VLM demo-项目复盘

1. 项目目标

1.1 核心任务

A. 结构化抽取

把多平台电商订单详情页截图（T/JD/XHS）抽取为统一结构的 5 个字段：

- order_outcome（SUCCESS/CLOSED）
- paid_amount（实付金额）
- original_amount（原始金额/商品总价/应付金额）
- discount_amount（优惠金额/立减/共减）
- order_time（下单/创建时间）

并对每个字段输出：

- value
- state（枚举：OK / MISSING / UNCLEAR）
- evidence（短原文证据，包含数值 + 语义锚点）

B. 一致性校验

当 paid/original/discount 都为 OK 时，校验：

> paid_amount = original_amount - discount_amount
>
> 输出 YES/NO；否则输出 SKIP 并说明跳过原因。

1.2 业务意义

把不可控的读图输出变成可控的数据资产，并对关键混淆数据进行全量校验。

- 抽取：解决多平台截图 → 归一化结构字段的通用能力（可用于报销/账务/个人财务助手/客服核对/风控核验等）

- 校验：提供自动 QA 信号，用于视觉定位错误来源（字段识别错、金额归一化错、模板漂移、可见性问题），指向下一轮数据策略/规则策略。

2. Demo 版本与迭代

2.1 V0产出：任务规格和评测集设计

1. 确定统一输出 json schema：5 个标签类型，每个标注：value + state + evidence 三大属性；1个校验判定，包含判定结果和判断evidence
2. 小规模样本管理与分桶：domain/page_state/info_completeness/missing_fields
3. 先用豆包1.8跑少量样本验证字段可抽取、口径可落地，进行初步评测
4. 评测中，对标签value和state双字段进行匹配，排除评测集数据中缺失带来的噪声；evidence字段不做评测统计，供追溯模型判定逻辑，进行错例归因迭代

2.2 V1产出：Goldset 1.0 + 标注指南 + 初版评测脚本 + Report 1.0评测报告

- 建立 Goldset（小规模人工真值）作为评测基线
- 编写标注指南：字段定义、判定逻辑、证据锚点规则、边缘案例等
- 跑模型输出 JSON并回填
- Python 脚本自动化批量评测，统计整体/分桶/错例

2.3 V2产出：Goldset 2.0（QC 后）+ Report 2.0评测报告

核心迭代点：对 Goldset 做 QC

- 原因：页面多处金额缺乏明确字段名，但视觉锚点稳定（商品旁灰字价格），字段数值重要、人眼视觉因果可稳定推断为原始总价，业务价值高；模型具备识别能力，有处理可能性，但需要进一步约束输出稳定性
- 风险控制：推断对错会在 reasoning 的一致性校验中暴露（YES/NO），因此该策略可通过 reasoning 做二次约束

3. 总体评测结果

- n = 12
- schema_ok_rate = 1.00，json输出合法性比例
- all_5_fields_acc = 0.75，全字段准确率，考察模型在一条完整样本上执行所有任务的端到端可用性
- reasoning_acc = 0.9167，校验结果准确率
- 各项单字段准确率（考察模型单项识别能力）：
  - order_outcome_acc = 0.833
  - paid_amount_acc = 0.833
  - original_amount_acc = 1.00
  - discount_amount_acc = 1.00
  - order_time_acc = 0.833

解读要点：

1. schema_ok_rate=1.00 说明输出结构稳定，适合进入自动化评测与迭代流水线
2. all_5_fields_acc 低于单字段 acc，错误往往集中在某一字段拖累整条样本全字段通过
3. reasoning_acc 较高，校验模块可作为有效 QA 信号，用于定位金额字段噪声与规则问题

4. 分桶评测结果

分桶逻辑：

- **domain**：TB / JD / XHS（模板差异）
- **page_state**：FULL / SHRINK / OCCLUDED 等（遮挡与缩放）
- **info_completeness**：COMPLETE / INCOMPLETE（字段缺失预期）
- **missing_fields**：NONE / ORDER_TIME 等（缺失类型）

对齐验收：不同桶允许不同的通过率门槛（红线桶更严、长尾桶更关注UNCLEAR策略是否正确）

指导采样：下轮补数据优先补高风险桶（遮挡/交易关闭/字段名漂移/多商品合计）

解读要点：

由于Demo评测集样本不足，分桶评测数据无法全量统计。根据report2.0，OCCLUDED遮挡桶准确率显著下降

5. 错例归因

类型1：字段语义锚点漂移

典型案例1：TB_2

交易关闭页没有实付款字段，只有应付款。模型因为没看到实付款三个字就判 paid_amount=MISSING，导致全字段失败与 reasoning SKIP。

根因：

同一个业务字段（用户关心的付款金额），在不同订单状态模板中语义锚点不同（实付款 vs 应付款）。

迭代动作：

- 规则层：将 paid_amount 定义升级为：
  - 优先匹配实付款；
  - 若订单 outcome 为 CLOSED 且页面存在应付款并处在付款金额位置，则允许用应付款填充 paid_amount，但 evidence 必须包含应付款；
  - 同时增加 paid_amount_source（仅作为内部诊断字段，不进最终5字段也可进 explain）
- 数据层：专门补采/合成交易关闭页样本，覆盖不同模板

类型2：可见性导致的状态误判

典型案例1：TB_17、TB_18

订单顶部的 outcome 被遮挡。真值标为 MISSING，但模型在读图时用局部文字（退款相关）猜 outcome：

- T17 看到退款成功 → 猜 CLOSED

- T18 只有极速退款 → 猜 SUCCESS
- 同一类遮挡问题导致输出逻辑横跳、不一致。

根因：

- outcome 的证据位置是页面顶部状态栏，遮挡后模型开始用局部文本做推断；
- 推断规则不稳定，导致模型跨样本不一致。

迭代动作：

- 规则层：
  - 业务安全策略，保证关键字段准确：outcome 必须引用顶部状态栏锚点；顶部不可见 → outcome_status = UNCLEAR，进入二轮复检，不允许模型进行推测
- 数据层：补充遮挡桶样本，对于明确指导模型在遮挡时输出 UNCLEAR + 证据说明遮挡

类型3：模板差异与字段名缺失

现象：

某些页面原始金额没有明确字段名，但通过视觉锚点可推断。不同平台/不同页面字段名、位置差异明显。

根因：

跨域泛化问题：字段名缺失、布局差异导致模型依赖视觉布局推断。

迭代动作：

- prompt：强化证据必须带语义锚点，减少只抓数字的幻觉
- 数据口径：明确哪些可推断被允许、哪些必须 UNCLEAR
- 分桶补数据：domain × page_state 组合补齐模板覆盖

6. 迭代策略

1. **业务口径/规则收敛**
   - outcome 遮挡：输出 UNCLEAR
   - paid_amount：交易关闭页允许应付款 fallack，标明UNCLEAR状态和证据锚点

2. **Prompt 结构优化**
   - 强化：每个字段 evidence 必须包含关键词+数值
   - 强化：outcome 必须引用顶部状态栏证据，否则 UNCLEAR

3. **数据策略**
   - 按桶补数据：交易关闭页、遮挡页、缩放页、多商品合计页
   - 增加红线桶比例，保证关键风险场景提升

4. **训练策略**
   - 小规模 SFT：针对风险桶（遮挡/复杂状态订单/模板漂移）做定向训练
   - 偏好对齐：让模型在看不清/遮挡时更倾向输出 UNCLEAR 而不是乱猜，对齐业务策略