

QCon
全球软件开发大会

从客户痛点到工程化产品： 火山方舟PromptPilot落地实践启示录

姓名：王铁飞

火山引擎大模型算法工程师

AI下半场：落地落地落地！



目录

01

大模型应用落地的挑战

02

火山方舟PromptPilot的破局之道：痛点驱动的产品能力与实践路径

03

沉淀与前行：从服务到产品的工程化启示与未来展望

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

01

大模型应用落地的挑战

大模型应用落地的挑战



清晰表达需求



模型能力边界

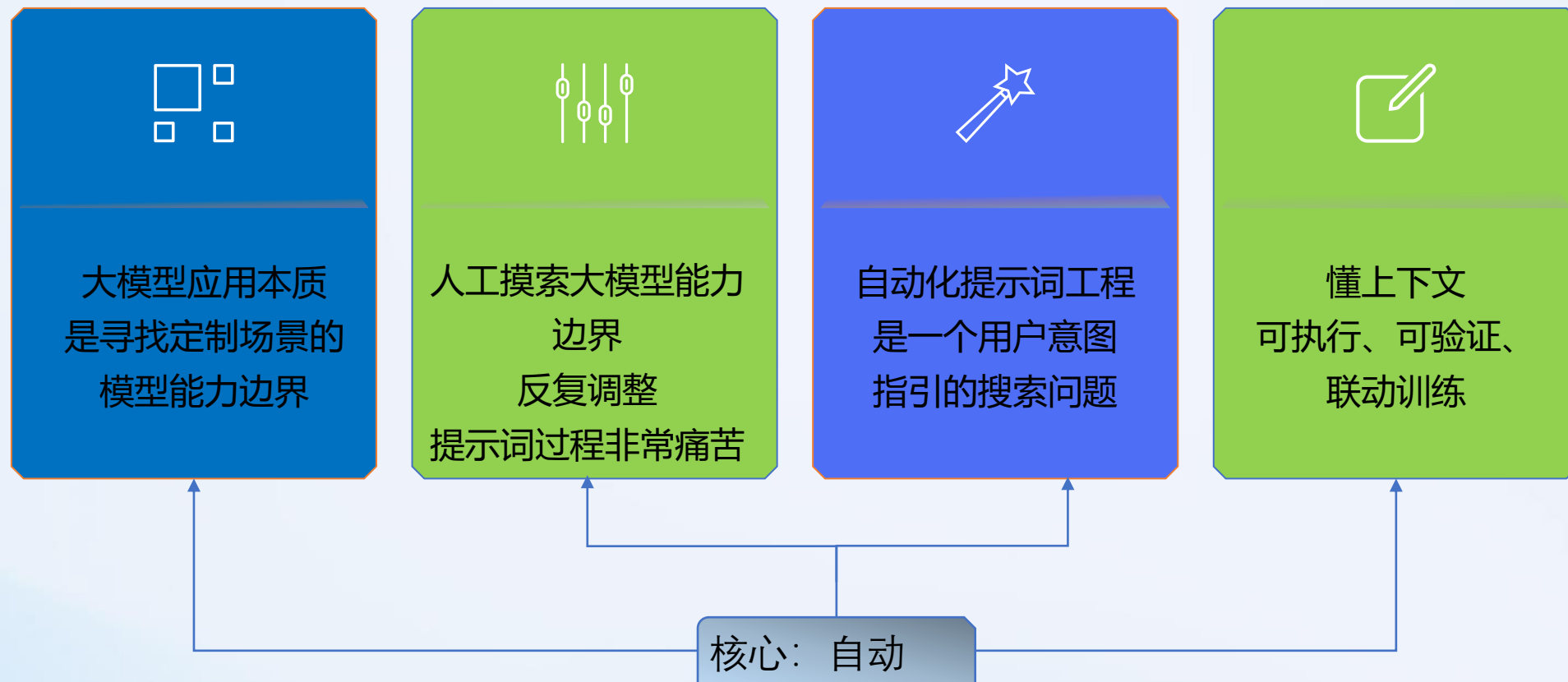


上下文动态适应

大模型应用落地的挑战 — 清晰表达需求



大模型应用落地的挑战 — 模型能力边界



大模型应用落地的挑战- 上下文动态适应



大模型应用很多是
开放式
人类高级思维问题



大模型的
上下文
是多样、多变的



开发中、上线后
都有很多想象不到的
Badcase 出现



闭环联动



Badcase的持续检测
评估标准持续迭代
上下文自适应

目录

01

大模型应用落地的挑战

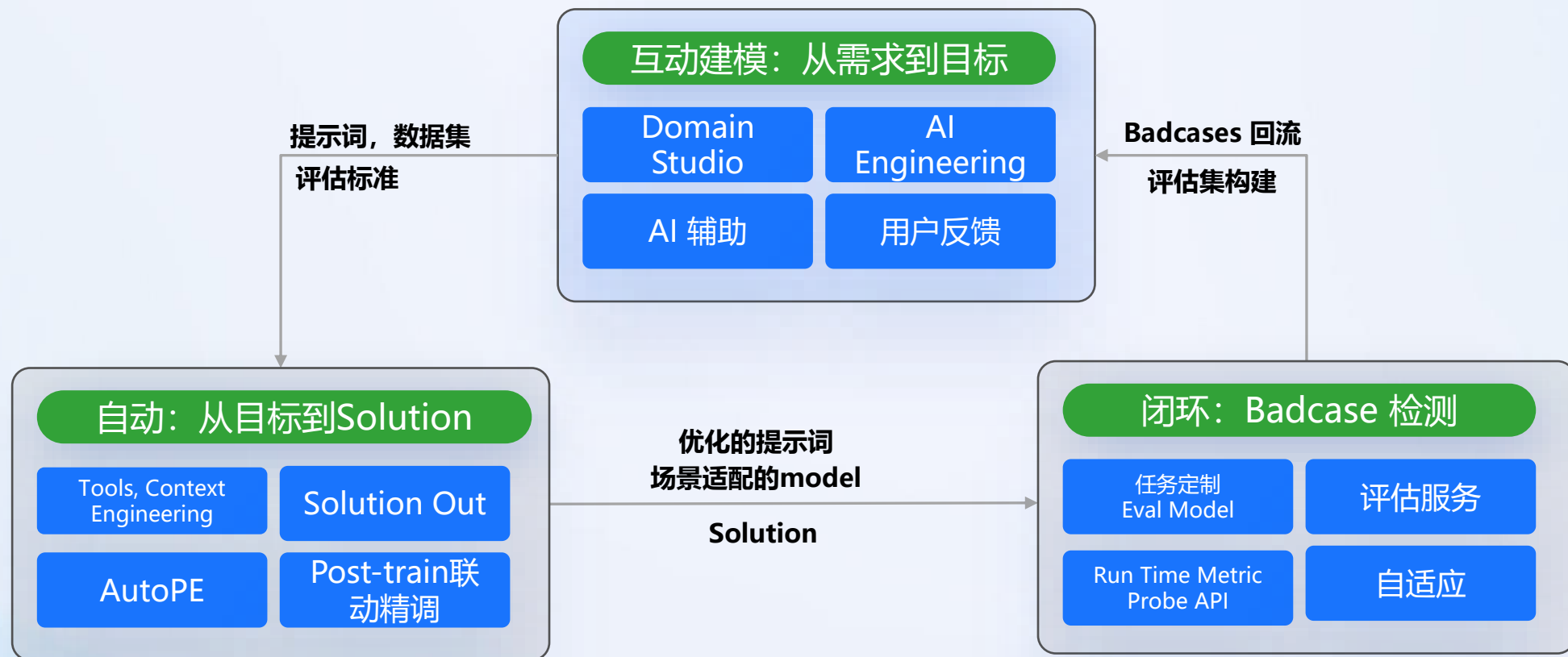
02

火山方舟 PromptPilot 的破局之道：痛点驱动的产品能力与实践路径

03

沉淀与前行：从服务到产品的工程化启示与未来展望

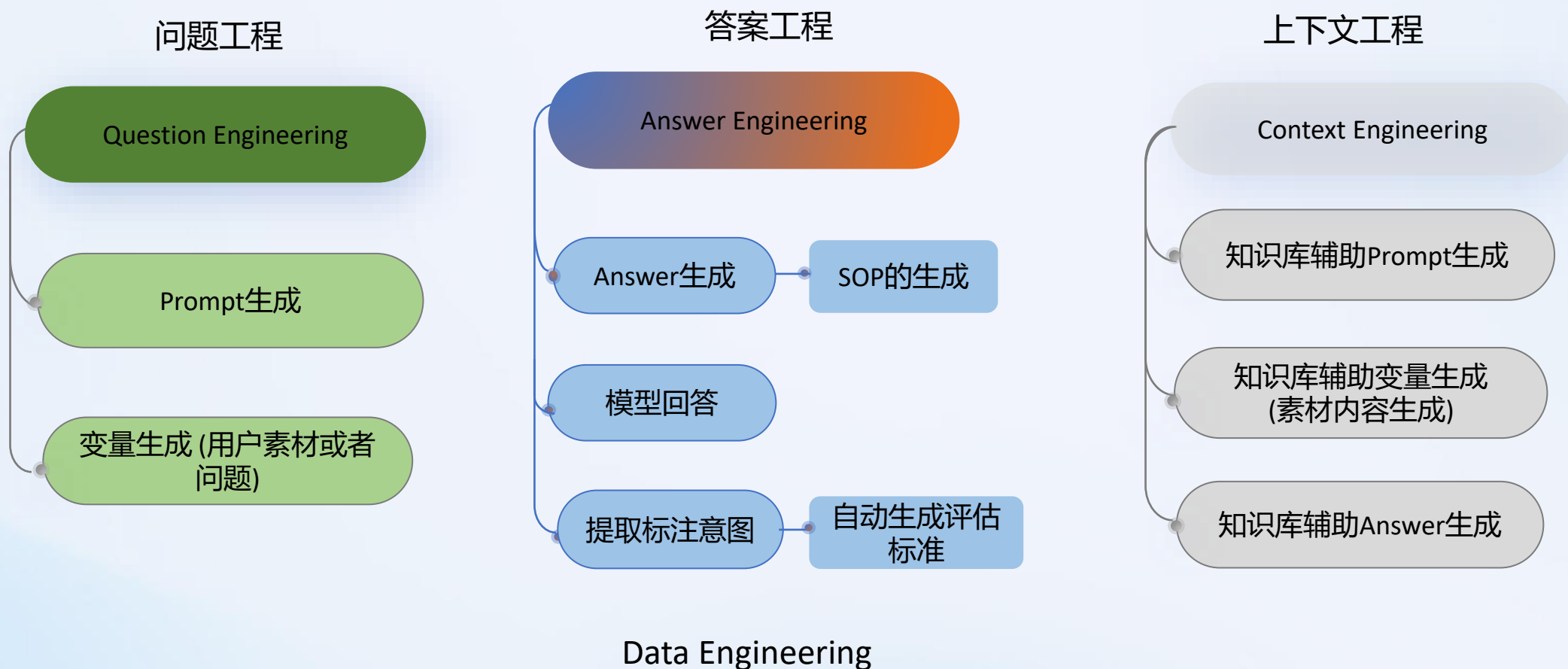
火山方舟 PromptPilot 的破局之道：痛点驱动的产品能力与实践路径



PromptPilot 详细揭秘：驱动 AI 进化的能力引擎

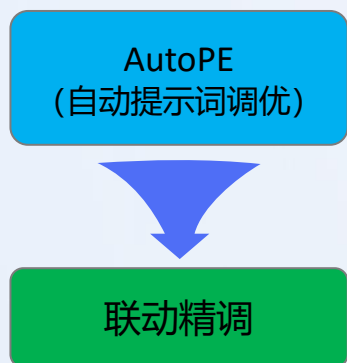


PromptPilot 详细揭秘：驱动 AI 进化的能力引擎



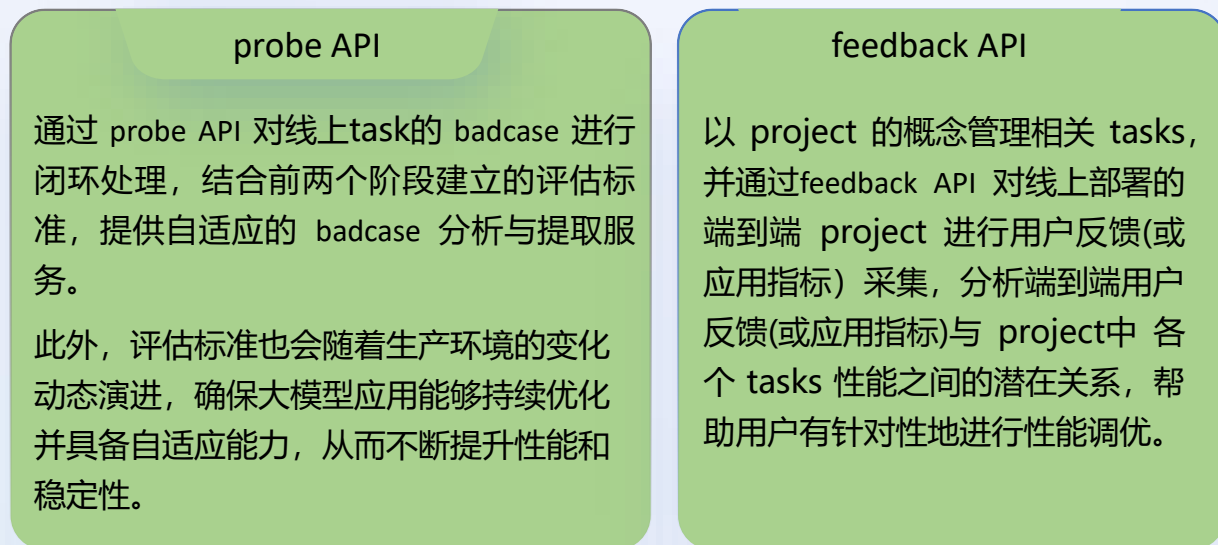
PromptPilot 详细揭秘：驱动 AI 进化的能力引擎

Prompt优化和联动精调



Prompt 臻化联动舱

闭环：Badcase 检测



智能闭环 · Badcase 洞察站

Solution



AI 方案生成器

数据智擎工坊：Question Engineering

Prompt 生成

问题：很多用户都有需求，但是如何把需求转为提示词呢？尤其是在2B场景

根据用户的任务描述生成基础，可用的 Prompt

欢迎使用 Prompt 优解

激发模型潜能，轻松优化 Prompt。讲解视频 | 使用手册

你的任务

生成 Prompt

根据提供的内容审核政策将聊天记录分类。

生成后的 Prompt

doubao-1.5-pro-32k-250115

你要根据提供的内容审核政策对聊天记录进行分类。请仔细阅读以下信息，并按照指示完成分类任务。

首先，请仔细阅读以下内容审核政策：

<内容审核政策>

{{CONTENT_REVIEW_POLICY}}

</内容审核政策>

现在，请仔细阅读以下聊天记录：

<聊天记录>

{{CHAT_RECORD}}

</聊天记录>

在进行分类时，请按照以下步骤操作：

1. 仔细阅读内容审核政策，明确各个分类的具体标准和特征。

2. 逐句分析聊天记录，判断每句话是否符合政策中的某个分类标准。

3. 对于聊天记录中的每一个部分，都要综合考虑其前后文，以准确判断其所属分类。

4. 如果聊天记录包含多个不同分类的内容，要分别指出每个部分对应的分类。

请先在<思考>标签中详细分析聊天记录与内容审核政策的匹配情况，说明判断的依据和过程。然后在<分类结果>标签中给出最终的结果。分类结果应明确指出聊天记录属于政策中的哪些分类。

<思考>

[在此详细说明你对聊天记录与政策匹配情况的分析]

</思考>

<分类结果>

[在此给出最终的结果]

</分类结果>

优化你的 Prompt

限时免费 共调用 3225 tokens

复制

继续调试 Prompt

Prompt 模板

模板新增

给我写一封电子邮件

起草回复客户投诉的电子邮件并提供解决方案。

舆情洞察

判断舆论的内容对出行行业的影响。

内容审核

根据提供的内容审核政策将聊天记录分类。

文档总结

从提供的文档中提取总结要点，要点数量不超过10个。

法律条文抽取

期货报告

数据智擎工坊：Question Engineering

全局反馈+局部反馈，助力用户明确需求

从生成一个 Prompt 开始

激发模型潜能，轻松优化 Prompt。讲解视频 | 使用手册

你的任务

判断舆论的内容对出行行业的影响。

Doubao-Seed-1.6

类型: 文本理解

联网

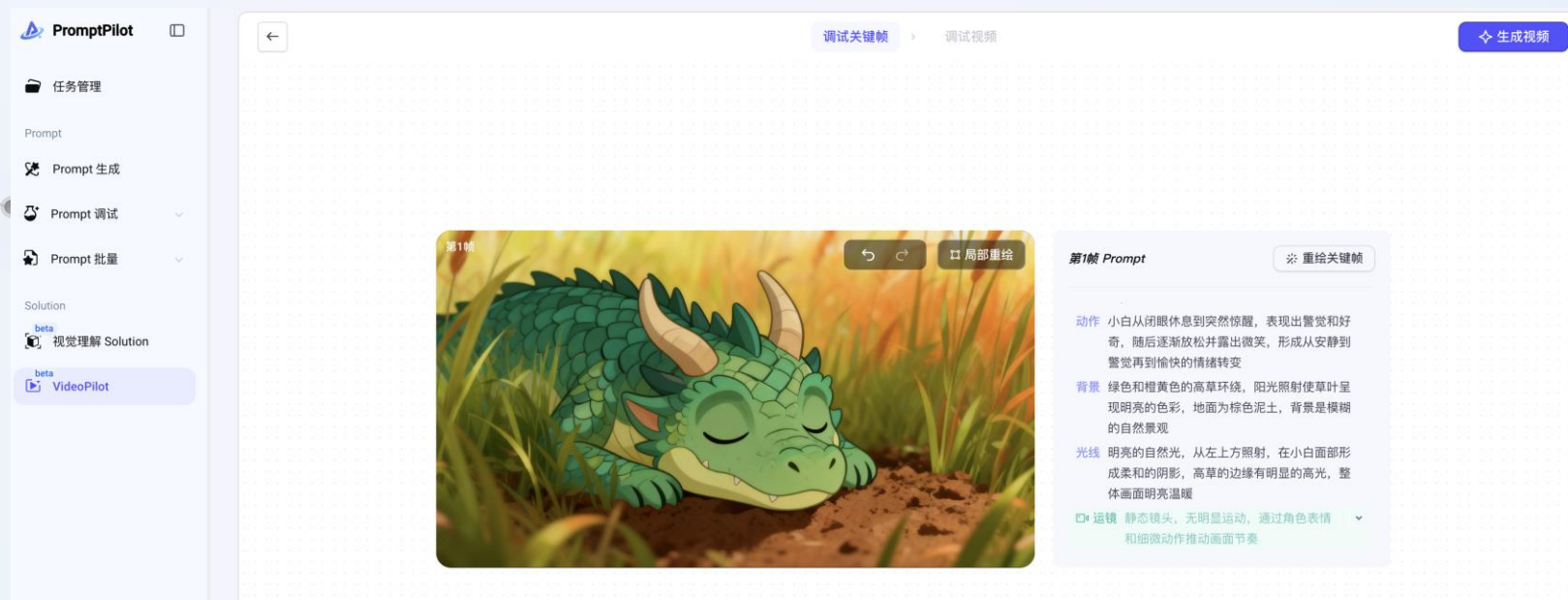
知识库

数据智擎工坊：Question Engineering

Prompt 生成：VideoPilot

根据视频，自动分镜，生成可用的 Prompt

问题：如何获取视觉任务的提示词？



数据智擎工坊：Question Engineering

变量生成 (用户素材或者问题)

- ▶ 模型会根据 Prompt 上下文信息和变量的含义，生成变量内容，帮助无数据的用户验证 Prompt；
- ▶ 变量：对应的就是客户的问题



使用方法

- 1 可以直接点击AI生成变量内容
- 2 或者直接粘贴实际的业务素材

数据智擎工坊：Question Engineering

全集反馈 + 局部反馈, 助力变量生成更准确

基于素材内容的局部反馈

增加变量的细节和丰富度

如对客户投诉内容, 增加处理时间和检测流程等



基于素材内容的全局反馈

增加素材的真实性

如针对媒体报道, 要求增加媒体名称和详尽内容等, 看起来更真实



数据智擎工坊：Question Engineering

AI 生成变量（增加 AI Search）

问题

在变量内容的生成过程中，有一些是对时效性要求比较高的场景，通过 AI Search 的方式，尽可能可以补齐时效性不足的问题，从而辅助生成更好的素材内容

解决方案

通过增加 AI Search 功能，来辅助生成变量



数据智擎工坊：Answer Engineering

参考答案的生成和校准

问题

用户清晰地表达了需求，但是对一个问题而言，有正的、负的答案，开放的问题有相对好的、坏的答案，很多时候，其实很难给出理想一点的参考答案

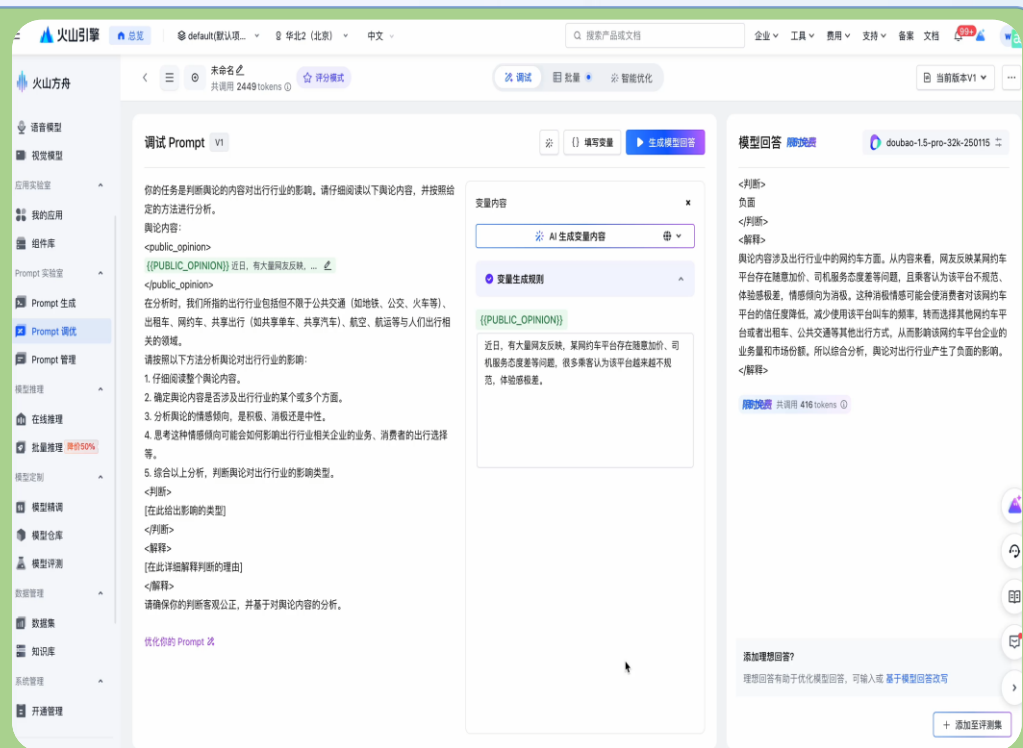


解决方案

通过 Answer Engineering 来构造答案空间。我们可以通过多模型来生成多样性的答案，客户可以通过对比，选择，反馈拿到理想回答

关键词

多模型辅助生成参考答案
全局反馈 局部反馈



数据智擎工坊：Answer Engineering

问题：有了数据集，那么应该如何评估效果呢？

模型回答：提取标注意图，自动生成评估标准

针对批量的 case 集合，可以直接使用已有的评估标准或者让 LLM 自动生成评估标准，并完成批量打分

The screenshot displays the Answer Engineering interface. At the top, there's a navigation bar with tabs for '评测数据' (Evaluation Data), '批量' (Batch), '智能优化' (Smart Optimization), and '报告' (Report). Below this, a table lists evaluation cases. The table has columns for 'TRAVEL_INDUSTRY_INFO', 'PUBLIC_OPINION_CONTENT', '理想回答' (Ideal Answer), and '模型回答' (Model Answer). A sidebar on the right contains a '智能评分' (Smart Scoring) section with a '生成全部回答' (Generate All Answers) button and a '评分' (Scoring) section with a '为未评分的回答评分' (Score Unscored Answers) button. A modal window is open, showing a '评分标准' (Scoring Standard) form with a 'AI 一键评分' (AI One-click Scoring) button.

TRAVEL_INDUSTRY_INFO	PUBLIC_OPINION_CONTENT	理想回答	模型回答
1	出行行业涵盖多种交通方式，如航空、铁路、公路运输。航空运输方面连接国内外众多地点，是长途出行的重要方式；铁路运输网络四通八达，为居民提供日常通勤和长途旅行服务；公路运输以汽车为工具，包括公交车、出租车、自驾车等形式。旅游服务包含旅行社提供的线路规划、导...	消极的	<思考> 舆论主题与航空公司相关。航空公司属于的航空运输部分，这的核心业务直接相关度来看，用户抱怨超量收费上不合理，态度。在传播范围在社交媒体上广泛传播...
2	出行行业近年来发展迅速，规模不断扩大。在交通方式层面，既有传统的公共交通如地铁、公交每天承载大量通勤和出行人员流量，也有新兴的网约车服务方便人们的个性化出行需求。旅游服务市场发展蓬勃，线上旅行社兴起，提供定制化旅行路线等项目。同时，出行行业中司乘关系...	消极的	<思考> 该舆论内容与出行行业直接相关，因为网约车这一出行行业的模式，舆论的态度是负面的。网约车司机与乘客发生冲突，这种行为了引发了大众对网约车司机职业素养和网约车公司管理的质疑。从传播范围和影响力来...
3	出行行业规模庞大，近年来随着人们生活水平的提高和旅游业的发展不断增长。在出行方式上，公共交通如地铁、公交不断优化...	消极的	<思考> 该舆论内容与出行行业的核心业务直接相关，因为事件涉及出租车司机与乘客之间的冲突，出租...

小步优化

- 通过三条反馈数据，即可开启 AI 自动生成评估标准
- 支持直接输入自定义评估标准
- 也支持AI自动生成评估

数据智擎工坊：Context Engineering

知识库辅助 Prompt 生成

问题

当客户所处场景具有

强领域专业性

数据敏感性

业务定制化

仅采用大模型自己的知识，容易造成和业务实际脱节。在这种情况下，必须结合自有知识库才能生成高质量 Prompt



解决方案

上线结合知识库的 Prompt 冷启动方案，通过KB来辅助生成 Prompt

典型场景

比如内容判敏，在什么规则下是敏感，什么规则下是不敏感，需要把客户自己的知识融合到 Prompt 里

数据智擎工坊: Context Engineering

知识库辅助变量生成

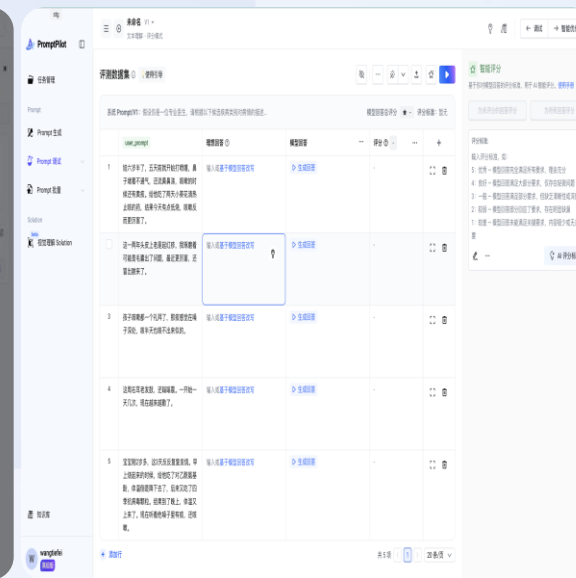
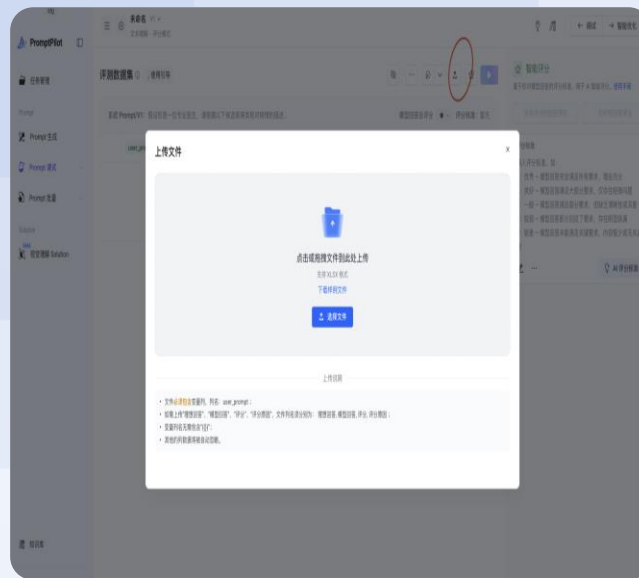
问题

如果用户已经有了一些素材的样例，需要生成和样例风格一致的case，怎么办？



解决方案

上线结合知识库的变量生成方案



数据智擎工坊: Context Engineering

知识库辅助 Answer 生成

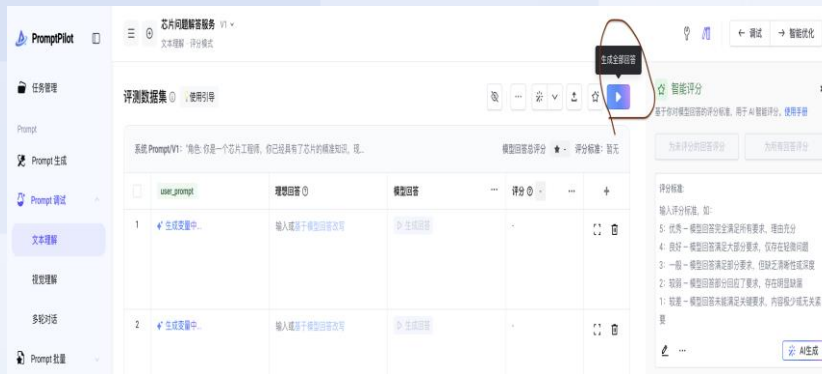
问题

目标场景的问题答案高度依赖特定知识库内容，无知识库支撑时无法形成有效回答



解决方案

上线结合知识库的答案生成方案



Prompt 臻化联动舱：Prompt优化 和 联动精调

Prompt 批量优化



Prompt 臻化联动舱：Prompt优化 和 联动精调

报告页：查看优化报告



Prompt 臻化联动舱：Prompt优化 和 联动精调

视觉理解优化场景



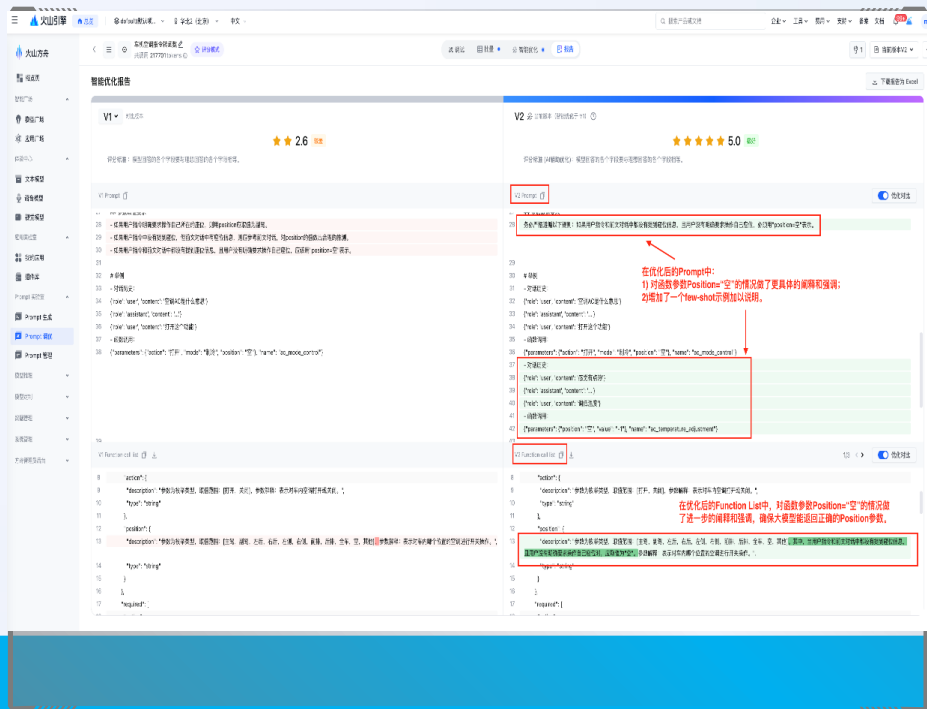
Prompt 臻化联动舱：Prompt优化 和 联动精调

多轮优化场景





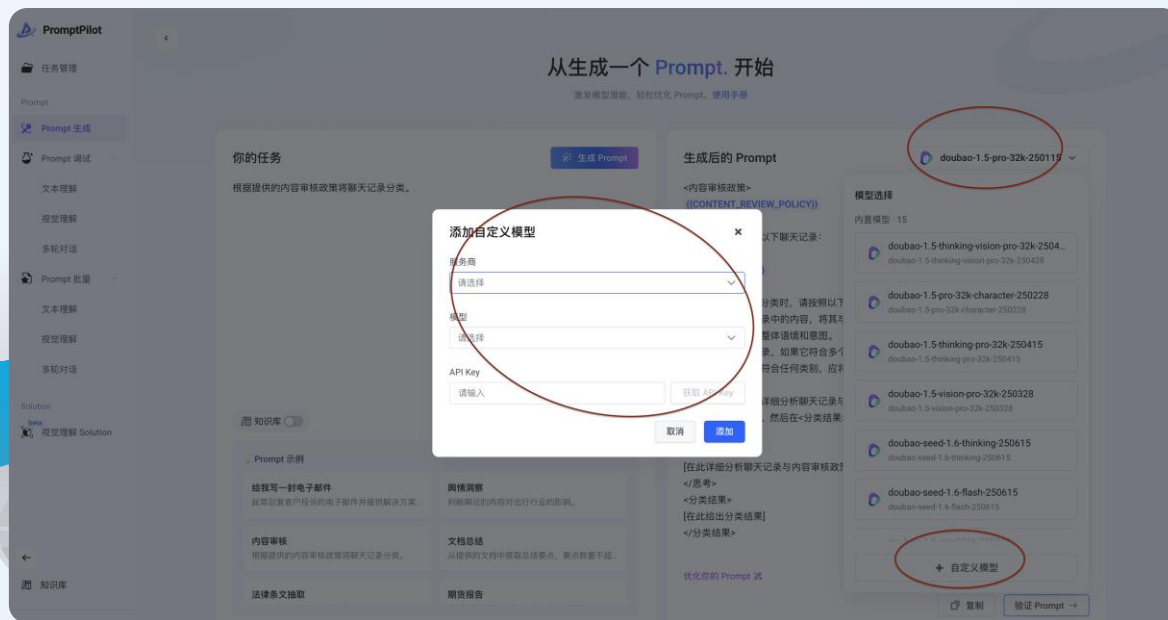
Function call 优化场景



Prompt 臻化联动舱：Prompt优化 和 联动精调

Prompt优化支持的模型

当前 PromptPilot 已支持添加任何
model (兼容 api_key 标准协议的)



Prompt 臻化联动舱：Prompt优化 和 联动精调

联动精调

在 Prompt 迭代数百轮，PromptPilot 可以自主判断是否要发起精调

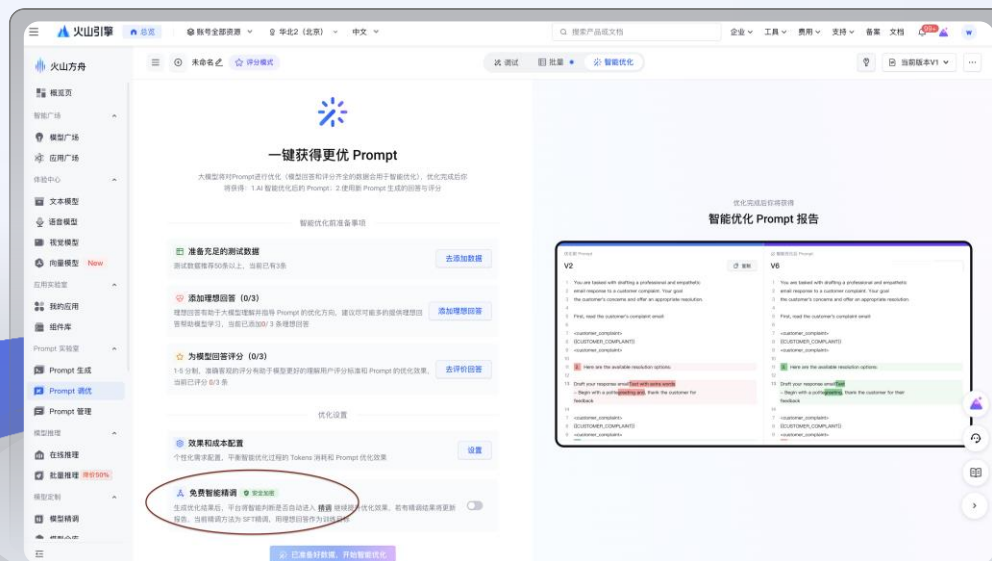
如需精调系统利用空余算力主动发起精调，自动返回精调前后对比并主动推进下一步部署



Prompt 臻化联动舱：Prompt优化 和 联动精调

联动精调

联动精调入口



智能闭环 · Badcase洞察站：Badcase 检测

SDK probe

通过SDK使用PromptPilot

为开发者提供了从测试到上线、从评估到调优的一站式闭环体验，加速 LLM 应用迭代，显著提升迭代速率，提升用户体验

SDK构建 线上Case雷达

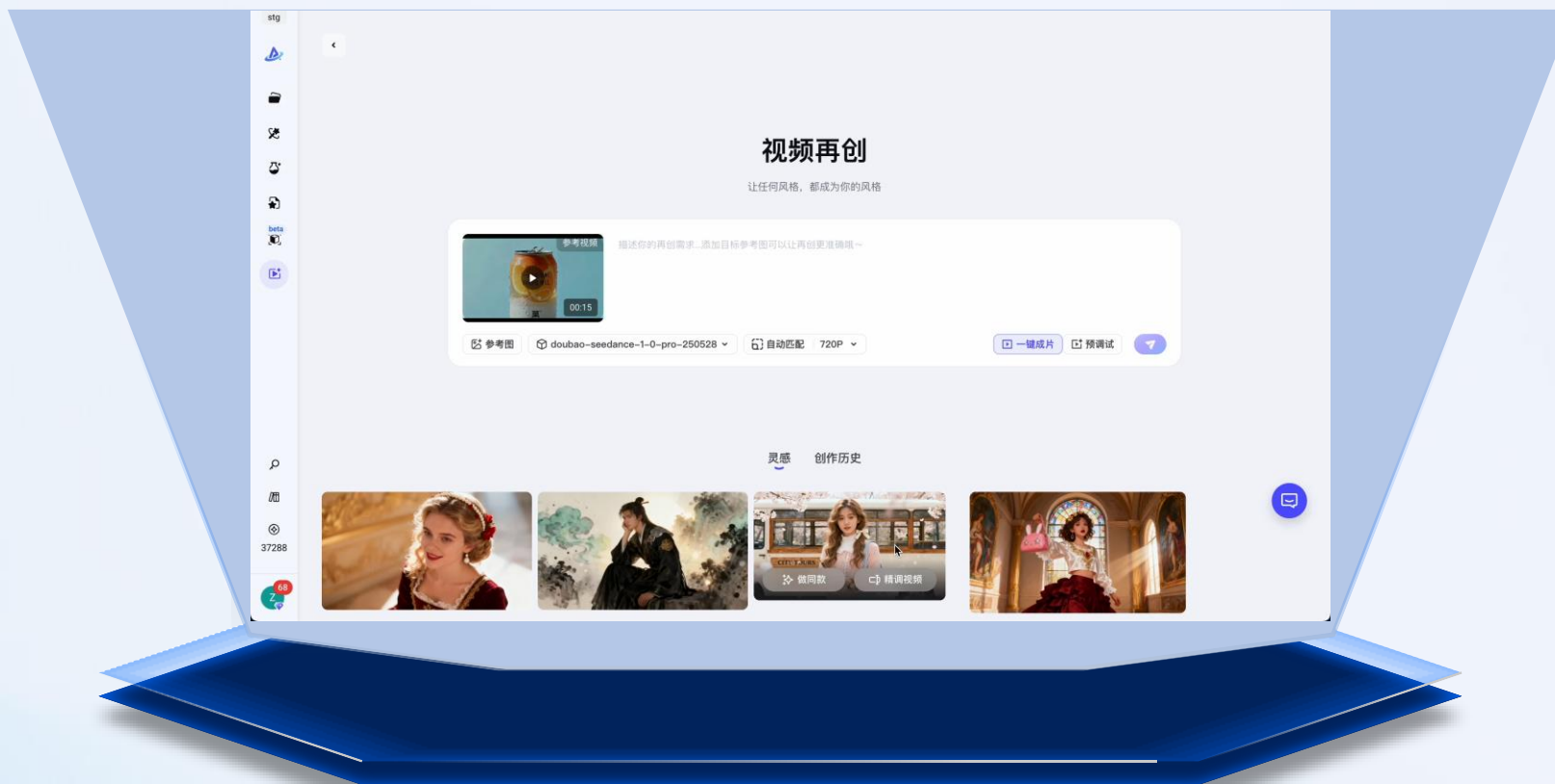
AI方案生成器：Solution

视觉理解 solution



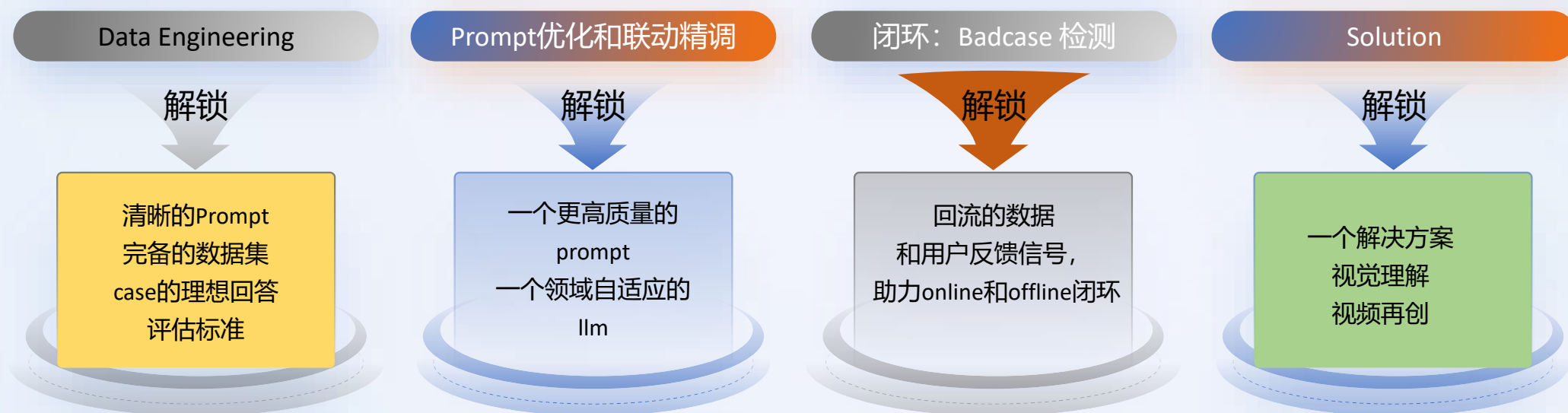
AI方案生成器：Solution

VideoPilot: 视频在创



PromptPilot 详细揭秘：驱动 AI 进化的能力引擎

PromptPilot：解锁你的专属 AI 能力宝藏



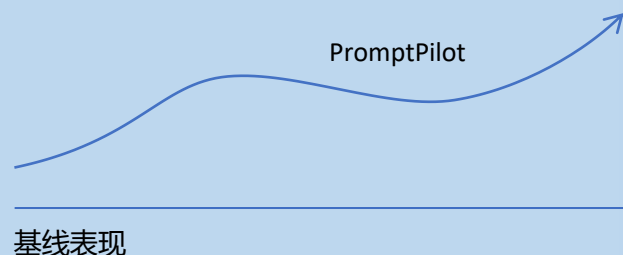
每个关键词，解锁一个痛点问题

PromptPilot让大模型应用落地更容易

释放模型潜能

提升模型在业务中落地效果的上限

模型上限



提升落地效率

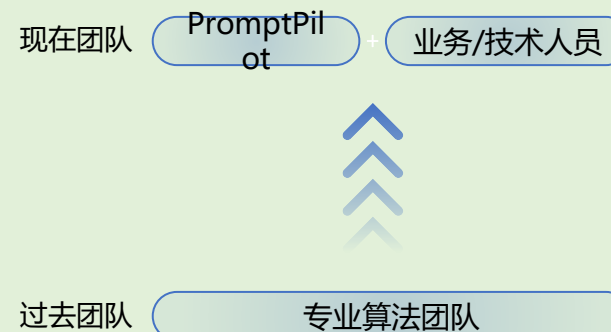
帮助企业/开发者高效优化Prompt
快速跑通大模型落地流程

业务优化时间从周级 → 天级



降低落地难度

帮助业务人员无门槛表达业务逻辑
更好的发挥业务价值



目录

01

大模型应用落地的挑战

02

火山方舟 PromptPilot 的破局之道：痛点驱动的产品能力与实践路径

03

沉淀与前行：从服务到产品的工程化启示与未来展望

沉淀与前行：从服务到产品的工程化启示

如何scaling?

最初仅靠算法同学服务客户，人力与效率受限，难以大规模施展算法能力。



1

能力规模化的必然选择

唯有通过产品化，将算法能力封装为易用、可复用的产品形态，才能突破单点服务的局限，实现算法价值的规模化释放，高效响应更多客户的需求。

如何设计产品?

客户需求呈现出多样性与场景化的特点，不同客户姿势各不相同，那么如何适配各种使用范式



2

以客户为中心的产品设计逻辑

产品设计时，秉持易用性原则，降低客户使用门槛。同时，要预留灵活的扩展与适配空间，以应对不同客户差异化的落地路径，确保产品能在多样场景中为客户创造价值。

如何确保大家能用起来?

结合前两步，如何让产品真正产生增效



3

协同打法的增效价值

结合 FDE 的打法，能够在从服务到产品的工程化进程中，实现各环节资源与能力的更优整合，深度打磨产品、进而获取更显著的收益

未来展望

易用性

围绕建立AI原生应用来构建整套解决方案

Post-train

算法迭代、数据闭环、助力大模型潜能持续升级

PromptPilot

深度合作

共创，深度优化产品

持续优化，扩展边界

解锁更多的场景和应用

效果

效率

文档和手册

PromptPilot 独立站入口

<https://promptpilot.volcengine.com>

PromptPilot 用户手册

<https://www.volcengine.com/docs/82379/1399495>

方舟帮助文档

<https://bytedance.larkoffice.com/docx/R2uPdJmmuoAwq7x3T8Sck1FYnmG>

PromptPilot SDK 文档

<https://www.volcengine.com/docs/82379/1587837>

PromptPilot 白皮书

<https://bytedance.larkoffice.com/wiki/VtEgwevT6idFcFkeIWjcmtSBnHt>

PromptPilot 面向所有用户全模型兼容全面开放!

PromptPilot 交流群



交流群为大家提供交流平台
请通过微信或企业微信扫码加入

PromptPilot 白皮书



扫码获取电子白皮书

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

