基于 Vertex AI 的智能产品推荐 Agent 技术白皮书

一、项目背景与目标

在数字化转型不断加速的当下，用户希望通过更自然、便捷的方式获取产品信息。本项目旨在构建一个基于 Google Cloud Vertex AI 的智能对话 Agent，帮助用户通过网站对话框实现高效的产品匹配与推荐。

二、技术架构概述

1. 架构总览

平台依托：Google Cloud Platform（GCP）

核心组件：

- Vertex AI Agent Builder
- Gemini 模型（Gemini 1.5/2.0/2.5均可）
- 向量检索知识库（基于产品知识构建）
- Web 嵌入式对话前端

2. 架构图（示意）

[用户浏览器] → [前端对话框] → [Agent Web Endpoint]

↓

Vertex AI Agent Builder（Gemini模型）

↓

自定义知识库（产品内容 Embedding + 索引）

三、核心功能说明

1. 知识库构建

数据来源：结构化/半结构化产品信息

处理流程：

- 数据清洗与标准化（如 Markdown 或 FAQ 格式）
- 自动拆分文档
- 使用 Vertex AI 内嵌的工具进行嵌入生成与索引构建

能力范围：仅基于知识库内容进行回答

2. Agent 能力设置

模型类型：Gemini 1.5/2.0/2.5均可

Agent 配置：

- 限定回答范围为产品知识库，不回答和产品无关的问题，这样可以避免输出有害信息。
- 启用 Grounding 机制（答案引用知识来源）
- 可以根据需要设置引导语

3. 对话式推荐体验

用户可自然发问，如：

- “有哪些产品可以选择”
- “我是餐厅老板，有合适的机器人推荐吗”

Agent 自动检索匹配内容并返回结果。

4. 过滤可能的有害信息

通过四层机制来避免agent输出有害信息。

阶段	防御目标	实施方式
知识库构建前	**源数据清洗**，剔除敏感内容	NLP过滤、人工复审、关键词排查
知识库构建时	向量生成前清洗，每段可打标签	文本分段级检测 + 标签辅助筛选
Agent设置	**避免输出任意内容**	设置为只针对知识库内容进行回答
LLM响应阶段	**过滤掉支持库中遗漏的有害信息**	通过提示词让大模型避免输出某些类型的信息

四、集成方式与部署

1. 前端集成

使用 <iframe> 或 JavaScript SDK 嵌入方式，支持移动端响应式体验

2. 安全性与权限控制

Agent 设置为私有部署，仅允许特定域名调用，所有请求通过 HTTPS，确保数据安全

五、技术优势

维度	描述
快速上线	基于 Vertex AI 平台，低代码创建 Agent
智能理解	使用 Gemini 大模型，理解复杂自然语言意图
精准推荐	仅基于内部知识库回答，避免幻觉现象
可溯源	回答引用知识条目，增强可信度

六、未来演进方向

- 支持多语言产品推荐（通过 Gemini 多语能力将产品介绍自动翻译为用户使用的语言）
- 接入实时库存或价格 API，实现“问答+交易”闭环
- 自动补充知识库能力（新增产品自动同步入库）

七、总结

本 Agent 实现了网站自然语言问答式的产品推荐体验，借助 GCP Vertex AI 平台与 Gemini 模型，完成了从结构化产品数据到智能推荐助手的闭环构建，极大提升用户获取信息的效率与准确度。