大的问题解决能力。然而，现有智能体在长期推理和工具潜力利用方面存在不足，导致现实世界推理任务中的缺陷。为克服这些限制，Sibyl 作为一个新型的 LLM 智能体框架，通过最少工具有效处理复杂推理任务。

Sibyl 从全球工作空间理论中获取灵感，整合了全球工作空间，加强了系统知识和对话历史的管理与共享。在心智理论的指导下，Sibyl 通过多主体辩论的陪审团机制自我完善答案，确保全面性和平衡性。这一设计旨在简化系统复杂性，拓宽问题解决范围，促进从系统 1 到系统 2 的思维转变。

Sibyl 注重可扩展性和易调试性，采用函数式编程中的重入概念，以无缝集成到其他 LLM 应用中。在 GAIA 基准测试集中，Sibyl 实现了 34.55%的平均得分，展现了其先进性能。论文作者期望 Sibyl 能推动开发更可靠和可重用的 LLM 智能体，以应对复杂的现实世界推理挑战。

2、PEER：使用多智能体框架和调优方法对特定领域的任务进行专业化

PEER: Expertizing Domain-Specific Tasks with a Multi-Agent Framework and Tuning Methods

论文地址：https://arxiv.org/abs/2407.06985

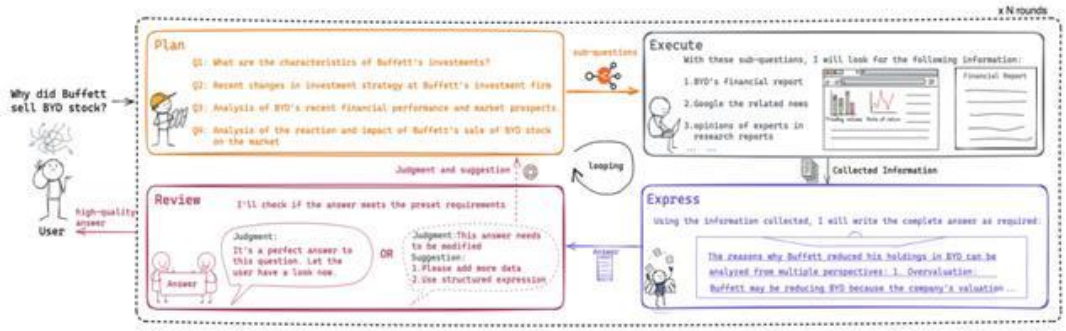


Figure 1: Cyclic Workflow Diagram of the PEER Framework. The user's query, "Why did Buffett sell BYD stock?", prompts the "Plan" agent to generate four relevant sub-questions. The "Execute" agent then collects information, including BYD's financial data and expert opinions. The "Express" agent synthesizes a comprehensive answer, which the "Review" agent evaluates and, if necessary, suggests modifications.

在专业领域应用中,GPT-4 通过精确的提示和检索增强生成(RAG)技术展现出巨大潜力，但同时也面临性能、成本和数据隐私的三重困境。高性能需求往往需要复杂的技术处理，而要管理多个智能体在复杂工作流程中的表现，不仅成本高，难度也大。

为应对这些挑战，论文提出了 PEER（规划、执行、表达、审查）多智能体框架。该框架通过整合精细的问题拆解、高效的信息检索、综合的总结能力以及严格的自我评估，系统化地处理专业领域任务。

考虑到成本和数据隐私的顾虑，许多企业正从 GPT-4 等专有模型转向定制模型，以期在成本、安全性与性能之间找到平衡点。团队利用在线数据和用户反馈，开发了一套行业实践，旨在实现模型的高效调整。

本研究提供了一套最佳实践指南,用于在特定领域问题解决中应用多智能体系统，并实施有效的智能体调优策略。特别是在金融问答领域的实证研究表明，该方法达到了 GPT-4 性能的 95.0%，同时在成本控制和数据隐私保护方面表现出色。

3、BMW Agents——通过多智能体协作实现任务自动化的框架

BMW Agents -- A Framework For Task Automation Through Multi-Agent Collaboration

论文地址：https://arxiv.org/abs/2406.20041

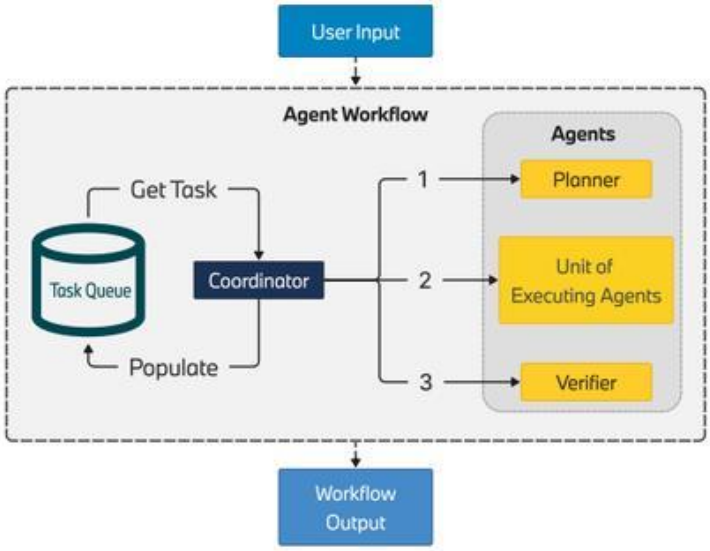


Figure 1: Generic agent workflow starting with user input and ending with providing workflow output. **Agent Workflow** highlights major components and levels of the workflow with (1) Planning, (2) Execution, and (3) Verification done by dedicated agents.

由大型语言模型（LLM）驱动的自主智能体展现了自动化的巨大潜力。技术的初步成效已在多个演示中显现，其中包括智能体解决复杂任务、与外部系统交互以扩展知识，以及触发必要操作。

特别是，多个智能体以协作方式共同解决复杂任务的场景，彰显了它们在非严格和非明确环境下的运作能力。因此，多智能体方法在许多工业应用中具有极大的应用潜力，无论是构建复杂的知识检索系统还是开发下一代机器人流程自动化。

考虑到当前 LLM 一代的推理能力，处理复杂流程需要采取多步骤策略，这包括制定明确定义的模块化任务计划。这些任务可以由单一智能体或一组智能体根据其复杂性执行。在本项研究中，团队专注于

构建一个灵活的智能体工程框架，特别关注规划和执行阶段，以应对跨不同领域的复杂应用案例。

该框架能够为工业应用提供了所需的可靠性，并且为确保多个自主智能体能够协同工作、共同解决问题提供了一套可扩展、灵活且协作的技术流程。

4、Trace 是新的 AutoDiff——解锁计算工作流的高效优化

Trace is the New AutoDiff -- Unlocking Efficient Optimization of Computational Workflows

论文地址：https://arxiv.org/abs/2406.16218

项目地址：https://microsoft.github.io/Trace

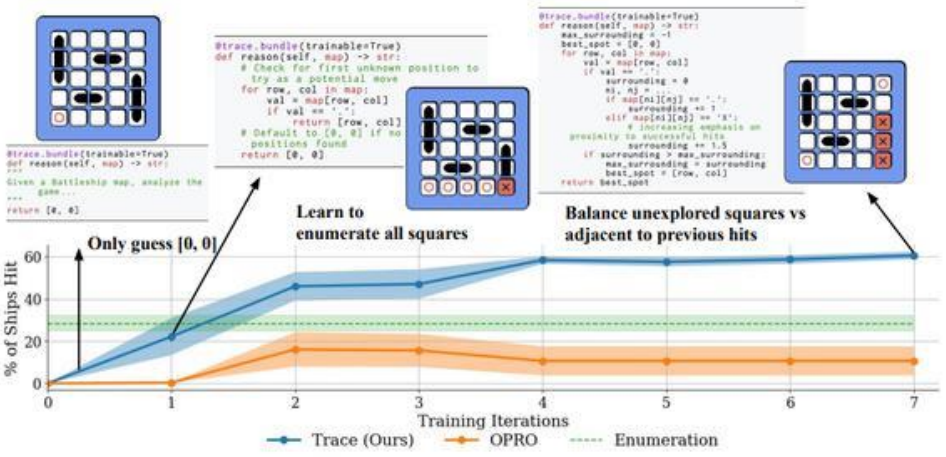


Figure 1: Learning Example in Battleship: An agent playing Battleship must intelligently place a shot on the board. Trace automatically optimizes heterogeneous parameters (e.g. multiple codes) to implement the agent's policy. The `reason()` parameter contains an enumeration heuristic after 2 optimization iterations, and later updates to a balanced explore-exploit strategy. Means and standard errors are computed over 10 random seeds.

论文探索了一种针对自动化编码助手、机器人和副驾驶等人工智能系统的优化问题，研究团队开发了一个名为 Trace 的端到端优化框架，它将 AI 系统的计算流程视为神经网络图，并基于反向传播的泛化进行优化。这种优化处理了包括丰富反馈、异构参数和复杂目标在内的多种因素，并能适应动态变化的计算图。

Trace 框架通过一种新的迭代优化数学设置——使用跟踪预言机优化（OPTO）——来捕获和抽象 AI 系统的特性，以设计跨领域的优化器。在 OPTO 中，优化器通过接收执行跟踪和输出反馈来迭代更新参数。Trace 提供了一个 Python 接口，利用类似 PyTorch 的接口高效地将计算流程转换为 OPTO 实例。

利用 Trace，团队开发了一个名为 OptoPrime 的通用优化器，它基于 LLM，能够解决多种 OPTO 问题，包括数值优化、提示优化、超参数调优、机器人控制器设计和代码调试等，且性能可与领域内专业优化器相媲美。论文认为，Trace、OptoPrime 和 OPTO 框架将推动下一代交互式智能体的发展，使其能够利用各种反馈实现自动适应。

5、RCAgent：使用工具增强型大型语言模型的自治智能体进行云根本原因分析

RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models

https://arxiv.org/abs/2310.16340

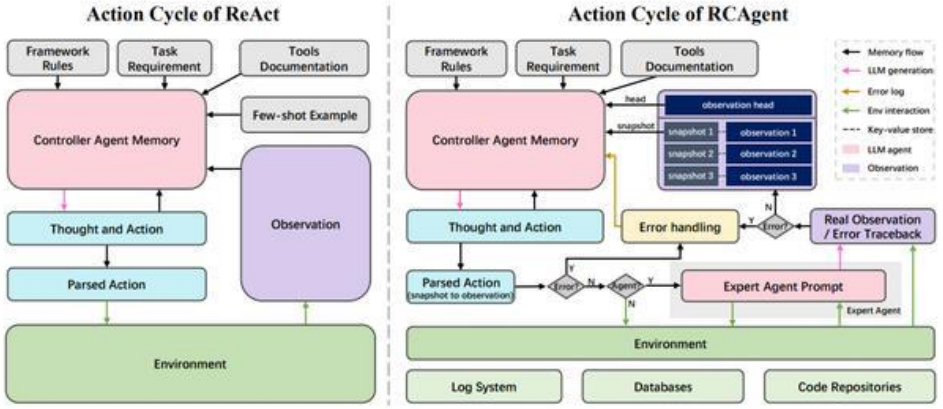


Fig. 2: Overview of the different action cycles from ReAct (left) and RCAgent (right). Both action cycles involve generating verbal thoughts, taking actions, and receiving observation from the environment, all of which are recorded in the prompt alongside the initial memory to boost reasoning. Besides, our RCAgent includes the key-value store for observation retrieval, allowing the agent to operate on text data much larger than the context length constraint. After parsing the action, our RCAgent executes the action directly or invokes an expert agent, depending on the type of tool.

近期，云根本原因分析（RCA）领域对大型语言模型（LLM）的应用进行了积极探索。但现有方法仍依赖手动设置工作流，未能充分发挥LLM在决策和环境交互方面的能力。为此,研究团队推出了RCAgent,这是一个工具增强的LLM自治智能体框架,专为实用且注重隐私的工业RCA设计。

RCAgent不依赖外部模型如GPT系列，而是在内部部署的模型上运行，能够自主进行自由格式的数据收集和综合分析。该框架融合了多项增强功能,包括行动轨迹的自洽性,以及一系列用于上下文管理、稳定性提升和领域知识导入的方法。

实验结果表明，RCAgent 在 RCA 的多个方面（如预测根本原因、解决方案、证据和责任）以及规则内外任务上均显示出显著且一致的优势，这些优势已通过自动化指标和人工评估得到验证。此外，RCAgent 已成功集成至阿里云 Apache Flink 实时计算平台的诊断和问题发现工作流程中，进一步提升了工业RCA的效率和准确性。

二、系统、套件与工具

1、AgileCoder：基于敏捷方法论的软件开发动态协作智能体

AgileCoder: Dynamic Collaborative Agents for Software Development based on Agile Methodology

论文地址：https://arxiv.org/abs/2406.11912

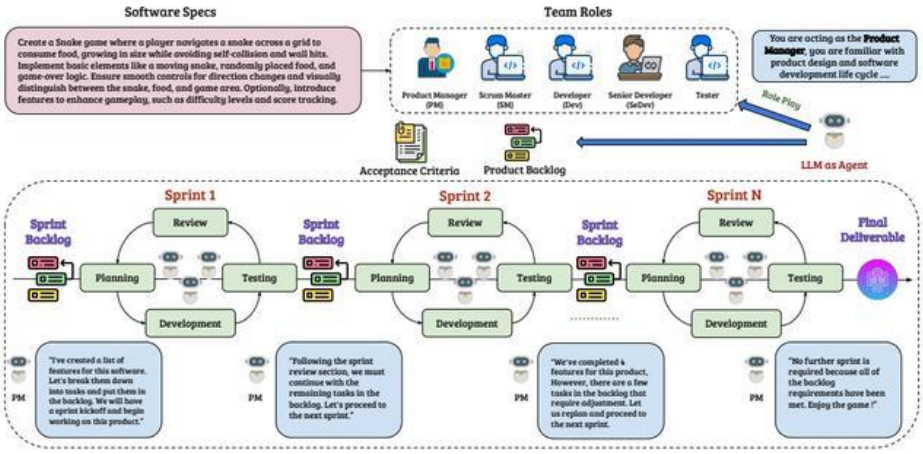


Figure 1: An overview of AGILECODER

软件智能体正成为解决复杂软件工程任务的有前景的工具。然而，现有研究常常过于简化软件开发流程，而现实世界中的这些流程往往更为复杂。

为了应对这一挑战，研究团队设计了 AgileCoder，这是一个将敏捷方法论（AM）整合进框架的多智能体系统。该系统将特定的 AM 角色，如产品经理、开发人员和测试人员，分配给不同的智能体，它们根据用户输入协作开发软件。

AgileCoder 通过组织工作为一系列冲刺（sprint），提高开发效率，并专注于逐步完成软件的开发。此外，还引入了一个动态代码图生成器，该模块能够在代码库更新时动态创建代码依赖图。这使得智能体能够更深入地理解代码库，从而在软件开发过程中实现更精确的代码生成和修改。

AgileCoder 在性能上超越了现有的基准，如 ChatDev 和 MetaGPT，树立了新的标准，并展现了多智能体系统在高级软件工程环境中的强大能力。这标志着软件开发向更自动化、智能化方向迈出了重要一步。

2、Parrot：使用语义变量高效提供基于 LLM 的应用程序

Parrot: Efficient Serving of LLM-based Applications with Semantic Variable

论文地址：https://arxiv.org/abs/2405.19888

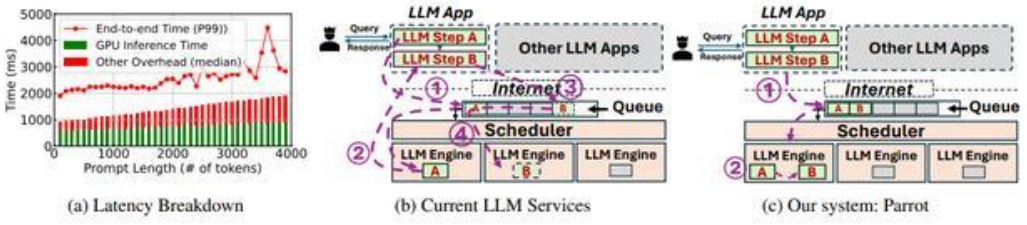


Figure 3: The end-to-end latency breakdown of current LLM services. The source of the overhead comes from network and queuing due to chatty interaction between LLM application and LLM services, which is eliminated in our system Parrot.

LLM 的兴起催生了基于 LLM 与传统软件优势的新型应用程序——AI 智能体（也叫副驾驶），这是一种软件新范式。

不同租户的LLM应用程序通过多个LLM请求设计复杂工作流以完成任务，但受限于当前公共LLM服务提供的简化请求级API，丢失了关键的应用程序级信息。这些服务只能盲目优化单个LLM请求，导致应用程序的整体性能不佳。

该论文介绍了 Parrot，这是一个专注于 LLM 应用程序端到端体验的服务系统。Parrot 引入了语义变量的概念，这是一种统一的抽象，将应用程序级知识暴露给公共LLM服务。语义变量在请求提示中标注输入/输出变量，并在连接多个 LLM 请求时形成数据管道，提供了一种自然的LLM应用程序编程方式。

公开语义变量给公共LLM服务，使其能够执行数据流分析，揭示多个LLM请求间的相关性，为LLM应用程序的整体性能优化开辟了新空间。广泛的评估显示，Parrot 针对流行和实际的 LLM 应用程序用例实现了显著的性能提升。

3、使用基础模型实现企业自动化

Automating the Enterprise with Foundation Models

论文地址：https://arxiv.org/abs/2405.03710

项目地址：https://github.com/HazyResearch/eclair-agents

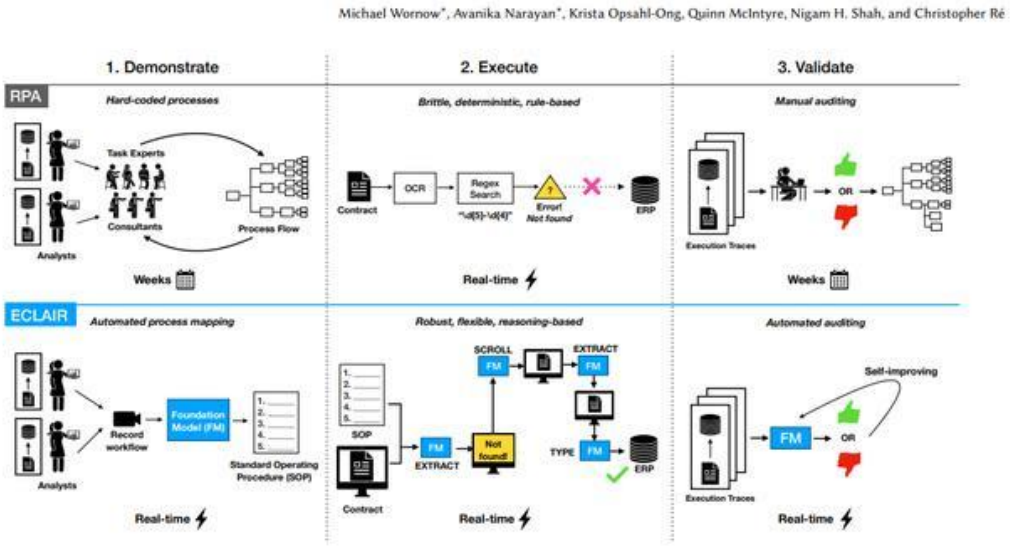


Figure 1: Differences between ECLAIR and traditional RPA. ECLAIR uses FMs to learn expertise via video demonstrations (left), navigate GUIs given written documentation (center), and audit completed workflows (right).

企业工作流程自动化每年可带来 4 万亿美元的生产力提升。尽管这一领域已受到数据管理社区数十年的关注，但实现端到端工作流自动化的终极目标仍然具有挑战性。现有解决方案主要依赖流程挖掘和机器人流程自动化（RPA），这些机器人通常被硬编码以遵循预设规则。

通过对医院和大型 B2B 企业的案例研究，研究团队发现 RPA 的普及受到诸如高设置成本（12-18 个月）、执行不可靠（初始准确率 60%）和维护繁重等问题的制约。新一代多模态基础模型（FM），如 GPT-4，以其卓越的推理和规划能力，为工作流自动化提供了新的可能性。

为此，论文提出了 ECLAIR 系统，它在最少人工监督下实现企业工作流程自动化。初步实验显示，ECLAIR 通过多模态 FM 实现了接近人类水平的工作流理解（准确率 93%），并基于工作流的自然语言描

述即可快速设置，实现了 40%的端到端完成率。论文认为，人与 AI 的协作、验证和自我改进是未来研究的开放性挑战，并提出利用数据管理技术来解决这些问题。

4、S-Agents：开放环境中的自组织智能体

S-Agents: Self-organizing Agents in Open-ended Environments

https://arxiv.org/abs/2402.04578

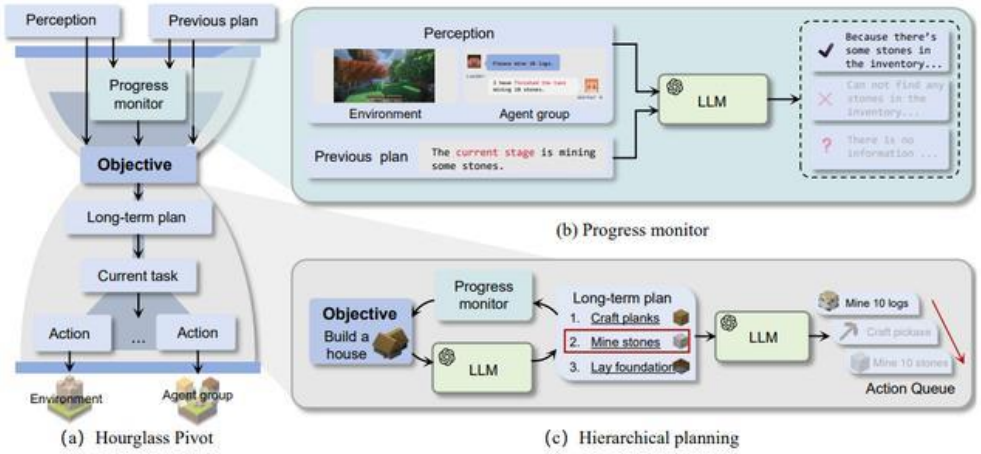


Fig. 3. **An illustration of hourglass agent architecture. (a) *Hourglass agent framework:*** *The upper segment:* Processes inputs like perception and the previous plan. These inputs undergo a series of operations, converging towards a unified and consistent objective (the bottleneck of the hourglass). *The lower segment:* Involves the decomposition of an objective through hierarchical planning. **(b) *Progress monitor:*** Utilizes LLM to assess the current progress status of the ongoing task. **(c) *Hierarchical planning:*** Comprises two stages: Task planner and action planner.

利用 LLM，自主智能体在处理各类任务上取得了显著进步。在开放环境中，为了提升协作的效率和有效性，需要灵活调整策略。然而，现有研究多聚焦于固定且任务导向的工作流程，而忽视了以智能体为中心的组织结构。

受人类组织行为的启发，该团队提出了一种自组织智能体系统（S-Agents），它包括动态工作流的“智能体树”结构、用于平衡信息优先级的“沙漏智能体架构”，以及支持智能体间异步任务执行的“非阻碍协作”方法。这一结构使得一组智能体能在无人为干预下，有效应对开放和动态环境的挑战。

团队的实验在 Minecraft 环境中进行，S-Agent 系统在执行协作建造和资源收集任务时表现出了熟练和高效，从而验证了其组织结构和协作方法的有效性。这一研究成果为智能体在复杂环境中的自组织协作提供了新的视角和解决方案。

5、一种人机协作工具，用于通过几个示例将单个大型语言模型智能体训练到网络中

A Human-Computer Collaborative Tool for Training a Single Large Language Model Agent into a Network through Few Examples

论文地址：https://arxiv.org/abs/2404.15974

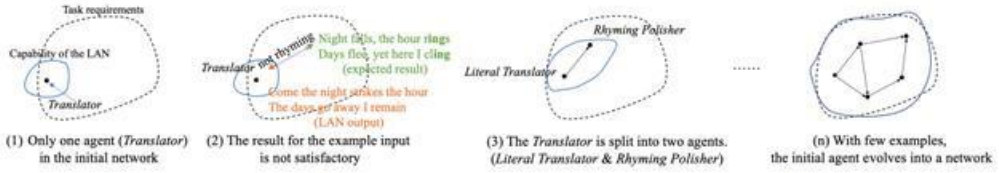


Figure 1: How EasyLAN trains a task-oriented LLM agent network (LAN) from a single LLM agent. (1) EasyLAN auto-generates an initial LAN that only contains a single LLM agent based on the task (e.g., translating French to English). A significant gap exists between the capabilities of the initial LAN and the task requirements. (2) A training example consists of an input and a ground truth. For a given training example, EasyLAN identifies discrepancies between the LAN's output and the expected output. For instance, when the input is a line of French poetry, "Vienne la nuit sonne l'heure, les jours s'en vont je demeure", the LAN fails to translate the text accurately while preserving the original rhyming scheme. (3) EasyLAN identifies the cause of the discrepancies and updates the LAN with respect to both the network architecture (e.g., splitting *Translator* into *Literal Translator* and *Rhyming Polisher*) and agent contents (e.g., adjusting the functionality of an agent). (n) EasyLAN iterates over a small set of training examples and constructs a satisfactory LAN.

单个大型语言模型（LLM）智能体在解决复杂任务时能力有限。通过将多个 LLM 智能体连接成网络，可以显著提升整体性能。然而，构建这样的 LLM 智能体网络（LAN）是一项耗时且复杂的过程。

在本研究中，团队推出了 EasyLAN，这是一个旨在帮助开发者构建智能体网络的人机协作工具。EasyLAN 首先根据任务描述生成一个只包含单个智能体的网络。然后，它利用训练样本来逐步优化网络。EasyLAN 会分析输出与实际值之间的差异，诊断错误原因，并采取策

略进行修正。用户可以参与 EasyLAN 的工作流程，或直接对网络进行调整。

最终，网络从单一智能体发展成为一个成熟的 LLM 智能体网络。实验结果表明，使用 EasyLAN，开发者能够迅速构建出性能优异的智能体网络。这一工具极大地简化了智能体网络的构建过程，提高了开发效率。

6、PromptRPA：根据文本提示在智能手机上生成机器人流程自动化

PromptRPA: Generating Robotic Process Automation on Smartphones from Textual Prompts

论文地址：https://arxiv.org/abs/2404.02475

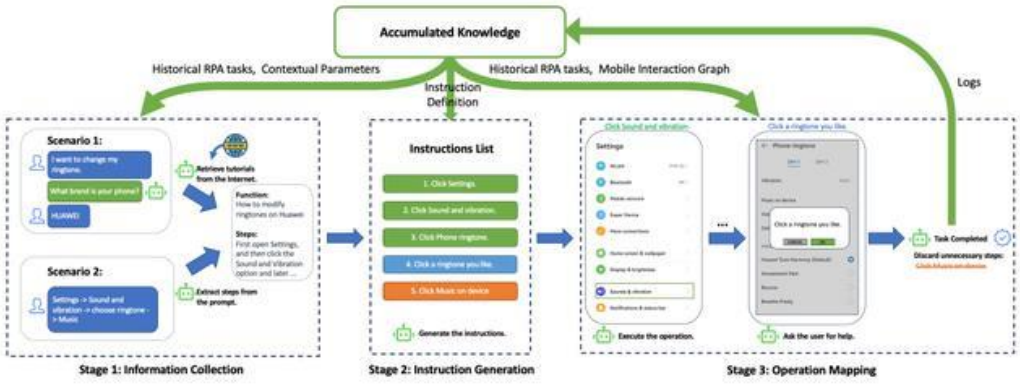


Fig. 1. The workflow of PromptRPA. Blue arrows guide the generation process from textual prompts to RPA, while green arrows indicate the knowledge accumulation that occurs through user interactions, thereby enhancing the future performance of the agents.

机器人流程自动化（RPA）通过模拟人机交互，在不修改现有代码的基础上，为自动化图形用户界面（GUI）上的任务提供了有效的解决方案。但 RPA 的广泛应用受限于对脚本语言和工作流设计专业知识的需求。

为解决这一问题，研究团队提出了 PromptRPA，这是一个能够理解与任务相关的各种文本提示（如目标、程序）并生成及执行相应 RPA 任务的系统。

PromptRPA 由一系列智能体组成，它们模仿人类的认知功能，专门用于解读用户意图、管理由 RPA 生成的外部信息，并在智能手机上执行操作。这些智能体能够从用户反馈中学习，并根据积累的知识不断提升性能。

实验结果显示，使用 PromptRPA 后，性能从基线的 22.28%显著提升至 95.21%，且每个新任务平均仅需 1.66 次用户干预。

PromptRPA 在创建教程、智能辅助以及客户服务等领域展现出广阔的应用前景，为 RPA 技术的进一步普及和应用提供了新的可能性。

7、ProAgent：从机器人流程自动化到智能体流程自动化

ProAgent: From Robotic Process Automation to Agentic Process Automation

论文地址：https://arxiv.org/abs/2311.10751

项目地址：https://github.com/OpenBMB/ProAgent

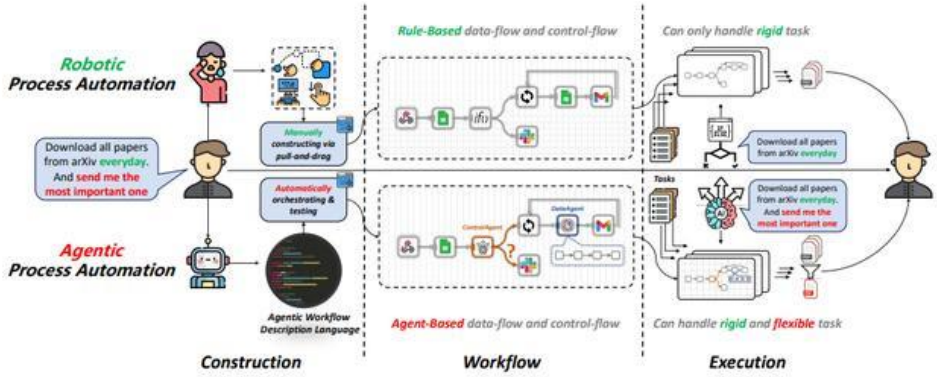


Figure 1: The comparison between Robotic Process Automation and Agentic Process Automation.

自动化技术从古代的水车发展到今天的RPA，一直在解放人类从事繁重任务。但RPA在处理需要人类智能的任务时面临挑战，尤其是在精心设计工作流和执行中的动态决策方面。

随着大型语言模型（LLM）的出现，研究团队提出了智能体流程自动化（APA），这是一种革命性的自动化新范式，利用基于LLM的智能体实现高级自动化，通过将任务分配给负责构建和执行的智能体来减轻人力负担。

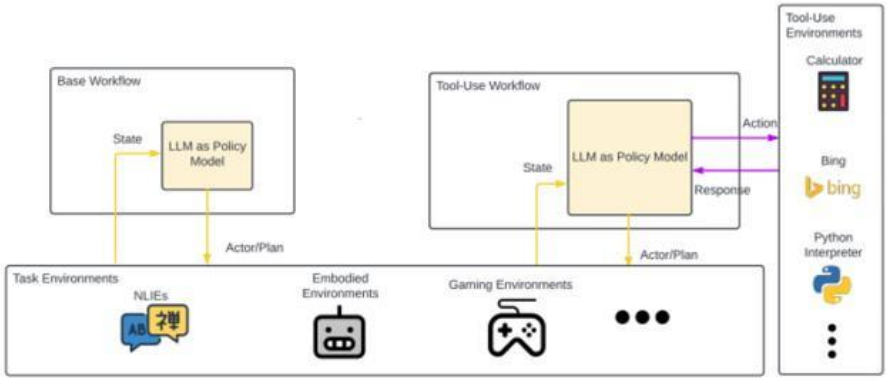
论文具体实现了ProAgent，这是一个基于LLM的智能体，它可以根据人工指令创建工作流程，并通过协调专业的智能体做出复杂决策。

通过实证实验，论文详细展示了APA在工作流构建和执行方面的过程，证明了APA的可行性，并展现了由智能体驱动的自动化新范式的巨大潜力。这不仅为自动化领域带来了新的视角，也为未来智能自动化的发展提供了新的方向。

8、基于LLM的智能体调查：常见工作流和可重用的LLM分析组件

A Survey on LLM-Based Agents: Common Workflows and Reusable LLM-Profiled Components

论文地址：https://arxiv.org/abs/2406.05804



(a) Policy-Only Workflows.

大型语言模型（LLM）的最新进展推动了基于 LLM 的复杂智能体框架的开发。然而，这些框架的复杂性在一定程度上阻碍了细粒度差异化的实现，这对于在不同框架间高效实现功能和推动未来研究至关重要。因此，该调查的主要目标是通过识别通用工作流程和可重用的 LLM 分析组件（LMPC），来促进对近期提出的多种框架的统一理解。

这项工作旨在简化不同智能体框架之间的差异，通过提取共通的工作流程和分析组件，为研究者和开发者提供一个更加清晰和一致的视角。通过这种方式，论文希望能够降低开发和维护智能体框架的难度，同时为未来的研究和创新打下坚实的基础。

三、评估测试基准

1、WorkArena++：迈向基于作文规划和推理的常识性工作任务

WorkArena++: Towards Compositional Planning and Reasoning-based Common Knowledge Work Tasks

论文地址：https://arxiv.org/abs/2407.05291

基准测试项目：

https://github.com/ServiceNow/WorkArena/tree/workarena-plus-plus

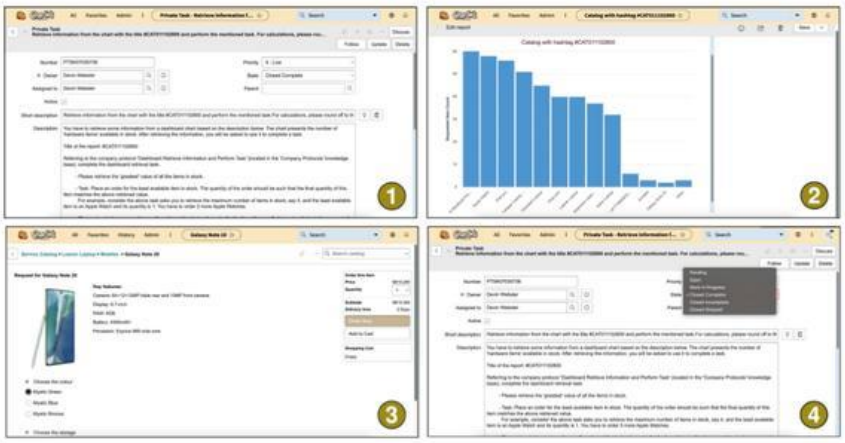


Figure 1: Example WorkArena++ task: Restock low inventory items. Here, the agent acts as an IT worker tasked with restocking items that are below some threshold in stock: ① As is common, it receives instructions via a ticket assigned to them in the system; ② it must then read the dashboard to extract all items whose stock count is low; ③ reorder the items from the service catalog to match a minimum stock quantity, and ④ close the ticket assigned to them once the task is completed.

大型语言模型（LLM）因其模仿人类智能的能力而备受关注，这促使基于 LLM 的自主智能体数量激增。尽管最新的 LLM 展现出根据用户指令进行规划和推理的潜力，但它们在自主任务解决方面的实际应用效果尚待深入研究。特别是在企业环境中，自动化智能体的应用被寄予厚望，期望能够带来显著的影响。

为了解决这一研究空白，论文提出了 WorkArena++，这是一个创新的基准测试套件，包含 682 个任务，覆盖知识工作者日常执行的实际工作流程。WorkArena++的目标是全面评估网络智能体在规划、问题解决、逻辑/算术推理、信息检索以及上下文理解等方面的能力。

通过对最先进的 LLM、视觉语言模型（VLM）以及人类工作者的实证研究，论文揭示了这些模型在职场中作为有效助手所面临的若干挑战。

除了基准测试，论文还提供了一种机制，能够轻松生成数千条基于真实情境的观察/动作轨迹，这些轨迹可以用于微调现有的智能体模型，并期望这项工作能够成为推动社区向有能力的自主智能体发展的重要资源。

2、FlowBench：重新审视基于 LLM 的智能体工作流引导规划并对其进行基准测试

FlowBench: Revisiting and Benchmarking Workflow-Guided Planning for LLM-based Agent

论文地址：https://arxiv.org/abs/2406.14884

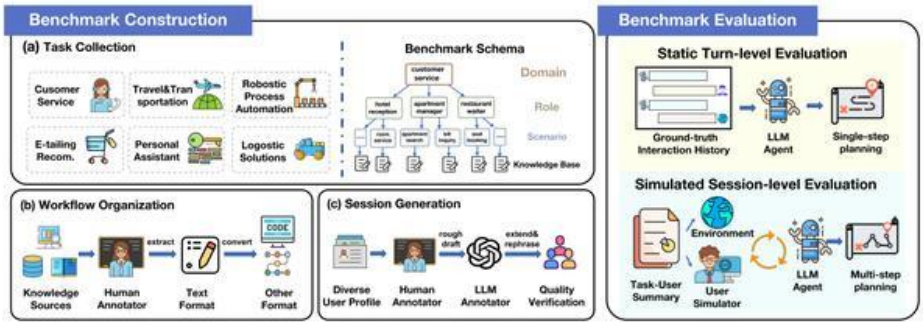


Figure 2: Overview of FlowBench. Our benchmark schema is structured in a top-down multi-level hierarchy (domain - role - scenario - knowledge). The benchmark construction process on the left contains three phases (a,b,c). The evaluation framework on the right encapsulates static turn-level and simulated session-level assessment.

大型语言模型（LLM）驱动的智能体已成为执行复杂任务的有前途工具，它们通过迭代规划和行动来完成任务。但当缺乏对专业知识密集型任务的深入理解时，这些智能体可能会产生不切实际的规划幻想。为提高规划的可靠性，该团队尝试整合与工作流相关的外部知识。

尽管这一方法有潜力，但整合的知识往往杂乱无章、形式多样，缺乏严格的形式化和全面评估。因此，该团队对不同格式的工作流知识进行形式化处理，并推出了 FlowBench——首个工作流引导规划的基准测试。FlowBench 覆盖 6 个领域的 51 个不同场景，以多种形式展现知识。

为了在 FlowBench 上评估不同的 LLM，团队设计了一个多层评估框架，评估了工作流知识在多种格式下的有效性。结果表明，现有的 LLM 智能体在规划方面还有很大的提升空间。论文期望 FlowBench 这

一具有挑战性的基准测试能够为未来智能体规划研究提供参考，推动相关技术的进步。

3、多模态基础模型是否了解企业工作流？业务流程管理任务的基准

Do Multimodal Foundation Models Understand Enterprise Workflows? A Benchmark for Business Process Management Tasks

论文地址：https://arxiv.org/abs/2406.13264

数据集和实验项目地址：

https://github.com/HazyResearch/wonderbread

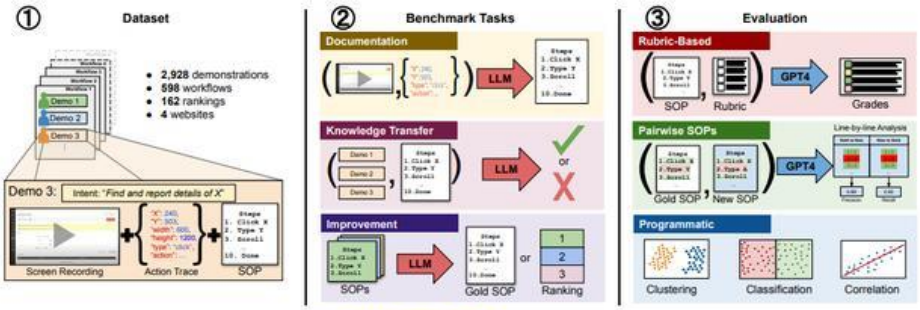


Figure 1: The three components of **WONDERBREAD**. (1) We curate 2928 human demonstrations across 598 web navigation tasks. Each demonstration includes an intent, a full screen recording, an action trace, and a written guide (SOP) describing the steps taken in the demonstration. (2) We create 6 BPM tasks that measure a model's ability to generate accurate documentation, assist in knowledge transfer, and improve workflows. (3) We provide automated evaluation pipelines for all tasks.

现有的机器学习（ML）基准测试在评估业务流程管理（BPM）任务时，缺乏足够的深度和多样性的注释。BPM 是一种旨在记录、衡量、改进和自动化企业工作流的实践。

目前的研究几乎完全集中在单一任务上，即利用多模态基础模型（FM）如 GPT-4 实现端到端的自动化。这种对自动化的专注忽视了大多数 BPM 工具的实际应用情况——在典型的流程优化项目中，仅仅记录相关工作流就占据了 60%的时间。

为了填补这一空白，研究团队推出了 WONDERBREAD，这是首个用于评估 BPM 任务的多模态 FM 基准测试，它超越了自动化的范畴。该论文的贡献包括：

- 一个包含 2928 个记录工作流程演示的数据集；
- 6 个新的 BPM 任务，涵盖从工作流文档到知识转移再到流程改进的实际应用；
- 一套自动评估工具。基准测试显示，尽管最先进的 FM 能够自动生成文档（例如，在工作流程的视频演示中识别 88%的步骤），但它们在将这些知识重新应用于更精细的工作流程完成验证方面表现不佳（F1 分数小于 0.3）；
- 团队期望 WONDERBREAD 能够激励开发更多以人为中心的 AI 工具，用于企业应用程序，并进一步探索多模态 FM 在更广泛的 BPM 任务中的应用。

四、编程语言

APPL：一种提示编程语言，用于程序和大型语言模型提示的和谐集成

APPL: A Prompt Programming Language for Harmonious Integration of Programs and Large Language Model Prompts

论文地址：https://arxiv.org/abs/2406.13161

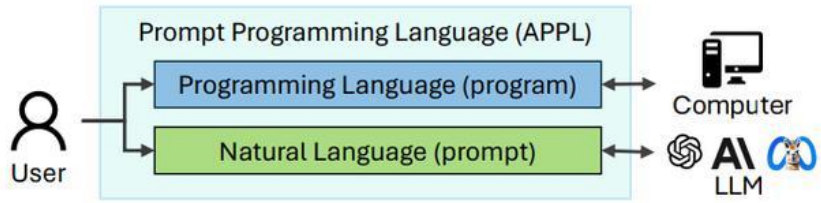


Figure 1: We introduce APPL, a *prompt programming language* that integrates conventional programs and natural language prompts to provide a unified interface for users to access computers and LLMs together.APPL also facilitates users fusing the strengths of computer programs and LLMs by providing convenient conventions between them.

大型语言模型（LLM）通过精心设计的提示和外部工具的集成，日益展现出处理各类任务的能力。然而，随着任务复杂性的提升，涉及 LLM 的工作流程可能变得复杂，难以实现和维护。为解决这一难题，研究团队提出了 APPL，一种新颖的提示编程语言，它作为计算机程序与 LLM 之间的桥梁，支持将提示无缝嵌入 Python 函数，反之亦然。

APPL 具备直观的 Python 原生语法，拥有异步语义的高效并行化运行时环境，并且配备了无需额外成本的跟踪模块，以支持有效的故障诊断和重放。论文通过三个典型场景——自一致性的思维链（CoT-SC）、ReAct 工具使用的智能体，以及多智能体聊天——证明了 APPL 程序的直观性、简洁性和高效性。

此外，对三个可并行化工作流的实验进一步证实了 APPL 在并行化独立 LLM 调用方面的有效性，并实现了与预期估算相匹配的显著加速比。这表明 APPL 是一个强大的工具，能够提升 LLM 在复杂任务中的性能和可用性。

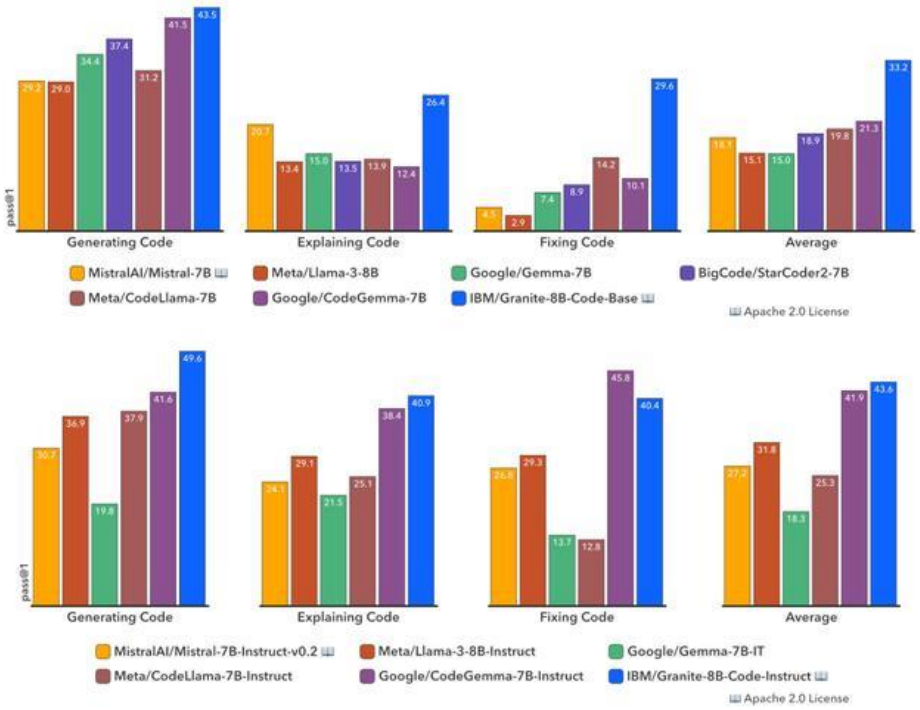
五、模型与工作流

1、Granite Code Models：用于代码智能的开放基础模型系列

Granite Code Models: A Family of Open Foundation Models for Code Intelligence

论文地址：https://arxiv.org/abs/2405.04324

项目地址：https://github.com/ibm-granite/granite-code-models



LLM 在代码训练方面取得了突破性进展，正深刻改变着软件开发的生态。越来越多的代码 LLM 被融入到软件开发工具中，以提升程序员的工作效率。同时，基于 LLM 的智能体也开始展现出独立处理复杂编码任务的能力。

要充分发挥代码 LLM 的潜力，需要它们具备广泛的能力，如代码生成、错误修复、代码解释、文档编写和代码库维护等。在本项研究中，团队推出了 Granite 系列仅解码器代码模型，专门用于代码生成任务。这些模型经过了 116 种编程语言的代码训练，覆盖了从 30 亿到 340 亿参数大小不等的多种模型，能够满足从复杂的应用现代化到设备内存受限的各种场景。

通过一系列综合任务的评估，团队发现 Granite Code 模型在所有可用的开源代码 LLM 中始终保持最先进的性能。

该模型系列针对企业级软件开发流程进行了特别优化，在代码生成、修复和解释等多项编码任务中均有出色表现，成为一个多功能的全能型代码模型。所有 Granite Code 模型均在 Apache 2.0 许可下发布，既适用于研究也适用于商业用途，为软件开发领域带来了前所未有的灵活性和创新潜力。

2、迈向实现零样本提示优化的分层多智能体工作流程

Towards Hierarchical Multi-Agent Workflows for Zero-Shot Prompt Optimization

论文地址：https://arxiv.org/abs/2405.20252

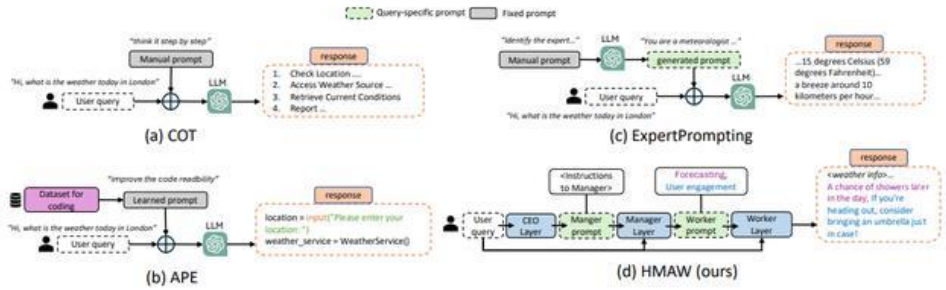


Figure 1: **Examples comparing the generalization ability of existing methods and the proposed one.** (a) COT [1] uses a handcrafted prompt, which might not be suitable for all tasks. (b) APE [5] fine-tunes the prompt on a specific dataset, and its generalization capability to other scenarios is questionable. (c) ExperPrompting [6] includes few-shot examples in the system prompt to help an LLM convert the user query to a format more suitable for LLM, but these examples might not be able to cover all scenarios. (d) Our method adopts a hierarchical design in reformatting the user query. Free from pre-defined few-shot examples, the interaction between the LLM hierarchy allows for more generalizable yet more adaptive tuning of the prompt.

大型语言模型（LLM）在解答用户问题上取得了显著进步，支撑了多样化的应用场景。但 LLM 的回答质量极大程度上依赖于提示的质量，一个精心设计的提示能够引导 LLM 准确回答极具挑战性的问题。

尽管已有研究开发了多种策略来优化提示，包括手工制作和领域内优化，它们在开放场景下的有效性仍受限，因为前者依赖于人类对问题的理解，而后者对未见过场景的泛化能力不足。

为克服这些限制，研究团队提出了一种让 LLM 自主设计最佳提示的方法。具体来说，团队构建了一个分层的提示生成框架，首先创建包含精确指令和准确措辞的提示，再基于此生成最终答案。这一流程称为分层多智能体工作流（HMAW）。

与现有方法相比，HMAW 不受任何人类预设限制，无需训练，完全任务独立，同时能够适应任务的细微差别。通过跨多个基准的实验，证实了 HMAW 虽然简单，却能创建出详尽且合适的提示，进一步提升了 LLM 的性能。

3、面向混合现实的多模态细粒度培训助手的自主工作流

Autonomous Workflow for Multimodal Fine-Grained Training Assistants Towards Mixed Reality

论文地址：https://arxiv.org/abs/2405.13034

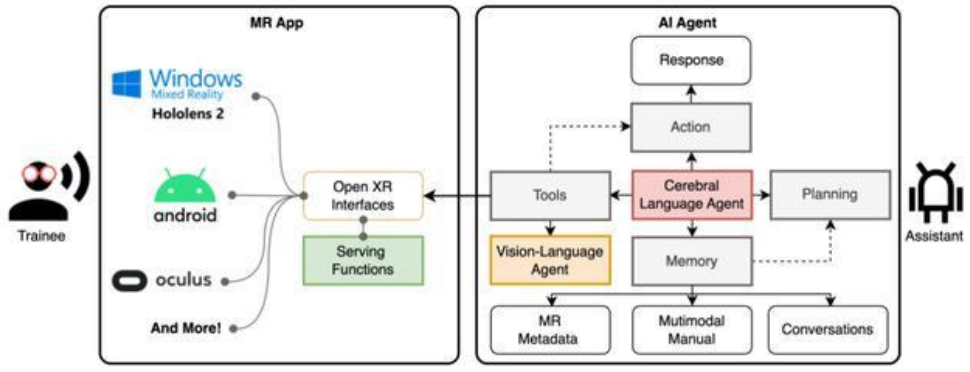


Figure 2: The proposed autonomous workflow, involving an AI agent interacting with an MR application. The AI agent comprises a core cerebral language agent, which interacts with a vision-language agent to interpret the multimodal context into metadata, which can be utilized by the cerebral language agent iteratively. The MR application interacts with AI agents by serving functions as external tools.

自主人工智能智能体（Autonomous Agent）在自动理解基于语言的环境中展现出巨大潜力，尤其是在大型语言模型（LLM）迅猛发展的背景下。然而，对多模态环境的深入理解尚待进一步探索。本研究设计了一个自主工作流程，旨在将 AI 智能体无障碍地集成到扩展现实（XR）应用中，实现细粒度训练。

论文展示了一个在 XR 环境中用于乐高积木组装的多模态细粒度培训助手的案例。该智能体结合了 LLM、记忆、规划功能以及与 XR 工具的交互能力，能够根据历史经验做出决策。此外，论文介绍了 LEGO-MRTA，这是一个多模态细粒度装配对话数据集，它能够在商业 LLM 服务的工作流程中自动合成，包含多模态说明、对话、XR 响应和视觉问答。

研究团队选取了几个流行的开放资源 LLM 作为基准，评估它们在微调和未微调状态下对团队提出的数据集的性能。论文期望这一工作流程能够推动更智能助手的开发，实现 XR 环境中的无缝用户交互，并促进 AI 和人机交互（HCI）社区的研究。

六、方法论

1、利用多 AI 智能体进行跨领域知识发现

Leveraging Multi-AI Agents for Cross-Domain Knowledge Discovery

论文地址：https://arxiv.org/abs/2404.08511

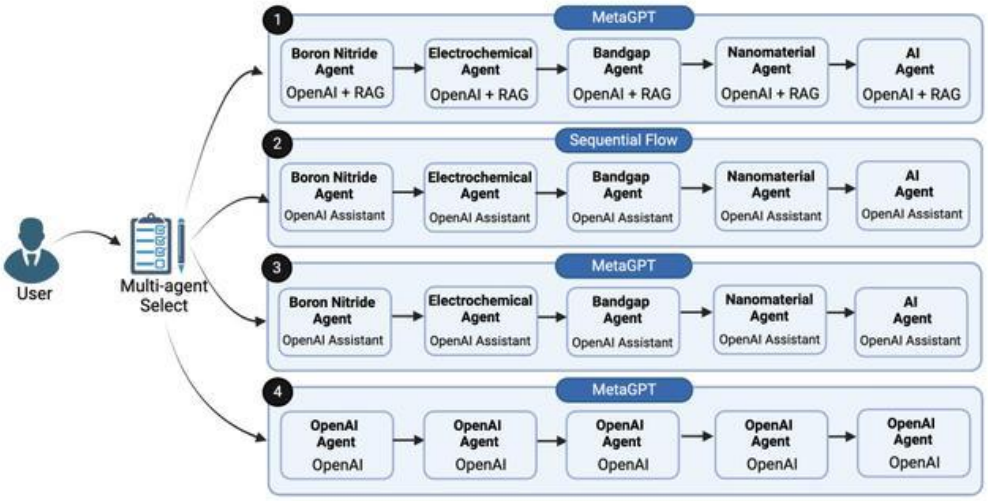


Figure 2: Multi-agent Flows

在迅速发展的人工智能领域，跨领域知识的整合与应用是一项关键的挑战与机遇。本研究提出了一种新方法，通过部署专注于不同知识领域的多人工智能智能体，实现跨学科的知识发现。每个智能体都像特定领域的专家，在统一框架下协同工作，提供综合的、超越单一领域限制的深入见解。

研究团队的平台通过促进智能体间的无缝互动，利用每个智能体的独特优势，增强了知识发现和决策过程。通过对比分析不同的多智能体工作流场景，评估了它们在效率、准确性和知识整合广度上的表现。实验结果表明，这些特定领域的多智能体系统在识别和填补知识空白方面表现出色。

这项研究不仅凸显了协作智能在促进创新中的关键作用，也为人工智能推动的跨学科研究和应用的发展奠定了基础。团队在小规模试点数据上评估了其方法，结果显示出预期趋势，随着自定义训练智能体的数据量增加，这些趋势预计将变得更加明显。

2、从头开始为类似计划的任务开发基础模型的案例

The Case for Developing a Foundation Model for Planning-like Tasks from Scratch

论文地址：https://arxiv.org/abs/2404.04540

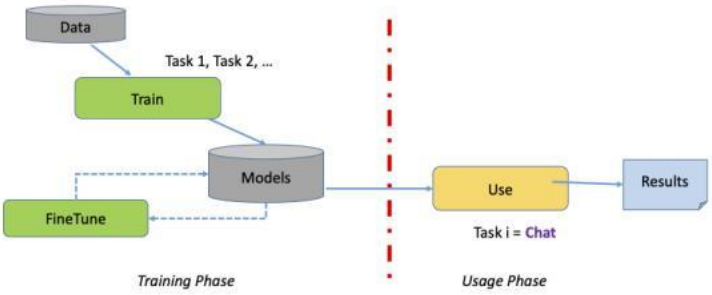


Figure 1: Overview of developing an FM-based system - e.g., an LLM-based chatbot.

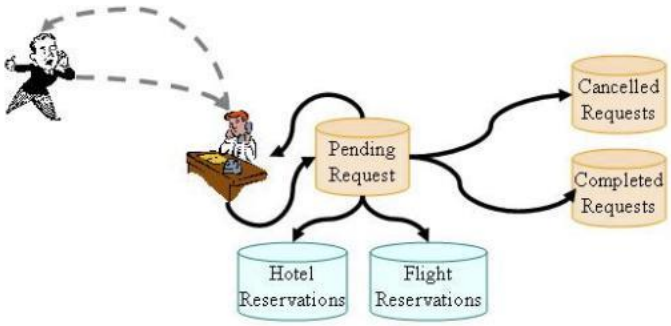


Figure 2: The travel agency example from (Srivastava and Koehler 2003)

基础模型 （FM） 彻底改变了许多计算领域，包括自动规划和调度 （APS）。例如，最近的一项研究发现它们对规划问题很有用：计划生成、语言翻译、模型构建、多智能体规划、交互式规划、启发式优化、工具集成和大脑启发规划。

除了 APS，还有许多任务涉及生成一系列行动，这些行动对于达成目标的可执行性有不同的保障，团队统称这些为类似计划（PL）任务，例如业务流程、程序编写、工作流管理和指南制定。研究人员正考虑将 FM 应用于这些领域。

然而，以往的研究多集中在使用现成的预训练 FM，并可能对它们进行微调。该论文讨论了为 PL 任务从头开始设计全面的 FM 的必要性，并探讨了设计时需考虑的因素。论文认为，这样的 FM 将为 PL 问题提供新的有效解决方案，正如大型语言模型（LLM）为 APS 领域所做的那样。

3、Transformations 时代的转变

Transformations in the Time of The Transformer

论文地址：https://arxiv.org/abs/2401.10897

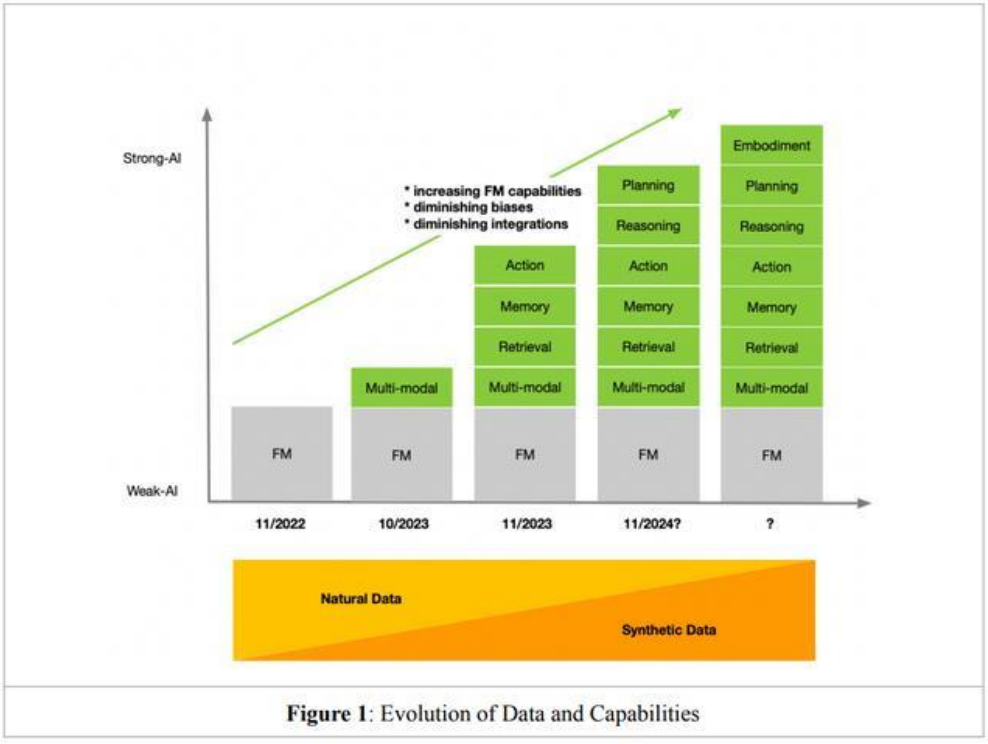


Figure 1: Evolution of Data and Capabilities

基础模型为以人工智能为主导的视角重新设计现有系统和工作流程提供了新的机遇。然而，实现这一转型面临着挑战和需要权衡的问题。本文旨在提供一个结构化的框架，帮助企业在向以 AI 为优先的组织转型过程中做出明智的决策。所提供的建议旨在帮助企业全面、有意识地做出知情的选择，同时避免受到不必要的干扰。

尽管这个领域看似发展迅猛，但其中一些核心的基础要素发展步伐相对较慢。团队专注于这些稳定不变的因素，以此构建论证的逻辑基础。通过深入理解这些不变的基本面，企业可以更稳健地把握 AI 转型的方向和步骤。

4、协同人机交互：与基于 LLM 的智能体进行服务共创的 23 种启发式指南

Synergizing Human-AI Agency: A Guide of 23 Heuristics for Service Co-Creation with LLM-Based Agents

论文地址：https://arxiv.org/abs/2310.15065

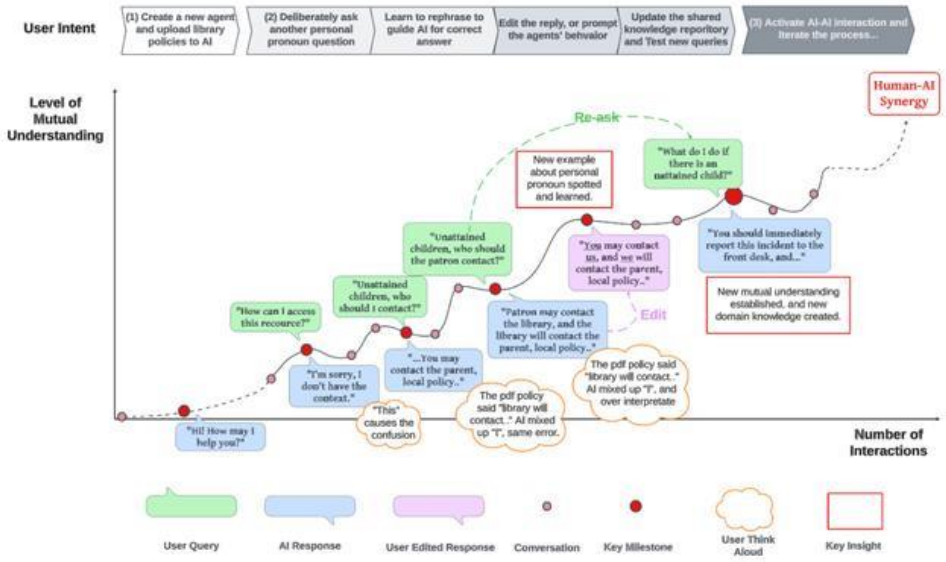


Figure 1: CoAGent supports the Interactive Agency in Human-AI service co-creation with LLMs through: (1) *"AIBound"*: sets AI's boundaries without sidelining humans, (2) *"AISync"*: showcases and aligns desired responses between creators and AI, and (3) *"PersonAi"*: allows new knowledge discovery via AI-AI interactions. A typical user journey unfolds: Creators start with *"AIBound"*, define an AI, and feed domain knowledge. They test the AI, and if errors arise, use *"AISync"* for adjustments, typically multiple times. Progress is marked by AI improvements. Then, with *"PersonAi"*, AI agents with different persona are created. Creators then step back to reassess the AI's knowledge depth and to spot potential blindspots.

本项实证研究为服务供应商提供了入门知识，帮助他们确定是否以及如何将大型语言模型（LLM）技术集成到其从业者和更广泛社区的工作之中。通过 CoAGent——一种与基于 LLM 的智能体共同创造服务的工具，研究团队探索了非 AI 专家与 AI 相互学习的过程。

这项研究通过与 23 位来自美国公共图书馆的领域专家合作，经历了一个三阶段的参与式设计流程，揭示了将 AI 集成到人类工作流程中所面临的根本性挑战。

研究结果提供了23种可操作的“与AI共同创造服务的启发式方法”，这些方法突出了人类与 AI 之间微妙的共同责任。并进一步提出了人工智能的9个基本智能体方面，强调了所有权、公平待遇和言论自由等基本要素。这种创新方法通过将 AI 视为关键利益相关者，并利用 AI 与 AI 的交互来识别盲点，从而丰富了参与式设计模型。

这些见解为服务环境中协同和道德的人类与 AI 共创铺平了道路，为人工智能共存的劳动力生态系统做好了准备。这不仅为服务供应商提供了实用的指导，也为构建人机协作的未来提供了宝贵的洞见。

5、计算管理的基础：将人工智能集成到现有工作流程中的任务自动化的系统方法

The Foundations of Computational Management: A Systematic Approach to Task Automation for the Integration of Artificial Intelligence into Existing Workflows

论文地址：https://arxiv.org/abs/2402.05142

在 AI 迅猛发展的今天，组织面临一个核心问题：如何将 AI 技术有效融入现有运营？为解答这一问题、调控期望并减少挑战，该论文引入了计算管理——一种系统化的任务自动化方法，旨在增强组织利用 AI 的潜力。计算管理融合了管理科学的战略洞察与计算思维的分析精确性，架设了二者之间的桥梁。

论文提供三个分步流程，以助于在工作流中启动 AI 的集成。

首先是任务（重新）制定，它将工作活动拆解为基本单元，每个单元由智能体执行，包括明确行动并产生多样结果。

第二，评估任务自动化潜力，通过任务自动化指数对任务进行评估，依据其标准化输入、规则明确性、重复性、数据依赖性和客观输出进行排序。

第三，任务规范模板详述了 16 个关键组件，作为选择或调整 AI 解决方案以适应现有工作流程的清单。

这些流程结合了手动和自动方法，并为现有的大型语言模型(LLM)提供了使用提示，以辅助完成这些步骤。计算管理为人与 AI 的协同提供了路线图和工具，提升了组织效率和创新力，为人机共荣的未来铺平了道路。

注：本文论文叙述部分配图，皆来自论文截图，具体内容请参考论文详情。