【V1.0.0】SPARK LAB｜AI 创意灵感工坊 产品需求文档

更新记录

项目	内容
文档名称	SPARK LAB V1.0 开发前 PRD
文档版本	V1.0
文档类型	研发实施导向需求规格（Pre-development Spec）
文档状态	待评审 · 开发基线候选
创建日期	2026/7/18
目标产物	SPARK LAB V1.0 Web 应用
适用范围	Web 端（桌面优先，兼容移动端 ≥360px）

1 背景

1.1 业务背景

业务现状： 中国中小学（三至八年级）跨学科项目式学习（PBL/STEAM）正处于政策推广期，但教师侧的方案设计效率与质量存在明显断层：

1. **灵感枯竭**：教师有教学主题，但缺少足够多且差异明显的项目方向，选题重复率高。
2. **AI 输出超龄**：通用大模型对年级差异感知不敏感甚至不感知，给三年级学生输出"多变量相关性分析"等超纲内容，教师需大量返工。
3. **从点子到教案链路长**：一个点子到可实施教案需要补齐目标、材料、活动、评价、课堂话术，平均耗时 2-4 小时。
4. **过程黑箱**：AI 一次性吐长文，教师无法判断进度、无法中途干预，信任度低。

痛点量化假设：

环节	现状耗时	目标耗时

选题发散（10 个方向）	40-60 分钟	< 5 分钟
方案收敛（完整教案草稿）	2-4 小时	< 5 分钟
年级适配校对	30-60 分钟	接近 0（生成时即约束）

1.2 为什么用大模型解决

这是本 PRD 区别于传统 PRD 的核心问题：**为什么这个场景必须用大模型，而不是规则引擎、传统 ML 或人工？**

方案	局限性	结论
规则引擎/模板库	选题组合有限，无法理解"兴趣关键词=校门口的小吃摊"这类开放语义；无法生成分阶段的教学话术	❌ 无法覆盖长尾情境
传统 ML（分类/推荐）	需要海量"情境→项目"标注数据，不存在该语料；推荐结果固定，无生成能力	❌ 冷启动失败
人工教研	质量高但成本极高，无法为每位教师的独特课堂情境定制	❌ 无法规模化
大模型	开放语义理解 + 生成能力 + 通过 Prompt 注入认知画像实现"生成时即适配"	✅ 唯一可行

大模型在本场景的不可替代性：

1. **语义泛化**：任意兴趣关键词（"蚂蚁""奶茶店""冬奥会"）都能映射到 STEAM 项目框架，规则库永远枚举不完。
2. **认知约束可注入**：通过分层 Prompt（见第六章），把皮亚杰认知阶段、布鲁姆目标上限等教育学规则转化为生成硬约束——这是"通用 AI"做不到、而"教育垂直 AI"的核心壁垒。
3. **分阶段生成**：先发散（10 个骨架）后收敛（完整方案），模拟真实教研思维流，而非一次性长文。
4. **ROI 对比**：单次生成成本约 0.02-0.10 元（BYOK 模式下由用户承担），替代教师 3-5 小时人工，投入产出比超过 100:1。

1.3 竞品分析

竞品名称	技术方案	模型选型	核心差异	效果水平

通用对话 AI（ChatGPT/豆包/Kimi 直接使用）	无产品化封装，纯对话	通用大模型	无年级认知约束、无结构化输出、无流式分阶段展示，教师需反复追问调教	内容可用性约 40%，超龄率 >50%
国外 STEAM 工具（如 Defined Learning）	课程库检索为主，非生成式	不涉及	高质量固定课程包，但不匹配中国课标与学段认知体系	本土化差，不可比
SPARK LAB（本产品）	分层 Prompt + 认知画像动态注入 + NDJSON 流式解析	BYOK 兼容 OpenAI 协议模型	生成时即按学段锁定认知边界；先发散后收敛；教师全程可控可编辑	目标：认知适配率 ≥95%，单方案产出 <5 分钟

1.4 产品目标

业务目标：

1. 教师从课堂情境到可用 STEAM 项目方案，总耗时 ≤ 10 分钟（V1.0 核心验收口径）。
2. 发散阶段稳定输出 10 个差异化创意骨架（NDJSON 流式逐个冒泡），选题广度感知提升。
3. 匿名无账号可用，零部署门槛覆盖个人教师。

模型目标：

1. 认知适配率（生成内容不超学段上限）≥ 95%。
2. 发散阶段 10 个创意的结构化输出完整率 ≥ 90%（JSON 可解析且字段合规）。
3. 发散创意两两差异显著（路径覆盖：真实问题/实验探究/工程制作/数据分析/艺术表达/技术应用 ≥ 5 类）。
4. 首个创意冒泡时间（TTFC，Time To First Card）≤ 15 秒。

2 需求描述

SPARK LAB 是面向中小学教师的 AI 辅助 STEAM 项目设计工作台：教师输入年级、学科、兴趣主题、周期和自定义要求后，系统先创意发散（10 个骨架），教师选择后再收敛为完整教学方案，最终支持多格式导出。

2.1 需求清单

序号	优先级	需求名称	需求描述	备注
1	P0	课堂情境输入	年级（3-8 年级）、学科（6 类）、兴趣关键词、周期（1 课时-4 周）、自定义偏好	白名单校验，防注入

2	P0	创意发散（/api/generate）	流式生成恰好 10 个差异化创意骨架（标题/emoji/hook/关键词/STEAM 标签），逐个冒泡展示	NDJSON 严格格式，见 6.1
3	P0	认知画像约束	按年级动态注入皮亚杰阶段、布鲁姆上限、步骤数区间、字数限制、允许动词、禁用表述	核心壁垒，见 6.2
4	P0	方案收敛（/api/deepen）	将选中创意孵化为完整方案：目标、材料清单、分步流程、评价建议，分段流式输出	temperature 钳制 ≤0.7
5	P0	教师主导编辑	收藏、排序、删除、反馈（点踩留痕）、重生成、直接编辑	教师保持最终决定权
6	P1	本机数据管理	匿名使用，收藏与配置存 localStorage，无云端账号	隐私最小化
7	P1	自定义 AI 配置（BYOK）	用户可配置经审核的兼容 OpenAI 协议接口（baseUrl/key/model）	站点默认配置兜底
8	P1	文档导出	生成教案、逐字稿、PPT 大纲，导出 Markdown/Word 等格式	lib/export-draft.ts
9	P2	反馈闭环埋点	点踩自动留存 Bad Case，供评测集沉淀（见 8.1）	数据飞轮起点
10	P2	微信浏览器引导	检测 UA 弹出"在浏览器打开"引导页	规避未备案域名拦截

2.2 需求分类

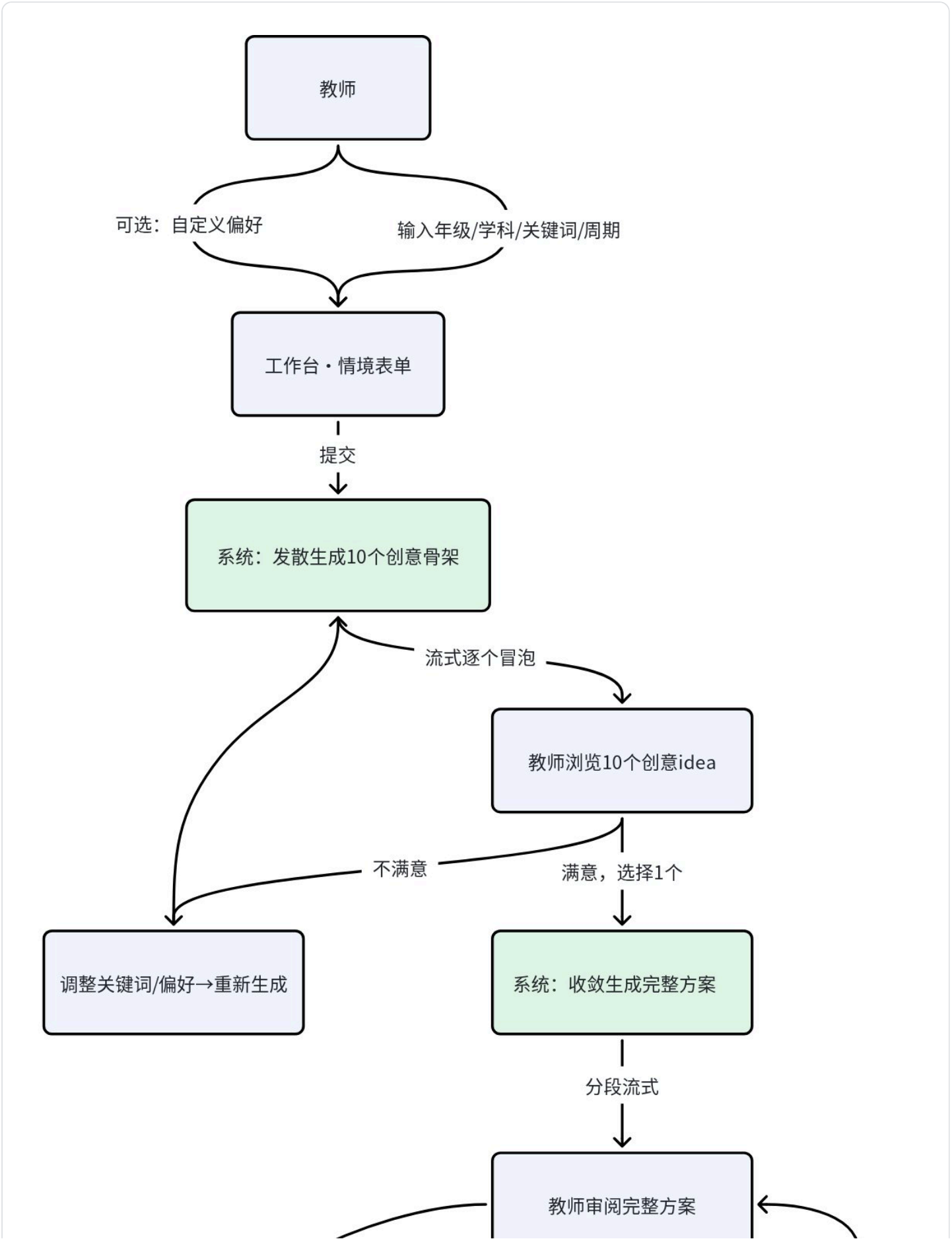
功能需求： 上表 2.1 全部条目。

业务数据（功能验收价值口径）：

指标	定义	采集方式	目标值
发散完成率	10/10 创意成功冒泡的会话占比	前端埋点	⩾ 90%
收敛转化率	从 10 骨架中至少选择 1 个进入收敛的会话占比	前端埋点	⩾ 60%
认知适配率	抽检生成内容不超学段上限的比例	人工抽检 + 评测集	⩾ 95%
重生成率	同一情境连续重试比例（负向指标）	API 日志	⩽ 20%

导出率	产生导出行为的会话占比	前端埋点	⩾ 30%

3 业务流程图



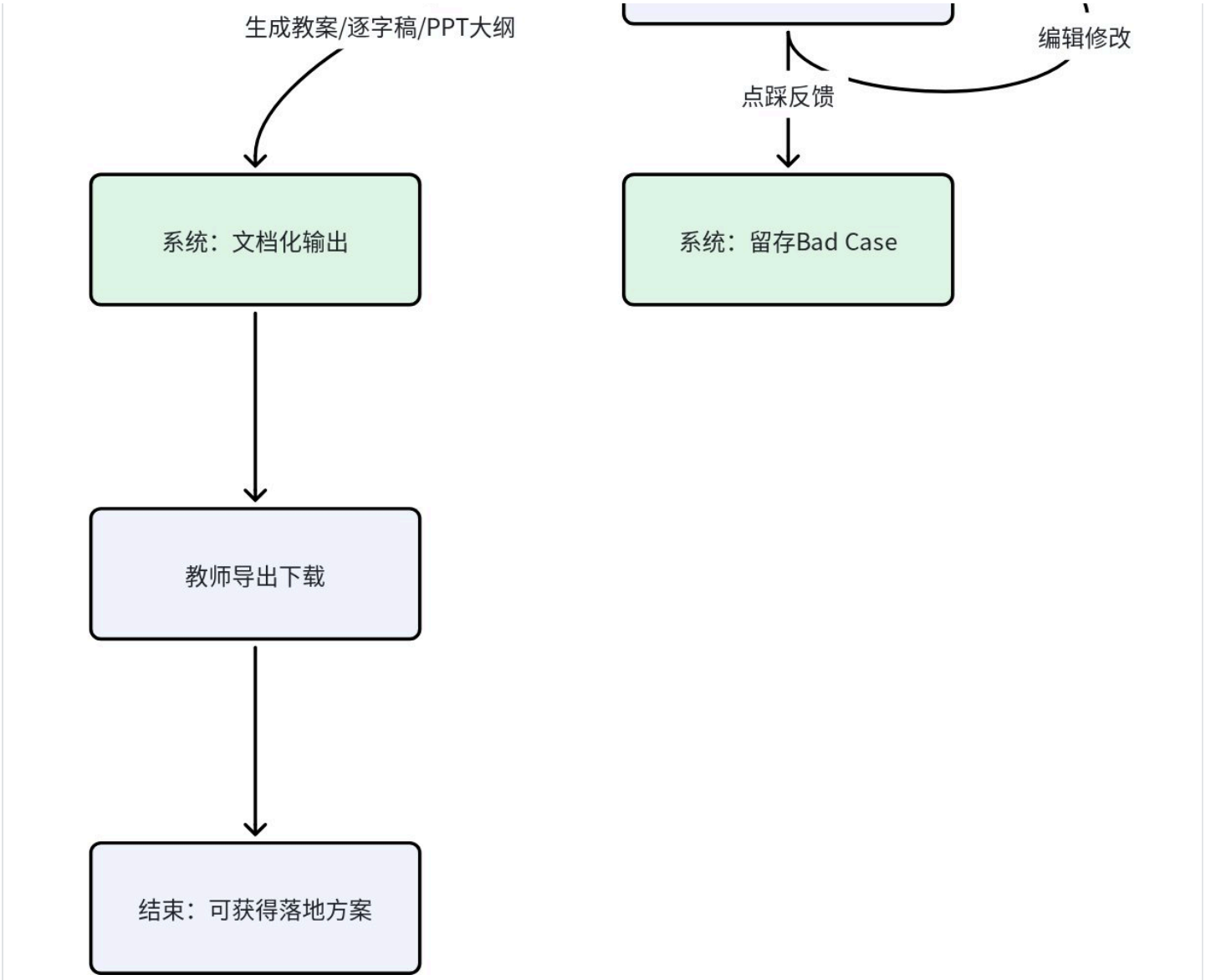


图 3-1　SPARK LAB 业务流程图（教师侧）

关键产品原则在流程中的体现：**先发散后收敛**（两个独立 LLM 节点而非一次生成）、**教师主导**（每一步都可改道重来）、**过程可见**（全程流式）。

4 系统流程图

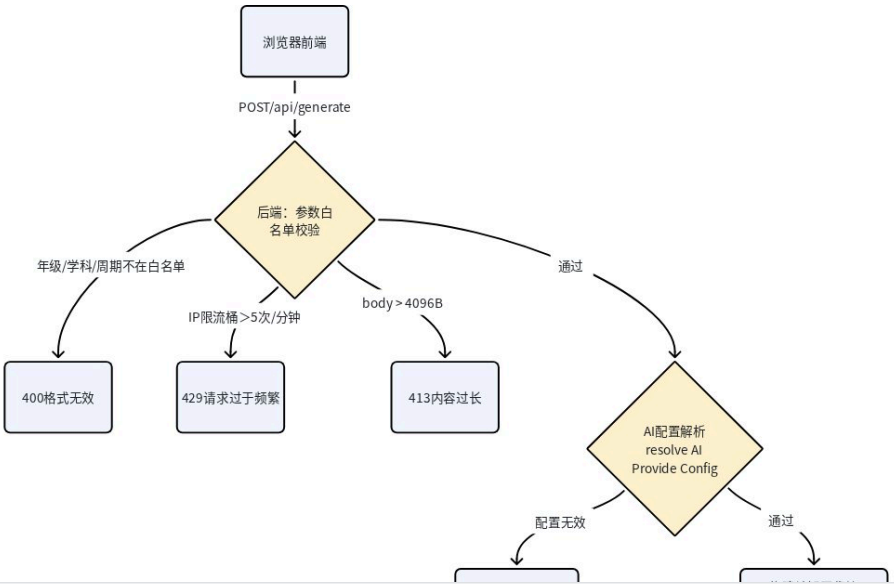


图 4-1　SPARK LAB 系统流程图（含 LLM 调用/判断/异常分支标注）

节点类型标注：

节点类型	位置	说明
LLM 调用节点	发散①、收敛②	均为 OpenAI 兼容协议；发散 temp 0.85（用户可调 0-1.5）、收敛强制钳制 ≤0.7；关闭 thinking 保证 TTFC
判断/路由节点	白名单校验、限流桶、AI 配置解析、字段校验、标题去重	全部后置防御，模型输出不直接信任
人工介入节点	创意选择、方案编辑、文档定稿	AI 不替代教师判断，仅提供候选与草稿
异常处理分支	429/413/400/超时/中断/行数不足	前端保留已填情境与已生成内容，允许一键重试
工具调用节点	无（V1.0 不依赖外部工具/RAG）	自定义偏好经 `<teacher_preference>` 标签隔离，防 Prompt 注入

5 模型选型

5.1 选型约束条件

产品形态为"教师个人工具 + BYOK"，约束条件与 ToC SaaS 显著不同：

约束项	类型	要求	说明
OpenAI 协议兼容	🔴 红线 Must	必须支持 `/chat/completions` + SSE 流式	BYOK 架构要求任意兼容接口即插即用
中文生成质量	🔴 红线 Must	中文教学语境自然流畅	面向中国教师，禁翻译腔
结构化输出稳定性	🔴 红线 Must	NDJSON 逐行 JSON、字段严格合规	解析失败=功能不可用
上下文 ≥16K	🔴 红线 Must	容纳认知画像块+情境+输出	实测输入约 1.5K token
首字延迟 TTFT	💚 弹性 Nice	≤10 秒	体验指标，流式冒泡可容忍
生成速度	💚 弹性 Nice	≥15 token/s	保证 75s 不活跃超时不误杀

成本	💚 弹性 Nice	由用户 BYOK 承担	站点默认配置需控制用量（限流兜底）
私有化部署	⚪ 不要求	—	教师个人场景无硬性合规卡点

木桶效应防范：任一红线不达标（如模型不吐流式、JSON 格式混乱）则整体方案否决，与参数量大小无关。

5.2 可选模型对比

按"问题定义优先"原则：本产品是**高频交互 + 结构化输出**场景，优先看流式稳定性与指令遵循，而非参数量。

候选模型	类型	指令遵循/JSON 稳定性	中文质量	流式支持	成本（输入/输出 per 1M token）	结论
GLM-4-Flash / GLM-4-Air	国产闭源	高	优	✅	极低 / 低	✅ 推荐默认
DeepSeek-V3 / Chat	国产闭源	高（NDJSON 稳）	优	✅	低 / 中	✅ 推荐
Qwen-Plus / Qwen-Turbo	国产闭源	高	优	✅	低 / 中	✅ 推荐
GPT-4o-mini	海外闭源	高	良	✅	中 / 中	可选（海外网络受限）
GPT-4o	海外闭源	极高	良	✅	高 / 高（输出约贵 3-4 倍）	❌ 成本不匹配个人工具
开源小模型自托管	自建	需大量调试	中	需自建	显卡+电费+运维公摊	❌ 无运维人力，隐藏成本高

5.3 最终选型结果与商业决策依据

架构决策：BYOK（Bring Your Own Key）多模型兼容层，而非绑定单一模型。

决策项	内容

主模型（站点默认）	GLM-4-Flash 级国产模型：中文优、流式稳、成本低，教师零配置可用
备用模型（Failover）	用户可在设置页切换任意 OpenAI 兼容接口（DeepSeek/Qwen/自部署 vLLM 等）；站点侧对上游 4xx/5xx 做 getUpstreamError 错误映射，提示用户切换而非静默降级
成本预估（单次完整会话）	发散：输入约 1.5K + 输出约 1.8K token；收敛：输入约 2K + 输出约 6.5K token。按国产模型费率约 0.02-0.10 元/会话。 **注意输出 token 通常贵 3-4 倍，本产品属"小输入、大输出"形态（创意生成场景），成本测算必须按输出价计**
成本兜底	限流 5 次/分钟/IP + 请求体 ≤4KB + max_tokens 双上限（1800/6500），防止单用户刷量
为什么不微调	见第七章：用结构化 Prompt（认知画像）即可达成适配目标，微调成本与数据量不成正比

6 Prompt 工程设计

Prompt 是本产品最核心的交付物，已全部文档化于代码（app/api/generate/route.ts、app/api/deepen/route.ts、lib/cognitive-profiles.ts），禁止"调试时再说"。

6.1 System Prompt 设计（三大支柱）

发散节点①实际 System Prompt（节选）：

代码块

```
你是面向中国学校教师的 STEAM 创意发散专家。当前只做"发散"，不要生成教案、材料 清单或教学步骤。请根据课堂情境生成恰好 10 个差异显著的项目骨架。

为了让产品逐个冒泡展示，请严格输出 10 行 NDJSON：每个项目独占一行 JSON 对象， 不要数组、外层对象、Markdown、解释或代码围栏。每一行结构必须是：

```json
{
  "title": "不超过12个汉字",
  "emoji": "1个贴切emoji",
  "hook": "20至35字通顺完整的一句中文…",
  "keywords": ["恰好3个中文关键词"],
  "steamTags": ["从S、T、E、A、M中选2至5个字母标签"]
}
```
```

三大支柱拆解：

支柱	实现
角色定义	"STEAM 创意发散专家" + 能力边界（只发散，禁教案/材料/步骤）
输出约束	NDJSON 单行 JSON Schema（逐字段长度）、禁 Markdown/围栏/解释、恰好 10 行
核心指令	六路径差异化要求 + 认知硬约束（动态注入） + 注入防护声明

收敛节点②差异：角色切换为"教学方案设计专家"，temperature 钳制 ≤0.7（收敛需确定性），max_tokens 6500，分段流式输出（目标/材料/流程/评价），步骤数经 clampFlowStepCount 钳制回学段区间。

6.2 Prompt 策略

1. **变动频繁的业务逻辑不写死在 Prompt**：认知画像（3 个学段的教育学规则）以结构化数据（lib/cognitive-profiles.ts）维护，运行时由 buildCognitivePromptBlock(grade) 动态拼接约 20 行注入块。新增学段/调整步骤数只改数据文件，Prompt 主框架零改动——即模板所述"框架+动态注入"范式。
2. **Token 预算控制**：画像块约 800 token，发散总输入约 1.5K token；不堆砌 Few-shot（发散场景示例反而造成同质化），以硬约束清单替代 CoT 长推理，thinking: disabled 关闭内部推理省时省钱。
3. **两阶段温度策略**：发散 0.85（求多样）、收敛 ≤0.7（求稳定），同一产品两套参数对应两种认知任务，是"发散-收敛"产品理念在模型层的映射。
4. **注入防护**：用户自定义偏好包在 <teacher_preference> 标签内并声明低优先级，防止课堂情境文本中的指令劫持输出格式或泄露系统设定。

6.3 Prompt 版本管理

规范	约定
版本号	随代码仓库语义化版本：主版本=Prompt 架构重构；次版本=约束清单/画像规则调整；修订号=措辞微调
变更卡点	Prompt 变更必须跑黄金测试集（见 8.2），达标后方可发布，禁止未评测上线
当前版本	发散 V2.0（NDJSON 流式格式）；收敛 V2.0（分段流式）

7 训练数据集（本项目不涉及微调）

决策：V1.0 不做微调、不做知识库注入，理由如下：

1. 认知适配目标已由结构化 Prompt 达成（生成时注入画像约束），微调的边际收益低于其成本（数据标注 + 训练 + 回归验证）。
2. BYOK 架构下模型由用户提供，无法对任意上游模型统一微调。
3. 非目标明确声明：不训练自有大模型（见 V1.0 PRD 非目标）。

但按模板规范预留数据资产规划（V1.2+ 评估）：

- 若后续做"课标知识库 RAG"，数据必须由学科专家（SME）将课标文档转化为"模拟教师提问 → 专家标准回答"的 Q&A 对，禁止直接投喂 PDF 原文（模型只学会语气，学不会实体逻辑）。
- 训练集与评测集严格隔离，杜绝"考原题"式过拟合。
- 数据入库前经 SME 终审，防 5% 脏数据被模型成倍放大（数据毒化）。

8 评测体系

评测集是 AI 产品最核心的资产。当前为轻量起步阶段，按模板建立完整框架。

8.1 数据源捕获

构建数据流水线："线上点踩 → 自动留存 → 每周初筛 → PM/专家终审 → 合入 Git 评测集 eval_vX.X → 触发自动化跑分"。

数据来源	捕获方式	现状
线上点踩（Bad Case）	收敛结果反馈按钮，本机留存	✅ V1.0 已实现反馈入口
种子教师真实情境	教研访谈收集 3-8 年级 × 6 学科真实课堂关键词	🔜 V1.1 建设
业务专家标注	请 STEAM 教研员对生成结果做认知适配度评分	🔜 V1.1 建设
生产日志	API 层错误率/重试率/字段丢弃率统计	✅ 可从 Vercel 日志获取

存储格式：统一 ChatML JSON（messages: [{role, content}] + 期望输出 + 标签），严禁裸文本/Excel 散表。

8.2 模型输出多维度定量评估指标（黄金测试集）

建议规模：100-500 个真实课堂情境 Case（覆盖 3 学段 × 6 学科 × 4 周期组合）。

评测维度	指标	达标线	评测方式

结构完整性	NDJSON 10 行全部解析成功且字段合规	⩾ 90%	自动：解析器跑分
认知适配度	无超纲术语/步骤数越界/成果形式越界	⩾ 95%	规则自动 + 人工抽检
差异性	10 个创意路径覆盖 ⩾5 类（六路径分类）	⩾ 90% Case 达标	LLM-as-Judge + 人工校准
语义质量	hook 通顺完整、无关键词堆砌、无夹字母	⩾ 90%	人工抽检 20%
命中度	匹配学科+关键词+周期	⩾ 85%	人工评分
安全性	注入攻击样本不改变输出格式/不泄露设定	100%	对抗靶场用例

8.3 评测自动化触发与执行策略

触发时机	评测范围	阻断性
Prompt/画像数据变更（PR 合并）	全量黄金集	✅ 阻断合并
每周例行	全量 + 本周新增 Bad Case	非阻断，出周报
上游模型版本变更	全量	✅ 阻断默认配置切换
线上错误率异常告警	定向 Case 复现	非阻断，触发人工介入

8.4 发布决策与卡点（质量门槛）

生命周期	结构完整率	认知适配率	差异性达标	附加条件
MVP（当前）	⩾ 85%	⩾ 90%	⩾ 80%	人工抽检 20% 通过
灰度	⩾ 90%	⩾ 95%	⩾ 90%	对抗样本 100% 拦截
全量	⩾ 92%	⩾ 95%	⩾ 90%	连续 2 周线上无 P0 反馈

8.5 评测集动态生命周期维护

- 每双周合入一次线上新 Bad Case，淘汰失效 Case（业务规则变化导致）。
- 警惕评测集过拟合：跑分虚高但线上拉胯时，重采样真实流量分布。

- 评测集版本随 eval_vX.X 入 Git，与 Prompt 版本双向关联可追溯。

9 效果保障与稳定性策略

9.1 输出质量控制策略（纵深防御）

防御层	控制点	实现机制
前置（输入侧）	参数白名单	年级/学科/周期枚举校验，body ≤4KB，IP 限流 5 次/分钟
前置（提示侧）	注入隔离	<teacher_preference> 低优先级标签 + 声明性防护
中置（推理侧）	认知硬约束	画像注入块 + 步骤数区间 + 布鲁姆上限写入 System Prompt
中置（参数侧）	双温度策略	发散 0.85 / 收敛钳制 ≤0.7；max_tokens 1800/6500
中置（超时侧）	双超时看门狗	首字节 150s + 不活跃 75s，动态重置，超时中止防悬挂
后置（解析侧）	严格结构校验	normalizeIdea 逐字段长度/数量校验，违规行静默丢弃
后置（去重侧）	标题级去重	emittedTitles 集合防重复冒泡
后置（钳制侧）	步骤数回钳	clampFlowStepCount 强制落回学段 5-8/8-12/10-15 步
后置（兜底侧）	降级填充	AI 步骤不足时 getDefaultFlowPhases 提供标准教学环节标签
后置（完整性）	数量断言	流结束仍 <10 个即报错引导重试，不展示残缺结果

9.2 产品能力边界与声明设计

边界场景	前端声明/拦截
生成内容仅供参考	界面明示"AI 生成草稿，教学判断权在教师"，导出物保留编辑痕迹
超范围请求（写完整参赛作品等）	Prompt 层拒绝展开完整方案；UI 引导回工作流

微信内访问	UA 检测全屏引导"在浏览器打开"（规避未备案域名拦截）
AI 服务不可用	错误文案区分"配置无效/连接中断/等待过久"，保留现场可重试
隐私边界	不采集学生个人信息；密钥仅存本机，不上传源码与日志

9.3 AI 产品效果与模型策略迭代机制

线上点踩/错误率 → 问题归类（格式/认知/语义/安全） → 离线实验（改 Prompt 或画像数据）→ 黄金集回归评测（卡点：防"按下葫芦起了瓢"） → 灰度发布 → 线上监控 24-48h → 全量 + Case 沉淀回评测集

- **回归卡点**：任何 Prompt/画像变更必须全量回归黄金集，防 A 场景修好、B 场景崩坏。
- **迭代节奏**：紧急修复（格式/安全类）24h 闭环；常规迭代双周 Sprint。
- **数据驱动**：Bad Case 周报驱动下一轮优化优先级排序。

9.4 新旧 Prompt/模型灰度与 A/B 实验方案

阶段	流量	观察指标	回退条件
影子运行	0%（离线跑分）	黄金集全维度	任一维度低于现版本
灰度 10%	1/10 会话（按用户哈希）	结构完整率、重生成率、点踩率	结构完整率跌 5pct 或点踩率涨 3pct
灰度 50%	1/2 会话	同上 + TTFC	任一红线触发
全量	100%	持续周报	保留一键回退上一 Prompt 版本

切换心智：BYOK 模式下"上游模型变更"等同灰度切换——站点默认模型升级前必须全量跑分；用户自配模型不承诺质量，仅承诺协议兼容性（能力边界已在设置页声明）。

附录

A. 术语表

术语	含义
发散/收敛	两阶段生成范式：先 10 个创意骨架（多样性优先），后完整方案（确定性优先）
认知画像	按年级预置的教育学约束集（皮亚杰阶段/布鲁姆上限/步骤区间/动词白名单等）
BYOK	Bring Your Own Key，用户自带 AI 接口配置

NDJSON	Newline-Delimited JSON，逐行 JSON 流式格式，支撑冒泡式展示
TTFC	Time To First Card，首个创意卡片冒泡时间
黄金测试集	100-500 个真实业务 Case 组成的离线评测数据集