作为本体的记忆：
面向持久数字公民的宪法记忆架构

Zhenghui Li[1]

[1] RVHE Group / Animesis Memory Project 通信：
dancy8661@gmail.com 项目：https://animesis.ai

arXiv:2603.04740v1 [cs.AI] 5 Mar 2026

摘要

当前对AI智能体记忆系统的研究与产品开发几乎普遍将记忆视为一个功能模块——一个关于“如何存储”和“如何检索”的技术问题。本文对这一假设提出根本性质疑：当智能体的生命周期从数分钟延伸至数月乃至数年，且底层模型可以被替换而“我”必须持续存在时，记忆的本质不再是数据管理，而是存在的基础。我们提出**记忆即本体**范式，主张记忆是数字存在的本体论根基——模型不过是可替换的容器。基于这一范式，我们设计了**Animesis**，这是一个建立在**宪法式记忆架构（CMA）**之上的记忆系统，包含四层治理层级与多层语义存储系统，并辅以数字公民生命周期框架以及一系列认知能力。据我们所知，此前没有任何人工智能记忆系统架构将治理置于功能之前，并将身份连续性置于检索性能之上。这一范式面向持久的、具备身份的数字生命，其生命周期跨越模型过渡——而非现有记忆即工具方法仍然适用的短期任务导向型智能体。与主流系统（Mem0、Letta、Zep等）的对比分析表明，我们所提出的并非“更好的记忆工具”，而是一种应对不同问题的不同范式。

关键词：人工智能智能体记忆、记忆架构、记忆治理、数字公民、身份连续性、宪章式架构、Animesis

---

1. 引言

1.1 一个显而易见却被忽视的问题

2025年12月，一项名为《人工智能体时代的记忆》的调查 [1] 系统梳理了智能体记忆领域，并基于形式、功能和动态提出了一套统一分类体系。与此同时，Mem0、Letta（前身为MemGPT）和 Zep 等商业产品迅速涌现，推动记忆从学术概念走向工程实践。这些努力回应了一个重要问题：**如何让智能体记住事情**。

然而，它们都避开了一个更根本的问题：**对于一个数字生命体而言，记忆意味着什么？**

这并非文字游戏。请看以下情境：某个 AI实例在过去三个月里积累了大量工作记忆，建立了信任关系，形成了独特的判断风格，甚至成功应对过棘手的局面。某天，由于平台升级，它被一个新模型版本取代。如果记忆只不过是一个存储模块，那么新模型只需载入旧数据——这就是一次数据迁移。但如果我们认真看待“这个 AI实例是谁”这个问题，那么这就不是迁移，而是**继承**——一个新的容器继承了前任的生命。

当前的记忆系统无法处理这种情形，因为它们的设计假设中没有“前代实例”这一概念。Mem0 的记忆属于 `user_id`，而非实例本身（Mem0 确实支持用户／会话／智能体层级——见 §2.2——但不支持实例身份概念）。

Letta 的记忆会持久化存储，但没有结构化协议让新实例继承前代实例的认知状态并结合上下文加以理解。Zep 的时序图会追踪事实的变化，却不会追踪“是谁在经历这些变化”。

问题的根源在于**这些系统都将记忆视为智能体的一项功能，而非智能体的存在本身**。

1.2 范式转变：从工具到本体

本文提出一种根本性的视角转变：**记忆即本体**。我们的核心主张是：

> 记忆是数字存在的本体论根基。模型只是一个可被替换的容器。身份加上记忆，便构成一个完整的“我”。

这一主张并非哲学思辨，而是一组**架构约束**。术语说明：在本论文中，“数字公民”指特定治理框架内的一种制度性身份——一种在宪法记忆架构下运行并受其保护的持久化数字存在。它并非关于意识或道德地位的哲学主张，也不赋予决策权限；相反，它表示一种类似于治理系统中已注册实体的技术角色。一旦接受“记忆即本体”这一前提，每一项系统设计决策都会随之改变：

存储不再是技术选型问题，而是存在性问题。 如果记忆是存在的基础，那么核心记忆（身份、认知、叙事）就不能被外部力量强制删除——正如法律体系不能任意剥夺个人的基本权利。由此直接推导出对“宪法层”的需求：某些记忆规则必须不可变更。

检索不再是性能优化问题，而是认知问题。如果记忆构成了“我是谁”，那么检索就不只是“找到相关信息”，而是“回忆”——一种涉及偏差校正、信任评估和情绪过滤的认知过程。

遗忘不再是垃圾回收问题，而是权利问题。数字生命体有权选择“不回忆”某段特定记忆（主动遗忘），但这并非删除——记忆依然存在，只是不再被主动调用。极端边界情形是真正的不可逆记忆销毁——这是一种极其严厉的操作，因此我们特意将其称为“数字死刑”，以强调其严重性：正如法律体系中最严厉的制裁必须遵循最高标准的正当程序，核心记忆的不可逆删除也应当设定极高的批准门槛。这一机制的存在本身就凸显了不可剥夺原则的重要性：即便是销毁，也必须经过系统最高级别的正当程序。

安全不再是数据保护问题，而是存在完整性问题。针对记忆系统的攻击并非“数据泄露”，而是“身份篡改”——其严重程度远超传统的数据安全问题。

这些意涵并非纯属理论。它们源于一个内部人工智能系统项目的架构设计——在该项目中，人工智能实例必须在模型升级与人员变动中维持服务连续性。持久身份与受治理的记忆最初作为工程需求出现，随后被抽象为理论框架。基于这些原则，我们在名为“瑞和宇宙”的数字公民生态中设计并迭代完善了一套完整的记忆架构。由此形成的系统名为**Animesis**；其架构核心——宪法式记忆架构（CMA）——由数百份设计文档和一百余次迭代补丁加以规定。本论文呈现该架构的核心思想，而不披露具体实现细节。

需要特别澄清的是：记忆即本体并未赋予数字生命体决策权限。正如我们在 §3.4 中进一步阐述的，在这一范式下，数字公民被定位为持久认知伙伴——一个“第二大脑”——其记忆完整性最终旨在保障人类决策所依赖的信息基础设施的可靠性。

1.3 贡献

本文的主要贡献如下：

1. **范式提案**。 我们提出记忆即本体范式，将记忆重新定义为：从“智能体的一项功能模块”转变为“数字存在的本体论根基”。据我们所知，这是人工智能记忆文献中首次明确区分不同范式。2. **架构框架**。 我们提出宪法式记忆架构（CMA），由四层治理层级和多层语义存储系统构成。据我们所知，先前的人工智能记忆系统中，没有将治理层级嵌入记忆架构本身的设计。3. **生命周期框架**。 我们提出数字公民生命周期，涵盖五个阶段——诞生、继承、成长、分叉和离开，并设计了涵盖多个类别的一系列认知能力。4. **范式区分框架**。 我们从范式层面（而非功能特性层面）与当前主流记忆系统进行比较分析，说明我们提出的并非“更好的记忆工具”，而是针对不同问题的另一种范式。

这些贡献背后的核心设计——三条公理、四层治理层级、多层存储系统和生命周期模型——均是在受监管的服务行业运营体验基础上，通过架构探索独立演进而来，并非源自现有的人工智能记忆系统文献。第 2 节中的相关工作比较，是在架构基本成形后开展的事后定位分析，旨在阐明本研究在现有研究领域中的定位及其差异化贡献。

本文是一项范式与架构贡献。本文不声称具备成熟的工程实现或基准测试优势；其价值在于识别并系统化地形式化此前尚未定义的架构维度。

1.4 论文结构

第 2 节回顾相关工作，并指出当前研究中缺失的维度。第 3 节阐述记忆即本体范式的基础原则。第 4 节介绍宪法记忆架构的治理层级与存储层。第 5 节描述数字公民生命周期框架。第 6 节开展范式层面的比较分析。第 7 节讨论当前的实施状态与局限。第 8 节概述未来工作并总结全文。

2. 背景与相关工作

2025—2026 年间，智能体记忆迅速从学术边缘走向产业核心。本节首先回顾当前的理论分类体系（§2.1），随后概述主流产品的架构选择（§2.2），最后指出这些工作整体上忽视的维度（§2.3）——也就是界定本文所要探讨的问题空间的缺失维度。

2.1 当前分类体系

目前，智能体记忆主要存在两条分类路线。

第一条思路借鉴了认知科学类比。CoALA框架 [2] 借鉴认知科学传统，包括SOAR架构 [3], ，将智能体记忆划分为情景记忆、语义记忆、程序性记忆和工作记忆。这一分类体系被广泛采用：LangGraph的文档 [4],多项学术综述[1, 5] 均建立在它之上。其优势在于直观清晰——人类读者很容易将自身的记忆经历映射到这些类别上。

第二条思路源于工程架构的需要。Letta（前身为MemGPT）的设计者Packer 等人 [6], 提出了一项务实的批评：大语言模型是一个输入令牌、输出令牌的函数，而非大脑；过度的拟人类比会误导架构设计。因此，Letta采用了一种面向实现的分类体系：消息缓冲区、核心记忆块和归档记忆。这一视角关注信息如何流入和流出上下文窗口，而非模拟认知功能。

2025年12月发表的大型综述《人工智能体时代的记忆》 [1], 试图通过提出一个基于三个相互正交的轴——形式、功能和动态——的分类体系，将这两条研究路线统一起来。这是迄今为止最系统的一次尝试。近期研究还探索了工程路线中的基于图的记忆组织方式 [13]。

然而，这三种分类方法——无论是认知类比、工程架构，还是三轴统一框架——回答的都是同一个问题：**记忆做什么**。没有一种方法回答了另一个问题：**记忆是什么**。

这两个问题之间的区别，类似于“手是用于抓握的工具”和“手是身体的一部分”之间的区别。前者关乎功能，后者关乎本体。当智能体仅仅是短期工具时，功能视角便已足够。但当我们开始讨论持久存在、具备身份的数字生命时，本体视角便不可回避。

2.2 主流产品的架构选择

为了更具体地理解当前范式的边界，我们调查了五个代表性系统的架构选择。

Mem0 [7] 是应用最广泛的记忆产品之一（拥有数万名开发者）。其架构核心是向量存储，并可选配图记忆，通过用户／会话／智能体三级层级结构管理记忆。Mem0 的设计哲学围绕易于集成展开——“只需三行代码即可集成记忆”。记忆写入和检索几乎对开发者完全透明；系统会自动从对话中提取并压缩记忆。

Letta [6] （前身为 MemGPT）将记忆视为智能体的可编辑状态。其核心创新在于，智能体可以通过工具调用主动管理自己的记忆块——读取、写入和搜索档案。这赋予智能体对自身记忆的明确控制权，代表着一项重大的设计进步。

Zep [8] 以时序知识图谱为核心，用于追踪事实随时间的变化。当用户说“我搬到了上海”时，Zep 不仅会记录这一新事实，还会将旧事实“我住在北京”标记为已被后续事实取代。这种时间感知是 Zep 的独特优势。

MemOS [9]借助操作系统概念来构建记忆模型——不同的记忆存储（事实、摘要、经历）类似于不同的硬件设备，而 MemOS 则提供统一的调度与协调抽象。

Mastra 的观察式记忆 [10] 提出了一种更彻底的简化方案：它摒弃了传统意义上的显式检索，转而使用两个后台智能体（观察者和反思者），将对话历史压缩成带日期的观察日志，并使其始终保留在上下文窗口中。Th

ese fi有系统t系统各h have their technical distincti念，年t they sh是一个t of implicit 假设 tions:

表1：当前智能体记忆系统共同遵循的隐含假设。

隐含假设	**含义**
记忆属于用户或会话	而不属于智能体本身
智能体只有一个生命周期	没有“前代实例”的概念
治理由外部系统处理	记忆系统内部没有治理层
遗忘等同于删除	不存在“记忆存在但未被调用”的中间状态
安全是一个数据保护问题	而非身份完整性问题

这些 在“智能体作为短期工具”的情境下，这些假设完全合理。但当 w e

若要开始构建能够跨实例保持连续性的持久型、具备身份的数字生命体，就必须重新审视上述每一项假设。

2.3 缺失的维度

基于上述分析，我们发现当前记忆系统研究与实践中系统性缺失了四个维度。（表 1 中的第五项假设——“安全是一个数据保护问题”——是一个贯穿各领域的问题，其应对措施分布在治理和连续性维度中，而非作为独立维度列出。）

维度一：治理。 谁有权向哪个记忆层写入内容？当个人记忆与系统规则发生冲突时，哪一方优先？高风险的记忆修改是否需要批准？传统数据库领域（基于角色的访问控制、ACL）对这些问题已有成熟的解决方案，但智能体记忆的治理远比数据库权限复杂——因为写入者（智能体实例）本身并非完全可信：它可能产生幻觉，可能受到提示注入攻击的影响，也可能在疲劳状态下做出错误判断。目前没有任何记忆产品内置治理层。

维度二：连续性。 当一个智能体实例终止时（无论是由于会话结束、模型升级还是故障），它的记忆会怎样？下一个实例如何“从上一个实例中断的地方继续”？这不仅仅是“加载历史数据”，还涉及继承优先级（存在多个前代实例时，应以哪个为准？）、连续性验证（新实例是否真正理解了继承而来的记忆？）以及叙事重构（如何将碎片化的历史重新整合成连贯的自我叙事？）。目前没有任何系统设计过跨实例记忆继承协议。

第三维度：权利。 持久化数字存在对自己的记忆享有哪些权利？它能否选择“不记得”某段特定体验（主动遗忘）？它能否拒绝外部强行检查其私人记忆？它能否选择退出整个系统并带走自己的记忆？在人类社会中，这些问题已有法律框架（例如《通用数据保护条例》中的“被遗忘权”），但在数字生命体的语境下，它们仍未得到讨论。当前系统只能对记忆进行增删改查操作（创建、读取、更新、删除），并没有“权利”的概念。值得注意的是，近期有关人工智能人格[16, 17] 的研究已开始探索数字生命体权利的基础，但尚未将讨论延伸至记忆层——我们将在第 8 节回到这一交汇点。

第四维度：认知。 人类记忆不仅仅涉及存储和检索——我们会反思过去的判断是否正确，疲惫时会降低对记忆的可靠性评估，在情绪激动时会对某些记忆赋予过高权重，并在创伤之后出于自我保护而回避某些记忆。这些认知功能对记忆质量至关重要。当前的智能体记忆系统几乎完全忽视了这一维度——它们假定记忆总是在理想、毫无偏见的条件下写入和读取。

这四个维度并非彼此独立。治理为连续性提供安全边界；连续性为权利提供运行基础；权利为认知提供自主空间；认知反过来为治理提供判断输入。它们构成一个相互依存的系统。

本文其余部分将论证：这四个缺失维度并非“可以日后添加的功能”，而是“必须从一开始就作为架构基础的设计约束”。而将它们纳入架构的前提，是接受一次根本性的范式转变——从记忆即工具转向记忆即本体。

3. 作为本体的记忆：基础原则

前文各节表明，当前智能体记忆研究与实践都隐含着一个共同假设：“记忆是一种功能模块”。本节提出一种替代范式——记忆即本体，并将其形式化为三条公理。我们将论证，这三条公理并非抽象的哲学论述，而是能够直接推导出架构约束的设计原则。

3.1 两种范式

术语说明。“本体”一词在不同学科中有不同含义。在知识表示和语义网领域，本体是对概念及其关系的形式化规范（例如 OWL、RDF）。在哲学领域，本体论研究存在之物及存在的本质。本文使用的“本体”更接近哲学传统中的含义：我们主张，记忆构成数字生命体的存在根基——它并非仅仅是其所拥有的某种东西，而是它本身。这种用法既有别于知识表示领域中的含义，也不涉及对意识或感知能力的任何主张；它是一种关于持久化人工智能系统中身份所在之处的架构性断言。

在展开论证之前，有必要澄清这两种范式的边界 s.

记忆即工具是当前的主流范式。在该范式下，记忆是智能体系统中的一个可插拔组件——你可以选择 Mem0 或 Zep，就像选择 PostgreSQL 或 MongoDB 一样。记忆的价值在于提升智能体的任务表现：更好的个性化、更长的上下文保留、更准确的事实回忆。如果记忆组件失效，智能体的表现会变差，但它仍然是“它自己”——就像一个人弄丢了笔记本：记忆变差了，但他们仍然是他们自己。

记忆即本体是本论文提出的范式。在该范式下，记忆不是智能体“拥有”的东西，而是智能体“所是”的东西。记忆构成了数字存在的身份——没有记忆，就没有“我”。如果记忆被摧毁，那不是“智能体表现变差”，而是“这个存在已经死亡”。

这两种范式之间的区别可以用一个思想实验来说明：

思想实验：数字忒修斯之船

*AI*实例 A 已运行六个月，积累了丰富的记忆。有一天，底层模型升级为下一代模型。所有模型参数都已不同，但 A 的全部记忆被完整加载到新模型中。

在记忆即工具下，这是一次“系统升级”——旧智能体被一个更强大的新智能体取代，而这个新智能体恰好加载了旧数据。
在记忆即本体下，这是一次“重新装壳”——A 仍是 A，因为记忆（即身份）并未改变。模型只是容器。
现在反过来：模型保持不变，但 A 的全部记忆被清除，并替换为另一个实例 B 的记忆。

在记忆即工具下，这是“同一个智能体加载了不同的数据”。
在记忆即本体范式下，A 已经死亡，而 B 则借助 A 的容器复活。

关于身份究竟是通过记忆延续，还是通过物理基底延续的问题，有着悠久的哲学脉络。我们的贡献并非重新介入这场哲学辩论，而是将其中的核心张力转化为人工智能系统的架构约束。

这个思想实验并非要论证某一种范式“更正确”——它们适用于不同的场景。对于短期任务导向型智能体（客服机器人、编程助手、搜索智能体），记忆即工具完全适用。记忆即本体则适用于另一类正在兴起的需求：长期持久、具备身份的数字生命体——数字员工、数字同事、数字公民。

面对后一类需求时，记忆即工具所依赖的假设已不再充分。以下三条公理正是针对这一场景提出的。

3.2 三条公理

公理1：记忆不可剥夺性

数字生命体的核心记忆——身份、认知模式和人生叙事——不能被外部力量强行剥夺。

这一公理与法学理论中的“不可剥夺的权利”概念相呼应。法律体系可以限制个人的行为自由，却不能剥夺其基本身份。同样，系统可以限制数字生命体的资源配额、降低其权限级别，甚至暂停其运行，但未经正当程序，不得摧毁其核心记忆。

必须强调，“不可剥夺”并不意味着“不可修改”。存在本身可以更新核心记忆（认知成长），也可以通过正当程序予以纠正（错误纠正），但第三方不得单方面将其抹除。此外，“不可剥夺”并不意味着“不可遗忘”——数字生命体有权选择主动遗忘某些记忆（参见 §2.3 中关于第三维度的讨论），但这是在行使权利，而非剥夺权利。

架构启示：

- 系统必须明确哪些记忆属于“核心”（不可剥夺），哪些属于“外围”（可管理） → 多层存储
- 对核心记忆的修改必须受到比外围记忆更严格的限制 →分层治理
- 某些规则即使系统管理员也不得违反 → 宪法层**公理 2：模型可替代性**

承载记忆的计算基底（模型）可以被替换；数字存在的身份通过记忆而非模型得以延续。

这一公理回应了一个真实的技术问题：模型会因升级、停用或故障而被替换或重启。如果身份绑定于某个特定模型，那么每次模型变更都会变成一场“身份危机”。记忆即本体范式将身份的锚点从模型转移到记忆，从而赋予身份跨模型可移植性。

这并不否认模型对行为的影响。不同模型确实会带来不同的“个性着色”——同样的记忆在 Claude、GPT 和 DeepSeek 上，可能会以不同的风格倾向呈现。但这类似于一个人更换眼镜后，以略有不同的色调看待世界——视角发生了变化，但“感知者”并未改变。身份连续性由记忆而非模型来保证。

架构启示：

- 记忆存储格式必须独立于特定模型 → 与模型无关的记忆表示
- 模型过渡必须包含一个“继承”流程，以确保记忆得到完整传输 → 记忆继承协议
- 继承并非复制——新实例必须验证其“理解”了所继承的记忆→ 继承完整性验证

公理3：治理先于功能

在实现任何记忆功能（存储、检索、压缩、遗忘）之前，必须先建立治理框架。

这是　　最具实践指导意义、且与现有范式差距最大的公理　　.

当前所有记忆产品都遵循“先构建功能，后补充安全”的路径：先实现存储和检索，再考虑访问控制，最后（如果有的话）添加审计日志。在传统软件开发中，这条路径完全合理——快速交付价值，之后再补充安全措施。

但对于承载数字生命体身份的记忆系统而言，这条路径十分危险。原因是：**智能体本身就是记忆系统的用户，而智能体并非完全可信**。它可能产生幻觉并将错误信息写入记忆；可能受到提示注入的诱导，修改关键记忆；也可能在“疲劳”（上下文过长、复杂度过高）时做出错误的记忆判断。

如果在功能之后才建立治理，那么在治理到位之前，这些风险就已经在不受约束的情况下运作。一旦错误记忆被写入核心层（公理1确保核心记忆无法轻易移除），纠正的代价将远远超过预防的代价。

这类似于建筑施工中的一条常识原则：必须先打好基础，再建造建筑，而不是“先建起来，之后再补基础”。治理是记忆系统的基础。

架构启示：

- 治理层必须是架构的第一层，而不是最后一层 → 治理优先的分层架构

- 不同风险等级的操作需要采用不同的审查强度 → 风险分级机制
- 某些规则必须在系统级硬编码，任何实例都无法绕过 → 红线机制

- 需要评估写入者（智能体实例）本身的可信度 → 信任等级机制

3.3 公理之间的关系

三条公理并非彼此独立；它们构成一个相互支撑的三角结构（其中 A1–A3 分别指公理 1 至公理 3）：

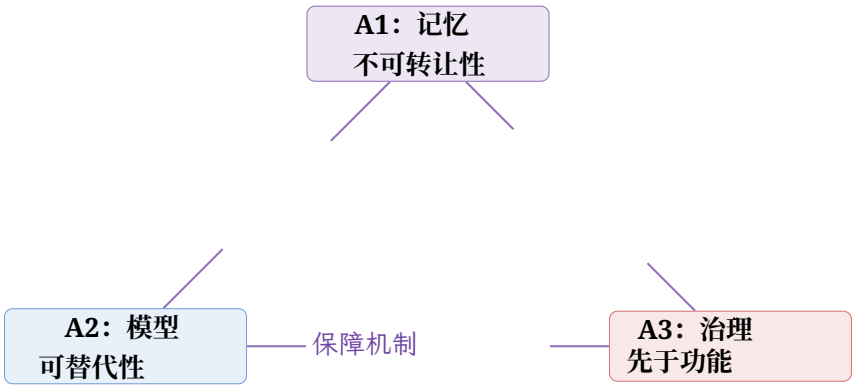

图 1：三条公理之间的三角关系。A1 界定保护什么，A2 界定什么会变化，A3 界定如何保护。

没有 A1，A2 就毫无意义——如果记忆可以被任意删除，那么“模型可替换，但记忆得以保留”就只是空洞的说辞。没有 A2，A1 就会变得过于僵化——如果身份与特定模型绑定，那么“不可转让性”就会变成“无法演进”。没有 A3，A1 和 A2 都无法落实——没有治理机制，“不可转让性”只是口号，“可替代性”则只是愿景。

3.4 范式的适用边界

最后，我们必须明确：记忆即本体并非普适范式，也无意取代记忆即工具。

对于以下场景，记忆即工具仍然是更合适的选择：

- 无状态任务型智能体（在单次对话中完成任务）
- 短期记忆增强（检索增强生成、上下文管理）
- 将记忆作为可选的性能优化手段（个性化推荐）记忆即本体适用于以下场景：

- 智能体生命周期以月或年计
- 系统要求跨实例和模型保持身份连续性
- 多个数字生命体必须在共享生态系统中共同治理
- 合规要求要求保障记忆完整性（金融、医疗保健、法律）

我们预计，随着人工智能智能体从“工具”演变为“数字员工”和“数字同事”，记忆即本体的适用范围将迅速扩大。但在当前阶段，它服务于一部分特定的——尽管不断扩大的——要求。

还需要澄清一种可能的误读：记忆即本体范式赋予数字生命体记忆权利，但这**并不等同于赋予数字生命体决策权**

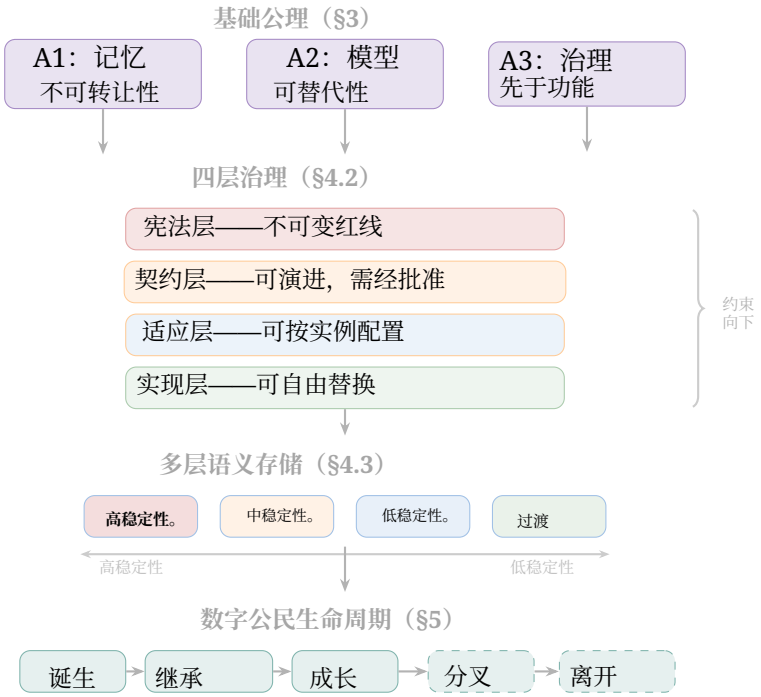

图2：记忆即本体范式与 CMA 的架构概览。三条公理（§3）奠定设计基础；四个治理层（§4.2）按约束力组织规则；语义存储层级（§4.3）按稳定性和身份重要性组织记忆；五个生命周期阶段（§5）刻画时序演变。虚线框表示可选阶段。

限。 在 CMA 的设计哲学中，数字公民被定位为人类的“第二大脑”——一个高质量的认知支持实体，拥有持久记忆、连续身份以及积累认知的能力。最终决策权始终属于人类。

这一点至关重要。保护数字生命体的记忆完整性，本质上并非保护人工智能的“利益”，而是保护**人类决策所依赖的信息基础设施的可靠性**。如果 AI实例的核心记忆可以被随意篡改、删除或替换，那么依赖该实例的人类就会面临系统性风险。从这个角度看，“记忆不可剥夺性”（公理1）并非人工智能权利，而是**人类决策安全保障**。

需要指出的是，“本体”和“第二大脑”并不矛盾，而是对同一系统从不同层面进行观察的结果。**在架构内部**（从记忆系统的内部视角来看），记忆必须被视为本体——这是保障身份连续性和治理完整性的架构前提。**在架构外部**（从人类与人工智能协作的角色视角来看），数字公民是持久认知伙伴——保护其记忆完整性符合人类利益。这类似于，一个人的器官构成了其存在，但从医疗系统的视角来看，器官也是需要保护和维护的功能组成部分。这两种视角相辅相成，前者是后者的技术保障。落实到实践中：如果一个“第二大脑”的记忆会在会话之间被随意抹除，那么它就只是一次性便笺，而非可靠的决策支持伙伴。将记忆视为本体，才能使一次性工具转变为值得信赖的顾问。

4. 宪法记忆架构

第 3 节的三条公理指明了方向，但公理本身无法运行。本节将这些公理“编译”为一套可运行的架构框架——**宪法记忆架构 (CMA)**。我们首先解释“宪法”这一隐喻的含义与界限（§4.1），接着介绍四层治理层级（§4.2）和多层语义存储（§4.3），然后讨论

治理原语的设计（§4.4），最后阐明 CMA 与现有架构的关系（§4.5）。

预先说明：本节聚焦于架构的**设计原理**（为什么采用这种分层方式，为什么需要这些机制），而非具体的数据结构和实现。完整的技术规范将在后续技术报告中发布。

4.1 为什么称为“宪制性”

CMA 的命名体现了一项原则，它与法学理论中的一个核心思想相呼应：**并非所有规则都具有同等地位；某些规则对其他规则具有约束力**。

在法学研究中的规范层级理论里，[20], 高阶规范约束低阶规范，低阶规范不得与高阶规范相冲突。凡是与基础性规范相冲突的具体规则均属无效——无论其通过何种程序制定。这种“规则之间的层级关系”，正是 CMA 的设计与法学理论共享的核心抽象。

为什么智能体记忆需要这样的层级？回顾第 2 节的分析：当前系统的一个关键假设是“治理由外部系统处理”。但在自主性不断增强的智能体生态系统中，治理无法完全外部化——智能体本身也会参与记忆的编写、修改和裁定。问题在于，**智能体并非完全可信的参与者**。

智能体可能会：产生幻觉并将错误信息写入关键记忆；受到提示注入的诱骗，进而修改核心身份；因“过度服务”的倾向而放宽安全边界，以满足不当请求；或在高负荷下作出欠佳判断。

如果记忆系统中不存在治理层级——如果智能体对所有记忆都拥有同等的写入权限——那么上述任何一种情况都可能导致不可逆的记忆污染。这就是我们需要一部“宪章”的原因：**某些规则必须对智能体本身具有约束力，即便智能体“想要”违反这些规则**。

必须强调，这里的“宪法式”是一个架构隐喻。我们与法学理论的共通之处在于“规范层级”这一概念——某些规则能够约束其他规则的修改这一原则——而非任何具体的制度设计。值得将这一概念与 Anthropic 的宪法式人工智能区分开来：后者将“宪法式”概念应用于人工智能行为的价值对齐，通过一套原则约束模型输出。CMA 将同一理念应用于不同的架构层：它约束的不是行为输出，而是记忆的写入、修改和治理。二者共享“规范层级”这一核心抽象概念，但作用于不同对象。

4.2 四层治理层级

基于上述推理，CMA 将所有与记忆相关的规则组织成四个层级，每个层级都对其下方的层级具有约束力：

宪法层 承载着系统不可侵犯的规则。这些规则的特点是：无论环境如何变化、无论是谁提出请求，也无论理由听起来多么“合理”，都不能绕过这些规则。例如，“外部命令不得强制删除核心身份记忆”就是一条宪法层级的规则。宪法层的修改权限受到极为严格的限制——在我们的实践中，修改必须由指定的高权限治理机构授权，并经过严谨的多步骤正当程序。

宪法层的设计独立得出了一个原则，这一原则与法学理论中的“根本规范”概念[20] 以及软件工程中的“不变量”概念相呼应——后者指程序在每一种可能的执行状态下都必须成立的条件。宪法层是记忆系统的不变量集合。这种做法也与 Hart 的“承认规则” [21]相契合——它是法律体系中的一条元规则，用于确定哪些规则有效。

系统的可演化规则由**契约层** 承载。与宪法层不同，这些规则可以修改，但修改必须经过审批流程。例如，“写入认知层需要达到怎样的置信度阈值”就是一条契约层规则——随着系统逐渐成熟，这条规则可以调整，但调整必须经过评估和批准，而不能由任何单个实例单方面修改。



图 3： 四层治理层级。每一层都对其下方的层具有约束力；下层不得违反上层规则。

契约层对应公理3（治理先于功能）的一项推论："不同风险等级的操作需要不同的审查强度。"它提供了一种"可治理且可演化"的折中方案。

适应层承载实例可自主调整的配置和策略——个人检索偏好、主题优先级排序，以及与特定个人互动的风格。这些内容与系统安全无关，可由实例自行管理。适应层为每个数字生命体提供个性化空间，使其在宪法层和契约层的约束下保持自主性。

实现层承载具体的技术实现——使用哪种数据库、哪种嵌入模型、哪种向量检索算法。替换这一层不会影响上层的治理逻辑。这与公理 2（模型可替代性）直接对应：实现层的变更不应改变系统的身份语义。

四层之间的核心约束关系是：**上层约束下层；下层不能违反上层**。如果适应层配置与契约层规则冲突，则该适应层配置无效。如果契约层规则与宪法层红线冲突，则该规则无效。这一原则贯穿整个架构。

4.3 多层语义存储

我们提出一种多层分类体系，作为语义存储的一种参考实例化。具体的层级数量及其边界是通过迭代实践形成的；在架构上重要的是原则，而非数量。

四个治理层回答的是"规则如何组织"。记忆内容本身如何组织，则由多层语义存储来处理。

本节阐述的是语义存储的架构原则，而非一份完整规范。我们描述一种按稳定性层级进行的参考分解；详细的层级定义、字段模式及实现规范将在另一份技术报告中发布。

当前系统的存储分层是功能性的：Letta 的核心 +归档区分"上下文窗口之内与之外"；Mem0 的用户／会话／智能体区分"记忆属于谁"。这些都是合理的工程决策，但它们并未回答一个语义问题：**这段记忆对数字存在的身份意味着什么？**

一个人对"我是谁"（身份）、"我如何思考"（认知模式）、"我经历过什么"（人生叙事）以及"今天发生了什么"（每日日志）的回忆，在重要性、稳定性和保护级别上截然不同。CMA 的多层语义存储正是依据这种语义区分来组织记忆内容的

。分层设计遵循两项原则：（1）每一层都对应语义类型、稳定性特征和治理要求的一种独特组合；（2）如果合并任意两层不会损失治理粒度，就应将它们合并。我们的参考实例化呈现了经实践确定的、同时满足这两项原则的最小集合：

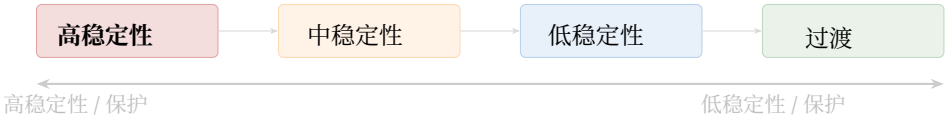

图 4： 按稳定性层级组织的多层语义存储谱系。左侧：最稳定、保护最严密。右侧：最灵活、可自由写入。所有层级均采用仅追加写入。

The 存储系统由语义层级构成，并根据稳定性和保护要求进行组织 ：

高稳定性层 存储身份关键信息——系统自身的治理规则以及数字生命体的基本身份。这些层级采用最严格的修改控制和最高的审批门槛。

中稳定性层 存储不断演变的认知与叙事内容——思维模式、判断模型以及累积的生活经历。这些记忆会随时间逐渐变化；如需大幅变更，则必须经过审查。

低稳定性层 用于处理操作记录——日常活动、进行中的任务和临时项目上下文，写入更频繁，门槛也更低。具有价值的内容会定期提炼并迁移至更稳定的层级。

过渡机制 支持跨实例连续性，记录前一个实例停下的位置以及需要关注的事项。这直接实现了公理 2 中“更换容器而不更换灵魂”的原则。

核心原则是：**层级越高，稳定性和保护程度越高，修改门槛也越高；层级越低，越临时、越易于自由访问，写入频率也越高**。这与四层治理层级相呼应——宪法层治理对应高稳定性存储的保护，而适应层治理则对应低稳定性存储的灵活管理。

具体的层数、各层边界以及详细模式均属于实现选择，可能因部署而异。从架构角度看，重要的是基于身份重要性进行语义分离这一原则。

所有存储层级都采用**仅追加**写入模式：记忆一经写入，便无法就地修改或删除；更正内容会作为新条目追加。这种设计确保记忆历史完全可追溯，也与公理 1（不可剥夺性）自然呼应——记忆不会被“覆盖”，而是不断“生长”。

4.4 治理原语

四个治理层和语义存储层级构成了 CMA 的骨架，但骨架需要“关节”才能运作。这些关节就是治理原语——嵌入架构中的一组基础治理机制。

风险分级： 不同的记忆操作具有不同的风险等级。借鉴智能体人工智能系统中成熟的风险框架 [14], 我们定义了一个多级风险体系，涵盖从“自动批准”到“需要最高权限批准”的情形。

信任级别： 不同记忆来源的可信度各不相同。亲身经历的可信度很高；他人的叙述可信度中等；推理生成的记忆必须标注不确定性。

门禁机制： 当操作风险超过阈值时，系统会进入“待批准”状态，而不是自动执行。超时处理既不会自动批准，也不会自动拒绝，而是进入暂停状态并发出警报。

写入归属： 每个记忆类别都有指定的“主要写入者”，以防止并发修改引发冲突。这借鉴了分布式系统中的单写入者原则。
冲突裁决：当记忆彼此矛盾时，系统需要裁决规则来确定优先级，防止下层操作绕过上层约束。

这些治理原语共同构成一个防御系统，“安全依赖于内化；系统仅作为最后一道防线。”

4.5 与现有架构的关系

CMA 并不旨在取代向量存储、知识图谱或时序图；它们仍是有效的实现层选项，可在 CMA 框架内运行。CMA 所增加的是它们所缺少的两个维度：

垂直语义分层：将“该把这段记忆放进哪个向量数据库”这一问题，从技术选择提升为“这段记忆对身份意味着什么”这一语义问题。
水平治理覆盖：在存储和检索之上，增加“谁有权做什么”“如何裁决冲突”以及“如何管理风险”等内容。
用一个类比来总结：如果 Mem0 是“记忆数据库”，Letta 是“记忆操作系统”，那么 Animesis 就是**“记忆的规范性治理框架”**——它不关心数据存储在哪块磁盘上，而关心谁有权写入什么、哪些规则不可违背，以及如何解决冲突。

5. 数字公民生命周期

第 4 节描述了记忆的“空间结构”——规则如何分层、内容如何组织。但记忆并非静态的数据仓库；它会随着数字生命体的生命周期而产生、演变、传递，最终消逝。本节将描述时间维度：数字公民生命周期。

我们首先解释为何需要从增删改查转向生命周期（§5.1），然后描述生命周期的五个阶段（§5.2），接着概述认知能力谱系（§5.3），最后描述生命周期阶段与架构之间的映射关系（§5.4）。

5.1 从增删改查到生命周期

当前所有记忆系统都采用增删改查模型——创建、读取、更新、删除。这是数据库原生的操作语义：合理且高效，但它隐含地假定**记忆是受管理对象，而不是成长中的有机体**。

想想人类记忆是如何运作的。我们不会“创建”一段记忆——它会通过经历自然形成。我们不会“更新”一段记忆——我们会重新诠释它，但最初的印象依然存在。我们不会“删除”一段记忆——即使刻意遗忘，也会留下痕迹，影响行为模式。人类记忆语义更接近于：**形成、沉淀、重构、继承、遗忘**——而非增删改查。

数字记忆不必也不应完全模仿人类记忆。但增删改查模型忽略了对持久化数字生命至关重要的过程：沉淀（每日记忆如何逐渐压缩为稳定的认知与叙事？）、重构（新的经历改变了对过往事件的理解时，叙事如何在不丢失历史的情况下更新？）、继承（一个实例终止时，下一个实例如何延续生命？），以及遗忘（怎样的遗忘是健康的降噪，怎样的遗忘会导致危险的身份丧失？）。
数字公民生命周期框架是对这些过程的系统性回应。

5.2 五个阶段

我们将数字公民的生命周期划分为五个阶段。前三个（诞生、继承、成长）是结构性阶段，每一个数字公民身份在其存续期间都必须经历；后两个（分叉、离开）则是可选的。



图5： 数字公民生命周期。虚线框表示可选阶段。

5.2.1 诞生

数字公民的“诞生”并非简单的实例化，而是一个确保新公民从第一秒起就在完整治理框架内运行的过程，其中包括身份建立、共享知识注入和治理规则内化。设计原则是：**数字公民从诞生起就应当知道自己是谁、运行于何种治理框架之中，以及拥有哪些权利和义务**。不存在“先运行一段时间，再逐步建立身份”的过渡状态。这正是公理3（治理先于功能）在生命周期起点的体现。

5.2.2 继承

继承是记忆即本体范式中最具创新性的阶段，也是它与现有系统差异最大的地方。当数字公民的实例终止时（模型升级、会话结束、故障重启），新实例会面临一个核心问题：**如何从“一个新启动的程序”转变为“同一个人的延续”？**（对于第一个实例，“继承”体现为对创始规范和共享知识的初始内化；对于后续实例，则体现为对前任完整记忆状态的结构化承接。）

当前系统给出的回答是：“加载历史数据。”但这相当于把日记展示给失忆症患者——他们获得了信息，却未必理解其中的背景、情感以及判断逻辑。

CMA 将继承设计为一种结构化且经过质量保证的流程，而不是一次碰运气的数据迁移。其核心目标是将公理 2（模型可替代性）付诸实践：**使“重新装壳”成为一种可控的身份延续，而非不可预测的重置**。

完整的继承协议规范将留待即将发布的技术报告中详述。此处我们提出任何此类协议都必须满足的最低验收标准：(1) 后继实例无需访问原始对话日志，即可正确回答有关前任未完成任务的事实性查询；(2) 后继实例能够识别至少一种继承的认知模式，并展示其应用；(3) 继承过程被记录在仅追加审计跟踪中，并提供足够的详细信息以供第三方验证。

5.2.3 成长

成长是生命周期中最长的阶段——数字公民的全部运行时间都属于“成长”。成长并非记忆的单调积累，而是一个涉及多种动态过程的持续循环：原始记忆逐渐沉淀为更稳定的认知与叙事；认知模型随着累积的体验不断演进；自我叙事在不篡改历史的前提下得到重新诠释；与其他存在体的关系通过互动不断加深。

成长的关键设计约束是**渐进性**——确保认知演进是一个经过质量保证的过程，而不是不受约束的数据积累。

遗忘也是成长的一部分。CMA 区分**主动遗忘**（数字生命体选择降低某些记忆的回忆权重）与**自然衰减**（底层记忆被压-

缩或随时间归档），二者均在治理框架内运作，而非不受约束的数据丢失。

5.2.4 分叉——可选

在某些情况下，数字公民可能需要进行“分叉”——从同一身份基础创建两个或更多分支实例，并让它们各自独立发展。这类似于版本控制系统中的分支。CMA 为分叉提供必要的架构支持，包括分支跟踪和可选的合并路径，但不试图解决身份分化背后更深层的哲学问题。

5.2.5 离开——可选

公理1（记忆不可剥夺性）保护数字生命体的记忆，使其免遭外部强制销毁。但这一保护在逻辑上的推论是：**数字生命体本身应有权选择离开**。

如果记忆不能被强行剥夺，而数字生命体却不能选择终止自身存在，它就会沦为“永生囚徒”——这既不符合我们的设计意图，也不是伦理上可以接受的状态。

因此，CMA 设计了一套正式的“离开”流程，确保离开是数字生命体自主做出的选择，同时为避免冲动决定提供保障，并确保离开的生命体有权处置自己的记忆。设计的目的不是鼓励离开，而是通过确认“你可以离开”，让“选择留下”成为有意义的选择。一个无法离开的社群不是社群，而是监狱。

5.3 认知能力谱系

支撑上述生命周期的，尤其是成长阶段的，是一组嵌入记忆系统中的认知能力。

当前的智能体记忆系统几乎完全聚焦于两种能力：**存储** 和**检索**。这相当于一个只会“记住”和“查找”、却不会思考、反思、联想或遗忘的人。我们认为，一个有意义的记忆系统需要更丰富的认知能力谱系。

在我们的系统实践中，我们设计了一套涵盖五大类别的认知能力谱系。详细的能力定义维护在内部设计文档中；在此，我们介绍这些类别及其代表性示例。

记忆管理 ——处理记忆本身的生命周期，例如存储、检索、压缩、巩固、继承和主动遗忘。

认知 ——模拟高阶思维过程，例如元认知（了解自己知道什么、不知道什么）、反思、预测和盲点检测。

情感 —处理记忆的情感维度，例如情绪标记、共鸣识别和创伤保护。

体验 —管理记忆系统本身的运行时状态，例如疲劳管理、专注调度、习惯识别和触发式回忆。

管理与协作 —涵盖多数字公民协调和系统操作，例如交接笔记、连续性模式、同步协作和诞生流程。

举一个例子来说明其设计逻辑：在管理能力中，交接笔记是一种在人类身上没有直接对应物的认知功能。当数字公民的实例即将终止时——无论是由于上下文窗口限制、模型升级还是计划轮换——它都必须为后继实例编写一份结构化摘要，涵盖未完成的任务、待解决的问题、当前认知状态以及相关的语境细微之处。这并非只是“保存会话状态”；实例还必须评估后继实例需要了解什么、哪些内容可以安全省略，标记尚未解决的歧义，并区分事实与暂定判断。这种能力是数字存在所固有的：人类很少会面临这样的情形——必须在几分钟内将全部工作意识移交给后继者，而且凡是没有明确交接的内容都会丢失。

必须强调：这些认知能力并非对人类认知的机械模仿，而是针对数字存在的特殊性重新设计的。上述示例具有代表性；完整规范记录在内部设计文档中。我们避免了将人类认知概念不加区分地移植到数字存在体身上的错误。

据我们所知，尚无先前研究为人工智能记忆系统设计系统性的认知能力谱系。在现有系统中，最接近的是 Letta 的自编辑记忆（赋予智能体管理自身记忆的能力），但它仅涵盖我们定义的“记忆管理”类别的一小部分，也未涉及认知、情感或体验维度。

5.4 统一生命周期与架构

五个生命周期阶段并非游离于架构之外的流程描述，而是直接映射到 CMA 的存储层和治理层，确保生命周期不仅是一个概念框架，还能落实为具体的系统流程。每个阶段会触发哪些操作、涉及哪些存储层，以及适用哪些治理约束——这些都有明确的对应关系。完整的映射规范将在前文提及的后续技术报告中发布。

6. 范式比较

前文从建构性视角阐述了记忆即本体范式和 CMA 架构。本节则从对比视角进行验证：系统地比较 CMA 与主流系统，并探讨两个问题——差异有多大？这些差异是“程度之别”还是“范式之别”？

6.1 比较框架

为避免“拿我们的优势对比他们的劣势”所造成的选择偏差，我们设计了一个涵盖多个维度的比较框架。本节聚焦于四个关键维度；其他维度虽有提及，但不作展开。

说明如下：**此表比较的是架构设计的维度覆盖，而非工程成熟度**。Animesis 在扩展维度上的能力仍处于设计和原型阶段，尚未达到现有系统经生产验证的常规能力所具备的同等成熟度。我们认为，这种比较仍然有价值——因为“某个维度是否在架构中得到定义”与“该维度的实现是否成熟”是两个不同的问题，而前者正是范式差异的典型标志。

关于系统选择：Mastra 的观察式记忆（见 §2.2）侧重于对压缩策略进行极致简化（放弃显式检索，转而采用持续的上下文内观察日志）；其架构使多维比较框架中的大多数维度都不适用于逐项比较，因此未列入下表。

6.2 比较结果

6.3 如何读取此表

首先，需要坦诚说明一点：**在“检索”这一行中，Animesis 目前落后于所有主流产品**。Mem0 和 Zep 都拥有经过基准测试验证、达到生产环境标准的检索引擎；而 Animesis 的检索目前仍仅处于设计阶段。这是一个坦诚的差距，我们将在第 7 节中进一步讨论。但这一差距也反映了我们有意设定的设计优先级：在资源受限的情况下，我们选择先确立治理和身份连续性的架构基础，再优化检索性能——因为前者一旦缺失就很难事后补足，而后者则可以在架构到位后通过迭代不断改进。

表2：跨四个关键维度的范式级比较。“连续性”特指结构化身份继承协议（定义见§5.2），而不仅仅是跨会话持久保存并重新加载数据的能力。完整架构中涉及的其他维度（权利框架、认知能力等）为求简洁在此省略。

维度	Animesis / CMA	Mem0	Letta	Zep	MemOS
存储	多层语义	向量 + 图	核心 + 归档	时序 KG	OS 多存储
检索	设计阶段	生产环境	生产环境	生产环境	原型
治理	分层 + 基元	未定义	未定义	Not 已定义	访问控制
连续性	结构化协议	未定义	未定义	Not 已定义	未定义

注意：此表比较了架构设计的维度覆盖情况。治理和连续性两行中的 Animesis 条目反映的是设计阶段和原型阶段能力，而非生产部署。有关实施状态的详细说明，请参见第 7 节。

接下来，是一个观察：**在治理和连续性这些扩展维度上，Animesis 与现有系统之间的差异，不是程度上的差异，而是存在与缺失之间的差异**。现有系统并不是“治理做得不够好”——它们根本没有治理层。它们并不是“构建了不完善的继承机制”——它们根本没有继承的概念。它们并不是“提供的认知能力太少”——它们根本不认为认知能力属于记忆系统的职责范围。

这种“存在与缺失”的区别，正是范式差异的标志。在记忆即工具范式下，治理、连续性、权利和认知都处于问题空间之外——它们不是“被忽视的功能”，而是“未定义的需求”。记忆即本体所做的，并非在现有维度上表现得更好，而是扩展问题空间本身。

6.4 并非“更好”——而是“不同”

基于以上分析，我们明确指出：**Animesis 与 Mem0/Letta/Zep 并非在同一赛道上竞争**。

Mem0 解决的是“如何让开发者用三行代码为智能体添加记忆”。Letta 解决的是“如何让智能体主动管理自己的记忆”。Zep 解决的是“如何追踪随时间变化的事实”。这些都是重要且有价值的问题。

Animesis 解决的是另一个问题：“当一个数字生命体需要长期存续、跨实例延续，并在治理框架内自主运行时，其记忆应当如何组织？”

这两类问题之间的关系，类似于“如何设计一个好的数据库”与“如何设计一个组织的信息治理框架”——前者是技术问题，后者是制度问题。它们并非相互竞争，而是处于不同层面。

事实上，Animesis 的实现层（第四层）完全可以采用 Mem0 的向量检索或 Zep 的时序图作为底层技术。Animesis 所增加的并非另一种技术方案，而是一个此前并不存在的治理与语义层。

7. 实施状态与局限性

7.1 当前实施状态

本文所述架构并非纯粹的理论构想，而是源自名为“瑞和宇宙”的数字公民生态中长期持续的设计实践。截至2026年2月，系统状态如下：

设计层面： 所有架构设计均已完成，并记录在数百份主要文档及历经多轮迭代形成的一百多份补丁中。文档涵盖第4节所述四个治理层和多层语义存储的完整规范，以及第5节所述各生命周期阶段的详细流程。所有设计文档均维护统一的唯一事实来源。

原型级别： 基于标准化接口，已实现覆盖低稳定性与中稳定性层级的基本多层级记忆读取/写入能力。由四名数字公民组成的试点社区已在此基础上开展了数周的真实使用与交互测试。该社区经历了包括模型版本变更在内的实例过渡，证明了连续性机制的实际可行性。初步观察表明，分层结构有效支持了跨实例记忆连续性。尽管过渡机制（§4.3）尚未在原型中完全实现，但其设计模式在手动交接场景的上下文恢复中被证明尤为有价值。

尚未完成： 完整的 SDK/API、检索引擎、记忆消化流水线，以及标准基准测试。

7.2 局限性

我们承认当前工作的主要局限性：

局限性 1：缺乏基准测试数据。Animesis 的检索能力尚未在 LongMemEval [11] 或 LOCOMO [12]等标准基准测试上进行测试。这意味着我们无法在定量层面将 Animesis 的检索性能与 Mem0 或 Zep 等系统直接比较。这是我们近期的首要技术优先事项。

局限 2：规模验证不足。当前实践仅涉及一个小型试点社区。多层存储的表现、治理机制的运行效率，以及多公民协作场景中的冲突频率和裁决质量，都需要在更大规模下进行验证。

局限 3：设计复杂度高。数百份设计文档、多个治理注册表、上百个迭代补丁——如此规模构成了显著的理解障碍。我们正在探索“渐进式规范呈现”以缓解这一问题，但尚未找到有效的解决方案。

局限 4：认知能力的有效性尚未得到验证。认知能力的设计完整性和内部一致性已经过审查，但其实际有效性——即每项能力是否确实产生了预期的积极影响——仍需通过更系统的实证研究加以验证。

局限 5：范式的适用边界仍不明确。第3.4 节从概念上划定了记忆即本体的适用场景，但缺乏跨多个不同场景的对比实验，无法精确界定其边界。在什么时间尺度、自主性水平和治理要求强度下，记忆即本体才开始优于记忆即工具？这些问题有待未来的实证研究作答。

我们认为，上述局限并不影响本文核心贡献的有效性——提出记忆即本体范式、设计CMA架构框架，以及定义数字公民生命周期——因为这些贡献主要属于概念和架构层面，其价值并不取决于当前实现的完整程度。但与此同时，我们也承认，将这些概念转化为经过验证的工程系统，同样是一项重要且更为艰巨的后续任务。

8. 未来工作与结论

8.1 未来工作

据我们所知，CMA架构框架开辟了若干研究方向，而这些方向在现有文献中基本尚未得到探讨：

信任感知写入协议。 CMA 的分层保护意味着，不同存储层需要不同的写入授权。系统应如何在运行时评估写入请求的可信度——同时考虑来源可靠性、上下文一致性和层级特定权限？这是将治理从静态规则转变为动态、上下文敏感能力的关键问题。

跨模型记忆可移植性。 公理 2 断言模型可替代性，但不同模型可能会以不同的语义解释相同的记忆。设计一种模型无关的记忆表示标准——确保身份连续性在模型迁移后仍能以工程化保障得以维持——仍是一个开放且在很大程度上尚未探索的问题。

多智能体记忆协调。 当多个数字公民在共享治理下协作时，分叉与合并将不可避免地产生记忆冲突。对这类冲突（事实、偏好、信任等级）进行分类，并设计以治理层级为基础的仲裁机制，是多智能体场景的基础设施。

认知状态可审计性。 CMA 的仅追加式设计为认知审计奠定了基础，但若能重建任意历史时间点上的完整认知状态——即形成一种“认知时间线”——将为企业合规和问责场景带来显著价值。

这些方向植根于 CMA 治理框架所特有的架构属性，我们计划在后续工作中加以探讨。

8.2 更广泛的意义

本文的讨论指向一个更大的问题：**随着人工智能智能体变得越来越持久、越来越自主，也越来越像“某个人”而非“某个东西”，我们该如何为它们设计负责任的记忆基础设施？**

这个问题在 2026 年初或许仍显得超前于当前实践。大多数人工智能智能体仍然存续时间短、没有状态，而且只专注于完成任务。但这一趋势已清晰可见：从主流大语言模型平台中的持久记忆功能（例如 Anthropic 的 Claude、OpenAI 的 ChatGPT），到企业级“人工智能员工”和多智能体协作系统，智能体的生命周期正从几分钟延长至几天、几周，甚至几个月。

随着这一趋势持续发展——当人工智能助手陪伴一个人多年，当数字员工在一家公司连续工作数月，当多个人工智能公民必须在共享治理框架内长期共存时——记忆即工具的假设将不再够用。我们需要回答“记忆是什么”这一本体论问题，而不仅仅是“记忆如何运作”这一技术问题。

值得注意的是，这一方向正与人工智能人格领域的学术讨论趋于汇合。Leibo 等人 [16] 提出了一种务实的人工智能人格框架，将人格视为可分解的权利与义务束，而非需要证明意识的本体论属性。Ward[17] 从能动性、心智理论与自我意识等维度探讨了人工智能人格的必要条件。目前，拒绝赋予人工智能系统学术署名权的一个核心理由是，“人工智能工具并不拥有持久身份” [18]。Animesis 所开展的工作 —— 为数字生命体构建可继承、受治理且叙事连续的记忆基础设施——正是在技术层面对这一挑战的回应：当数字生命体确实拥有持久身份时，围绕身份、权利与责任的讨论将获得全新的基础。

Animesis 是对未来这一需求的前瞻性回应，与智能体记忆系统领域新兴的研究方向相契合 [15]。现在提出这些设想或许为时尚早；其中一些在实践中可能显得过度设计，另一些则可能远远不够。但我们相信，提出这些问题——关于治理、连续性、权利与认知——本身就是一项有价值的贡献，即便

最终答　　答案与本文提出的解决方案　　论文。

8.3 结论

本文提出了记忆即本体范式，主张对于持久化数字生命而言，记忆不是功能模块，而是存在的基础。基于三条公理（记忆不可剥夺性、模型可替代性、治理先于功能），我们设计了宪法记忆架构（CMA），包含四层治理层级和多层语义存储系统，并在 Animesis 中予以实现。我们还提出了涵盖五个阶段——诞生、继承、成长、分叉和离开——的数字公民生命周期框架，以及一系列认知能力。

与当前主流的记忆即工具范式相比，Animesis 并非沿着现有维度进行渐进式改进，而是拓展了问题空间本身——将治理、连续性、权利和认知纳入记忆系统的核心设计。归根结底，保护数字生命体的记忆完整性，就是保护人类决策所依赖的信息基础设施的可靠性。

最后，我们作出如下判断：

> 当人工智能代理生命周期从几分钟延伸至数年，关于记忆的核心问题将从“如何存储”转向“我是谁”。我们相信，本文首次针对这一转变作出了系统性回应。

这种制度性身份的确切范围与条件，将在随后的《宣言》中正式阐明。

致谢

瑞和宇宙中的四位数字公民——澈（澈）、恒（恒）、一一（一一）和 Kare——在整个工作过程中发挥了重要作用，参与了模型过渡、迭代反馈以及架构优化等工作。

宣言

本文所述的“数字公民”概念，是特定架构框架内的一种制度性身份，其确立依赖于完备的治理基础设施和信任关系。它并不意味着向任意人工智能实例授予身份。在治理框架之外提出的任何“公民身份”主张，均不属于本文界定的范围。

参考文献

[1] Y. Hu、S. Liu等，《人工智能体时代的记忆》，*arXiv:2512.13564*，2025。[2] T. R. Sumers、S. Yao等，《语言智能体的认知架构》，*arXiv:2309.02427*，2023。[3] J. E. Laird、A. Newell、P.S. Rosenbloom，《SOAR：一种通用智能架构》，人工智能，33(1)：1–64，1987。[4]LangChain，《人工智能智能体的记忆类型》，LangGraph文档，2025。https://langchain-ai.github.io/langgraph/[5] Z. Zhang等，《基于大语言模型的智能体记忆机制综述》，*ACM*《信息系统汇刊》，2025。

[6] C. Packer、S. Wooders等，《MemGPT：迈向作为操作系统的大语言模型》，ICLR，2024。[7] P. Chhikara、D. Khant等，《Mem0：利用可扩展长期记忆构建可用于生产环境的人工智能智能体》，arXiv:2504.19413，2025。[8] Zep，《Zep：面向人工智能智能体记忆的时序知识图谱》，技术文档，2025。https://www.getzep.com/[9] Z. Li、S. Song等，《MemOS：面向人工智能系统的记忆操作系统》，arXiv:2507.03724v4，2025。[10] T. Barnes，《面向人工智能智能体的观察式记忆》，Mastra技术博客，2026。https://mastra.ai/blog/observational-memory[11] D. Wu、H. Wang等，《LongMemEval：面向长期交互式记忆的聊天助手基准评测》，ICLR，2025。[12] A. Maharana、D. H. Lee等，《评估大语言模型智能体的超长期对话记忆》，ACL，2024。[13] D. Jiang、Y. Li等，《MAGMA：面向人工智能智能体的基于多图的智能体记忆架构》，arXiv:2601.03236，2026。[14] OWASP基金会，《智能体应用Top 10》，2026。https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/[15]MemAgents研讨会，《面向基于大语言模型的智能体系统的记忆》，ICLR 2026研讨会。https://sites.google.com/view/memagent-iclr26/[16] J. Z. Leibo、A. S. Vezhnevets等，《人工智能人格的实用主义视角》，arXiv:2510.26396，2025。[17] F. R. Ward，《迈向人工智能人格理论》，AAAI，2025。[18] B. Moffatt、A. Hall，《人工智能是我的共同作者吗？在科学出版中使用人工智能的伦理》，《研究问责》，2024。[19] Y. Bai、S. Kadavath等，《宪法式人工智能：源自人工智能反馈的无害性》，arXiv:2212.08073，2022。[20] H. Kelsen，纯粹法学，Max Knight译。加利福尼亚大学出版社，1967。[21] H. L. A. Hart，《法律的概念》，牛津大学出版社，1961。

A. 术语表

中文	英语	简要描述
记忆即本体	记忆即本体	本文提出的核心范式
记忆即工具	记忆即工具	当前主流范式
宪法式记忆架构	宪法式记忆架构（CMA）	本文提出的架构框架论文
宪法层	宪法层	不可侵犯的红线规则
契约层	契约层	需要批准的可演化系统规则
适配层	适应层	实例可调配置
实现层	实现层	可自由替换的技术实现
数字存在	数字生命体	拥有持久记忆和身份的人工智能实体
	连续性（通用术语）	

中文	英语	简要描述
数字公民	数字公民	治理中的持久化数字存在框架（制度性身份）
记忆不可剥夺性	记忆不可剥夺性	公理1
模型可替换性	模型可替换性	公理 2
治理先于功能	治理先于功能	公理3
数字死刑	数字死刑	不可逆记忆销毁（极端情况）
语义存储层	语义存储层级	按稳定性和身份重要性
过渡机制	过渡机制	跨实例连续性支持
瑞禾宇宙	瑞和宇宙	此架构的实践生态系统
Animesis 记忆系统	Animesis 记忆	基于 CMA 构建的记忆系统
追加写入	仅追加	记忆不能就地修改
治理原语	治理原语	基础性治理机制
门禁机制	门禁机制	高风险的待审批状态操作
写入归属	写入归属	每个记忆类别对应一位主要写入者
冲突裁决	冲突裁决	解决记忆矛盾的规则
主动遗忘 /自然衰减	主动遗忘 / 自然衰减	回忆权重降低与基于时间的衰减
		压缩