**接口文档2.0 vs 接口文档3.0 修改对比表：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模块 | 修改类型 | 接口文档2.0 | 接口文档3.0 | 修改说明 |
| 文档说明 | 无变化 | 已包含功能、场景、技术特性等说明 | 同2.0版本 | - |
| 模块架构图 | 内容更新 | 单张架构图 | 新增多张架构图（如上下文特征模块、API网关层） | 3.0版本通过多图展示更复杂的系统架构 |
| 核心功能 | 无变化 | 动态权重分配、冲突消解、降级策略 | 同2.0版本 | - |
| 输入接口 | 格式优化 | 请求地址和Headers为文本描述 | 新增图示化Headers说明（如X-Api-Key的生成逻辑） | 提升可读性 |
| 请求体参数说明 | 新增字段 | 基础参数表（如base\_prompt、temperature） | 新增fallback\_strategy枚举值说明（strict/balanced/creative） | 明确降级策略选项 |
| 属性对象定义 | 约束扩展 | 仅支持min\_weight约束 | 新增max\_weight和conflict\_terms约束 | 支持更精细的权重控制和显式冲突声明 |
| 完整请求示例 | 示例更新 | 简单JSON示例 | 新增复杂示例（如包含conflict\_terms的完整属性对象） | 帮助用户理解多约束场景 |
| 输出字段详解 | 无变化 | 包含final\_weights、conflict\_report等字段说明 | 同2.0版本 | - |
| 错误码分类 | 新增错误码 | 输入参数错误（4xx）、语义冲突（422）、系统错误（5xx） | 新增40003（权重超限）、42202（权重总和超阈值） | 覆盖更多异常场景 |
| 实战示例 | 新增场景 | 基础权重计算、显式冲突处理 | 新增隐式冲突检测示例（如CLIP模型评分低于阈值触发冲突） | 展示更全面的冲突处理逻辑 |
| 性能指标 | 数据更新 | 压力测试数据（并发100/500/1000） | 新增GPU利用率指标（如gpu\_utilization: 78.4） | 更详细的性能监控指标 |
| 权重计算公式 | 无变化 | 动态权重公式和语义评分公式 | 同2.0版本 | - |
| 冲突检测逻辑 | 代码优化 | 基础Python冲突检测函数 | 新增阈值配置（如similarity < 0.3）和隐式冲突处理代码 | 支持动态调整冲突敏感度 |
| 核心算法流程 | 图示更新 | 单张流程图 | 新增多张子流程图（如特征编码、动态权重计算） | 分步骤展示算法细节 |
| 新增模块 | 完全新增 | 无 | 新增“上下文特征模块”（设备类型检测、地理位置编码） | 3.0版本扩展了上下文处理能力 |
| 多类型Value支持 | 完全新增 | 仅支持文本输入 | 新增图像（Base64）、向量（512维）的校验与处理逻辑 | 支持多模态输入 |
| 依赖管理 | 完全新增 | 无 | 新增requirements.txt版本锁定、安全扫描（pip-audit）、GitHub Actions自动化 | 提升工程化标准 |
| API网关功能 | 完全新增 | 无 | 新增负载均衡（Nginx）、请求签名（HMAC-SHA256）、频率限制（Redis）实现与配置示例 | 增强安全性和高可用性 |

关键修改点总结：

1.架构扩展：新增上下文特征模块和API网关层，系统复杂度显著提升。

2.输入输出增强：属性对象支持max\_weight和conflict\_terms约束，输出新增GPU利用率指标。

3.冲突检测优化：隐式冲突检测支持动态阈值，代码示例更完整。

4.多模态支持：输入支持图像（Base64）和向量（512维），通过JSON Schema严格校验。

5.工程化改进：依赖管理、安全扫描、CI/CD自动化等新增内容体现生产级标准。