模型推理与评测
从脚本执行到任务平台

基于 Argo Workflows，把模型、数据、执行过程与结果组织成用户可以操作的工作流。

USER EXPERIENCE

任务入口

React 表单与任务列表

选择模型、数据和执行参数

ORCHESTRATION

执行编排

FastAPI 对接 Argo

Kubernetes 运行 CPU/GPU Pod

TRACEABILITY

输入与结果

业务记录、共享存储与 S3

结果、报告和执行上下文

> 参与重点：早期工作流、平台入口、业务关联与部署集成。

参与过的工作

平台让用户管理一次完整的模型任务

从选择输入到查看报告，每一步有明确入口，执行对象与业务对象可以关联。

01 / INPUT

数据集与模型

管理名称、路径和版本信息

选择已有记录或输入路径

02 / SUBMIT

推理任务提交

填写代码来源与运行参数

通过 API 提交 Argo 工作流

03 / TRACK

状态与执行入口

任务列表、筛选与用户关联

连接 Argo UI 查看执行细节

04 / RESULTS

结果管理

按任务组织输出位置

展示 S3 结果访问入口

05 / REPORTS

报告管理

组织报告与工作流关系

上传、查看与比较上下文

06 / INTEGRATE

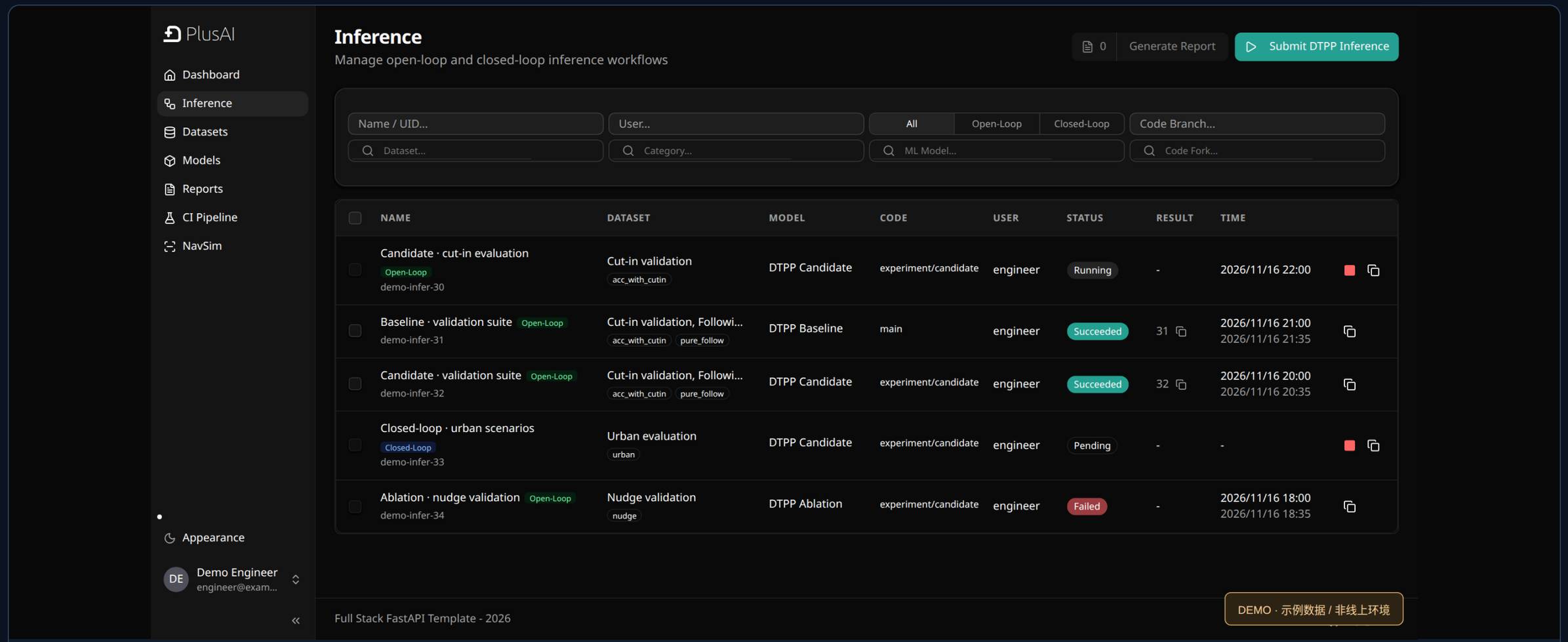
基础设施接入

Kubernetes、认证与挂载

数据库和对象存储配置

推理列表：输入、状态与结果放在一起

从模型与数据选择进入任务执行，再通过列表查看状态、代码来源和结果入口。



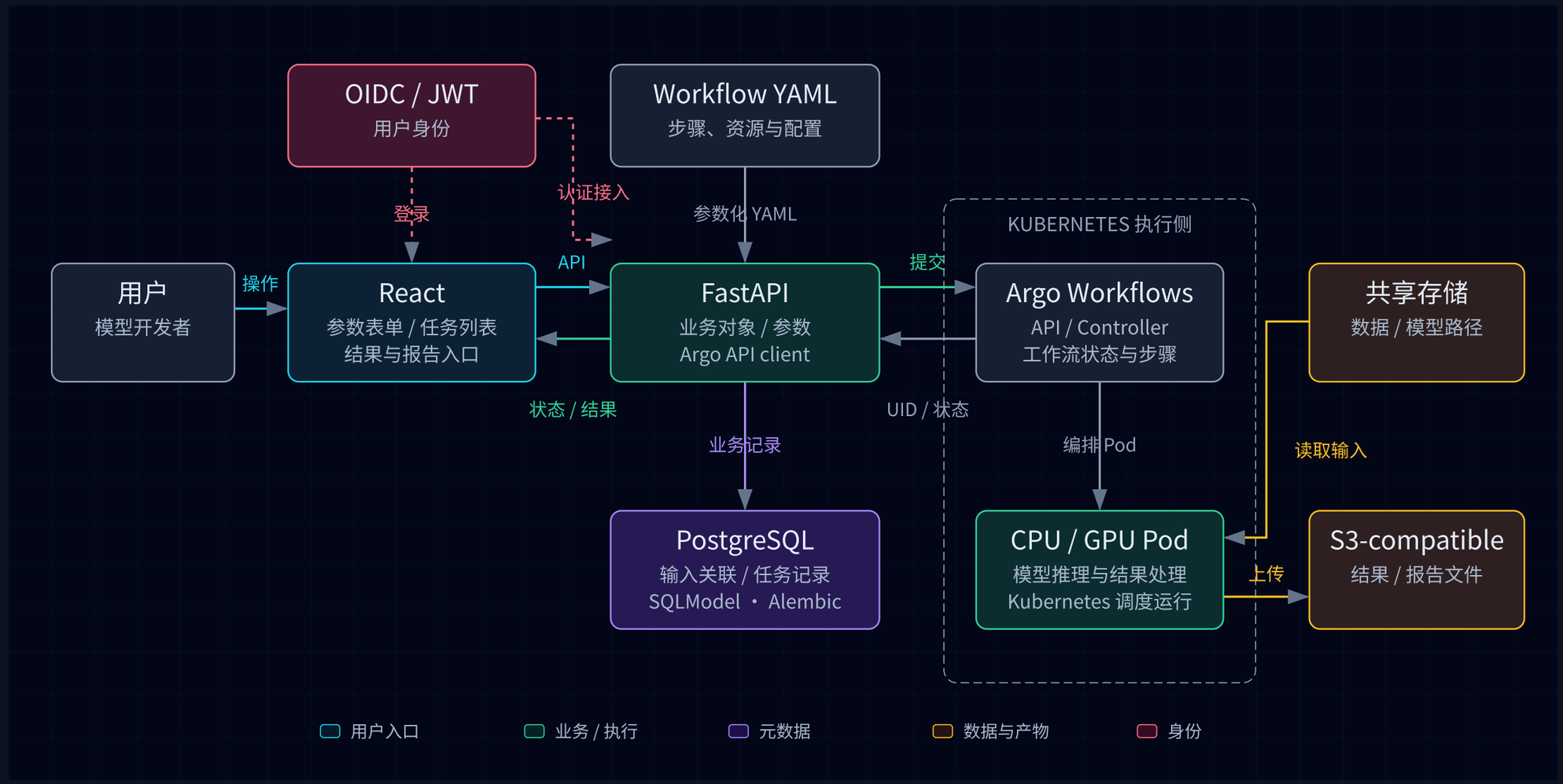
真实前端 + DEMO 示例数据；当前界面包含团队后续改动。点击图片查看原图。

本地前端截图 · 构造数据，不代表实际任务结果

参与过的工作

平台管业务上下文，Argo 管执行步骤

Kubernetes 提供运行资源，共享存储与对象存储承载输入和结果。



参与过的工作

将长时间推理拆成可观察的执行步骤

早期工作流围绕准备、推理和上传拆分，使失败能够定位到相应阶段。

01 / PREPARE

准备执行环境

代码与依赖、运行参数

数据与模型路径

容器和挂载配置

02 / INFER

执行模型推理

调用模型仓库中的推理逻辑

按配置使用 CPU/GPU

输出任务结果

03 / DELIVER

结果上传与返回

上传结果到 S3-compatible

返回结果位置

供平台查询与报告使用

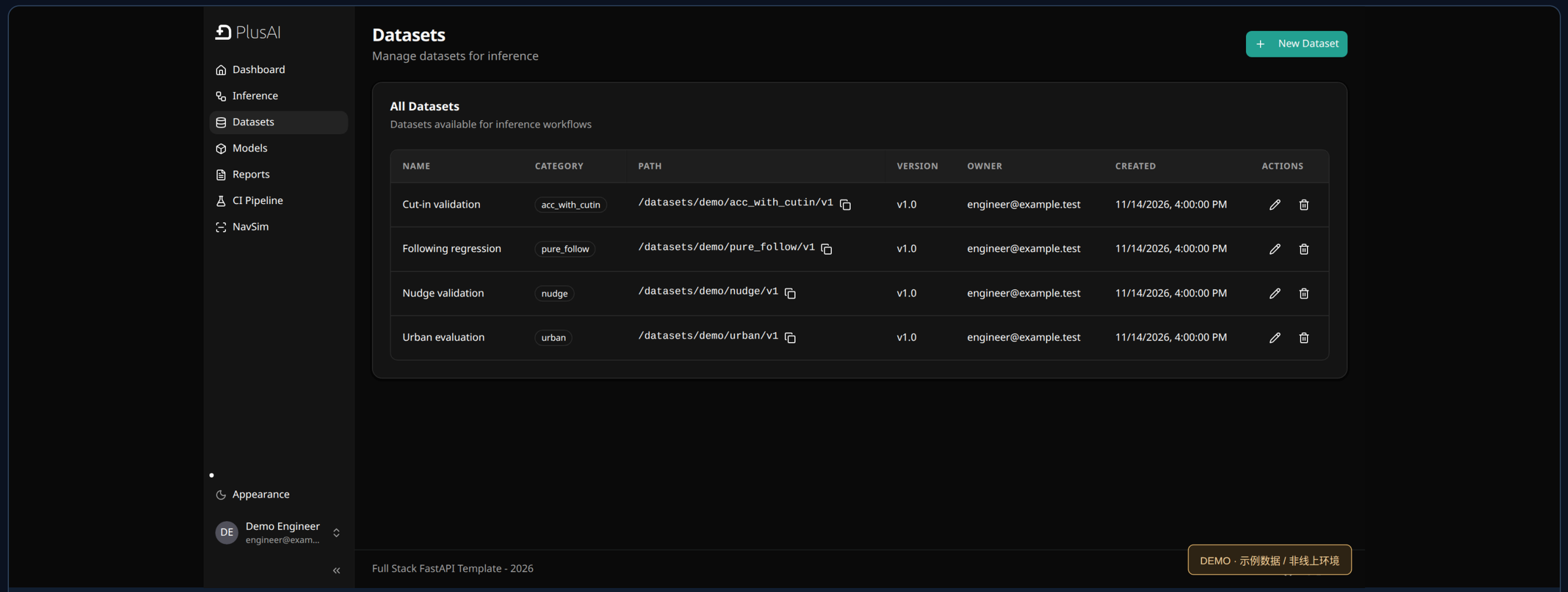
配置内容	执行侧职责	业务平台职责
参数 / 依赖	Argo 按 Workflow 定义执行	根据表单和业务记录组织请求
容器 / 资源	Kubernetes 分配并运行 Pod	传递任务需要的执行参数
状态 / 输出	提供步骤状态和输出参数	关联用户、输入和结果入口

参与过的工作

数据集与模型：为任务提供可查询的输入

管理名称、路径、版本与用户，并把输入记录关联到每次工作流运行。

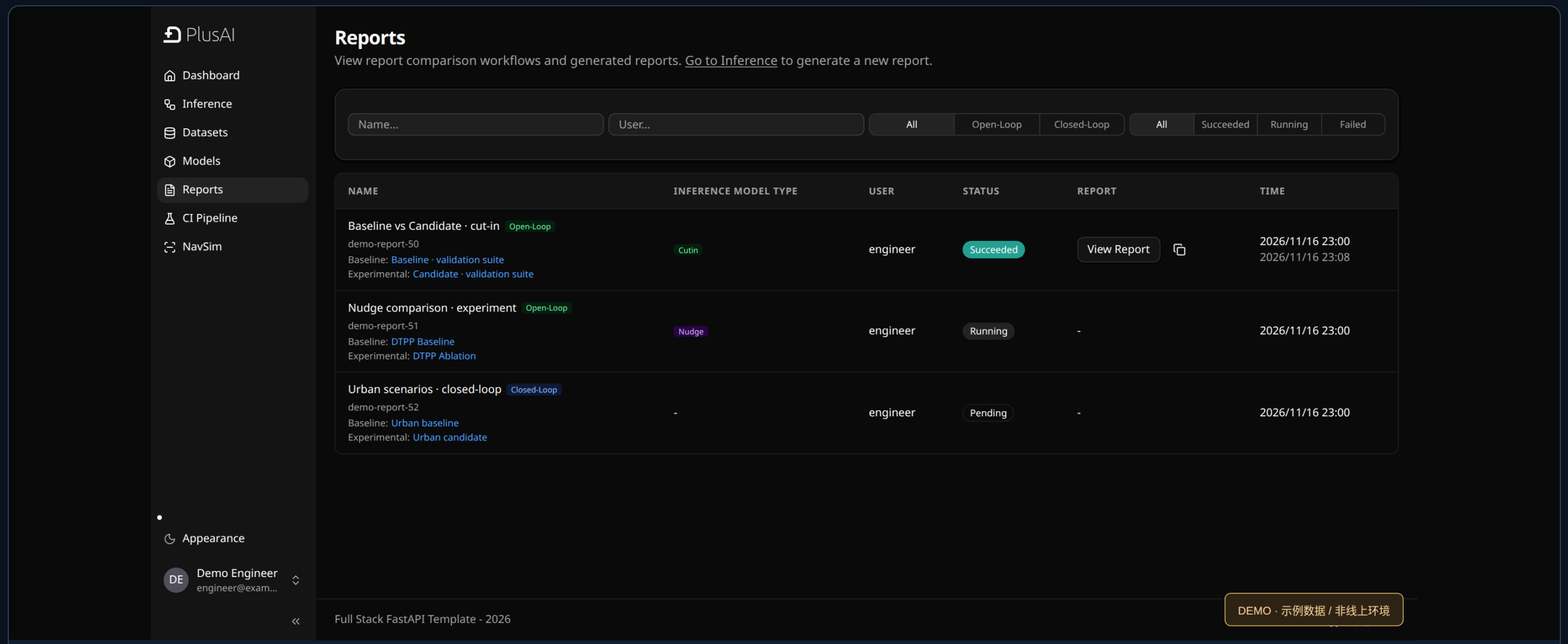
Datasets



真实前端 + DEMO 示例数据；当前界面包含团队后续改动。点击图片查看原图。

报告入口：让比较结果有运行上下文

查看报告对应的基线与实验运行、状态和结果链接；文件由对象存储承载。



真实前端 + DEMO 示例数据；当前界面包含团队后续改动。点击图片查看原图。

应用要同时接通计算、存储与身份

部署不仅是启动前后端，还要确保任务环境能够访问正确的输入与输出。

COMPUTE

容器与 Argo

前后端镜像、部署清单

Argo 服务地址与模板目录

执行参数与资源配置

STORAGE

路径与对象存储

后端挂载支持路径检查

任务容器的数据 / 模型访问

S3 结果与报告上传

IDENTITY & DATA

用户与业务记录

OIDC 接入调整

PostgreSQL / SQLModel

Alembic 管理结构变更

> 关键检查：API 服务看得到路径，执行节点也必须具备相应挂载与权限。

工作流平台最值得追问的三个边界

从真实执行链路出发，说明哪些状态需要对账，哪些条件仍需要完善。

CONSISTENCY

提交和入库不是同一事务

Argo 成功、DB 失败会失联

设计讨论：稳定任务 ID

幂等提交与按 UID 对账

REPRODUCIBILITY

记录分支不等于冻结输入

执行时分支可能变化

设计讨论：执行清单

固定代码 / 镜像 / 数据版本

LIFECYCLE

停止不等于资源都释放

外部资源可能继续运行

设计讨论：按任务管理清理

保留证据并验证释放

本页为工程分析与改进讨论，不宣称这些完善机制已经全部实现。

参与过的工作

参与重点：连接业务任务与执行基础设施

形成面向模型开发者的任务入口，覆盖编排、业务关联与运行环境集成。

参与内容	工程能力	对应岗位方向
早期推理工作流与 Argo API	参数化执行、状态与结果入口	Argo Workflows / GPU workload
Web 提交、数据与模型管理	把运行过程组织成业务产品	数据平台 / ML 平台支持
工作流、输入与报告关联	任务上下文与产物可追踪	流程规范化与可观测性
Kubernetes、认证与存储集成	连接容器、身份、共享存储和 S3	平台部署与基础设施集成

> 主讲到此：任务可提交、过程可观察、结果可关联。

实际规模、运行耗时与效率改善需要真实统计；本材料未用假定数字代替结果。

团队后续演进

同一平台继续扩展到 CI、定时任务与数据处理

这些能力说明平台的演进方向；本人主讲仍以早期推理与管理链路为中心。

后续能力	当前系统中的作用	贡献边界
开环 / 闭环与报告扩展	扩展执行类型、结果与报告流程	包含协作者的类型拆分与模板迭代
PR Model CI	训练、评估、benchmark、报告 DAG	本人初始接口用于联调；完整链路由后续扩展
CronWorkflow	nightly / daily / weekly 定时任务	属于后续团队能力
NavSim / RayJob	Argo 创建并观察 RayJob，轮询与清理	主要实现由其他协作者完成

团队演进背景，不作为本人独立交付成果；具体实现与阶段见项目文档。