个人简历 | PERSONAL RESUME

个人信息

姓名：孙晨曦　　**年龄：**25

电话：19350524359　　**入职时间：**离职-随时到岗

邮箱：sunchenxi@wuqiai.cn　　**个人网站：**www.wuqiai.cn

个人优势

3年头部大厂（阿里巴巴 + 字节跳动 + 智谱）AI 产品经理经验，覆盖 C 端 AI 应用与底层大模型两端,我关注 AI 技术与业务之间的连接，**擅长用刚好够用的技术方案、以更低的复杂度和成本推动 AI 产品高效落地。**在淘天 AI 购物助手项目中,我以较低技术成本落地「商品对比推荐」模块，**对比后下单转化率 +5.7%。**我习惯从「值不值得」出发做技术取舍,并通过 badcase 归因与离线评测建立数据闭环,让每次迭代都低成本、有指向。

工作经历1

2025.07-2026.06　　**阿里巴巴 · 淘天集团**　　**AI产品经理**

淘天 AI 购物助手_【购物车商品AI对比推荐】

项目概述：

千问与淘天购物场景打通后，淘宝上线 **AI 购物助手**能力。**购物车场景下的「商品对比推荐」模块**覆盖全品类，不同品类对比逻辑差异较大（3C 比参数/性能/续航/适配，服饰比材质/版型等），由不同品类 PM 分别负责。我**主导 3C 品类**对比从需求、方案、AI 能力落地到上线的产品工作。功能上，用户在**购物车点击「对比」**，从浏览、购物车、收藏夹中选择多个商品，由 **AI 自动生成结构化对比与购买建议**，帮助用户在相似商品间**快速决策、降低比价成本**。

核心职责：

3C品类需求分析与产品方案设计：

- 拆解 3C 购物车多商品决策链路，明确「信息散落、结论缺失」的核心痛点，**完成竞品与场景分析，输出流程图、交互原型与 PRD**
- **设计「对比入口 → 多来源选品 → AI 自动对比 → 购买决策」的端到端流程**；针对 3C 各子品类（手机壳/耳机/充电器/数码配件等）梳理对比维度（芯片/性能/续航/适配/参数）与专属边界（型号命名混乱、适配关系复杂、参数专业度高）
- 系统定义商品失效/下架、跨店跨类、规格不一致等边界规则，保障 3C 复杂商品状态下的体验稳定

AI应用工作流（Workflow）设计与产品化落地：

- 设计 **3C 对比完整 workflow：**规则层（到手价计算、字段归一化）→ 结构化入参构造 → Prompt 约束 → 模型生成 → 结构化输出解析 → 兜底（guardrail），六环节全链路职责定义
- **入参构建策略：**3C 商品字段（芯片/续航/适配/参数）取舍与组织方式，**控制输入质量**；到手价经规则校验后入参，**字段命名不统一**（如「续航时间」/「电池使用时长」）用规则层映射表归一化，规避比价错误与信息缺失风险
- **定义 AI 输出结构与约束：**（总述→相同点→核心区别→选购建议→小提示→参数对比表→追问入口的七段式结构，措辞克制、可解析），将模型能力塑形为可用、可渲染、结论直达行动的 3C 对比结果

项目迭代数据闭环与评测：

- 制定商品对比任务的**数据标注规范（labeling guidelines）**，统一标注口径，为模型迭代提供稳定数据信号
- **建立 badcase 归因体系**，按信息抽取 / 结论判断 / 风险措辞三类错误归因并结构化反馈算法侧，驱动定向优化
- 设计覆盖 3C 跨子品类、多规格、复杂适配等长尾边界场景**离线评测集与评测维度(evaluation set / rubric)**，以评测驱动版本迭代

工作成果：

- 功能已在安卓端上线，iOS 端处于版本/灰度覆盖阶段
- 建立 3C 板块对齐主链路的漏斗指标与数据口径（入口点击率、AI 对比生成成功率、推荐采纳率、对比后下单转化率等）
- 通过 **A/B 实验（实验组 vs 对照组）度量功能增量**：**入口点击率（模块总计）提升 4.5%**；AI 对比生成成功率达 98%；**对比后下单转化率提升 5.7%**；用户决策时长下降 25%
- 沉淀线上 badcase 收集 → 归因 → 评测 → 反馈闭环，支撑模型效果持续迭代

工作经历2

2024.06-2025.05　　**北京字节跳动科技有限公司**　　**AI产品实习生**

Seedream_【Seedream 系列文生图模型数据构建与效果迭代】

项目概述：

于 **Seedream 文生图方向**担任产品实习生，参与 Seedream 系列文生图（Text-to-Image）**模型的数据构建与效果迭代。**聚焦提升模型的**图文对齐（image-text alignment）能力与生成美感**，深度参与高质量图文对（image-caption pair）**语料构建、生成效果评测与竞品 benchmark**，支撑 Seedream 3.0（2025 年发布）在中文语义理解、指令遵循（prompt following）与美学表现上的迭代。

核心职责：

高质量训练语料构建与数据标准设计：

- 参与**摄影、设计、艺术、动画 4 大品类**图文对语料的**美学标注体系设计**，定义**数据清洗（data filtering）规则**与**「结构化描述 + 整体描述」的分层 caption 框架**，从光影、构图、色彩体系、艺术风格、绘画技法等**多维度提升 caption 信息密度与准确性**，强化模型图文对齐
- 基于 badcase 归因分析持续迭代标注 SOP，推动标注流程标准化，**提升语料质量与标注一致性**

生成效果评测与体验归因：

- 搭建多维评测标准，从美感、指令遵循度、中文字符渲染、风格保真度等维度对模型输出进行系统性评估，兼顾主观评测与 badcase 分析
- **对生成缺陷进行归因**，将高频问题结构化反馈至算法团队，**推动「评测—反馈—迭代—验证」闭环**

竞品 Benchmark 与跨团队协作：

- 对**国内外主流文生图模型**开展**横向 benchmark**，从生成质量、prompt 遵循、中文能力等维度输出差距分析，为迭代方向与差异化定位提供输入
- 结合数据看板监测核心指标，衔接产品、算法、标注团队，在需求评审与效果对齐中推进协作落地

工作成果：

- 参与构建 4 大品类美学标注体系，累计迭代标注规范 8 版、处理图文对数据 10 万+，标注一致性提升至 90%+
- **输出效果评测/竞品 benchmark 报告10+份**，沉淀结构化 badcase 与优化建议，部分被采纳应用于模型迭代
- 相关工作支撑 Seedream 3.0 在图文对齐、中文渲染与美学表现上的提升

工作经历3

2023.05-2024.04　　北京智谱华章科技股份有限公司　　AI产品实习生

智谱 AI — 【AgentTuning：GLM 系列 Agent 能力训练数据构建与评测环境搭建】

项目概述：

AgentTuning 是智谱 GLM 系列大模型在 Agent 方向的底层数据预研项目。当时开源基座模型在工具调用、多步 Agent 任务上明显弱于闭源模型，项目通过**构建高质量 Agent 交互数据**、**搭建统一评测体系来提升模型的 Agent 能力**，同时保证其在训练时未见任务上的泛化能力与通用语言能力不退化。我作为 AI 数据实习生，负责 **AgentInstruct 训练数据**的采样、清洗与质量过滤流水线，以及 **held-out 评测环境的适配与搭建**，相关工作支撑了 ChatGLM3、GLM-4 将工具调用与 Agent 任务逐步做成模型的原生能力。

核心职责：

AgentInstruct 数据构建与质量过滤：

- 在 Agent 环境中**批量调用 GPT-4 采样 ReAct 格式**交互轨迹，每步动作**附带思维链（Chain-of-Thought）**
- 实现轨迹奖励过滤流水线，按任务奖励设阈值剔除推理错误与无效动作样本，最终筛出 1,866 条高质量交互数据
- 执行数据泄露检查，比对训练轨迹与 **held-out 评测集**，确保无分布重叠

Held-out 评测环境适配：

- **独立完成 SciWorld、ReWOO 两个开源框架接入**，参与其余 4 个环境（MiniWob++、HotpotQA、WebArena、Digital Card Game）的适配
- 编写适配层抹平各框架动作空间与观测格式差异，使同一模型可用统一协议跑通
- **基于 TGI（Text Generation Inference）搭建评测推理服务**，支持多实例并行跑分

微调实验与错误归因：

- 参与 Llama2-chat 7B/13B/70B 三档基座的**混合微调(AgentInstruct + ShareGPT)**，负责多组混合比例的消融跑测与结果整理
- **对 ALFWorld、WebShop 做错误分析**，规则化归类 invalid action、repeated generation 等错误类型

工作成果：

- **交付 AgentInstruct 数据的采样与过滤流水线**，产出全量 1,866 条训练数据，作为项目唯一训练数据源支撑全部微调实验；其中数据泄露检查确保训练集与 held-out 评测集无重叠，是项目「泛化性」结论得以成立的前提
- **交付 SciWorld、ReWOO 两个环境的适配层及 TGI 评测推理服务，使 6 个异构 held-out 框架**可在统一协议下跑通并支持多实例并行，将全套评测执行周期从人工逐个调试压缩为一次性批量跑完
- **产出错误分析的量化结论**：定位到原始 Llama2 的失败集中在 invalid action、repeated generation 等低级错误，而非任务理解本身，从数据层面说明开源基座模型的 Agent 短板更多来自训练数据与评测缺失，为后续数据构建方向提供直接依据