袁　梦

AI 模型评测 / AI 训练师 · 多模态方向

24 岁 · 现居北京 | 13910616825 | yume825@yeah.net

个人主页

核心优势

多模态视频评测：独立承接算法需求，从零设计评测体系与标注规范，沉淀 Bad Case 库**归因回流推动模型迭代**；熟练用 Excel 与数据看板定位异常、产出结构化报告。

影视视听语言 × AI 视频生成：懂景别、运镜、构图、光影与镜头调度，把**主观审美拆成可量化、可复现的评分维度**；熟用 Midjourney、SD、即梦、可灵，掌握 ComfyUI 与 LoRA，创作者视角反哺评测。

大模型训练数据全链路：熟悉 SFT、奖励模型、RLHF、DPO 等后训练流程；有 **SFT 对话生产、RLHF 偏好排序、RAG 相关性标注**实操经验，重度对比使用 ChatGPT、Claude、豆包、GLM。

AI 工程与 Prompt 工程：深度使用 Dify、Cursor、Claude Code 搭自动化链路；用评分锚点注入、Few-shot 与思维链构建评测提示词，搭「**大模型初评＋人工复核存疑**」人机协同流水线。

工作经历

热热科技　AI 训练师助理（多模态方向）

2025.04 — 至今

多模态视频生成场景的评测与数据标注质检：从真实业务请求构建评测集与指标体系，产出多模型横评报告支撑选型；把镜头语言、角色一致性、表演与视频描述的判断标准沉淀成训练数据规范。

网易　AI 训练师（实习）

2024.09 — 2025.04

游戏 NPC 角色扮演对话与游戏内搜索问答的训练数据建设：SFT 对话编写改写、RLHF 偏好排序、RAG 检索相关性标注；推动规则制定与迭代、日常质检和 badcase 归因回流。

项目实践

多模态短剧生成智能体评测　评测组长助理

2026.05 — 2026.07

系统级评测：判断甲方自研智能体能否从一句话创意端到端生产完整短剧——覆盖故事扩写、分镜、视频、转场配乐到成片全链路。

- **评测集构建：**50 条单句创意按「题材 × 子类 × 代表创意」三层组织，覆盖 **7 大题材 17 个细分子类**，再用线上高频请求、失败样本和平台爆款补充，兼顾真实分布与能力边界。
- **评测体系 0→1：**通用视频质量＋短剧专项**五分制打分卡**（开场与结尾钩子、冲突升级、剧情连贯、角色与环境连续性、配乐转场），每条附错误归因标签和证据说明。
- **质检与归因：**样本匿名随机展示，重要样本双评＋分歧交第三人仲裁；失败按环节归因（故事、提示词、首帧、视频、转场、配乐、组装），输出算法侧可直接使用的优化方向。
- **横评交付：**多系统同题横评（自研全量、对照小云雀等抽样），累计评估**约 450 分钟成片及中间产物**，拆解为上千个评测单元（数百个片段、相邻片段对及中间产物）。

> ★ **成果：**沉淀短剧智能体评测方法论与五段式报告模板，输出**三档准入结论**（可无人审试运行、可人机协作生产、暂不具备完整生产能力），为上线、选型与算法迭代提供量化依据。

短剧多镜头视频生成模型横评　评测执行与报告

2026.02 — 2026.04

底座选型：智能体立项前的多模型横评——判断哪个基座模型撑得住「跨镜头不串脸、环境稳定、运镜可控」的多镜头生产。

- **分层评测集：**20 个短剧故事三层脚本（角色设定与参考图、场景、镜头级提示词），共 **100 个镜头提示词、50 张参考图**，覆盖重生复仇、古风权谋、悬疑探案等题材。
- **四维打分卡＋红线：**视觉质量、脚本对齐、跨镜头一致性、运动质量各五分制；角色串脸、结构崩坏、明显穿模直接判不可用；每条样本归入满意、可用、不可用三档。
- **横评执行：**Seedance、可灵、Veo、grok、HappyHorse 同题生成 **500 条镜头片段**，另评 **400 组相邻镜头对**；分歧样本双评＋分差达 2 分仲裁；模型裁判按同一打分卡交叉校验。

> ★ **成果：**交付五模型横评报告，覆盖 **5 模型 × 100 提示词共 500 条镜头片段＋400 组相邻镜头对**跨镜头一致性专项评估。

项目实践

文生视频模型训练标注 标注组长助理

2025.04 — 2025.10

训练数据 0→1：为视频生成模型生产高质量训练数据，提升镜头语言、角色一致性与表演表现；高峰期引入众包团队，先把规则立起来。

- **规则 0→1 迭代 3 版：**试标验证 → 小批对齐 → 落地规则文档；设计**镜头级六维标注**（色调、角度、构图、景别、光影、镜头移动）＋必填与单多选约束的结构化打标框架。
- **主观审美锚定：**把「画面惊不惊艳」拆成清晰度、光影质感、构图美感等可独立打分的子维度，每档配公认样例作锚点（5 分惊艳、3 分达标、1 分崩坏），关键样本双评＋分差 ≥2 仲裁，**评分一致性 85%→96%**。
- **Caption 规范：**只做客观陈述、按「主体→环境→光影→镜头运动→风格质感」固定顺序、方位颜色具体化、禁止画面外信息；视频须额外写动作先后变化与运镜变化。
- **众包质控：**试标筛人、培训与争议样例库；**四道质检关卡**（自检 → 组长复核 → 专职抽检 20–30% → 负责人终审）；负责每日数据分发与抽检调度，badcase 案例库归因回流规则。

> ★ **成果：**全程交付约 **5 万条**标注数据，各批验收**准确率 96% 以上**；众包团队培训后一致率达标；收尾后延续带一期图生视频描述标注专项（2025.11 — 2026.01）。

游戏 NPC 对话与搜索问答数据建设 · 网易 AI 训练师（实习）

2024.09 — 2025.04

三类训练数据生产：NPC 回答要同时满足人设、语言风格、好感度、任务进度与世界观约束；搜索问答用 RAG 检索增强保证事实性。

- **SFT 对话编写与改写：**依据人设卡与游戏状态编写标准回答；按角色一致性、世界观、多轮记忆、状态一致性、语言风格**五维逐条判定**，清洗为「角色设定＋游戏状态＋问题＋标准回答」统一格式并控制场景配比。
- **RLHF 偏好排序＋RAG 标注：**安全性为底线、依次比较角色一致、事实、风格、流畅并写明理由，供奖励模型训练；相关性分档＋**模糊样本构建**＋答案改写降幻觉。
- **规则迭代与质检归因：**试标校准、争议维度补正反锚点样例；错误按 OOC 出戏、世界观错误、记忆丢失、知识越界等类型归因，回流补充 SFT 训练数据。

> ★ **成果：**团队累计交付 **4.6 万+ 条**（个人生产与质检约 1 万条），各批验收**准确率 95% 以上**；标注规范迭代 **3 版**，支撑角色扮演模型人设一致性与世界观关联性迭代。

自动化工作流构建 项目助理 · 与各项目并行

2025.08 — 至今

自发提效：标注质检中发现高频重复环节与口径漂移，自发用 AI 编程（vibe coding）搭提效工具，与日常作业并行。

- **机标初评链路：**搭建「关键帧抽取 → 大模型按评分表初评 → 人工精修终审」链路，每个评测维度配专用提示词（角色设定、评分锚点注入、Few-shot、思维链约束）；定期校验机评与人工一致率，下滑即回调提示词。
- **人效质量管理插件：**完成量、有效量、质量分、作业时长拖入即自动规则计算，识别产能偏低、质量异常、疑似刷量等风险，一键导出；末端接大模型自动生成日报、周报与复盘分析。

> ★ **成果：人均产能从 60 条/天提升到 70 条/天（约 17%）**，单批交付周期明显缩短；插件与 Dify 工作流模板沉淀为跨项目复用的提效工具箱。

教育背景

曼彻斯特大学（University of Manchester） · 创意文化产业 硕士文凭（PGDip） 2023.09 — 2024.11

曼彻斯特大学（University of Manchester） · 教育学 学士 2020.09 — 2023.06

工具与技能

评测：打分卡设计 · 双评仲裁 · 模型裁判交叉校验 | **质检：四道关卡 · 锚点校准 · Badcase 归因闭环**

视频生成：即梦 · 可灵 · Seedance · Veo | AIGC 创作：Midjourney · Stable Diffusion · ComfyUI · LoRA

提效：Dify · Cursor · Claude Code · Python（数据处理） | 后训练认知：SFT · RM · RLHF · DPO · RAG

数据分析：Excel · 数据看板 · 飞书多维表格 | 语言：雅思 6.5 · 英语工作语言