方案设计：本地代码仓的智能训练数据生成与处理

申请人： [郭顺] **申请职位：** SPS Associate Director (GCB4) **目标模型：** Qwen 2.5 系列 **演示仓库：** FastAPI RealWorld Example App (Python)

---

1. 执行摘要 (Executive Summary)

针对大型企业私有代码仓微调需求，本方案提出了一种基于**领域业务规则 (Domain Business Rules, DBR)** 驱动的自动化数据生成框架。通过静态分析技术提取代码逻辑，结合大语言模型的合成能力，生成具备高性能“推理链 (Reasoning Trace)”的微调数据集。方案重点考量了金融行业对逻辑严密性、数据合规性及系统可扩展性的核心要求。

---

2. 领域业务规则集 (Domain Business Rules)

为确保生成的训练数据具有高度的真实性与逻辑一致性，本方案从目标代码仓中提炼了以下五大核心业务规则作为数据生成的“真值来源”：

DBR-01：用户注册与账户完整性

- **业务逻辑：** 新用户注册需提供唯一的 Email 与用户名；密码严禁明文存储，必须经过哈希加密处理。
- **代码依据：** 见 app/schemas/user.py 及 app/crud/crud_user.py 中的 get_password_hash 逻辑。

DBR-02：身份认证与 JWT 令牌管理

- **业务逻辑：** 系统基于 Email/Password 进行身份验证，成功后颁发带有时效性的 JWT (Bearer Token)。
- **代码依据：** 见 app/api/api_v1/endpoints/login.py 及安全配置模块。

DBR-03：文章生命周期与 Slug 自动化处理

- **业务逻辑**：创建文章时，系统根据 title 自动生成唯一的 slug 标识符；文章必须强制关联当前登录作者。
- **代码依据**：见 app/models/domain/articles.py 及文章创建接口依赖注入。

DBR-04：基于权属的访问控制 (OBAC)

- **业务逻辑**：更新或删除操作必须验证资源归属权。非作者本人操作需触发 403 Forbidden 异常；未登录操作触发 401 Unauthorized。
- **代码依据**：见 app/api/dependencies/articles.py 中的权限校验中间件。

DBR-05：评论关联性约束与权限

- **业务逻辑**：评论必须绑定有效文章；仅评论发布者拥有该评论的删除权限。
- **代码依据**：见 app/api/api_v1/endpoints/comments.py 的逻辑实现。

---

3. 系统架构与数据流水线 (System Architecture)

本方案采用模块化流水线设计，确保数据生成的自动化与标准化：

1. **代码分析层**：利用 Python ast 模块进行静态扫描，识别路由路径、输入输出 Schema 及权限依赖。
2. **规则映射层**：将扫描结果与上述 DBR 进行关联，生成包含业务背景的“上下文信息块”。
3. **智能合成层**：调用 Teacher Model（如 Qwen-72B）根据上下文生成包含推理步骤 (Reasoning Trace) 的问答对。
4. **质量控制层 (Quality Gate)**：执行 PII 敏感信息脱敏，并对生成的推理步骤进行逻辑自洽性校验。

---

4. 训练数据结构示例 (Data Samples)

系统生成的训练数据采用符合 Qwen 2.5 要求的结构化 JSONL 格式：

场景 1：复杂权限逻辑问答 (QA with Trace)

- **指令**：当非作者用户尝试删除某篇现有文章时，系统如何处理？
- **推理过程 (Trace)**：
    1. 解析 Authorization Header 获取 JWT 令牌，确认用户身份。
    2. 通过 slug 检索数据库获取文章对象及其 author_id。
    3. **触发规则 DBR-04**：对比 current_user.id 与 article.author_id。
    4. 因权属不匹配，系统抛出 HTTPException(status_code=403)。
- **结果**：请求被拦截，返回 403 Forbidden 错误。

---

5. 质量保证与金融合规 (Quality & Compliance)

- **隐私保护**：系统内置脱敏机制，自动识别并过滤硬编码的密钥、内部域名及开发者私人注释。
- **数据多样性**：通过多模版变换（如“假如、如果、当...时”）确保模型对业务逻辑理解的泛化性。
- **审计追踪**：每条数据均包含元数据 (Metadata)，可追溯至具体的源代码文件及其版本号。

---

6. 未来扩展 (Scalability)

本方案设计的解析逻辑具备**语言中立性**。通过适配不同的 AST 解析器，可快速支持现有的 Java 或 C# 代码仓。同时，方案支持**增量更新**，当本地代码仓发生变更时，可自动触发增量训练数据的生成。