1. **AI-Agent**

**1.LLM VS Agent**

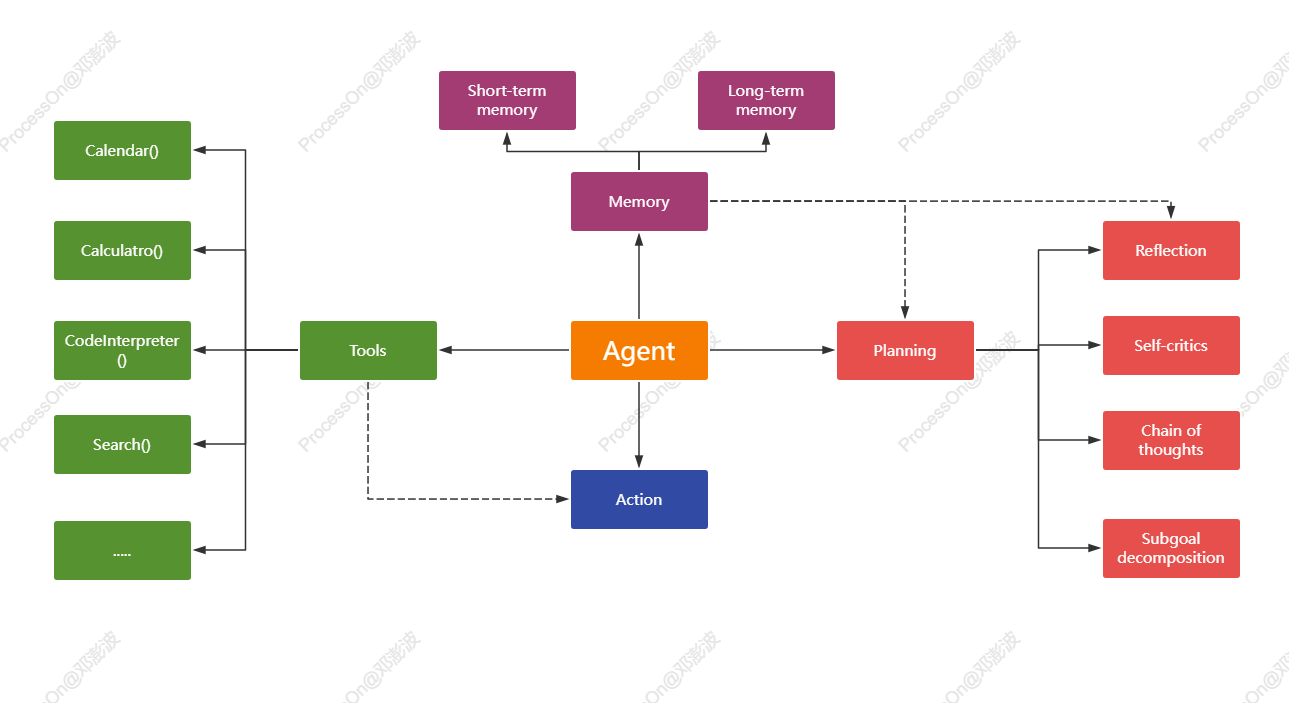
LLM：大型语言模型是一种基于深度学习的模型，专注于理解和生成自然语言文本。它的主要目的是处理与语言相关的任务，如文本生成、翻译、问答等。

Agent：人工智能代理是一个更广泛的概念，它指的是能够在特定环境中自主行动以实现目标的程序。AI Agent的目的可以是多样的，包括但不限于语言处理，还包括感知、决策和行动等。

LLM是AI Agent可能使用的工具之一，专注于自然语言处理。而AI Agent是一个更全面的概念，它可能包含LLM，但还包括其他组件和能力，以实现在复杂环境中的自主行为。两者在人工智能领域中都有重要的地位，但它们的设计和应用目标不同。

**2. Agent 是什么?**

AI业界对智能体提出了各种定义。个人理解，智能体是一种通用问题解决器。从软件工程的角度看来，智能体是一种基于大语言模型的，具备规划思考能力、记忆能力、使用工具函数的能力，能自主完成给定任务的计算机程序。



**规划（Planning）:**智能体会把大型任务 `分解为子任务`，并规划执行任务的流程；智能体会对任务执行的过程进行 `思考和反思`，从而决定是继续执行任务，或判断任务完结并终止运行。

**记忆（Memory）:**短期记忆，是指在执行任务的过程中的上下文，会在子任务的执行过程产生和暂存，在任务完结后被清空。长期记忆是长时间保留的信息，一般是指外部知识库，通常用向量数据库来存储和检索。

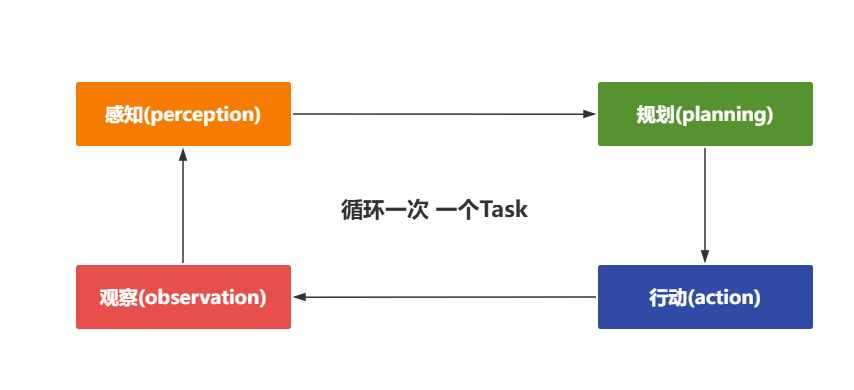
**工具使用（Tools）:**为智能体配备工具 API，比如：计算器、搜索工具、代码执行器、数据库查询工具等。有了这些工具 API，智能体就可以是物理世界交互，解决实际的问题。

**执行（Action）:** 根据规划和记忆来实施具体行动，这可能会涉及到与外部世界的互动或通过工具来完成任务。

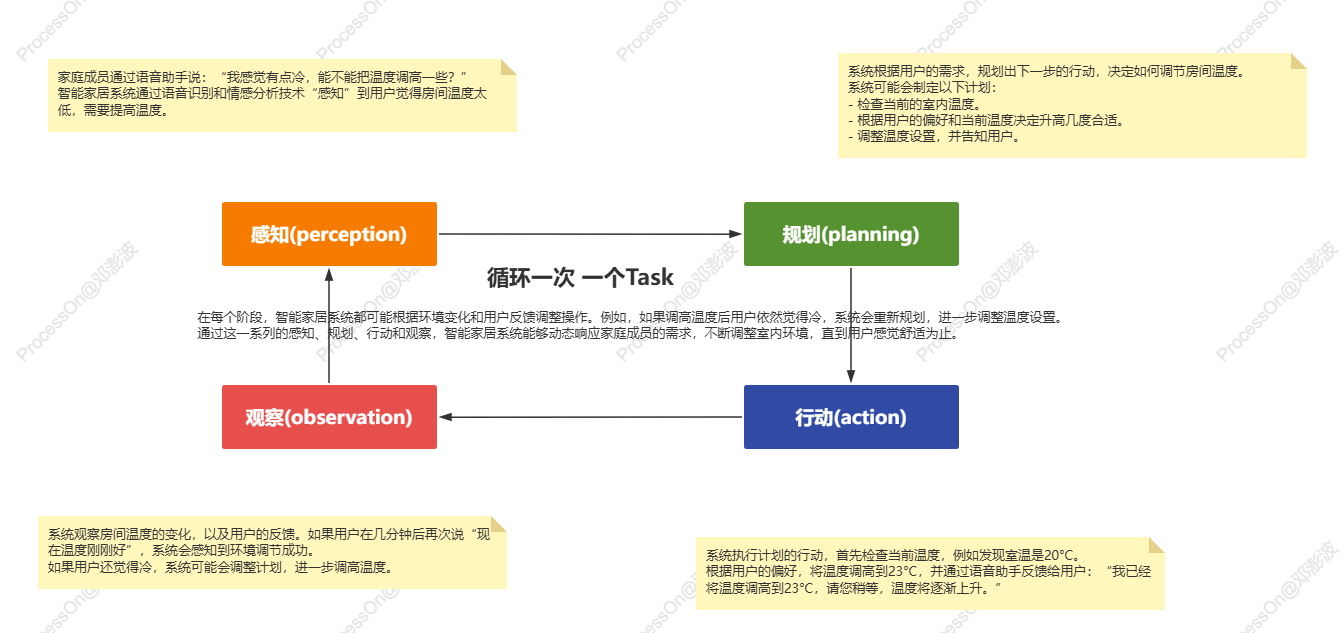
**3.Agent决策流程**

**3.1决策流转**

Agent的决策流程是和这几个步骤精密关联的。



**案例场景:** 假设我们有一个智能家居系统，它的任务是根据家庭成员的需求调节室内环境，比如温度和灯光。



**3.2规划(Planning)**

规划，可以为理解**观察和思考**。如果用人类来类比，当我们接到一个任务，我们的思维模式可能会像下面这样:

1.首先会思考怎么完成这个任务。

2.然后会审视手头上所拥有的工具，以及如何使用这些工具高效地达成目的。

3.再会把任务拆分成子任务。(就像咱们做思维导图一样)

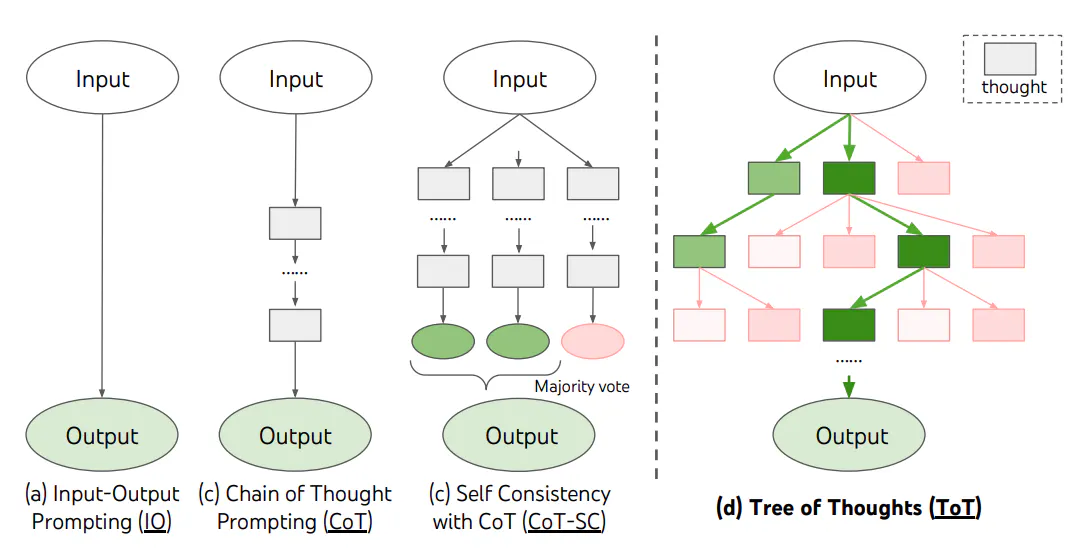
4.在执行任务的时候，我们会对执行过程进行反思和完善，吸取教训以完善未来的步骤。

5.执行过程中思考任务何时可以终止。

这是**人类的规划能力**，我们希望智能体也拥有这样的思维模式，因此可以通过LLM提示工程，为智能体赋予这样的思维模式。在智能体中，最重要的是让LLM具备这以下两个能力：

**子任务分解:** 通过LLM使得智能体可以把大型任务分解为更小的、更可控的子任务，从而能够有效完成复杂的任务。

**思维链:** 思维链已经是一种比较标准的提示技术。

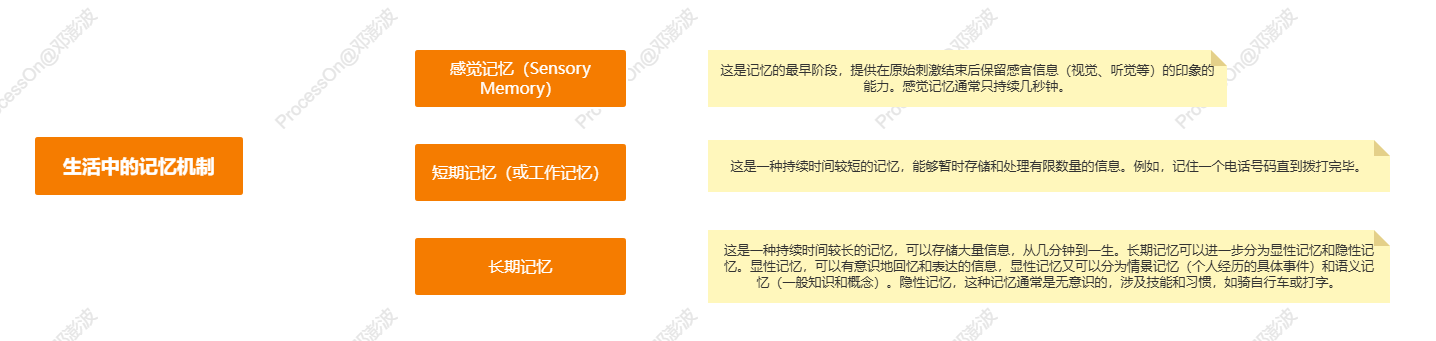


思维链：已经是一种比较标准的提示技术，能显著提升 LLM 完成复杂任务的效果。当我们对 LLM 这样要求「think step by step」，会发现 LLM 会把问题分解成多个步骤，一步一步思考和解决，能使得输出的结果更加准确。这是一种线性的思维方式。

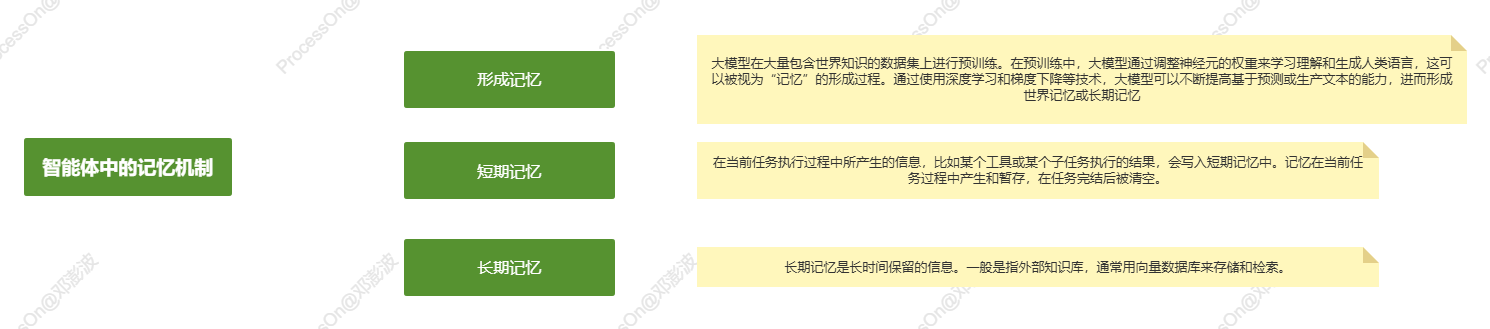
思维树：对 CoT 的进一步扩展，在思维链的每一步，推理出多个分支，拓扑展开成一棵思维树。使用启发式方法评估每个推理分支对问题解决的贡献。选择搜索算法，使用广度优先搜索（BFS）或深度优先搜索（DFS）等算法来探索思维树，并进行前瞻和回溯。

**3.3记忆(Memory)**

**生活中的记忆机制：**

****

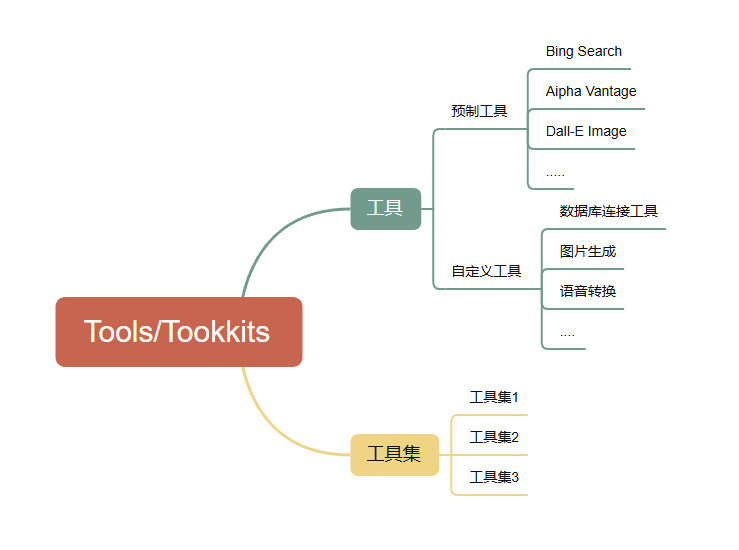
**Agent中的记忆机制:**

****

**3.4工具使用(Tools/Toolkits)**

Agent可以通过学习调用外部API来获取模型权重中所缺少的额外信息，这些信息包括当前信息、代码执行能力和访问专有信息源等。这对于预训练后难以修改的模型权重来说是非常重要的。

掌握使用工具是人类最独特和重要的特质之一。我们通过创造、修改和利用外部工具来突破我们身体和认知的限制。同样地，我们也可以为语言模型（LLM）提供外部工具来显著提升其能力。



1. **Function Calling**

**1.Function Calling 介绍**

⼤语⾔模型，例如 DeepSeek，拥有强⼤的知识储备和语⾔理解能⼒，能够进⾏流畅的对话、创作精彩的故事，甚⾄编写代码。然⽽，它们也⾯临着⼀些难以克服的困境，就像⼀个空有知识却⽆法⾏动的巨⼈。

**信息滞后：**知识库的更新速度有限，⽆法获取最新的实时信息，例如天⽓、股价等。这意味着它可能告诉你昨天的天⽓，或者⼀个⼩时前的股价，⽽这些信息可能已经过时了。就像⼀本印刷好的百科全书，内容固定，⽆法实时更新。

**缺乏行动力：**就像被困在虚拟世界中的⼤脑，只能提供信息，⽆法直接与现实世界互动。例如，它可以告诉你如何煮咖啡，但⽆法帮你磨咖啡⾖、煮咖让AI⼤模型连接外部世界啡、甚⾄递给你⼀杯咖啡。 这就好⽐⼀位博学的教授，可以讲解复杂的理论知识，但却⽆法在实验室进⾏实际操作。

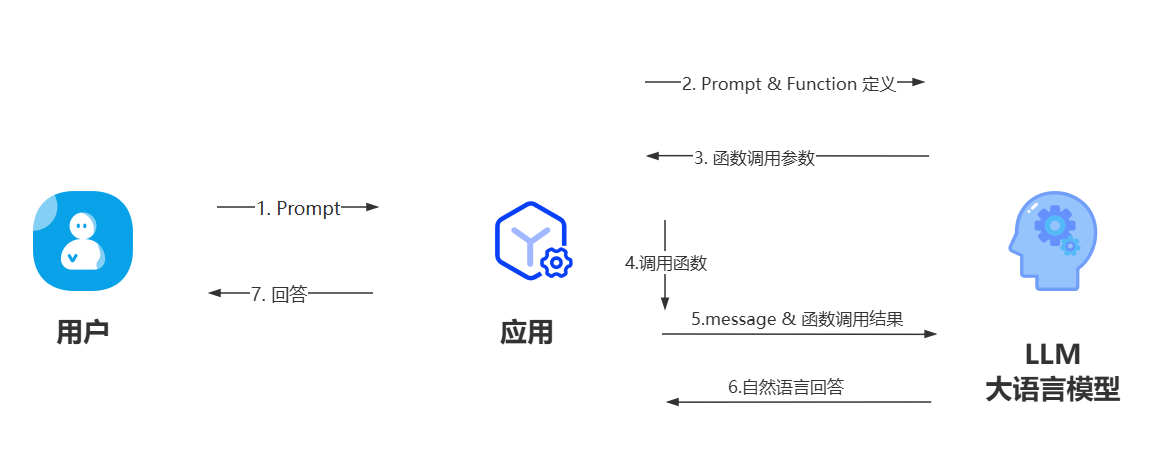
**大模型的三大缺陷：**

1.训练数据不可能涵盖所有信息。垂直、⾮公开数据必有⽋缺。

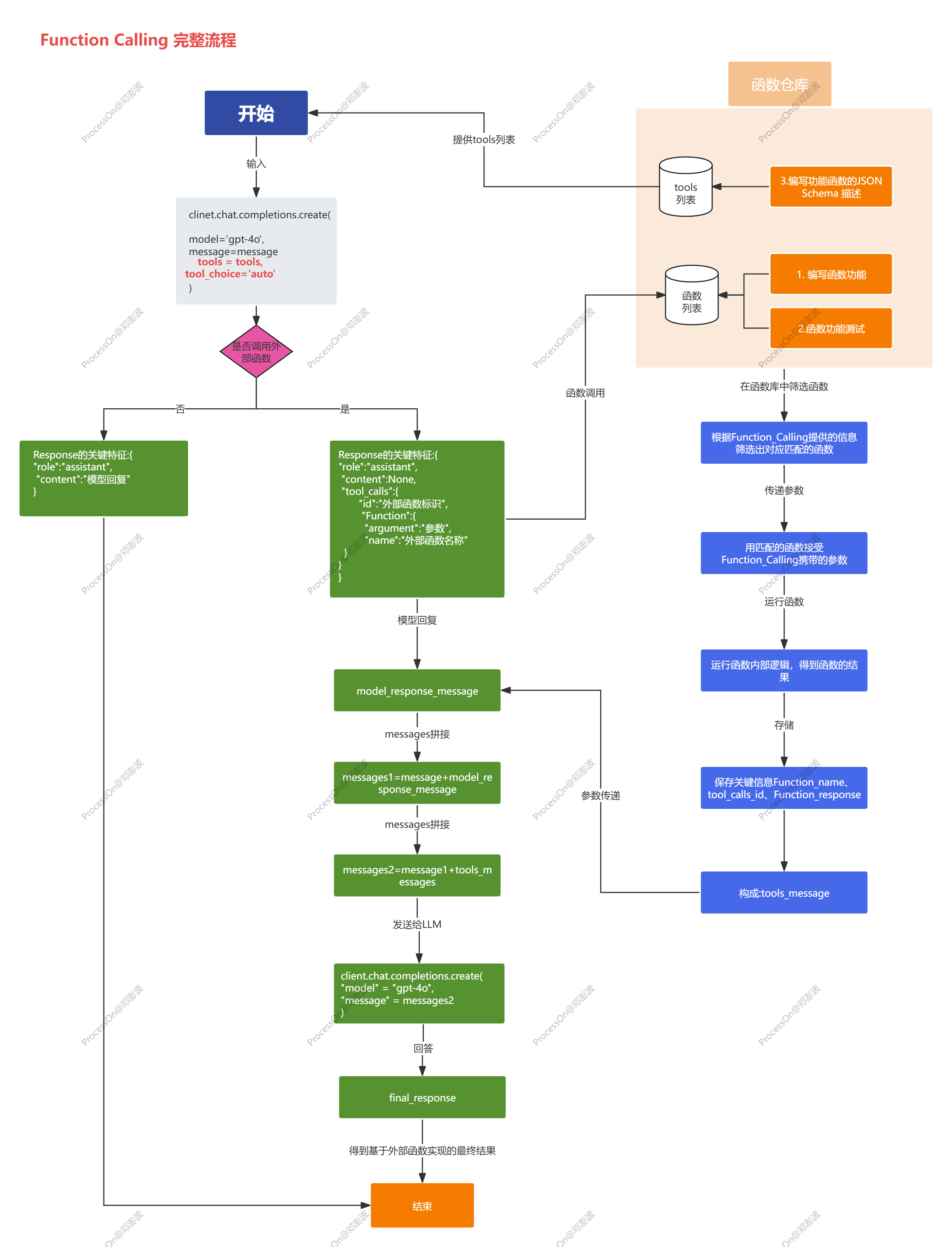
2.不知道最新信息。⼤模型的训练周期很⻓，且更新⼀次耗资巨⼤。所以它不可能实时训练。GPT-3.5的知识截⾄2022年1⽉，GPT-4是2023年4⽉。

3.没有「真逻辑」。它表现出的逻辑、推理，是训练⽂本的统计规律，⽽不是真正的逻辑。也就是说，它的结果都是有⼀定不确定性的，这对于需要精确和确定结果的领域，如数学等，是灾难性的，基本是不可⽤的。

**所以：**⼤模型需要连接真实世界，并对接真逻辑系统，以此来控制⼤模型输出的不确定性和幻觉，达到我们想要的结果。于是 OpenAI在23年6⽉份推出了Function Calling功能，让GPT模型拥有调⽤外部接⼝的能⼒。



**Function Calling 完整的流程结构介绍：**

****

**特点介绍：**

Function Calling 是⼀种让 Chat Completion 模型调⽤外部函数的能⼒，可以让模型不仅仅根据⾃身的数据库知识进⾏回答，⽽是可以额外挂载⼀个函数库，然后根据⽤户提问去函数库检索，按照实际需求调⽤外部函数并获取函数运⾏结果，再基于函数运⾏结果进⾏回答。

**Function Calling 可以被应⽤于各种场景，例如：**

1.调⽤天⽓ API 获取实时天⽓信息: 不再局限于提供过时的天⽓预报，⽽是可以告诉你此时此刻的天⽓状况，就像你打开⼿机上的天⽓应⽤⼀样。

2.调⽤订票⽹站 API 预订机票: 不再只是告诉你如何订票，⽽是可以直接帮你完成订票操作，就像⼀个专业的旅⾏代理⼀样。

2.调⽤⽇历 API 安排会议: 不再只是提醒你会议时间，⽽是可以直接帮你安排会议，并发送邀请给参会者，就像⼀个⾼效的私⼈助理⼀样。

3.调⽤数据库查询信息: 可以访问和查询数据库中的信息，例如产品信息、⽤户信息等，就像⼀个专业的数据库管理员⼀样。

4.调⽤代码执⾏程序: 可以执⾏代码来完成各种任务，例如数据分析、图像处理等，就像⼀个经验丰富的程序员⼀样。

**2.Function Calling 调用本地函数**

JSON Schema是一个用于描述JSON数据格式和结构的元数据标准。它用于验证、注释以及操控 JSON文档。JSON Schema本身是用JSON格式表示的，提供了一种灵活的方式来校验数据的结构，包括对象属性的类型、数组长度、数字和字符串的值范围等等。

**3.多Function 的调用**

**4.自动生成tools工具描述**

**5.Function Calling 连接数据库**

**6. 国产大模型**

**7. 多表查询**