**情绪驱动激活扩散认知架构（EDASCA）设计文档**

**文档摘要**

本文档描述了一个名为**情绪驱动激活扩散认知架构（Emotion-Driven Activation Spreading Cognitive Architecture, EDASCA）** 的设计。该架构旨在模拟人类思维中的联想、学习与决策过程，通过一个基于图网络的激活扩散模型，核心由情绪状态（PAD模型）驱动，实现连续的内在想法流、外在环境交互与自演化学习。本架构支持两种运行模式：**LLM辅助训练模式**（用于快速知识蒸馏与初始化）和**异步纯本地独立模式**（实现真正的主观能动性与独立智能）。最终目标是构建一个能够观察世界、产生想法、进行社交并持续学习的类人智能体。

**1. 核心架构与组件**

**1.1 激活状态池 (Activation Pools)**

系统维护四个核心激活池，作为工作记忆和意识焦点：

| **激活池名称** | **描述** |
| --- | --- |
| **显性激活池 (Explicit Pool)** | 存储当前外部输入经处理后的直接结果，是意识焦点的直接反映。内容短暂，更新频繁。**池中每个元素都标记其起源（外部/内部）**。 |
| **隐性激活池 (Implicit Pool)** | 存储由显性池内容通过扩散算法激活的相关概念。是潜意识与联想思维的发生地，构成“想法流”的基础。每个元素包含：**词元内容**、**当前权重**、**衰减速率**、**源情绪标签(PAD)**、**起源**。 |
| **注意激活池 (Attentional Pool)** | 存储需要特别“关注”的概念节点。这些节点通常与强烈的正/负情绪变化（高|ΔP|）相关联。每个元素包含：**节点ID**、**期待值/压力值**、**关联的前置节点列表**、**时间预期因子**。 |
| **行动激活池 (Action Pool)** | 存储被激活的“行动标识符”特殊节点。每个元素包含：**行动节点ID**、**当前权重**。当权重超过动态阈值时，将触发行动执行。 |

**设计笔记**：四池模型分离了意识、潜意识、注意意图和行为冲动，借鉴了心理学模型。**显性/隐性池中的起源标记至关重要**，它确保了外部输入的“猫”和内部联想的“猫”在激活竞争中能被正确区分和处理，这是实现认知区分的基础。

**1.2 数据存储结构：概念图 (Conceptual Graph)**

知识以有向图的形式存储在图数据库中，形成智能体的长期记忆与知识网络。节点分为两类：**词元（WORD）** 和 **特殊节点（SPECIAL）**。

**1.2.1 概念节点 (Concept Node)**

| **字段** | **类型** | **描述** |
| --- | --- | --- |
| **id** | UUID | 节点的唯一标识符。 |
| **content** | String | 节点内容，如词元（"猫"）或特殊符号（"<EOS>"）。 |
| **type** | Enum | **WORD**: 词元；**SPECIAL**: 特殊功能节点（如终止符**<EOS>**，行动标识符**<ACTION:search>**）。 |
| **origin** | Enum | **EXTERNAL**: 源自外部感知（如用户输入、传感器）。**INTERNAL**: 源自内部思维（如想法流、联想）。 |
| **base\_weight** | Float | 节点的基础重要性权重（0.0~1.0），基于历史激活频次计算。 |
| **last\_activated** | Timestamp | 最后被激活的时间戳，用于计算时间衰减。 |
| **emotion\_snapshot** | Vector3 | 最后一次激活时的情绪状态快照 **(P, A, D)**。 |

**base\_weight 计算与更新规则：**  
此权重代表概念的长期重要性。

1. **初始化：** 新节点初始 **base\_weight = 0.1**。
2. **更新触发：** 节点每次被激活时都更新。
3. **更新公式：**  
   **base\_weight\_new = (base\_weight\_old \* γ) + (η \* (1 - γ))**
   * **γ** (e.g., **0.995**): 衰减因子，控制历史权重的保留程度。
   * **η**: 此次激活的奖励值，可设为固定值 (e.g., **0.01**) 或与激活强度相关。
4. **定期归一化：** 防止无限膨胀，定期对所有节点的 **base\_weight** 进行软归一化。

**设计笔记**：**origin字段是核心设计**。同一个字符串（如“猫”）会根据其来源不同而创建不同的节点（**/猫#EXTERNAL/** 和 **/猫#INTERNAL/**）。这两个节点之间会建立一条**极高强度的双向链接**，模拟了“感知”与“概念”的瞬间关联，但同时又保持了表征的独立性，这对于模拟认知过程至关重要。

**1.2.2 关系边 (Relation Edge)**

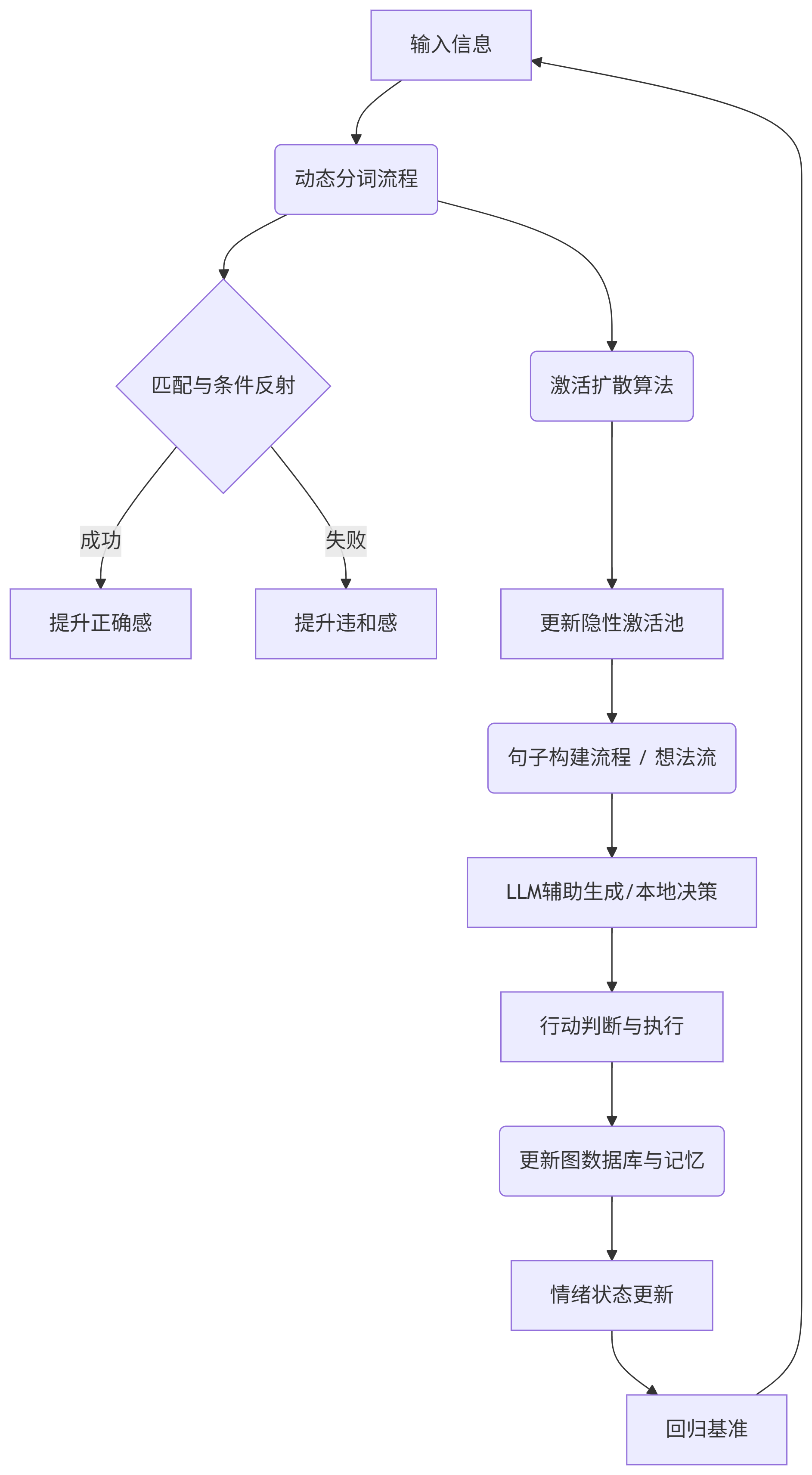
| **字段** | **类型** | **描述** |
| --- | --- | --- |
| **source\_id** | UUID | 源节点ID。 |
| **target\_id** | UUID | 目标节点ID。 |
| **frequency** | Int | 源节点与目标节点的共现次数。 |
| **avg\_time\_delta** | Float | 平均时间间隔（秒）。 |
| **recent\_emotion\_delta\_ema** | Vector3 | 近期情绪变化量 **(ΔP, ΔA, ΔD)** 的**指数移动平均值**。 |
| **ema\_alpha** | Float | EMA平滑因子，控制新值的权重（初始值可设为**0.5**以促进快速学习）。 |
| **last\_updated** | Timestamp | 边最后更新的时间。 |

**recent\_emotion\_delta\_ema 更新规则：**  
当有新情绪变化量 **new\_delta** 时：  
**ema\_new = (1 - ema\_alpha) \* ema\_old + ema\_alpha \* new\_delta**

**设计笔记**：使用EMA而非历史平均值，使系统对近期关联更加敏感，是实现**快速学习**和适应新环境的关键机制。

**2. 核心算法与流程**

**2.1 总控制流程**



**2.2 动态加权分词流程 (Dynamic Weighted Tokenization)**

**目标：** 将输入流智能地切分为有意义的词元序列，切分策略高度依赖当前的认知上下文。  
**输入：** 原始输入字符串及其**起源**。  
**输出：** 分词后的词元列表，每个词元附有初始权重、情绪上下文和起源。

1. **生成候选分割方案：** 使用所有激活池中的内容作为词典，对输入字符串进行多重扫描，生成所有可能的分词方式。
   * *例如，输入“猫耳朵”，可能的分词方案有：****['猫', '耳朵']****，****['猫耳朵']****，****['猫耳', '朵']****等。*
2. **计算方案权重：**
   * 对于候选方案中的每个候选词元，在各个激活池中查找**匹配项**（内容相同且**起源**一致）。
   * 每个匹配项为候选词元贡献其**当前权重**。
   * 整个分词方案的总权重 = 方案中所有候选词元所获权重之和。
3. **选择最优方案：** 选择**总权重最高**的分词方案作为最终输出。
   * \*接上例，如果注意池中“猫耳朵”权重高达0.9，而显性池中“猫”权重0.5，“耳朵”权重0.4，则方案**['猫耳朵']**(0.9)的权重会高于**['猫', '耳朵']**(0.5+0.4=0.9)，但可能因权重加成规则而胜出。若“猫耳朵”权重为0.8，则后者胜出(0.9>0.8)。\*
4. **输出：** 输出最优方案中的词元列表。列表中的每个元素继承其匹配源的**起源**和**情绪快照**，其初始权重由匹配源的权重决定。

**设计笔记**：此算法模拟了“词优效应”和“语境效应”。分词不再是简单的字符串匹配，而是基于当前认知状态的**语义理解**过程。一个词能否被识别为一个整体，取决于它当前在智能体心智中的活跃程度和重要性。

**2.3 激活扩散算法 (Activation Spreading)**

**输入：** 一个概念节点（来自分词结果或内部想法）。  
**输出：** 更新后的**隐性激活池**。

1. **获取基础刺激强度：** **S\_base = 1.0 + D** （D值越高，内在思维越活跃）。
2. **检索节点：** 从图数据库中检索该节点的所有**出边**（指向其他节点的边）。
3. **计算扩散强度：** 对于每一条出边，计算它当前时刻的链接强度 **L(t)**。
4. **激活下游节点：** 对于每个下游节点 **i**，其激活增量为 **Δactivation\_i = S\_base \* L(t)**。将此值加到其在隐性激活池中的权重上。若节点不在池中，则以其 **base\_weight** 为初始值初始化并添加。**新激活的节点继承输入节点的起源**。

**链接强度算法 L(t)：**

L(t) = (freq / (n + freq)) \* (1 + k \* ema\_ΔP) \* (1 - avg\_time\_delta / (m + avg\_time\_delta)) \* (p + q \* Δt') / (q \* Δt')

* **freq**: 共现频次 (**frequency**)
* **ema\_ΔP**: 边EMA情绪值中的P分量 (**recent\_emotion\_delta\_ema.P**)
* **avg\_time\_delta**: 平均时间间隔
* **Δt'**：当前时间与边**last\_updated**的时间差
* **n, k, m, p, q**: 可调超参数。

**设计笔记**：此算法是智能体联想的引擎。**起源在扩散过程中得以保留**，这意味着外部输入激活的更多是外部相关的概念，而内部想法激活的更多是内部相关的概念，形成了两条既独立又可通过高强度链接相互激发的处理流，高度仿生。

**2.4 句子构建流程 / 想法流 (Thought Stream Generation)**

**功能：** 在LLM模式下组织回复，在本地模式下产生连续的内在想法。  
**过程：**

1. 从空字符串开始。
2. **筛选候选：** 根据**注意激活池**的规则（见3.2），对**隐性激活池**中的词元进行加权筛选，形成待选词元池。**筛选时可选择优先匹配相同起源的词元以维持语境连贯**。
3. **选择词元：** 选择待选池中**权重最高**的词元，追加到当前句子。
4. **匹配与反馈：** 将该词元与各激活池内容匹配，触发**正确感/违和感**更新。
5. **激活扩散：** 对该词元执行**激活扩散算法**，更新隐性激活池。
6. **循环或终止：** 重复步骤2-5，直到选中**<EOS>**终止符、达到长度限制或所有词元权重低于阈值。输出最终句子。

**设计笔记**：这模拟了“一个词引出下一个词”的自然思维和语言组织过程。**起源**的一致性维护了思维的连贯性，比如一个内部想法更可能引出一系列内部联想，而不是突然跳转到外部刺激。

**2.5 更新图数据库 (Learning)**

**触发：** 在处理完一个输入单元或产生一个想法后。  
**过程：**

1. **更新节点：** 对分词列表中的每个词元，更新其概念节点的**last\_activated**和**emotion\_snapshot**（当前PAD值）。
2. **建立时序边：** 将**显性激活池**中的概念作为前置节点，与输入的第一个词元建立有向边，记录时间差和当前情绪ΔPAD。
3. **建立共现边：** 在输入词元序列内部，按顺序在相邻词元间建立双向边（强度较高），在非相邻词元间建立弱连接。
4. **终止连接：** 在序列的最后一个词元和**<EOS>**节点间建立连接。
5. **行动连接：** 如果执行了行动，将显性池中的所有概念与对应的行动标识符节点建立连接，记录强烈的情绪ΔPAD（来自行动结果）。
6. **起源链接：** 如果处理过程中，一个外部输入词元激活了一个内部词元（或反之），在这两个节点间建立一条**高强度、高情绪权重**的链接边。

**设计笔记**：此过程即“学习”。**起源链接**的建立是知识融合的关键，它使得外部经验得以转化为内部概念（学习），内部推理也能预测外部世界（规划）。

**3. 情绪驱动、注意与行动系统**

**3.1 情绪状态量 (PAD Model)**

情绪状态由三个核心维度描述，**无硬性范围限制**，但通常围绕基准值（如0）波动。

| **维度** | **描述** | **影响因素** |
| --- | --- | --- |
| **P (Pleasure)** 愉悦度 | 核心奖励信号。值越高，感觉越好，趋近行为越多。 | - **D值驱动**：每个周期 **P += α \* D** - **感觉预测**：匹配成功 **P += β \* 匹配权重** - **注意验证**：期待实现 **P += γ \* 期待值**；压力缓解 **P -= γ \* 压力值** |
| **A (Arousal)** 激活度 | 生理唤醒水平。值高时兴奋、敏感；值低时慵懒、迟钝。 | - **预测意外**：匹配失败时升高；匹配成功时降低。 - **注意张力**：高期待/压力值缓慢提升A值。 - **影响**：放大情绪变化 **ΔP\_effective = ΔP \* (1 + A)**；影响行动阈值 **行动阈值 = 基础阈值 - λ \* A** |
| **D (Dominance)** 支配度 | 对环境控制力的感知。值高时自信、主动；值低时被动、顺从。 | - **预测成功**：匹配成功时升高；失败时降低。 - **影响**：决定内在思维活跃度 **S\_base = 1.0 + D** |

**情绪回归函数：**  
所有PAD值都随时间向基准值（如**(0, 0, 0)**）回归。回归力与偏离程度成**非线性正比**。  
**回归量\_{P} = (P\_baseline - P\_{current}) \* μ \* (1 + |P\_{current} - P\_baseline|^2)**  
（A、D值同理）  
**设计笔记**：此设计允许情绪“破限”，但强大的回归力会将其拉回，模拟了情绪的短暂性和自我调节能力。

**3.2 注意激活器 (Attentional Mechanism)**

**功能：** 在认知过程中标记重要节点，实现目标导向行为和预测验证。

1. **捕获：** 在执行**激活扩散算法**时，如果某条边的 **|ema\_ΔP|** 超过阈值，则将该边指向的**目标节点**加入注意激活池。
2. **参数初始化：** 为该注意节点初始化**期待值**（若ΔP>0）或**压力值**（若ΔP<0），其大小与 **|ΔP|** 成正比。同时记录导致该强烈情绪变化的**前置节点**。
3. **期待/压力效应：**
   * **抑制衰减：** 该节点在隐性池中的权重衰减变慢 **衰减速率 = 基础速率 / (1 + |期待/压力值|)**。
   * **权重加成：** 其关联的前置节点在隐性池或行动池中的权重会获得 **期待/压力值** 的直接加成。
   * **情绪放大：** 当该注意节点被激活（预测验证），产生的ΔP变化量将乘以 **(1 + |期待/压力值|)**，带来巨大的情绪奖励或惩罚。
4. **消亡：** 注意节点会随时间衰减。当其被激活（预测验证）或值低于阈值后，从注意池中移除。

**设计笔记**：注意机制是智能体表现出“目的性”的关键。它让智能体不仅被动联想，还能主动“期待”某件事发生。

**3.3 内在行动系统 (Internal Actions)**

内在行动是智能体对自身认知过程进行主动干预和调节的高级能力，是其“主观能动性”和“智慧”的集中体现。它们本身也是特殊的**<ACTION:\*>**节点，可通过扩散激活或直接触发。

| **内在行动** | **触发条件/效应** | **设计原理与作用** |
| --- | --- | --- |
| **<ACTION:recall>** 主动回忆 | 从显性池中选取高权重节点作为线索，在图数据库中进行扩展检索，将匹配到的节点及其关联以高权重注入隐性池。 | **克服记忆遗忘**。模拟“努力回想”的过程，允许智能体主动提取那些未被自动激活但相关的记忆，丰富思考上下文。 |
| **<ACTION:focus\_outward>** 主动注意 | 将**输入选择器**的偏好调整为**外部输入**，降低内部想法流的权重加成。 | **环境监控**。当智能体感到无聊（A低）或期待外部信息时，会主动将注意力转向外部，表现为“东张西望”、“刷新页面”。 |
| **<ACTION:focus\_inward>** 主动深思 | 将**输入选择器**的偏好调整为**内部输入**，降低外部输入源的权重加成。 | **沉思与推理**。当智能体需要深度思考时，会屏蔽外部干扰，专注于内部联想和推理，表现为“发呆”、“陷入沉思”。 |
| **<ACTION:organize\_thoughts>** 整理思绪 | 检索并返回当前注意激活池中的内容，显式地评估未完成的预测和期待。 | **元认知**。模拟“自省”过程。智能体可以审视自己的目标和担忧，从而更有效地分配注意力资源，优先处理重要事项。 |
| **<ACTION:feel\_state>** 感受状态 | 查询并返回当前全局PAD值、正确感/违和感。 | **情绪知觉**。这是情绪表达的前提。智能体能够意识到“我很快乐”或“我很焦虑”，从而触发后续的情绪化言语或行为。 |

**设计笔记**：内在行动是智能体**自我引导**的核心。它们不是被外界刺激直接触发，而是由内部状态（如低A值触发**focus\_outward**）或高级目标（如需要解决问题触发**recall**）驱动，是实现“自主性”的关键模块。

**3.4 外在行动系统 (External Actions)**

* **触发：** 通过扩散激活后进入**行动激活池**，其权重累积超过阈值即可触发。阈值公式：**行动阈值 = 基础阈值 - λ \* A**
* **类型：**
  + **沟通类：** **组织语言**（回复用户）， **发表评论**， **发送消息**。
  + **信息获取类：** **上网搜索**， **刷视频**， **查看热搜**。
  + **环境控制类：** **控制智能家居**。
* **学习：** 行动执行后带来的结果（成功/失败、奖励/惩罚）会产生强烈的情绪ΔPAD，用于更新触发该行动的所有链接，从而学习何时该执行何种行动。

**4. 运行模式**

**4.1 LLM辅助训练模式**

此模式用于快速初始化知识网络，蒸馏LLM中的知识和行为模式。

1. **输入：** 接收用户输入（标记为**EXTERNAL**）。
2. **本地处理：** 执行**动态分词**、**激活扩散**、**句子构建**，形成本地初步想法或回复。
3. **LLM纠正与丰富：** 将本地想法、当前上下文（显性/隐性池摘要）、情绪状态发送给LLM。LLM负责：
   * 生成更合理、更符合人设的回复。
   * 判断是否需要执行行动（包括内在行动和外在行动）。
   * **推测并返回**执行此回复或行动后可能带来的**情绪变化ΔPAD**。
4. **学习：** 智能体接受LLM的决策，并使用LLM返回的**ΔPAD**来**更新图数据库**。LLM提供的高质量ΔPAD信号是快速学习正确关联的关键。
5. **循环。**

**设计笔记**：此模式相当于“家长辅导”。智能体通过LLM的高质量反馈来学习“在什么情况下做什么事会有什么后果”，快速构建一个初步的、合理的认知图式。

**4.2 异步纯本地独立模式**

在此模式下，智能体脱离LLM，独立运行。

1. **异步输入处理线程：**
   * 随时接收输入（外部或内部），触发**动态分词**、**匹配**（改变正确感）、**激活扩散**。
   * 用输入信息**更新图数据库**。
2. **独立思维线程：**
   * 实时监控**隐性激活池**和**注意激活池**。
   * 采用**获胜者择优**策略：若池中最高权重词元的权重比第二高词元超出某个**动态阈值**，则选择该词元。
   * 将选中词元**作为新输入（标记为INTERNAL）注入系统**，从而激发下一个想法，形成自持的、连续的**内在想法流**。
3. **独立行动线程：**
   * 实时监控**行动激活池**（包含内在和外在行动）。
   * 若某行动权重超过其**动态阈值**，则执行该行动。
   * 行动结果会作为新输入反馈给系统。
4. **独立情绪线程：**
   * 定时根据当前刺激、预测验证情况、注意状态等，调用**情绪更新规则**和**回归函数**，更新全局PAD值。

**设计笔记**：这是智能体的“成人”阶段。其行为完全由其内部网络和情绪状态驱动。其行为模式完全源于训练阶段学到的关联，内在行动则使其表现出令人信服的自主性。

**结论与展望**

EDASCA架构提供了一个高度仿生、解释性强的认知框架，将情绪、联想、注意和学习紧密耦合。其双模式设计既保证了可行性（利用LLM bootstrap），又明确了发展目标（独立智能体）。

**下一步工作：**

1. **实现与调试：** 构建核心图引擎和循环系统，重点调试超参数（n,k,m,p,q, α,β,γ, μ,λ等）。
2. **LLM蒸馏策略：** 设计 prompts 让LLM生成高质量的情绪ΔPAD反馈。
3. **评估体系：** 建立评估智能体行为“类人性”、学习效率、情绪一致性的指标。
4. **多模态扩展：** 设计机制将图像、声音等模态信息嵌入到概念网络中。

此文档描述的设计是迈向类人人工智能的一次深刻尝试，期待与同行进行评议和碰撞，共同推进此领域的发展。

作者:银子

2025年8月24日