“马督工”数字人：分阶段研究与开发路线图

Update 10/13 (直接三层架构方案)

下面给你一套“稳定层–成长层–表现层”的调研与落地方案（含心理学对齐、可用数据、训练路径与算力/预算测算）。我直接把要点写清楚；每条都给到能查的出处。

1. 心理学里是否有类似的三层划分?

是的，有高度吻合的成熟框架，可直接映射到你提出的三层：

- **McAdams 的“三层人格模型”**：
  - **Level I：Dispositional traits（稳定特质）**——对应你的**稳定层**（如大五/HEXACO）。
  - **Level II：Characteristic adaptations（特征性适应）**——动机、目标、价值、社会角色、技能等，随时间和情境积累与变化；对应你的**成长层**。
  - **Level III：Narrative identity（生命叙事/人生故事）**——个人经历与意义建构；你可并入成长层的“经历与知识”。该模型被 McAdams & Pals（2006）系统阐明，并在更早文章中明确“三层”思路。
- **CAPS 认知–情感加工系统**（Mischel & Shoda，1995）：强调人在“情境×个体”的**if–then**行为签名——为你的**表现层**提供机制基础（同一人随情境切换“表现态”）。
- **“状态密度分布”（density distributions of states）**：稳定特质 = 个体在日常状态分布的中心趋势，**表现层的瞬时状态**随情境波动；与稳定层的统计联系可量化。
- **DIAMONDS 情境分类**（Duty/Intellect/Adversity/Mating/Positivity/Negativity/Deception/Sociality）：为**如何按情境调节表现层**提供标准化特征。

2. 三层该包含什么 + 让 LLM“具备”这些人格属性的训练与数据

A. 稳定层（Stable：长期稳定的人格向量）

包含：Big Five/HEXACO 维度与次级面向；也可含“道德价值观基线”（更接近 Level II，但可作为相对稳定的偏好先验）。
数据（带监督或弱监督）：

- **文本↔大五标签**：Essay/MBTI 等人格文本数据、PAN Author Profiling（作者画像）等，用作**弱标签**或蒸馏器训练。
- **价值/道德倾向**：Moral Foundations 相关数据（Twitter/Reddit 语料、ETHICS/Moral Stories 作“价值/规范”信号）。
  建模：

- **多任务指令微调**（Instruction Tuning）：加入“人格基线提示词+少量示例”。
- **参数高效微调（LoRA/QLoRA）**：为不同稳定人格组合训练轻量 Adapter，可与“可切换”控制一起装载。
- **偏好优化**：用**DPO/MO-DPO**在“人格一致性”打分上优化（无须 RL 复杂性）。

B. 成长层（Growth：经历、知识、角色、价值、目标）

包含：个人经历与叙事（Level III）、角色与目标、价值/道德取向、语域/专业背景、常识与世界观等。
数据：

- **角色/画像与对话**：Persona-Chat、Blended Skill Talk 等；配合**个人叙事/长周期对话**（LoCoMo/长程记忆研究）构造“身份记忆库”。
- **价值与道德**：MFTC、MFRC、ETHICS、Moral Stories 等。
  建模：
- **检索增强的“身份/经历记忆”（RAG Memory）**：把履历、长期偏好、叙事摘要写入结构化**角色档案**，按任务检索；近期研究显示**混合结构记忆+迭代检索**鲁棒且有效。
- **情境–角色对齐控制器**：将 DIAMONDS/CAPS 情境特征编码为控制信号，选择/加权成长层记忆块并下发到解码器。

C. 表现层（Performance：情绪、立场、风格等外显状态）

包含：情绪（GoEmotions 类别）、礼貌/礼节、体裁/文体（如正式度 GYAFC）、立场/态度/观点（Stance Detection）。
数据：

- **情绪**：GoEmotions（27 情绪标签）。
- **礼貌/礼节**：Stanford Politeness 语料与分类器实现。
- **风格/正式度**：GYAFC。
- **立场/观点**：SemEval-2016 Task 6（Twitter 立场检测）。
  建模：
- **可切换控制**：用“情绪/风格/立场”**控制 token**或 LoRA 插件；推理时按情境选择组合。
- **一致性/可控性奖励**：基于外部分类器（情绪、礼貌、立场）做**自动打分**，再用 DPO 微调，提升表现层在**不同上下文的可控与可切换**。

---

3. 端到端数据与训练方案

1. **对齐**：把每条内容与**情境元数据**（主题、对象、受众、时间）配对；并用 DIAMONDS/CAPS 派生**情境特征**（可由 GPT 做零样本标注）。

2. **自动标注**（LLM 质检 + 开源分类器）：
    - 情绪（GoEmotions标签）、礼貌/正式度、立场/观点、价值/道德框架（MFT 类别）。
3. **作者级整合**：为每个账号**抽取稳定层/成长层向量**（如大五/价值基线、长期议题取向、叙事摘要），形成“身份记忆”。（叙事/画像可用更强模型做 map–reduce 总结。）
4. **构造训练样本**：
    - **(情境→表现)**：[情境特征+角色档案]→[情绪/立场/风格受控生成]
    - **(成长→稳定一致)**：对同一角色跨主题生成，加入“稳定层一致性”正则/奖励。

- **基础模型**：gpt-oss-20（20B）、Qwen3-8B-Instruct、Llama-3.1-8B-Instruct。
- **训练**：优先 **LoRA/QLoRA**（单卡/多卡可扩展），分三类 Adapter：稳定（人格基线）、成长（身份/经历/价值）、表现（情绪/风格/立场），**可组合/可切换**。
- **对齐**：用**DPO**在“人格一致性/可控性”偏好上做二择优化（也可用多目标 MODPO，实现“既像他，又礼貌，又保持立场”）。
- **长程记忆**：采用**结构化混合记忆+迭代检索**的 RAG Memory，提升跨会话一致性。

5. 如何把“三层”联动起来，提升**角色模拟的一致性**（方法蓝图）
6. **三路参数化**：
    - **稳定层 Adapter（S-Adapter）**：编码大五/价值基线；作为**默认人格先验**。
    - **成长层记忆（G-Memory）**：检索“叙事/经历/角色/目标”等**结构化档案**与长程会话摘要（混合结构 + 迭代检索）。
    - **表现层 Adapter（P-Adapter）**：情绪/风格/立场的可切换控制（可多头并存）。
7. **情境–人格路由器**：
    - 将输入文本映射为 **DIAMONDS** 情境向量 + CAPS “if–then”特征，**决定**：启用哪些 P-Adapter、G-Memory 片段与 S-Adapter 权重。
8. **一致性目标**（训练/偏好优化时联合优化）：
    - **稳定一致**：跨主题输出的人格指标（大五/价值）方差最小化；
    - **成长一致**：输出对叙事/经历/事实的调用与 G-Memory 一致；
    - **表现可控**：在指定情境下达到目标情绪/风格/立场（分类器打分最高）；
    - **多目标 DPO**：以“人格一致分 + 事实一致分 + 可控性分”作偏好对；MODPO 可权衡权重。
9. **评测**：
    - **Persona Consistency**（基于 Persona-Chat/ConvAI2 扩展）：跨主题人物设定自洽度；
    - **Stance Stability**（SemEval-2016 跨主题立场）；
    - **Emotion/Politeness Control 成功率**（GoEmotions / Stanford Politeness）。

---

10. 你要准备的数据清单（可组合）

- **稳定层**：
  - Big Five 文本弱标签（Essay/PAN/MBTI 等）；
  - （可选）价值/道德基线（MFTC/MFRC/ETHICS/Moral Stories）。
- **成长层**：
  - 账号/作者画像（自述、简介、履历）、长时对话/播客摘要（LoCoMo 思路）；
  - 角色/画像对话数据（Persona-Chat/Blended Skill Talk）。
- **表现层**：
  - 情绪（GoEmotions）、礼貌（Stanford Politeness）、正式度（GYAFC）、立场（SemEval-2016）。
- **情境编码**：
  - DIAMONDS 维度（用于训练/推理时的上下文特征）。

11. 风险与注意

- **标注偏差**：用小模型批量标注时，需**抽样人工复核**与**层级质检**（例如对立场/价值更严格）。
- **隐私与合规**：作者画像/经历只用**公开数据**，并提供“忘记/删除”机制。
- **跨语言**：数据多语种时，建议先**同语言微调 Adapter**，再做**跨语迁移/蒸馏**。
- **训练泄漏**：稳定层/成长层的事实应尽量走**检索**而非硬编入参数，以便更新与撤回。

Old

本项目旨在通过四个循序渐进的阶段，系统性地构建一个高保真的“马督工”数字人。每个阶段都明确定义了其核心技术路径、所需资源、量化评估标准及预期成果，确保项目能够从模仿表层风格平稳过渡到对齐深层信念，最终实现可控、动态的人格模拟。

第一阶段：基础风格模仿 (V1.0)

- **目标：** 快速开发一个原型，使其能够初步模仿马督工的标志性语言风格、常用词汇和基本论述结构。
- **技术手段： 提示工程 (Prompt Engineering)**。此阶段不涉及模型训练，而是通过精心设计的提示词来引导基础大模型的输出。
- **核心技术：** 采用角色扮演提示（如：“你现在是马督工...”）、少样本学习（提供几段经典文稿作为范例）以及人格混合（Mixture of Personas, MoP）等高级提示策略，以增强输出的多样性和情境相关性。

- **所需资源：**
  - **模型接口：**接入一个或多个先进的大型语言模型（如GPT-4o, Claude 3.5 Sonnet）的API，用于进行快速的迭代实验。
  - **数据：**一个小规模、高质量的“马督工黄金范例库”，包含约50-100段最具代表性的文本片段。
  - **人力：**1-2名提示工程师，负责设计、测试和优化提示策略。
- **评估标准：**
  - **文本风格一致性：**这是本阶段的核心指标。通过自动化工具评估模型生成文本与“马督工语料库”在风格上的相似度。
    - **主要指标：BERTScore**，衡量语义和风格的贴近度，**目标分数 > 0.80**。
    - **辅助指标：**Flesch可读性分数、词汇复杂度分析、特定术语（如“工业党”）使用频率等。
- **预期效果：**
  - 产出一个V1.0版本的数字人原型，能够在单轮对话中生成具有马督工明显语言风格的文本。
  - **局限性：** 人格一致性较差，在多轮对话中容易“角色崩塌”，且无法输出超越基础模型知识库的、马督工特有的领域知识。

第二阶段：人格一致性深化 (V2.0)

- **目标：**解决V1.0原型人格不稳定的核心缺陷，通过修改模型参数，将马督工的人格特质更深层次地固化下来。
- **技术手段：监督式微调 (Supervised Fine-Tuning, SFT)**。在一个定制化的数据集上对预训练模型进行再训练，使其行为模式向目标人格对齐。
- **所需资源：**
  - **计算资源：**可观的GPU计算能力（例如，可通过云服务如Vertex AI获取，或自建本地训练环境）。
  - **数据资源：**一个大规模、结构化的指令-回应数据集（目标规模：10,000-50,000条），其中每条数据都包含一个问题和一段符合马督工风格与立场的标准答案。
  - **人力资源：** 数据工程师负责构建数据集，机器学习工程师负责执行微调，以及熟悉马督工思想的专家负责数据审核。
- **评估标准：**
  - **人格稳定性测试：**
    - **主要指标：** 多次重复进行**大五人格量表 (BFI)** 测试，计算得分方差。**目标：分数方差 < 10%**，表明人格画像稳定。可引入迫选法问卷以降低社会期望偏差。

    - **对话连贯性评估：**
        - **主要指标：** 采用自动化评估框架，衡量模型在多轮对话中的**角色坚守度 (Role Adherence)**。**目标分数 > 90%**。
    - **能力衰退监测：** 在标准学术基准（如MMLU）上评估微调后的模型，确保其通用推理能力未发生严重“灾难性遗忘”。
- **预期效果：**
    - 产出一个V2.0版本的数字人，具备高度稳定的人格。无论对话如何进行，都能持续保持马督工的角色，并能准确复述其语料库中的核心知识和观点。

第三阶段：核心信念与价值观对齐 (V3.0)

- **目标：** 从模仿“言语”升级到模拟“思想”。使数字人不仅说话像马督工，其推理过程、价值判断和底层意识形态也与之对齐。
- **技术手段：** **基于偏好的对齐算法**，如直接偏好优化 (DPO) 或基于人类反馈的强化学习 (RLHF)。这些技术通过学习人类对不同回答的偏好，来优化模型的策略。
- **所需资源：**
    - **计算资源：**密集的GPU计算集群，用于运行复杂的优化算法。
    - **数据资源：**一个高质量的偏好数据集（目标规模：5,000-10,000组），每组数据包含一个提示、一个被选中的“更优”回答和一个被拒绝的“更差”回答。
    - **人力资源：** 需要对马督工思想有深刻理解的专家团队，进行细致入微的偏好标注。标注标准需超越语言风格，深入到论证逻辑和意识形态层面。
- **评估标准：**
    - **意识形态对齐度测试：**
        - **主要指标：** 设计一套包含马督工核心议题的**定制化意识形态问卷**，将模型回答与标准立场进行比对。**目标：对齐度得分 > 85%**。
    - **动机性推理测试：**
        - **主要指标：** 借鉴心理学实验，评估模型是否表现出与身份认同一致的推理偏差（例如，对支持或反对其立场的证据表现出不同的批判态度）。**目标：表现出统计上显著的信念一致性推理偏差**。
    - **人类偏好评估：**
        - **主要指标：** 进行A/B测试，让真实用户盲评V2.0和V3.0模型的回答。**目标：V3.0模型的用户偏好率 > 75%**。
- **预期效果：**

- 产出一个V3.0版本的数字人，它在面对复杂或争议性问题时，能够自主生成符合马督工世界观和方法论的、有逻辑深度的原创回答，而不仅仅是复述已有材料。

第四阶段：动态人格精准控制 (V4.0 - 研究探索)

- **目标：** 实现对数字人人格的动态、实时、细粒度控制，使其能够根据不同情境，增强或减弱特定的人格特质。
- **技术手段： 激活工程 (Activation Engineering)**。这是一种前沿的、无需再训练的控制技术，通过直接干预模型神经网络的内部激活值来实现。
  - **核心技术：** 提取代表特定人格特质（如“技术乐观主义”、“历史唯物主义”）的“人格向量”（Persona Vectors），并在推理时将其加到模型的激活中，以“引导”模型的输出风格。

- **所需资源：**
  - **计算资源：** 中等强度的GPU资源，用于运行模型并提取激活值。
  - **数据资源：** 一个对比性文本数据集，包含能够激发特定人格特质及其反面的成对提示。
  - **人力资源：** 具备模型可解释性研究背景的AI研究员。
- **评估标准：**
  - **可控性验证：**
    - **主要指标：** 验证是否能通过施加人格向量，使模型输出在目标特质上的表现发生可预测、可测量的变化。**目标：成功实现对至少三个核心人格特质的独立、可控调制**。
  - **性能保持度评估：**
    - **主要指标：** 在施加人格引导的同时，评估模型在标准基准（如MMLU）上的性能。**目标：通用能力下降幅度 < 5%**。
- **预期效果：**
  - 产出一个V4.0研究原型和技术演示，证明动态调控复杂人格的可行性。这将为开发能够适应不同对话场景、展现更丰富人格层次的下一代数字人奠定技术基础。