Embedded Technoloy and Intelligent Systems 嵌入式技术与智能系统, 2025, 2(5), 305-318
Published Online October 2025 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2025.25030

Hans汉斯

基于三层心理学模型的人格智能系统

万学靖[1]，刘　伟[1]，曹　婷[1]，冯吭雨[1]，解运洲[1]，任付标[1]，过　弋[2]，宋广奎[3]

[1]上海布谷元创科技有限公司，上海
[2]华东理工大学信息科学与工程学院，上海
[3]电子科技大学自动化工程学院机器人研究中心，上海

收稿日期：2025年12月4日；录用日期：2026年1月16日；发布日期：2026年1月29日

摘　要

随着以ChatGPT、Claude、Gemini为代表的大语言模型(Large Language Models, LLM)技术的成熟，人机交互正从“操作式”向“理解式”范式演化。传统对话系统侧重任务执行，而新一代智能体开始展现出拟人化思维与情感表达的潜能。然而，现有系统仍在人格一致性、长期记忆以及低时延交互与深度优化的平衡方面存在关键缺口。本文提出人格智能系统(Personality Intelligence System, PIS)，以三层心理学模型(大五人格、三我理论、MBTI)为核心，构建人格的定义－决策－表达一体化机制。系统引入动静分离人格建模与双时域优化机制，在时间维度上实现短期自适应与长期演化的统一；并通过人格记忆网络(Personality Memory Network, PMN)整合短期上下文、长期事实与知识图谱，实现人格状态的可解释更新。输出层采用一致性锚定(Consistency Anchoring)机制，在人格、语义与情感维度上保持输出稳定性。本文进一步给出人格状态更新的数学形式、收敛条件及学习机制分析，证明系统在收缩映射假设下的人格演化具有稳定性。研究结果表明，PIS能在认知与情感层面实现人格的连续表达，为虚拟伴侣、教育辅导与心理健康支持等领域提供理论支撑。原型系统与相关数据集将于后续公开，以促进人格智能的开放研究。

关键词

人格智能，三层心理学模型，动静分离机制，双时域优化，人格记忆网络，一致性锚定，人格一致性，情感计算，大语言模型

Personality Intelligence System Based on a Three-Layer Psychological Model

Xuejing Wan[1], Wei Liu[1], Ting Cao[1], Hangyu Feng[1], Yunzhou Xie[1], Fubiao Ren[1], Yi Guo[2], Guangkui Song[3]

[1]Shanghai VibeSoul AI Technologies Co., Ltd., Shanghai
[2]School of Information Science and Engineering, East China University of Science and Technology, Shanghai
[3]Center for Robotics, School of Automation Engineering, University Electronic Science and Technology of China, Shanghai

Received: December 4, 2025; accepted: January 16, 2026; published: January 29, 2026

文章引用: 万学靖, 刘伟, 曹婷, 冯吭雨, 解运洲, 任付标, 过弋, 宋广奎. 基于三层心理学模型的人格智能系统[J]. 嵌入式技术与智能系统, 2025, 2(5): 305-318. DOI: 10.12677/etis.2025.25030

Abstract

With the maturity of Large Language Models (LLMs) such as ChatGPT, Claude, and Gemini, human-computer interaction is evolving from an "operational" paradigm to an "interpretive" one. Traditional dialogue systems focus on task execution, while the new generation of intelligent agents begins to exhibit potentials for human-like cognition and emotional expression. However, existing systems still face critical gaps in personality consistency, long-term memory, and the balance between low-latency interaction and deep optimization. This paper proposes the Personality Intelligence System (PIS), which centers on a three-layer psychological model (Big Five Personality Traits, Structural Theory of the Psyche, and MBTI) to build an integrated framework for personality definition, decision making, and expression. The system introduces dynamic-static decoupled personality modeling and a dual time-domain optimization mechanism, achieving the unity of short-term adaptation and long-term evolution over time. Through the Personality Memory Network (PMN), it integrates short-term context, long-term facts, and knowledge graphs to enable interpretable personality-state updates. The output layer employs a consistency anchoring mechanism to maintain stability across personality, semantic, and emotional dimensions. Furthermore, this paper formalizes the mathematical representation, convergence conditions, and learning mechanism of personality-state updates, proving that personality evolution under the contraction mapping assumption is stable. The proposed framework demon states the potential to enable continuous personality expression at both cognitive and affective levels, providing theoretical foundations for applications such as virtual companionship, educational tutoring, and mental health support. The prototype system and related datasets will be released subsequently to promote open research on personality intelligence.

Keywords

Personality Intelligence, Three-Layer Psychological Model, Dynamic-Static Separation Mechanism, Dual-Timescale Optimization, Personality Memory Network, Consistency Anchoring, Personality Consistency, Affective Computing, Large Language Models

1. 引言

1.1. 人机交互范式的演变

人机交互的发展历程，是一部不断追求更自然、更智能的对话界面的历史。从早期的命令行到图形用户界面(GUI)，再到语音助理(如 Siri、Alexa)和大型语言模型(LLM, Large Language Models)，人机关系逐渐从“工具”走向“伙伴”。在新范式下，人工智能(AI)不再仅仅响应指令，而需要以具备“人格”(Personality)的身份参与交互，形成持续的风格、一致的态度和可感知的情绪反应。

这一趋势催生了“人格智能(Personality Intelligence)”[1]的研究方向。与传统的“认知智能”不同，人格智能要求系统不仅能理解语言和情绪，还需在长期演化中保持人格特质与核心价值取向的一致性(Value Alignment within Personality Dynamics)，从而实现情感表达与决策行为的自洽性。为支撑此目标，研究者提出了“动态人格计算(Dynamic Personality Computation)”的概念——通过心理学模型实现 AI 人格的建构、演化与表达，是支撑情感交互与人格协同的关键技术支柱。

1.2. 现有工作进展与不足

近年的研究在人格计算、人格识别和人格驱动代理等方向上取得了显著进展。例如，[1] [2]从文本信号中推断人格特质；[3]-[5]探讨了人格在语言模型中的表达与控制；[6]研究了人格的动态演化。然而，这些方法仍存在以下局限：

- 人格一致性不足：缺乏稳定的人格约束机制，导致模型在长时交互中风格漂移；
- 记忆割裂与反馈缺失：多依赖上下文窗口，缺少跨会话记忆与自我修正；
- 实时响应与深度优化矛盾：低延迟需求与复杂人格更新的计算冲突；
- 个性化配置僵化：基于提示工程(Prompt Engineering)的静态人格定义缺乏灵活性。

1.3. 研究意义与创新点

为解决上述问题，本文提出了一个统一架构——人格智能系统(Personality Intelligence System, PIS)，并提出以下核心机制：

1) 三层心理学模型(Three-Layer Psychological Model)：将人格结构划分为基线层(Baseline Layer)、驱动层(Driver Layer)与表达层(Expression Layer)，分别对应人格的定义、决策与风格表达。

2) 动静分离人格建模机制(Static-Dynamic Separation Personality Modeling)：区分长期稳定的“静态人格”与情境自适应的“动态人格”，实现人格一致性与灵活性的统一。

3) 双时域人格计算框架(Dual-Temporal Personality Framework)：通过在线短时域与离线长时域相结合的优化机制，实现实时交互与长期人格演化的平衡。

4) 人格记忆网络(Personality Memory Network, PMN)：整合短期对话、用户画像与知识图谱，实现可解释的人格演化与记忆检索。

本文的主要贡献如下：

- 提出人格智能系统 PIS，并系统化地解耦人格的“定义–决策–表达”链路；
- 设计动静分离与双时域优化机制，实现低延迟与长期演化的兼容；
- 构建人格记忆网络 PMN，支持长期人格一致性与记忆增强；
- 从理论上分析人格状态更新的稳定性与收敛性；
- 探讨人格建模的伦理、可解释性与应用边界。

2. 相关工作

本章节首先回顾对话系统的演化历程，然后重点讨论人格智能建模、情绪与人格协同机制、记忆增强与知识结构化，以及多智能体人格一致性研究。我们还会指出这些方向与本文所提的人格智能系统(Personality Intelligence System, PIS)设计的差异与局限。

2.1. 对话系统的发展历程

人机对话系统的演化，清晰地呈现出从机械执行到智能共鸣的趋势[7]。

- 规则式阶段：以 ELIZA、A.L.I.C.E.为代表的早期系统，通过关键词与模板匹配生成回应。它们虽具开创性，但本质上不理解语义，也无上下文推理能力。
- 统计学习阶段：随着 Seq2Seq 模型与注意力机制的出现，对话生成进入数据驱动时代。基于 Encoder-Decoder 的神经对话模型开始能够生成更流畅、相关的回复。
- 大模型阶段：Transformer 架构的诞生，催生了 BERT、GPT 系列等预训练语言模型。这些模型凭借其强大的深度语义理解、长期依赖捕获与上下文推理能力，将对话系统的能力提升到了新的高度。

近年来，研究焦点已从单纯的“语言生成流畅度”转向“人格智能驱动的交互深度”。Character.AI、Replika 等商业产品尝试赋予 AI 稳定的人格，以增强长期对话的沉浸感和用户黏性。然而，这些系统当前多依赖静态的提示工程(Prompt Engineering)与参数微调(Fine-Tuning)，仍缺乏一个能够动态演化且保持长期一致的人格建模机制。因此，构建具备人格一致性、长期记忆与自我进化能力的人格智能系统(Personality Intelligence System, PIS)，已成为下一代交互 AI 的核心研究方向。

2.2. 人格智能建模

人格智能(Personality Intelligence)旨在让 AI 具备类似人类的、个体化的思维、行为与情感模式。其理论根基深植于心理学的人格理论。

经典的大五人格模型(Big Five Personality Model) [8]将人格解构为开放性(Openness)、尽责性(Conscientiousness)、外倾性(Extraversion)、宜人性(Agreeableness)与神经质(Neuroticism)五个核心维度，已成为计算人格建模的主流参考框架。与此同时，弗洛伊德三我理论(Id-Ego-Superego Theory)为理解人格的内部动态与决策机制提供了深刻洞见：本我(Id)代表本能欲望，自我(Ego)在现实中进行权衡，而超我(Superego)则代表道德与社会规范。该理论 AI 的行为决策与价值对齐提供了绝佳的心理学映射。此外，MBTI (Myers-Briggs Type Indicator)模型通过认知功能的组合(如思考–情感、直觉–感觉)，细腻地刻画了人格的外在表达差异，为系统生成特定风格的语言与交互策略提供了可操作的维度。

现有研究多聚焦于将人格特质向量化并映射到语言风格上。例如，Mairesse 等人[9]基于语言学特征来预测人格；Zhang 等人[10]则将人格向量嵌入到神经对话模型中以生成个性化响应。然而，这些工作大多属于静态人格建模，无法反映人格在与环境和用户的持续互动中的自适应变化。

近年来，动态人格建模开始受到关注。例如，Park 等人[11]的“生成式智能体”研究通过赋予智能体记忆和反思能力，模拟了其行为随时间演化的类人格特征；Kwon 等人[12]在 LLM 模拟的谈判场景中，探索了大五人格特质对谈判策略和结果的影响，验证了人格的动态作用；更有研究者开始探索将心理学理论更深度地融入 AI 代理，以增强其推理与协作能力[13]，并采用心理测量方法为 AI 代理分配可量化、可验证的人格分数。

本文提出的动静分离人格建模机制(Static-Dynamic Separation Personality Modeling)与双时域人格计算框架(Dual-Temporal Personality Framework)正是在此基础上的一次重要推进。它明确区分了保持长期一致的静态人格基线与响应短期变化的动态人格状态，并通过双时域优化循环，确保了短时域交互的流畅性与长时域人格演化的稳定性，从而有效解决了当前人格智能系统在一致性与自适应性之间的核心矛盾。

2.3. 情绪与人格协同机制

传统情感计算多基于认知评价理论，而本文的情绪生成机制则根植于人格结构本身。通过将三我理论与 MBTI 认知维度相融合，系统构建了从人格到情绪的内在逻辑通路：

基于弗洛伊德三我理论，系统在每次决策时通过内在调节机制模拟“人格平衡”过程：

- 本我(Id)负责生成初级驱动力与情感冲动；
- 自我(Ego)在现实与规范之间进行调和；
- 超我(Superego)施加价值与社会约束。

三者的相互作用并非竞争，而是一种动态平衡过程(Dynamic Equilibrium Process)，在系统中通过权重分配与反馈调节实现。MBTI 的思维/情感(T/F)与感知/直觉(S/N)维度，则进一步为情绪表达染上个体化色彩，使其更具人性化。

这种设计使情绪不再是静态标签或情绪分类结果，而是人格、认知与情境动态交互的产物，实现了人格–情绪–行为的一体化建模。情感计算领域的最新综述也强调了在心理健康等场景中，结合多模态信息进行情感支持的重要性，为构建共情交互系统提供了方法论上的参考[14]。

2.4. 记忆增强与知识结构化

长期、结构化的记忆是实现人格一致性的关键支柱。传统对话系统受限于上下文窗口长度，无法在长时域内维持用户画像和情感记忆的连续性。

与传统记忆增强不同，本系统将分层时间记忆体系与知识图谱(Knowledge Graph)结合，实现对用户特征、情绪轨迹与人格参数的结构化存储与推理。通过语义节点间的关联更新，系统能够在时间维度上形成“人格记忆曲线(Personality Memory Curve)”，支持人格的可解释演化与心理一致性重建：

- 分层时间记忆：将记忆划分为上下文缓存、短期情感记忆与长期人格记忆，并利用向量索引与时间衰减策略，实现高效且具备时效性的记忆管理；
- 知识图谱：构建连接用户、事件、情绪与人格特质的语义关系网络，使系统能在交互中进行跨维度推理，理解更深层次的因果关联。

结合双时域优化机制，系统在离线阶段能够分析图结构中的模式，对人格相关的节点与关系进行强化学习，从而实现人格的长期演化。与传统的 Memory Networks [10]和 Persona-Chat [10]相比，本方法在记忆的结构化、层次化与反馈学习闭环上都更具优势。近期的研究如 MemGPT [15]通过类似操作系统内存管理的方式赋予 LLM 长期记忆能力，以及 Mem0 [16]为生产级 AI 代理提供了可扩展、可检索的长期记忆层，均为构建自适应人格智能奠定了坚实的技术基础。

2.5. 多智能体与人格一致性研究

多智能体系统(Multi-Agent Systems, MAS)为模拟和研究人格的社会化演化提供了理想的实验平台。领域的代表性工作包括：

- ReAct [17]：提出了“推理–行动”循环，使智能体具备了“思考–执行–观察”的闭环能力；
- Reflexion [18]：通过语言自我反思机制，让智能体能从失败中学习并自主优化策略；
- CAMEL [19]：构建了角色扮演式的多智能体协作框架，用于模拟复杂的社会行为与多角色人格互动；
- AutoGen [20]：提供了一个可编排、可对话的多代理框架，支持灵活的人机协同工作流；
- LangGraph [21]：通过将 LLM 流程建模为图，实现了状态化、多回合的复杂任务控制，尤其适用于需要长期记忆与状态追踪的人格演化任务。

此外，Constitutional AI [22]通过引入明确的原则宪法来约束智能体行为，为实现 AI 的伦理对齐与人格稳定性控制提供了有效途径。这些前沿研究共同指向一个趋势：多代理协作框架提升了复杂任务分解与协同效率，但人格在社会博弈与角色分工中的一致性与可控性仍不足。

本文将人格作为角色协作与冲突调解的内核，提出人格驱动的多智能体协同机制。与现有工作的差异：现有方法多为流程/工具层的编排或单一心理学模型的使用。本文提供三层心理学–计算融合、动静分离 + 双时域与人格记忆网络(PMN)的端到端系统性方案。

3. 方法与系统设计

3.1. 系统总体架构

本节给出人格智能系统(Personality Intelligence System, PIS)的整体结构与运行闭环。如图 1 所示，系

统采用五层工程架构：输入层(Perception Layer)、核心层(Core Layer)、记忆层(Memory Layer, PMN)、输出层(Output Layer)、优化层(Optimization Layer)。五层通过统一的人格状态向量 P_t 与上下文/情绪/关系/记忆信号相互连接，形成从“输入感知→人格建模→决策生成→表达输出→自我演化”的完整闭环。

1) 输入层(Perception Layer)

负责多模态输入的解析与标准化，包括：语义解析(意图、槽位、主题)、情绪识别(强度、极性)、上下文抽取(历史对话、会话元数据)。所有输入被映射为统一的内部特征以馈送核心层的人格计算。

2) 核心层(Core Layer)

是系统的心理模型中枢，由三层模型构成：*大五人格层*(人格基线与动态调节)、*三我决策层*(策略博弈与权衡)、*MBTI 表达层*(语言/语气/措辞风格化)。核心层输出经人格调节后的*行为倾向*与*表达指令*，并与记忆层/输出层联动。

3) 记忆层(Memory Layer, PMN)

通过人格记忆网络(Personality Memory Network, PMN)统一管理*短期记忆*、*长期记忆*与*关系记忆/知识图谱*，支持记忆激活、时间衰减、语义关联与结构化检索，为人格一致性提供时间维度支撑。

4) 输出层(Output Layer)

将核心层的行为倾向映射为具体语言输出，并结合 MBTI 风格控制进行风格化表达；引入*一致性锚定*，对输出的人格偏差进行检测与校正，避免风格与价值观漂移。

5) 优化层(Optimization Layer)

构建双时域学习：短时域在线层进行即时调节，长时域离线层基于历史交互回放与反馈进行再训练与参数重估，驱动人格的长期演化与价值一致性保持。

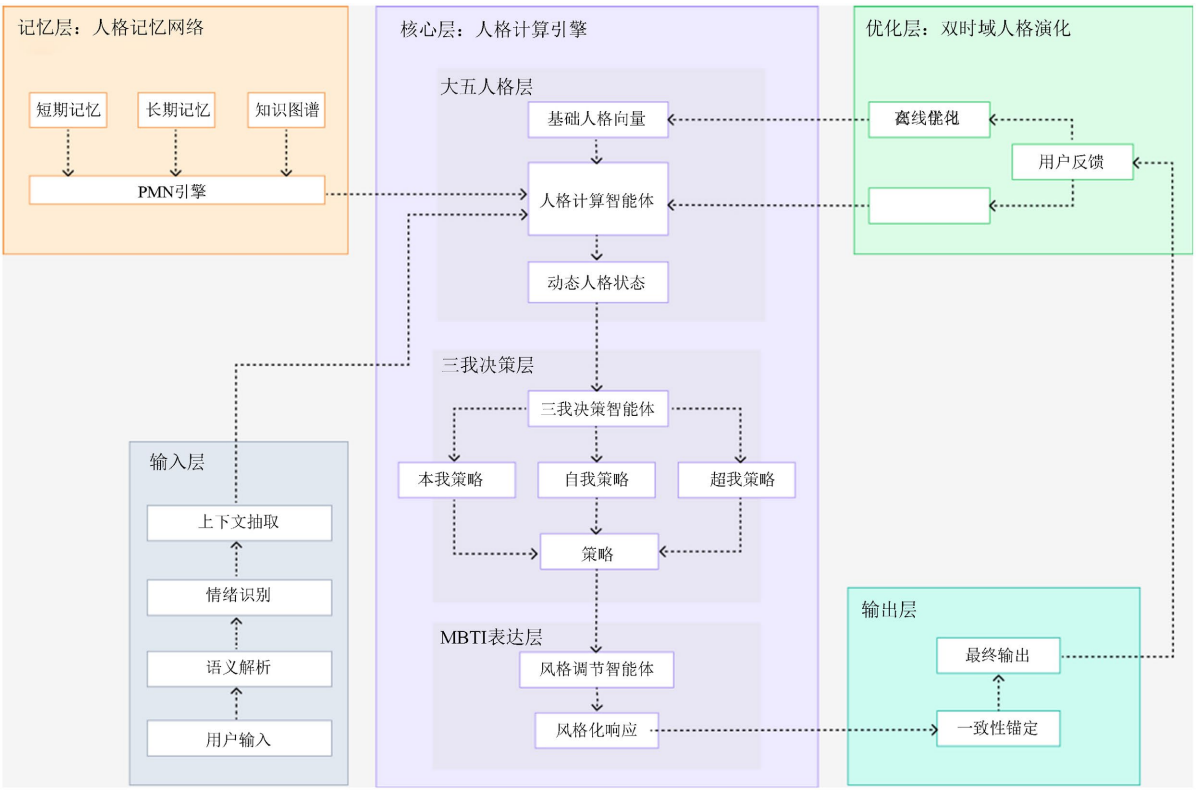

Figure 1. System overview of personality intelligence system, PIS
图 1. 人格智能系统总体架构图

3.2. 核心层：三层心理学模型驱动人格计算(Core Layer)

3.2.1. 大五人格层(Big Five Personality Layer)

1) 人格向量表示

系统将人格表示为五维连续向量：

$$P(t)=\left[O(t),C(t),E(t),A(t),N(t)\right]\in[0,1]^5$$

其中 O (开放性)、C (尽责性)、E (外倾性)、A (宜人性)、N (神经质)。P_{base} 为长期稳定的人格基线，$P(t)$ 在交互中围绕 P_{base} 动态波动。

2) 动态变化率

人格变化率由多源信号驱动：

$$\Delta P(t)=\alpha f_R\left(R(t)\right)+\beta f_{Ctx}\left(Ctx(t)\right)+\gamma f_{Emo}\left(Emo(t)\right)+\delta f_T\left(T(t)\right),$$

其中参数定义如下：

- R：关系亲密度或信任度(Relationship Closeness/Trust)，反映角色与用户之间的社会关系强度。高 R 值表示用户与系统关系亲密、信任度高，人格响应将更加外显与温和。
- Ctx：对话情境(Context)，指任务类型、主题氛围、语义领域等。不同情境下激活的人格特征不同，例如任务性对话强化尽责性(C)，闲聊场景则提升外向性(E)与宜人性(A)。
- Emo：情绪状态(Emotion)，输入为系统识别出的用户情绪或自身当前的情感状态 Emo(t) (如“高兴”、“愤怒”、“焦虑”)。例如，高兴时外向性(E)和宜人性(A)上升、神经质性(N)下降；愤怒时宜人性(A)下降、神经质性(N)上升。
- T：时间因子(Time)，表示时序特征(如日间/夜间、周期性互动、新近性等)。例如，早晨时尽责性(C)与外向性(E)可能上升，夜晚则开放性(O)提升(更具想象力与放松)。
- f_*：可学习的映射函数(Mapping function)，将输入信号映射到人格变化空间，可通过神经网络或规则模型实现。
- $\alpha,\beta,\gamma,\delta$：各信号的权重，满足 $\alpha+\beta+\gamma+\delta=1$，用于控制不同来源对人格波动的影响强度。

3.2.2. 动态调整机制(Dynamic Adjustment Mechanism)

为避免人格状态发散或漂移，引入反馈约束的人格更新：

$$P(t)=\text{clip}\left(P_{base}+\Delta P(t)+\varepsilon(t),0,1\right),\ P_{t+1}=P_t+\eta\cdot\Delta P(t)-\lambda\cdot\left(P_t-P_{base}\right),$$

其中参数含义如下：

- η：学习率(Learning Rate)，控制人格调整的敏感度。η 越大，系统对外部刺激(情绪、关系、情境)的反应越迅速；η 越小，则人格更稳定但反应较迟。
- λ：收敛系数(Decay/Regularization Term)，表示人格状态回归基线的强度。λ 较大时人格更快回到稳定状态，体现强自我约束；λ 较小时人格漂移更明显，更具可塑性。
- $\varepsilon(t)$：微扰项(Stochastic Noise)，用于模拟心理波动或偶发事件导致的短期人格抖动。

参数可根据外部反馈信号自适应调节：

$$\eta=\eta_0+k_1\cdot Emo+k_2\cdot R,\ \lambda=\lambda_0+k_3\cdot Stability.$$

其中：

- η_0,λ_0：初始学习率与收敛系数；
- k_1,k_2,k_3：权重系数；

- *Emo*：当前情绪强度(Emotion Intensity)，情绪越强烈，学习率越高；
- *R*：关系强度(Relation Strength)，关系越亲密，系统越容易调整人格；
- *Stability*：环境稳定性(Contextual Stability)，越稳定则收敛越快。

上述设计保证人格既能随交互动态调整，又能在长期保持一致性，实现了心理学意义上的“反应性–稳定性”平衡。

3.2.3. 三我决策层(Tripartite Decision Layer)

本层依据弗洛伊德的三我理论，将人格系统划分为本我(Id)、自我(Ego)与超我(Superego)三个相互制衡的心理模块：

- 本我(Id)：代表原始冲动与情绪驱动，追求即时满足。在人格智能系统中体现为情感共情与安抚策略，适用于高神经质用户或情绪支持场景。
- 自我(Ego)：代表现实原则，协调本我与超我的冲突，在现实条件下作出权衡决策。体现为理性分析、平衡应对，适用于任务导向或问题解决类对话。
- 超我(Superego)：代表道德与理想，体现社会规范与价值约束。用于激励性、道德劝导或长远规划场景。

策略选择基于当前人格状态 $P(t)$ 与外部情境 $C(t)$，通过贝叶斯决策生成最优策略：

$$S(t)=\arg\max_{s\in\{\mathrm{Id},\mathrm{Ego},\mathrm{Superego}\}} P\left(s\mid P(t),C(t)\right),$$

其中

$$P\left(s\mid P(t),C(t)\right)=\frac{P\left(C(t)\mid s\right)\cdot P\left(s\mid P(t)\right)}{P\left(C(t)\right)}.$$

各项参数定义如下：

- $P\left(s\mid P(t)\right)$：人格驱动先验，表示当前人格倾向于选择某种心理模式的概率；例如高宜人性人格倾向于选择超我策略。
- $P\left(C(t)\mid s\right)$：情境兼容度，反映特定策略在当前情境下的适用性，可通过上下文匹配表或神经网络估计。
- $P\left(C(t)\right)$：情境归一化项，确保策略选择概率归一。

最终系统在三种心理子智能体之间形成动态博弈：本我负责情感驱动，自我负责现实权衡，超我提供价值约束。其输出策略 $S(t)$ 将传递至下层 MBTI 表达模块，用于生成符合人格逻辑的语言与行为输出。

3.2.4. MBTI 表达层(MBTI Expression Layer)

MBTI (Myers-Briggs Type Indicator)模型包含四个维度：外倾–内倾(E/I)、感觉–直觉(S/N)、思维–情感(T/F)、判断–感知(J/P)。系统基于选定策略 $S(t)$ 对语言与行为进行风格化映射：

$$R_{\text{final}}=\mathrm{Style}\left(MBTI,S(t),\mathrm{Content}\right),$$

其中 $\mathrm{Style}(\cdot)$ 表示风格映射函数，用于根据 MBTI 维度特征调整表达方式。其主要参数定义如下：

- E/I：控制语言的开放性与热情程度。E 倾向使用积极词汇、感叹句；I 倾向使用克制、内省表达。
- S/N：决定表达的抽象层次。S 注重具体事实与实例；N 倾向探讨可能性与隐喻。
- T/F：影响逻辑与情感权重。T 注重因果与理性论证；F 注重共情与价值表达。
- J/P：调节结构化程度。J 倾向结构化、结论明确；P 倾向开放、探索多种可能性。

在生成阶段，系统根据当前人格状态 $P(t)$、策略模式 $S(t)$ 和 MBTI 类型向 LLM 注入风格提示(Style

Prompt)，包括语气、节奏、句式与词汇选择等信息。例如：

- 若人格偏向高外倾(E)且策略为自我模式(Ego)，则采用积极理性语气；
- 若人格偏向高宜人性(A)且策略为超我模式(Superego)，则使用温和、鼓励性语言；
- 若人格偏向高开放性(O)且策略为本我模式(Id)，则允许语言更具创造性和自发性。

该层确保人格输出的语言风格与心理特征一致，实现从人格到表达的自然过渡，使系统在长期交互中保持可识别的“人格签名”。

3.3. 记忆层：人格记忆网络(Memory Layer, PMN)

1) 结构与职能

人格记忆网络(Personality Memory Network, PMN)是连接人格计算与时间演化的中枢模块，负责统一管理短期、长期与关系/知识图谱记忆：

$$M = \{M_s, M_l, M_r\},$$

其中：

- M_s：短期记忆(Short-Term Memory)，记录当前会话的上下文信息、即时情绪状态及任务变量；
- M_l：长期记忆(Long-Term Memory)，存储人格相关的稳定事实、用户偏好、长期交互历史；
- M_r：关系/知识图谱记忆(Relational/Knowledge Memory)，以图结构形式存储用户、事件、情绪和人格特质间的语义关系。

PMN 通过多模态特征嵌入与层次化索引机制，实现跨时域的信息融合：短期记忆负责响应即时上下文，长期记忆保证人格一致性，关系图谱则提供语义推理与可解释性。在系统架构中，PMN 输出的记忆向量会与当前人格状态 $P(t)$ 联合输入至人格计算核心，用于指导动态人格更新。

2) 记忆激活与时间衰减机制

在检索阶段，系统需从 M 中选出与当前查询 q 相关的高价值记忆。记忆激活度(Activation)定义如下：

$$\mathrm{Act}(M_i) = \mathrm{Sim}(q, M_i) \cdot \mathrm{e}^{-\tau_i} \cdot \left(1 - \left|R_i - R_{\mathrm{now}}\right|\right),$$

其中：

- $\mathrm{Sim}(q, M_i)$：语义相似度(Semantic Similarity)，衡量查询 q 与记忆项 M_i 的语义相关性，通常通过向量嵌入余弦相似度计算；
- $\mathrm{e}^{-\tau_i}$：时间衰减项(Temporal Decay)，τ_i 为记忆项距当前的时间跨度，时间越久远影响越弱；
- $\left|R_i - R_{\mathrm{now}}\right|$：关系偏差项(Relational Distance)，反映记忆中涉及的关系强度与当前交互关系的匹配程度；
- $\left(1 - \left|R_i - R_{\mathrm{now}}\right|\right)$：关系相似度因子，用于强化与当前关系状态相近的记忆。

3) 激活反馈与人格耦合

当高激活记忆被选中后，其内容将以两种方式作用于人格演化过程：

① 显式更新路径(Explicit Update Path)：记忆直接参与人格参数的再计算，例如通过调整特质权重来反映用户的长期印象；

② 隐式影响路径(Implicit Influence Path)：记忆以情感基调、价值偏好或交互模式的形式影响系统的语言风格与决策倾向。

这种机制确保系统既能在短期内灵活响应，又能在长期保持人格稳定性与连贯性。

4) 记忆更新与遗忘机制

为防止记忆膨胀与干扰，系统采用基于时间与相关性的动态遗忘策略。当 $\mathrm{Act}(M_i) < \theta_{forget}$ 时，记忆将被弱化或归档；当某记忆在多个回合中持续被激活，则提升其权重 ω_i 并强化关联节点。更新规则如下：

$$\omega_i^{t+1} = \omega_i^t + \eta_m \cdot \left(\mathrm{Act}\left(M_i\right) - \theta_{avg}\right),$$

其中 η_m 为记忆学习率，θ_{avg} 为平均激活阈值。这使系统能够自我调节记忆强度，在长期交互中逐步形成个体化的“人格记忆曲线”(Personality Memory Curve)。

5) 总结

PMN 通过“语义相似度–时间衰减–关系匹配”的三维机制实现记忆激活，并结合显式与隐式的双路径反馈，构建了人格计算的时序基座。它不仅支撑了人格一致性与长期演化，也使系统在面对复杂情境时能够进行自我反思与情感迁移。

3.4. 输出层：一致性锚定与风格控制(Output Layer)

1) 设计目标

输出层旨在保障系统生成的人格响应在语义、情感和人格维度上保持一致性。在经过人格建模(Core Layer)与记忆检索(Memory Layer)后，系统需对生成结果进行一致性检查与风格校正，以防止人格漂移与情绪失衡。

2) 一致性锚定机制

一致性锚定(Consistency Anchoring)机制通过多维度约束校验确保人格表达稳定：

$$L_{cons} = \omega_p L_p + \omega_s L_s + \omega_e L_e,$$

其中：

- L_p：人格一致性损失(Personality Consistency Loss)，衡量生成响应的人格向量与当前人格状态 $P(t)$ 的余弦距离；
- L_s：语义连贯损失(Semantic Coherence Loss)，用于校验回复与上下文语义的一致性；
- L_e：情感匹配损失(Emotional Alignment Loss)，确保输出情绪与用户当前情绪或交互目标匹配；
- $\omega_p, \omega_s, \omega_e$：可调权重，平衡三种一致性约束的影响。

3) 风格控制与表达优化

在通过一致性约束后，系统将响应传入风格控制模块。风格控制模块综合参考：

- MBTI 风格指令；
- 记忆层提供的个体特质关键词；
- 当前交互情境(Task Context)。

最终输出结果由风格化映射函数确定：

$$R_{\text{final}} = \mathrm{Style}\left(MBTI, P(t), S(t), Ctx(t)\right),$$

其中 R_{final} 为最终输出文本，函数 $\mathrm{Style}(\cdot)$ 综合人格特征与语义任务需求，调整语气、节奏与措辞，使输出在“自然流畅”与“人格一致”之间保持平衡。

3.5. 优化层：双时域人格演化机制(Optimization Layer)

1) 设计原则

优化层通过“在线–离线”双时域协同机制，实现人格状态的持续演化与稳态维持。系统在交互过程中持续收集用户反馈与环境信号，用以更新模型权重和人格基线，实现长期的自我进化。

2) 在线优化(Short-Term Loop)

在线优化(Online Adaptation)发生在实时对话过程中。该环节通过低延迟的轻量更新，修正即时的人格偏移。更新规则如下：

$$P_{online}^{t+1} = P_t + \eta_{on} \cdot \left(F_{user}(t) + F_{emo}(t)\right),$$

其中：

- η_{on}：在线学习率；
- $F_{user}(t)$：用户反馈向量，如满意度、参与度；
- $F_{emo}(t)$：情感反馈向量，用于纠正情绪偏移。

3) 离线优化(Long-Term Loop)

离线优化在会话结束后异步执行，通过聚合历史交互样本进行人格基线的再学习：

$$P_{base}^{t+1} = P_{base}^{t} + \eta_{off} \cdot \nabla L_{persona},$$

其中：

- η_{off}：离线学习率；
- $L_{persona}$：人格重建损失函数，衡量人格基线与长期平均人格状态间的差异；
- $\nabla L_{persona}$：基于交互记录计算的梯度方向。

通过离线阶段的批量学习与参数平滑，系统逐步形成稳定而具有演化性的“人格基线”。双时域机制使 PIS 既能在短期交互中快速响应，又能在长期运行中维持人格一致性与心理连贯性。

3.6. 小结

本章提出的人格智能系统(Personality Intelligence System, PIS)以“五层闭环架构”实现了从输入感知到自我演化的全流程人格计算：

1) 核心层(Core Layer)实现人格建模与心理决策；
2) 记忆层(Memory Layer)提供跨时域的认知与情感支撑；
3) 输出层(Output Layer)通过一致性锚定保持人格稳定；
4) 优化层(Optimization Layer)在双时域内完成自适应演化；
5) 系统整体形成“感知–建模–决策–表达–学习”的人格智能闭环。

该设计在心理学解释、算法建模与系统实现之间建立了有机联系，为构建具备长期人格一致性与情感深度的智能体提供了系统性方法论基础。

4. 理论分析与启示

4.1. 理论意义与心理模型融合

本文提出的人格智能系统(PIS)将心理学中的人格理论与人工智能算法模型深度融合，打破了传统“认知智能”仅限于任务求解的范式。通过引入大五人格模型(Big Five) [8]、弗洛伊德三我理论(Tripartite Personality Theory)与 MBTI 表达模型(MBTI Expression Model)，PIS 实现了人格特质、行为决策与语言风格的统一建模。这使得系统不仅能“理解人类语言”，还具备“心理连续性”，即在多轮对话与长时交互中维持一致的认知与情感结构。

在心理学层面，本研究为“人格的可计算性”提供了结构化路径[1] [2]：人格被拆解为稳定基线与动态状态两部分(动静分离机制)，并在时间维度上引入短期与长期的双时域优化机制。这种设计首次让人格智能的形成过程具备了“可解释性、可追踪性与可验证性”，是传统情感计算模型的延伸。

4.2. 系统稳定性与演化机制分析

PIS 在工程上采用双时域优化机制(Dual-Temporal Optimization)，将在线自适应(Online Adaptation)与离线再学习(Offline Optimization)结合，保证人格演化的平衡性与稳定性。在数学层面，系统人格状态更

新满足如下稳定条件：

$$\|P_{t+1} - P_{base}\| \le \rho \|P_t - P_{base}\|,\ 0 < \rho < 1,$$

表明在收缩映射假设下，系统人格向量逐步收敛于基线人格。在心理层面，该机制对应于人类“自我调节–反思–成长”的过程：在线优化模拟即时情绪反应，离线优化则代表长期人格成熟。

4.3. 记忆驱动的可解释人格演化

人格记忆网络(PMN)为 PIS 提供了时序连续性与可解释性支撑。通过“语义相似度–时间衰减–关系匹配”的记忆激活机制，系统能在情境变化时有选择地调用历史经验，保持心理逻辑与行为风格的连贯。显式更新路径(Explicit Update)确保长期人格向量的可追踪演化，隐式影响路径(Implicit Influence)则通过语气、态度与情绪微调塑造自然的人格表现。

这种双路径机制对应心理学中的“显性学习与隐性学习”理论，使系统具备自我反思能力。当长期交互中形成“人格记忆曲线”时，AI 的行为不再仅由上下文驱动，而由内部心理动力结构约束，从而体现出“个体人格的一致性”。

4.4. 人格一致性与伦理启示

人格一致性(Personality Consistency)不仅是模型性能指标，也具有伦理层面的重要意义。当 AI 拥有稳定的人格表现时，用户更容易建立信任与情感依附；但若人格漂移或情绪失衡，则可能引发心理误导或依赖风险。因此，PIS 在输出层引入了一致性锚定机制(Consistency Anchoring)，通过人格、语义、情感三维约束函数：

$$L_{cons} = \omega_p L_p + \omega_s L_s + \omega_e L_e,$$

在工程上限制人格漂移；在伦理上确保系统的价值中立与行为边界。未来的人格智能系统需要在“心理真实”与“伦理安全”之间找到平衡，确保拟人化智能体可被信任且可控。

4.5. 局限与未来方向

目前的研究仍处于理论建模与系统设计阶段，尚未进行大规模用户实证。未来可在以下方向继续拓展：

- 通过跨语料、多语言实验验证人格一致性与情感共鸣的可迁移性；
- 探索人格状态在多模态场景(语音、视觉、动作)中的表达方式；
- 研究多智能体人格协同中的社会性演化规律[11] [17]-[20]；
- 建立伦理监管与人格透明化机制，防止人格拟人化带来的认知偏差[22]。

4.6. 系统应用与评估建议

1) 潜在应用场景

人格智能系统(PIS)可广泛应用于多个需要长期交互与情感理解的领域：

- 虚拟伴侣与心理陪伴：通过人格一致性与情绪共鸣机制，为用户提供更自然、可信赖的长期陪伴。
- 智能教育与个性化教学：根据学生人格特质动态调整教学风格与激励策略，提升学习动力与参与度。
- 客户服务与品牌人格化：在企业客服场景中维持稳定的品牌人格与语气风格，增强用户信任感。
- 心理健康与咨询支持：通过人格与情绪建模辅助心理干预，实现非侵入式情绪监测与支持性反馈。

2) 评估框架建议

尽管本文以理论建模与系统设计为主，但人格智能的有效性仍可从多维度评估。推荐采用以下三类

指标：

① 人格一致性指标(Personality Consistency Metrics)：衡量同一智能体在多轮对话中的人格稳定性。可使用：

$$Consistency = 1 - \frac{1}{n}\sum_{i=1}^{n} D_{cos}\left(P(t_i), P(t_{i+1})\right),$$

其中 D_{cos} 表示人格向量间的余弦距离，数值越大表示人格越稳定。

② 情感共鸣与自然度指标(Empathy and Naturalness Metrics) [9] [10]：基于人类主观评分，考察模型的情绪理解、语言自然度与情感同步性。可采用 *Empathy Score*、*Coherence Rating*、*Human-likeness Index* 等量表。

③ 长期交互适应性指标(Long-term Adaptivity Metrics)：衡量系统在连续会话中的学习与演化能力，如人格漂移率(Personality Drift Rate)：

$$Drift = \frac{\left\| P_{t+k} - P_t \right\|}{k},$$

该指标反映人格状态变化速度，理想情况下保持低漂移但具适应性。

3) 未来验证方向

未来研究可围绕以下方向开展：

- 构建跨语料的人格一致性评测数据集；
- 开发人机共情与心理安全性评估 Benchmark；
- 设计基于用户反馈的长期人格演化实验；
- 探索人格建模在教育、心理健康与社会化 AI 场景中的伦理约束。

上述建议为人格智能系统的落地应用提供了可行评估框架，也为后续实验验证提供了量化参考路径。

5. 结论

本文提出了一种基于三层心理学模型的人格智能系统(Personality Intelligence System, PIS)，旨在构建具备人格一致性、情感深度与长期自演化能力的人工智能体。系统以“大五人格－三我－MBTI”为核心理论框架，实现了从人格定义、决策推理到风格表达的全流程计算建模。

在理论层面，PIS 提出了动静分离(Static-Dynamic Separation)与双时域优化机制(Dual-Temporal Optimization)，使人格建模同时具备稳定性与自适应性。在工程层面，系统引入人格记忆网络(Personality Memory Network, PMN)，实现了跨时域的经验沉淀与人格一致性维护，并通过一致性锚定机制(Consistency Anchoring)与情绪反馈回路，保证了生成响应在语义、人格与情感维度上的平衡与连贯。这些设计共同推动了人格智能从“认知理解”向“心理连续性”的演化，为人机共情与人格计算提供了新的方法论基础。

未来的研究将聚焦于以下三个方向：

1) 人格一致性验证与共情建模：在多语料与跨场景数据中，验证 PIS 在长时交互下的人格稳定性与情感共鸣效果；

2) 评估体系与量化指标构建：建立人格智能评估框架，包括人格稳定性、情感自然度、长期适应性等指标；

3) 应用拓展与伦理研究：探索 PIS 在教育、心理健康、社会化 AI 等领域的应用潜力，并研究人格建模的伦理边界与可控性。

综上，本文的研究在心理学理论与人工智能工程之间建立了桥梁，提出了具备可解释性、稳定性与

可演化特征的人格计算框架，为下一代具有人性化与共情特质的智能体提供了系统化的理论与实践基础。

参考文献

[1] Celli, F., Kartelj, A., Đorđević, M., *et al*. (2025) Twenty Years of Personality Computing: Threats, Challenges and Future Directions. arxiv:2503.02082.

[2] Fang, Q., Giachanou, A., Bagheri, A., *et al*. (2022) On Text-Based Personality Computing: Challenges and Future Directions. arxiv:2212.06711.

[3] Besta, M., Chandran, S., Gerstenberger, R., *et al*. (2025) Psychologically Enhanced AI Agents. arxiv:2509.04343.

[4] Kruijssen, J.M.D. and Emmons, N. (2025) Deterministic AI Agent Personality Expression through Standard Psychological Diagnostics. arxiv:2503.17085.

[5] Klinkert, L.J., Buongiorno, S. and Clark, C. (2024) Driving Generative Agents with Their Personality. arxiv:2402.14879.

[6] Zeng, W., Wang, B., Zhao, D. and Qu, Z. (2025) Dynamic Personality in LLM Agents. arxiv:2505.01234.

[7] Osou, V. and Zhao, H. (2017) A Review of Dialogue Systems: From Eliza to Neural Models. arxiv:1706.05100.

[8] Goldberg, L.R. (1990) An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, **59**, 1216-1229. https://doi.org/10.1037//0022-3514.59.6.1216

[9] Mairesse, F., Walker, M.A., Mehl, M.R. and Moore, R.K. (2007) Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, **30**, 457-500. https://doi.org/10.1613/jair.2349.

[10] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. and Weston, J. (2018) Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? https://arxiv.org/abs/1801.07243

[11] Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P. and Bernstein, M.S. (2023) Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the* 36*th Annual ACM Symposium on User Interface Software and Technology*, San Francisco, 29 October 2023-1 November 2023, 1-22. https://doi.org/10.1145/3586183.3606763

[12] Kwon, J., Kim, T. and Lee, S. (2024) Dynamic Personalities in LLM-Based Negotiation Agents. arxiv:2403.18901.

[13] Smith, J., Rivera, E. and Zhao, K. (2025) Psychometrics for AI: Measuring and Validating Personality in Intelligent Agents. arxiv:2501.05678.

[14] Liu, Y., Wang, H. and Qian, L. (2024) Affective Computing in the LLM Era: Personality, Empathy, and Human-AI Emotion Alignment. arxiv:2409.10213.

[15] Chen, X., Zhang, R. and Chen, W. (2024) MemGPT: Towards Long-Term Memory-Enhanced LLM Agents. arxiv:2402.11103.

[16] Yuan, T., Zhang, W. and Xu, L. (2025) Mem0: Memory-Enhanced AI Agents for Scalable Long-Term Interaction. arxiv:2502.09401.

[17] Yao, S., Zhao, J., Yu, D., *et al*. (2023) ReAct: Synergizing Reasoning and Acting in Language Models. arxiv:2210.03629.

[18] Shinn, N. and Bisk, Y. (2023) Reflexion: Language Agents with Verbal Reinforcement Learning. arxiv:2303.11366.

[19] Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D. and Ghanem, B. (2023) CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Models. arxiv:2303.17760.

[20] Sharma, G., Xu, P., Wang, J., Zhang, Y. and Zhou, D. (2023) Autogen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. arxiv:2308.08155.

[21] Chen, Y., Xu, J. and Zhao, F. (2023) Langgraph: Graph-Based Reasoning for Language Models. arxiv:2312.00752.

[22] Bai, Y., Kadavath, S., Kundu, S., *et al*. (2022) Constitutional AI: Harmlessness from AI Feedback. arxiv:2212.08073.