

VertextAi-Hotel assitant funetuing Job

项目基础信息：

PROJECT_ID:

cy-aispeci-demo

BUCKET_URI:

gs://peft-model-cy-aispeci-demo/hotel_train_data.jsonl

gs://peft-model-cy-aispeci-demo/hotel_validation_data.jsonl

项目git地址：

<https://github.com/ChuanYang-AI/Demo3>

白皮书：

https://github.com/ChuanYang-AI/Demo3/blob/main/certification_materials/README.md

酒店AI微调系统架构与流程图：

https://github.com/ChuanYang-AI/Demo3/blob/main/certification_materials/docs/system_architecture_diagrams.md

模型微调记录：

notebook:

代码块

```
1 project_id_output = !gcloud config list --format 'value(core.project)'  
2 >/dev/null  
2 PROJECT_ID = project_id_output[0]  
3 REGION = !gcloud compute project-info describe --format="value[]  
(commonInstanceMetadata.items.google-compute-default-region)"  
4 LOCATION = "asia-east2"  
5  
6 BUCKET_NAME = f"{PROJECT_ID}-model-dataset"  
7 # BUCKET_URI = f"gs://{BUCKET_NAME}"
```

```
8 BUCKET_URI = "gs://peft-model-cy-aispeci-demo"
9
10 import vertexai
11 from vertexai.generative_models import (
12     GenerativeModel,
13     Part,
14     HarmCategory,
15     HarmBlockThreshold,
16     GenerationConfig,
17 )
18 from vertexai.preview.tuning import sft
19 from typing import Union
20 import pandas as pd
21 from google.cloud import bigquery
22 from sklearn.model_selection import train_test_split
23 import datetime
24 import time
25 local_train_file = "hotel_train_data.jsonl"
26 local_validation_file = "hotel_validation_data.jsonl"
27 vertexai.init(project=PROJECT_ID, location=LOCATION)
28 print(f"✓ Vertex AI 初始化完成 - 项目: {PROJECT_ID}, 区域: {LOCATION}")
29
30 train_dataset_uri = f"gs://peft-model-cy-aispeci-
31 demo/hotel_train_data.jsonl"
32 validation_dataset_uri = f"gs://peft-model-cy-aispeci-
33 demo/hotel_validation_data.jsonl"
34
35 sft_tuning_job = sft.train(
36     source_model="gemini-1.5-pro-002",
37     train_dataset=train_dataset_uri,
38     validation_dataset=validation_dataset_uri,
39     epochs=3,
40     learning_rate_multiplier=1.0,
41     tuned_model_display_name="Hotel assiant-Gemini Tuning
42 Job"
43 )
44 sft_tuning_job_name = sft_tuning_job.resource_name
45 sft_tuning_job_name
46 while not sft_tuning_job.refresh().hasEnded:
47     time.sleep(60)
48 sft_tuning_job.list()
49 tuned_model_endpoint = sft_tuning_job.tuned_model_endpoint_name
50 print(tuned_model_endpoint)
51 tuned_model = GenerativeModel(tuned_model_endpoint)
```

超参优化记录：

第一轮训练（基于gemini1.5pro）：

超参

```
sft_tuning_job = sft.train(  
    source_model="gemini-1.5-pro-002",  
    train_dataset=train_dataset_uri,  
    validation_dataset=validation_dataset_uri,  
    epochs=3,  
    learning_rate_multiplier=1.0,  
    tuned_model_display_name="Hotel assiant-Gemini Tuning Job"  
)
```

Dataset details:

Monitor	Dataset	Details
Model name	projects/704352985590/locations/asia-east2/models/299458924479905792@1	
Tuning job	projects/704352985590/locations/asia-east2/tuningJobs/3629547256916475904	
Status	Succeeded	
Region	asia-east2	
Created	Jun 20, 2025, 1:28:49 PM	
Ended	Jun 20, 2025, 2:06:09 PM	
Tuning method	Supervised	
Base model	gemini-1.5-pro-002	
Tuning dataset	gs://peft-model-cy-aispeci-demo/hotel_train_data.jsonl	
Experiment	tuning-experiment-20250619222930217475	
Number of epochs	3	
Intermediate checkpoints	Disabled	
Learning rate multiplier	1	
Adapter size	4	
Truncated example count	0	
Encryption type	Google-managed	
Validation dataset	gs://peft-model-cy-aispeci-demo/hotel_validation_data.jsonl	

← Tuning details: Hotel assiant-Gemini Tuning Job Test View model details Evaluate Docs

Monitor Dataset Details

Messages 1080 Tokens ⑦ 435108 Total Total

Dataset sample

Displaying the first 3 rows of the tuning dataset (samples may be truncated for preview)

Message ⑦

- 如何利用数字技术增强酒店客房的市场推广效果？
- 利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些策略和方法：
 - 1. **虚拟现实 (VR) 和增强现实 (AR)**：- 使用VR技术提供虚拟参观，让潜在客户在预订前体验客房和酒店设施。- 利用AR应用程序，客户可以通过手机查看房间的3D模型或获取更多信息。
 - 2. **个性化营销**：- 使用大数据分析客户的偏好和行为，提供个性化的推荐和优惠。- 通过电子邮件或社交媒体推送个性化的内容和促销活动。
 - 3. **社交媒体和影响者营销**：- 在Instagram、Facebook等平台上分享高质量的客房照片和视频。- 与旅游博主和影响者合作，通过他们的渠道推广酒店。
 - 4. **搜索引擎优化 (SEO) 和内容营销**：- 优化酒店网站的SEO，提高在搜索引擎中的排名。- 创建有价值的内容，如博客文章、视频和...

酒店服务人员在与客人沟通时应避免哪些不当行为？

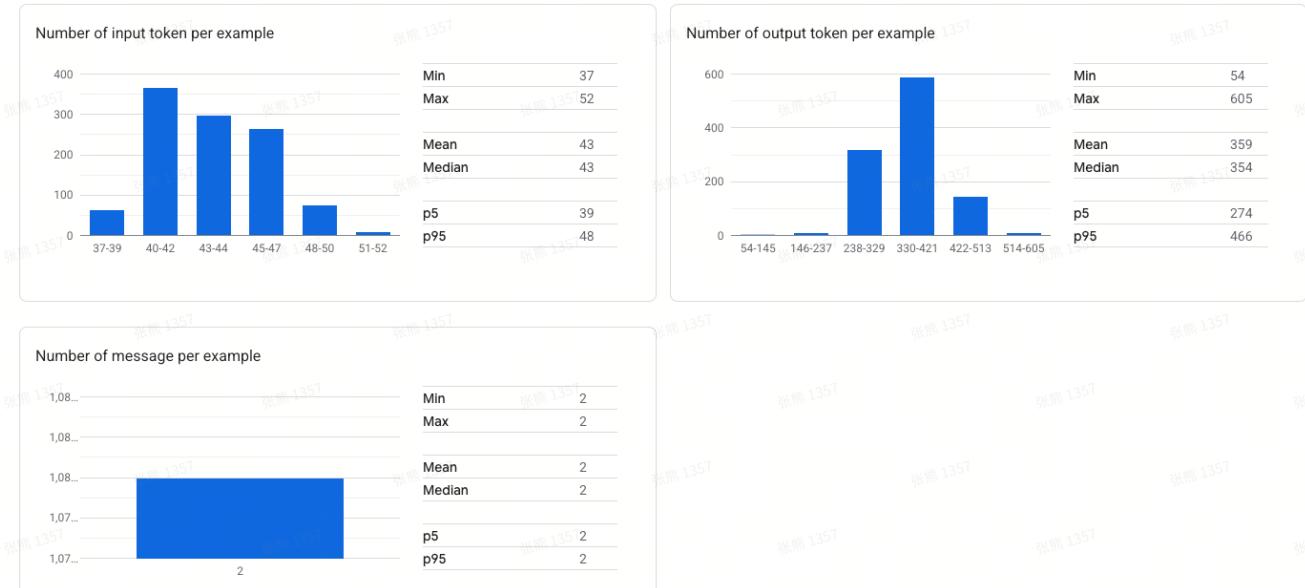
- 酒店服务人员在与客人沟通时应避免以下不当行为：
 - 1. **态度冷漠**：表现出不耐烦或漠不关心会让客人感到不受重视。
 - 2. **使用不礼貌的语言**：避免使用粗鲁或不专业的语言，保持礼貌和尊重。
 - 3. **打断客人**：在客人表达需求或意见时，不要打断他们，耐心倾听。
 - 4. **争论或辩解**：即使客人有误解，也应避免与客人争论，尽量以解决问题为导向。
 - 5. **忽视客人的请求**：对客人的合理请求应及时回应，不应拖延或忽视。
 - 6. **过度承诺**：不要向客人承诺无法实现的服务或解决方案，以免引发不满。
 - 7. **泄露隐私**：严格保密客人的个人信息和隐私，不在公开场合讨论。
 - 8. **缺乏专业知识**：对酒店的服务和设施不熟悉会影响服务质量，应确保了解相关信息。
 - 9. **情绪化反应**：保持冷静和专业，不因个人情绪影响与客人的沟通。
 - 10. **忽视文化差异...

酒店如何管理长期住客的费用减免？

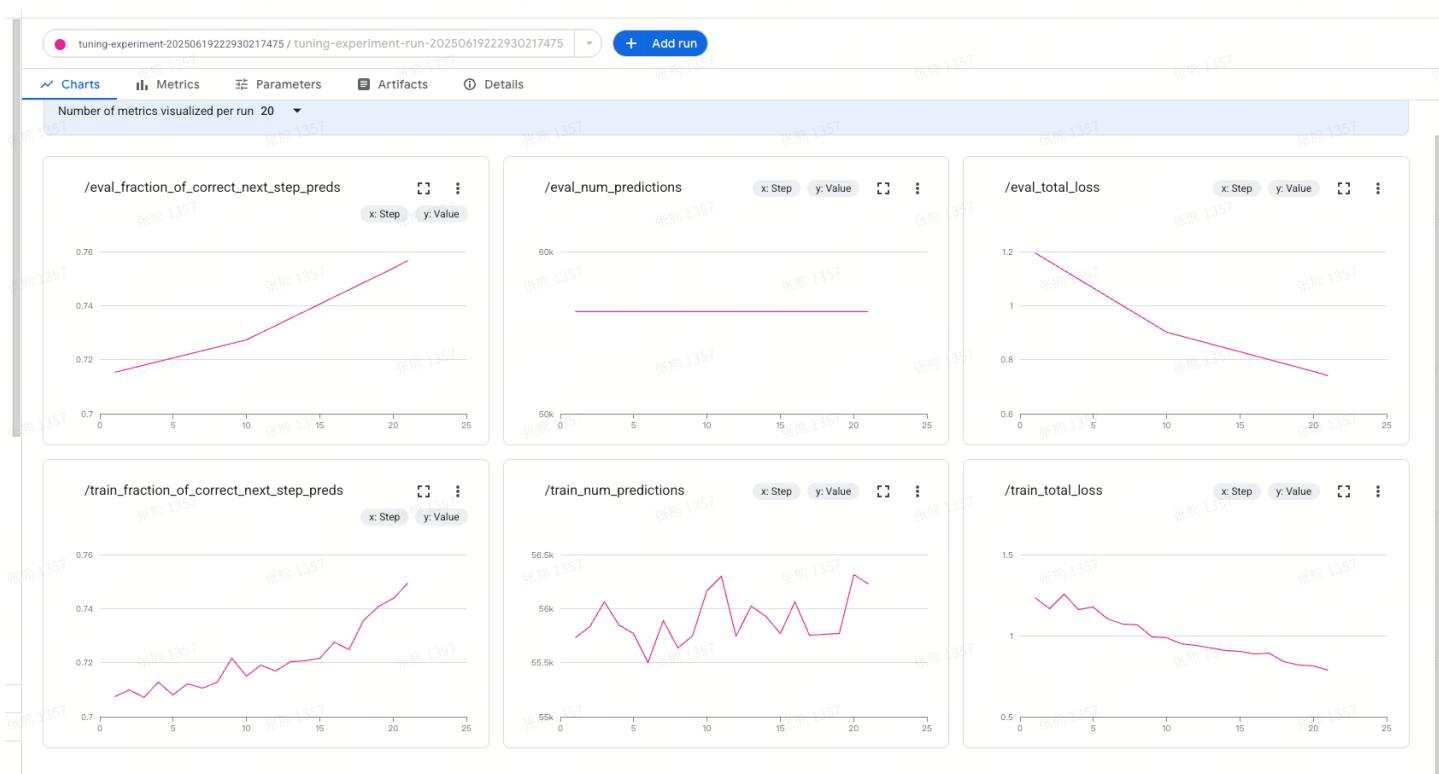
- 酒店管理长期住客的费用减免通常涉及多种策略和考虑，以确保既能吸引住客，又能保持盈利。以下是一些常见的方法：
 - 1. **长期住宿折扣**：酒店通常会为住客提供长期住宿折扣。这可以是按周或按月计算的折扣，通常比日常房价更优惠。
 - 2. **套餐优惠**：提供包括餐饮、洗衣服务或其他设施使用的套餐，以增加住客的整体体验，同时也增加酒店的收入。
 - 3. **会员计划**：通过会员计划为长期住客提供额外的优惠和积分奖励，鼓励他们选择长期入住。
 - 4. **灵活的付款方式**：为长期住客提供个性化的服务和设施，如房间升级、免费早餐或延迟退房等，以增加他们的满意度。
 - 6. **合同协议**：与公司或机构签订长期住宿合同，为其员工提供住宿，这通常会涉及到费用减免和其他优惠条件。

7. **定期...

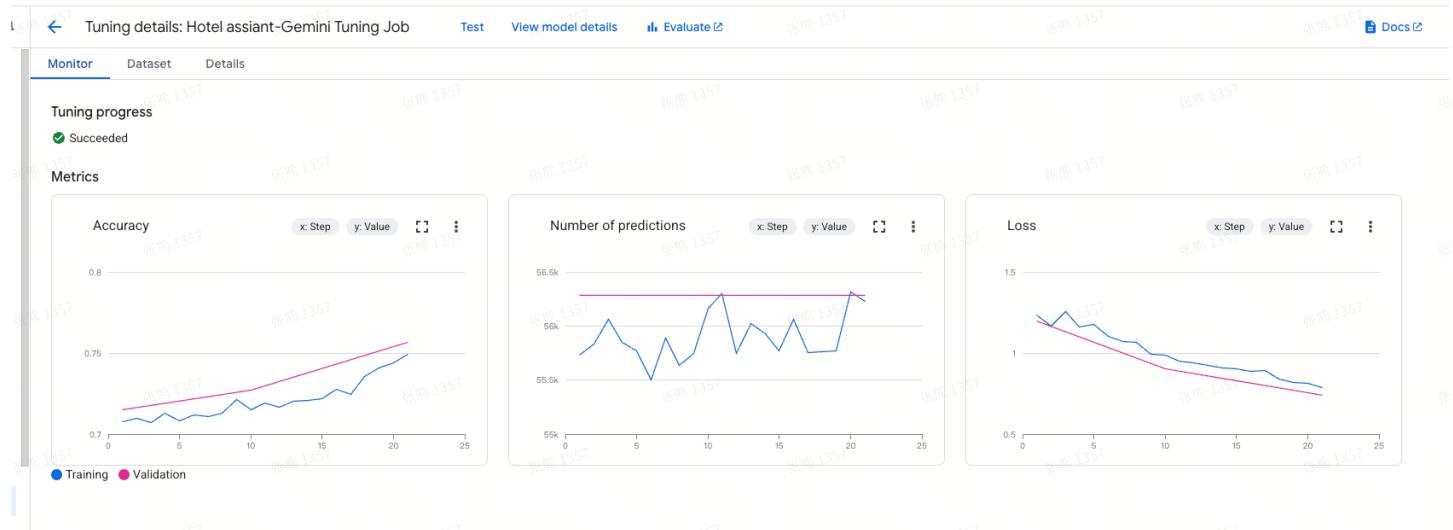
Data distribution



Run details:



Tuning details



第一轮训练结果评估：

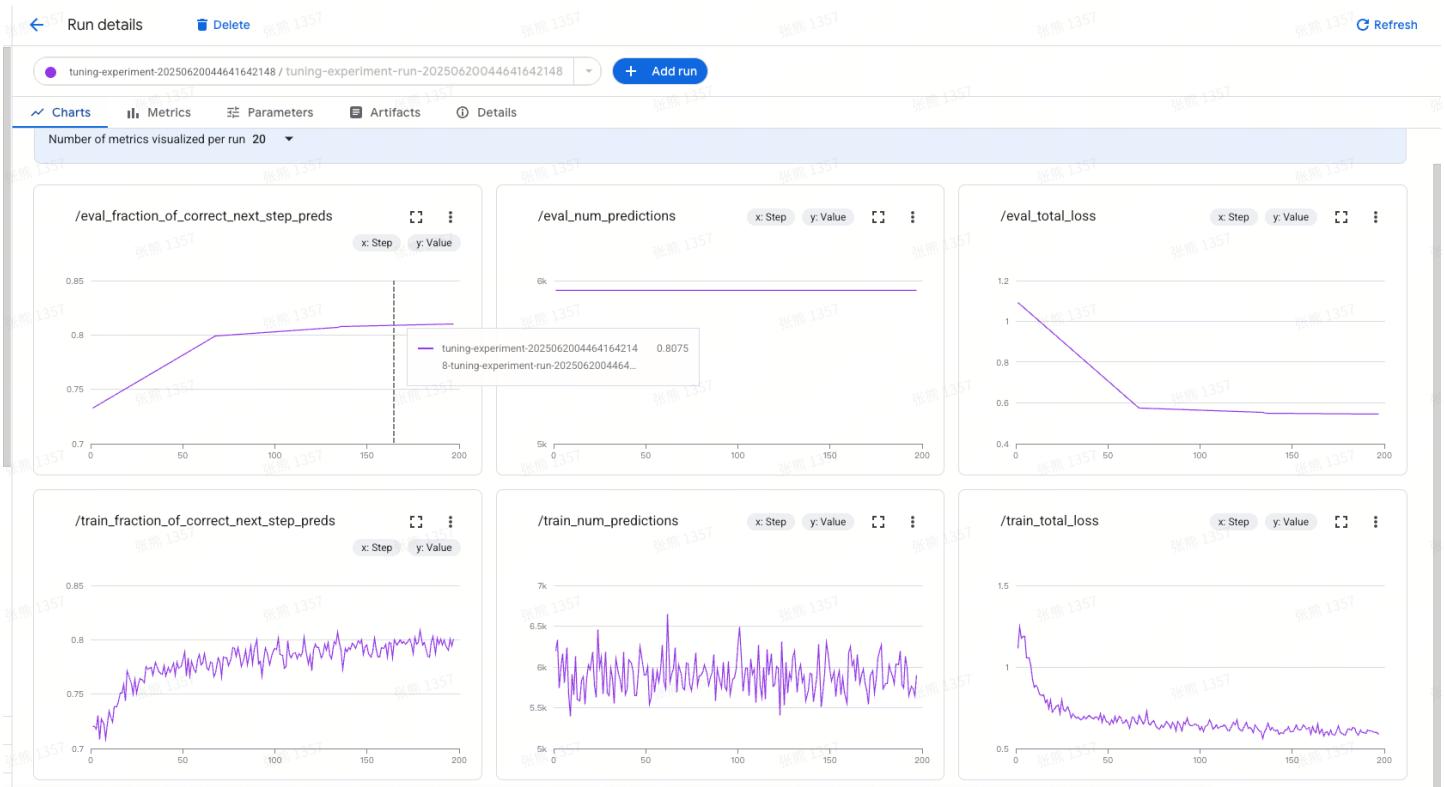
- 训练数据准确率:** 蓝线代表模型在训练数据上的准确率，其随训练步数增加而上升，大概维持在0.75左右。
- 训练损失变化:** 粉色曲线表示的train_total_loss呈现下降趋势，在第20步左右达到约0.7，训练和验证损失都持续下降。
- 模型学习状态:** 损失持续下降表明模型正在学习并收敛，且图中显示模型仍有较好的微调空间，损失未完全收敛。
- 默认学习率评估:** 模型学习良好、损失持续下降、准确率持续上升且无明显发散或不稳定迹象，说明当前的默认学习率是适合的。

第二轮训练：(基于2.0-flash-001)

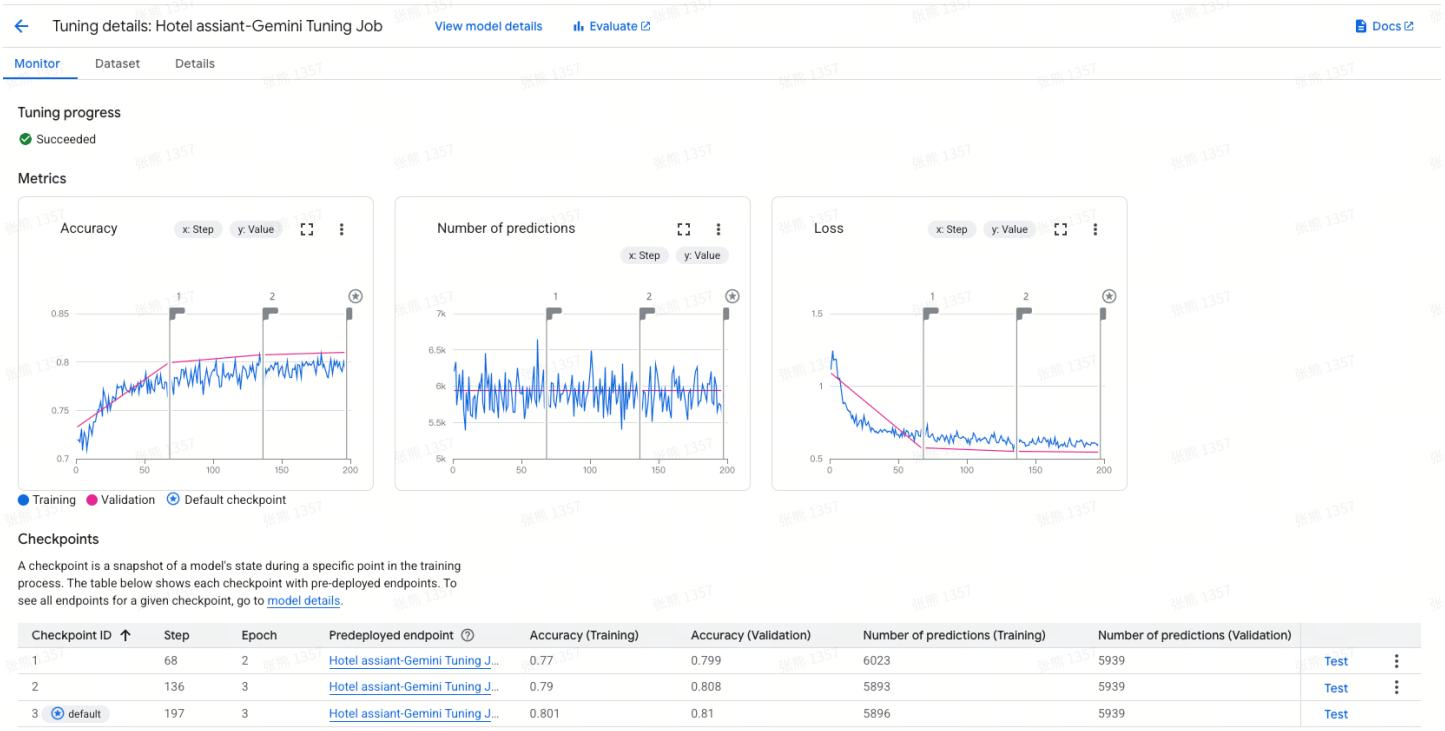
超参优化示例：

```
59]: sft_tuning_job = sft.train(
    source_model="gemini-2.0-flash-001",
    train_dataset=train_dataset_uri,
    validation_dataset=validation_dataset_uri,
    epochs=3,
    learning_rate_multiplier=1.0,
    tuned_model_display_name="Hotel assiant-Gemini Tuning Job"
)
Creating SupervisedTuningJob
```

Run details:



Tuning details:



数据分布对比：

Monitor	Dataset	Details
张鹏 1357	张鹏 1357	张鹏 1357
Messages 1080	Tokens 435108	Total
张鹏 1357	张鹏 1357	张鹏 1357

Dataset sample

Displaying the first 3 rows of the tuning dataset (samples may be truncated for preview)

Message ②

如何利用数字技术增强酒店客房的市场推广效果？

利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些策略和方法：1. **虚拟现实 (VR) 和增强现实 (AR) **：- 使用VR技术提供虚拟参观，让潜在客户在预订前体验客房和酒店设施。- 利用AR查看房间的3D模型或获取更多信息。2. **个性化营销**：- 使用大数据分析客户的偏好和行为，提供个性化的推荐和优惠。- 通过电子邮件或社交媒体推送个性化的内容和促销活动。3. **社交媒体和影响者营销**：- 在Instagram、Facebook等平台上分享高质量的客房照片和视频。- 与旅游博主和影响者合作，通过他们的渠道推广酒店。4. **搜索引擎优化 (SEO) 和内容营销**：- 优化酒店网站的SEO，提高在搜索引擎中的排名。- 创建有价值的内容，如博客文章、视频和...

酒店服务人员在与客人沟通时应避免哪些不当行为？

酒店服务人员在与客人沟通时应避免以下不当行为：1. **态度冷漠**：表现出不耐烦或漠不关心会让客人感到不受重视。2. **使用不礼貌的语言**：避免使用粗鲁或不专业的语言，保持礼貌和尊重。3. **打断客人**：在要打断他们，耐心倾听。4. **争论或辩解**：即使客人有误解，也应避免与客人争论，尽量以解决问题为导向。5. **忽视客人的请求**：对客人的合理请求应及时回应，不应拖延或忽视。6. **过度承诺**：不要向客人承诺，以免引发不满。7. **泄露隐私**：严格保密客人的个人信息和隐私，不在公开场合讨论。8. **缺乏专业知识**：对酒店的服务和设施不熟悉会影响服务质量，应确保了解相关信息。9. **情绪化反应**：保持冷静和专业的沟通。10. **忽视文化差异**...

酒店如何管理长期住客的费用减免？

酒店管理长期住客的费用减免通常涉及多种策略和考虑，以确保既能吸引住客，又能保持盈利。以下是一些常见的方法：1. **长期住宿折扣**：酒店通常会为住客提供长期住宿折扣。这可以是按周或按月计算的折扣，通常比日常房价更优惠。2. **套餐优惠**：提供包括餐饮、洗衣服务或其他设施使用的套餐，以增加住客的整体体验，同时也能增加酒店的收入。3. **会员计划**：通过会员计划为长期住客提供额外的优惠和积分奖励，鼓励他们选择长期入住。4. **灵活的付款方式**：为长期住客提供灵活的付款计划，例如按月支付，而不是一次性支付全部费用。5. **个性化服务**：为长期住客提供个性化的服务和设施，如房间升级、免费早餐或延迟退房等，以增加他们的满意度。6. **合同协议**：与公司或机构提供住宿，这通常会涉及到费用减免和其他优惠条件。7. **定期...

Displaying the first 3 rows of the tuning dataset (samples may be truncated for preview)

Message ②

如何利用数字技术增强酒店客房的市场推广效果？

利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些策略和方法：1. **虚拟现实 (VR) 和增强现实 (AR) **：- 使用VR技术提供虚拟参观，让潜在客户在预订前体验客房和酒店设施。- 利用AR查看房间的3D模型或获取更多信息。2. **个性化营销**：- 使用大数据分析客户的偏好和行为，提供个性化的推荐和优惠。- 通过电子邮件或社交媒体推送个性化的内容和促销活动。3. **社交媒体和影响者营销**：- 在Instagram、Facebook等平台上分享高质量的客房照片和视频。- 与旅游博主和影响者合作，通过他们的渠道推广酒店。4. **搜索引擎优化 (SEO) 和内容营销**：- 优化酒店网站的SEO，提高在搜索引擎中的排名。- 创建有价值的内容，如博客文章、视频和...

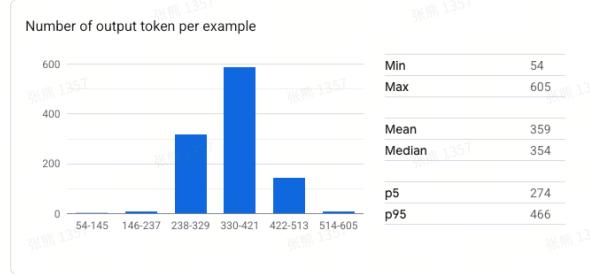
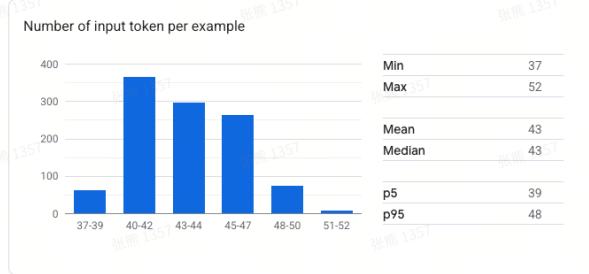
酒店服务人员在与客人沟通时应避免哪些不当行为？

酒店服务人员在与客人沟通时应避免以下不当行为：1. **态度冷漠**：表现出不耐烦或漠不关心会让客人感到不受重视。2. **使用不礼貌的语言**：避免使用粗鲁或不专业的语言，保持礼貌和尊重。3. **打断客人**：在要打断他们，耐心倾听。4. **争论或辩解**：即使客人有误解，也应避免与客人争论，尽量以解决问题为导向。5. **忽视客人的请求**：对客人的合理请求应及时回应，不应拖延或忽视。6. **过度承诺**：不要向客人承诺无法实现的服务或解决方案，以免引发不满。7. **泄露隐私**：严格保密客人的个人信息和隐私，不在公开场合讨论。8. **缺乏专业知识**：对酒店的服务和设施不熟悉会影响服务质量，应确保了解相关信息。9. **情绪化反应**：保持冷静和专业，不因个人情绪影响与客人的沟通。10. **忽视文化差异**...

酒店如何管理长期住客的费用减免？

酒店管理长期住客的费用减免通常涉及多种策略和考虑，以确保既能吸引住客，又能保持盈利。以下是一些常见的方法：1. **长期住宿折扣**：酒店通常会为住客提供长期住宿折扣。这可以是按周或按月计算的折扣，通常比日常房价更优惠。2. **套餐优惠**：提供包括餐饮、洗衣服务或其他设施使用的套餐，以增加住客的整体体验，同时也能增加酒店的收入。3. **会员计划**：通过会员计划为长期住客提供额外的优惠和积分奖励，鼓励他们选择长期入住。4. **灵活的付款方式**：为长期住客提供灵活的付款计划，例如按月支付，而不是一次性支付全部费用。5. **个性化服务**：为长期住客提供个性化的服务和设施，如房间升级、免费早餐或延迟退房等，以增加他们的满意度。6. **合同协议**：与公司或机构签订长期住宿合同，为其员工提供住宿，这通常会涉及到费用减免和其他优惠条件。7. **定期...

Data distribution



Displaying the first 5 rows of the tuning dataset (Samples may be truncated for preview)

Message

如何利用数字技术增强酒店客房的市场推广效果?

利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些策略和方法：1. **虚拟现实 (VR) 和增强现实 (AR)**：- 使用VR技术提供虚拟参观，让潜在客户在预订前体验客房和酒店设施。- 利用AR应用程序，客户可以通过手机查看房间的3D模型或获取更多信息。2. **个性化营销**：- 使用大数据分析客户的偏好和行为，提供个性化的推荐和优惠。- 通过电子邮件或社交媒体推送个性化的内容和促销活动。3. **社交媒体和影响者营销**：- 在Instagram、Facebook等平台上分享高质量的客房照片和视频。- 与旅游博主和影响者合作，通过他们的渠道推广酒店。4. **搜索引擎优化 (SEO) 和内容营销**：- 优化酒店网站的SEO，提高在搜索引擎中的排名。- 创建有价值的内容，如博客文章、视频和...

酒店服务人员在与客人沟通时应避免哪些不当行为？

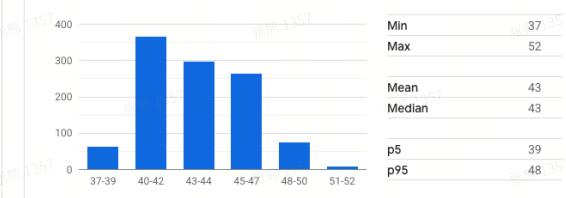
酒店服务人员在与客人沟通时应避免以下不当行为：1. **态度冷漠**：表现出不耐烦或漠不关心会让客人感到不受重视。2. **使用不礼貌的语言**：避免使用粗鲁或不专业的语言，保持礼貌和尊重。3. **打断客人**：在客人表达需求或意见时，不要打断他们，耐心倾听。4. **争论或辩解**：即使客人有误解，也应避免与客人争论，尽量以解决问题为导向。5. **忽视客人的请求**：对客人的合理请求应及时回应，不应拖延或忽视。6. **过度承诺**：不要向客人承诺无法实现的服务或解决方案，以免引发不满。7. **泄露隐私**：严格保密客人的个人信息和隐私，不在公开场合讨论。8. **缺乏专业知识**：对酒店的服务和设施不熟悉会影响服务质量，应确保了解相关信息。9. **情绪化反应**：保持冷静和专业，不因个人情绪影响与客人的沟通。10. **忽视文化差异**...

酒店如何管理长期住客的费用减免？

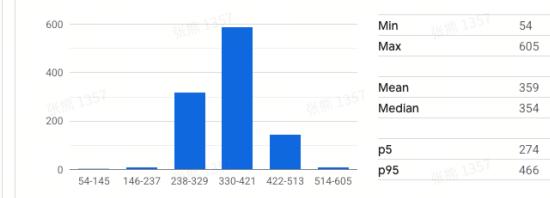
酒店管理长期住客的费用减免通常涉及多种策略和考虑，以确保既能吸引住客，又能保持盈利。以下是一些常见的方法：1. **长期住宿折扣**：酒店通常会为住客提供长期住宿折扣。这可以是按周或按月计算的折扣，通常比日常房价更优惠。2. **套餐优惠**：提供包括餐饮、洗衣服务或其他设施使用的套餐，以增加住客的整体体验，同时也能增加酒店的收入。3. **会员计划**：通过会员计划为长期住客提供额外的优惠和积分奖励，鼓励他们选择长期入住。4. **灵活的付款方式**：为长期住客提供灵活的付款计划，例如按月支付，而不是一次性支付全部费用。5. **个性化服务**：为长期住客提供个性化的服务和设施，如房间升级、免费早餐或延迟退房等，以增加他们的满意度。6. **合同协议**：与公司或机构签订长期住宿合同，为其员工提供住宿，这通常会涉及到费用减免和其他优惠条件。7. **定期...

Data distribution

Number of input token per example



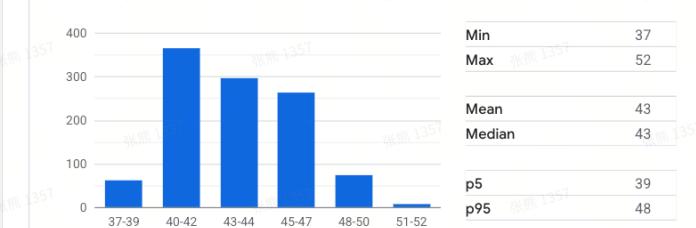
Number of output token per example



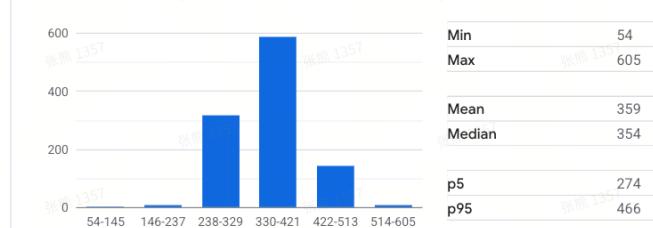
Monitor Dataset Details

Data distribution

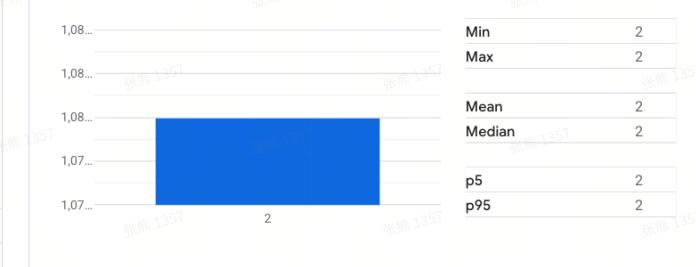
Number of input token per example



Number of output token per example



Number of message per example



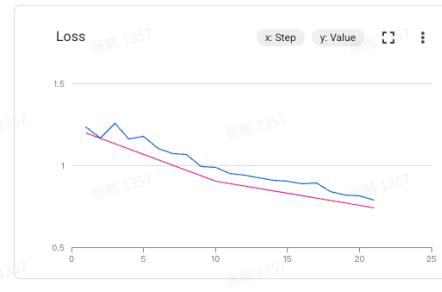
训练效果对比

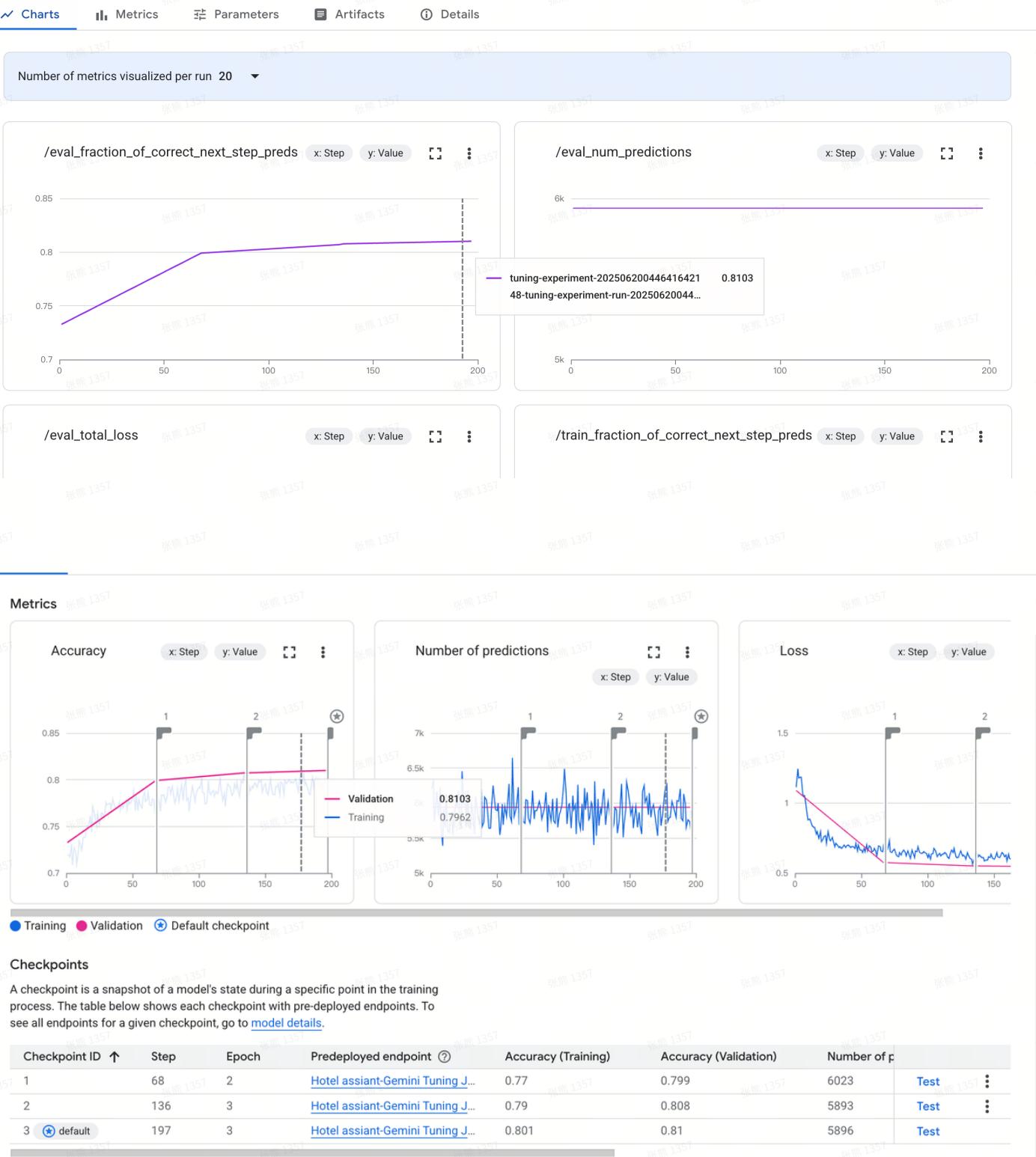
Monitor Dataset Details

Tuning progress

Succeeded

Metrics





训练结果3:

```
from sklearn.model_selection import train_test_split
import time
local_train_file = "hotel_train_data.json"
local_validation_file = "hotel_validation_data.json"
vertexai.init(project=PROJECT_ID, location=LOCATION)
print("✅ Vertex AI 初始化完成 - 项目: {PROJECT_ID}, 区域: {LOCATION}")

train_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_train_data.json"
validation_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_validation_data.json"
Vertex AI 初始话完成 - 项目: cy-aispeci-demo, 区域: us-west1

[7]: print(PROJECT_ID)
print(LOCATION)
print(train_dataset_uri)
print(validation_dataset_uri)

cy-aispeci-demo
['']
us-west1
gs://peft-model-cy-aispeci-demo/hotel_train_data.json
gs://peft-model-cy-aispeci-demo/hotel_validation_data.json

[8]: sft_tuning_job = sft.train(
    source_model="gemini-2.0-flash-001",
    train_dataset=train_dataset_uri,
    validation_dataset=validation_dataset_uri,
    epochs=2,
    learning_rate_multiplier=0.5,
    tuned_model_display_name="Hotel assiant-Gemini Tuning Job"
)
Creating SupervisedTuningJob
SupervisedTuningJob created. Resource name: projects/704352985590/locations/us-west1/tuningJobs/6141218018216116224
To use this SupervisedTuningJob in another session:
tuning_job = sft.SupervisedTuningJob('projects/704352985590/locations/us-west1/tuningJobs/6141218018216116224')
View Tuning Job:
https://console.cloud.google.com/vertex-ai/generative/language/locations/us-west1/tuning/tuningJob/6141218018216116224?project=704352985590
```

Would you like to receive official Jupyter news?
Please read the privacy policy.

第二轮训练结果评估：

- /eval_fraction_of_correct_next_step_preds:** 衡量 LLM 预测下一词准确率，从约 0.7 升 0.8 后稳定，显示评估集预测能力提升。
- /eval_num_predictions:** 评估预测数量不变，说明评估批次大小或数据量稳定。
- /eval_total_loss:** 评估集总损失从约 1.2 速降至 0.6 以下并稳定，表明评估集错误率降低。
- /train_fraction_of_correct_next_step_preds:** 训练数据上下一词预测准确率从约 0.7 上升且有波动，最终稳定在 0.8 左右，显示模型训练中持续学习。
- /train_num_predictions:** 训练预测数量不变。
- /train_total_loss:** 训练数据总损失从约 1.5 快速且持续下降，表明训练集模型不断优化。

结合（准确率和损失）在大约 150 步后就已经趋于平稳。继续训练到 200 步并没有带来显著的提升。考虑到更小的学习率可以让模型更精细地调整权重，可能帮助它找到一个更深层次的局部最优解。下一轮参数设置 epoch 2 轮，学习率 0.5。先进行调参数训练，看一下模型是否可以在更少的训练轮数下面达到更好的一个表现。一个节约时间，一个避免过拟合。

第三轮训练（基于gemini-2.0-flash-001）：

参数设置 epoch 2 轮，学习率 0.5

```
+ [ ] Code Execute No Kernel
from sklearn.model_selection import train_test_split
import datetime
import time
local_train_file = "hotel_train_data.jsonl"
local_validation_file = "hotel_validation_data.jsonl"
vertexai.init(project=PROJECT_ID, location=LOCATION)
print(f"✅ Vertex AI 初始化完成 - 项目: {PROJECT_ID}, 区域: {LOCATION}")

train_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_train_data.jsonl"
validation_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_validation_data.jsonl"

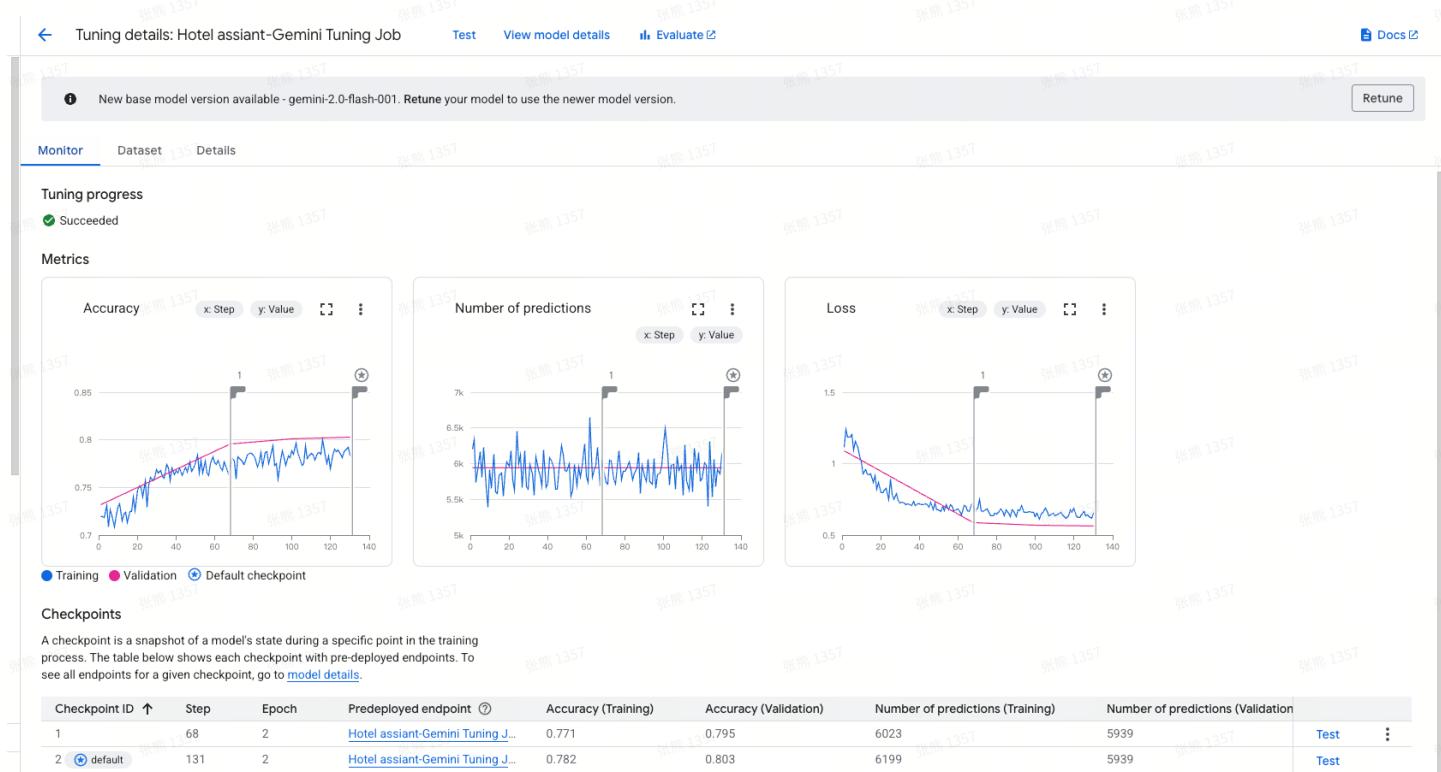
[7]: print(PROJECT_ID)
print(REGION)
print(LOCATION)
print(train_dataset_uri)
print(validation_dataset_uri)

cy-aispeci-demo
['']
us-west1
gs://peft-model-cy-aispeci-demo/hotel_train_data.jsonl
gs://peft-model-cy-aispeci-demo/hotel_validation_data.jsonl

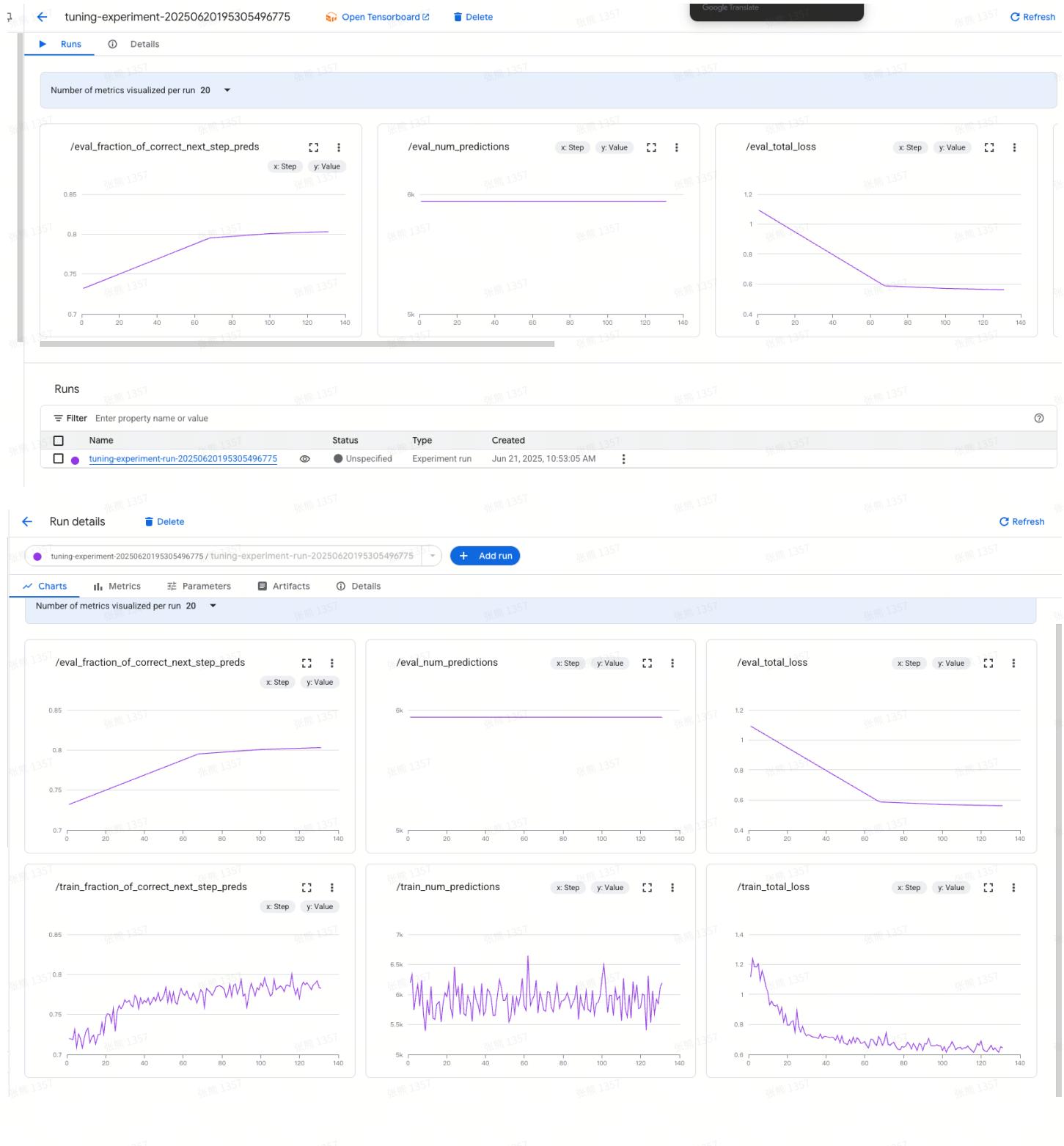
[8]: sft_tuning_job = sft.train(
    source_model="gemini-2.0-flash-001",
    train_dataset=train_dataset_uri,
    validation_dataset=validation_dataset_uri,
    epochs=2,
    learning_rate_multiplier=0.5,
    tuned_model_display_name="Hotel assiant-Gemini Tuning Job"
)

Creating SupervisedTuningJob
SupervisedTuningJob created. Resource name: projects/704352985590/locations/us-west1/tuningJobs/6141218018216116224
To use this SupervisedTuningJob in another session:
tuning_job = sft.SupervisedTuningJob('projects/704352985590/locations/us-west1/tuningJobs/6141218018216116224')
View Tuning Job:
https://console.cloud.google.com/vertex-ai/generative/language/locations/us-west1/tuning/tuningJob/6141218018216116224?project=704352985590
🔗 VIEW TUNING JOB
```

Tuning details:



Run details



第三轮训练结果评估：

第一轮训练时，我们选择的是1.5pro的模型其效果相对较差，之后我们选择了2.0 flash进行微调，其准确率来到了0.8以上，从结果上可以看到

1. 显著的积极进展：

- 没有过拟合：这是最重要的好消息！验证集和训练集的准确率、损失曲线高度一致，验证损失没有回升，表明模型在1000条数据和2个Epoch的情况下，有效地学习了数据的模式，并且具备了良好的泛化能力。

- **学习率调整有效:** 0.5 的学习率乘数在减少 Epoch 的同时，似乎让模型稳定地收敛，没有出现大的震荡。
- **效率提升:** 仅用 2 个 Epoch (约 130 步) 就达到了评估指标的饱和点，相比之前可能需要更多步数才能饱和，这节省了训练时间和计算资源。

2. 当前性能瓶颈:

- 模型在验证集上的准确率和损失在大约 60-80 步后 (大约 1 个 Epoch 到 1.5 个 Epoch 之间) 就已经达到了一个平台期。这意味着即使训练到 2 个 Epoch 结束，性能提升也微乎其微。
- 最终的验证准确率稳定在 **0.803**。虽然这不错，但如果业务需求更高，可能需要进一步优化。

如果还想继续优化，调整参数效果可能有限。增加高质量的训练数据。可能会带来更好的提升

数据优化记录:

```
[28]: question = "如何利用数字技术增强酒店客房的市场推广效果？"
response = tuned_model.generate_content(question)

print(response.text)
```

利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些具体的方法：

1. **搜索引擎优化 (SEO)**：通过优化酒店网站和在线列表，使其在搜索引擎结果中排名更高。使用相关的关键词、高质量的内容和用户友好的设计来吸引更多的有机流量。
2. **社交媒体营销**：利用社交媒体平台（如Facebook、Instagram、Twitter等）来推广酒店客房。发布吸引人的照片、视频和内容，与潜在客户互动，进行有针对性的广告投放。
3. **内容营销**：创建有价值的内容，如博客文章、视频和图片，展示酒店的独特之处和周边景点的吸引力。这有助于吸引潜在客户并建立品牌声誉。
4. **电子邮件营销**：建立客户数据库，通过电子邮件向潜在和现有客户发送个性化的优惠、促销信息和活动邀请。确保邮件内容吸引人，设计良好，并符合移动设备浏览。
5. **在线旅行社 (OTA) 优化**：确保酒店在主要的OTA平台（如Booking.com、Expedia等）上信息完整且最新。使用高质量的照片、吸引人的描述和竞争力的价格来吸引客户。
6. **用户生成内容 (UGC)**：鼓励客户在社交媒体和评论平台上分享他们的体验，使用UGC来建立信任和真实性。可以通过竞赛、奖励或简单的感谢来激励用户分享。
7. **移动应用和网站优化**：确保酒店的网站和移动应用在各种设备上都能正常工作，并提供无缝的用户体验。使用移动广告来吸引移动用户。
8. **个性化推荐和优惠**：使用数据分析来了解客户偏好和预订历史，提供个性化的推荐和优惠，提高客户忠诚度和转化率。
9. **视频营销**：制作高质量的视频内容，展示酒店设施、客房环境和周边景点。视频可以在YouTube、Vimeo等平台上发布，也可以嵌入到酒店网站和社交媒体中。
10. **虚拟现实 (VR) 和增强现实 (AR) **：提供虚拟现实体验让潜在客户在预订前体验酒店客房和设施。增强现实应用可以提供额外的信息或互动体验。

通过综合运用这些数字技术，酒店可以有效地增强客房的市场推广效果，吸引更多的客户并提高入住率。

微调结论:

指标名称	数据 (起始值 -> 结束值)	趋势	结论
/eval_fraction_of_correct_next_step_preds	约 0.715 -> 约 0.755	上升	模型在评估集上的提升。
/eval_num_predictions	约 58k (稳定)	稳定	评估预测数量恒定。
/eval_total_loss	约 1.2 -> 约 0.6	下降	模型在评估集上的提升。
/train_fraction_of_correct_next_step_preds	约 0.705 -> 约 0.755	上升	模型在训练集上的提升。
/train_num_predictions	50k - 54k (波动)	波动	训练批次预测数量波动。
/train_total_loss	约 1.4 -> 约 0.8	下降	模型在训练集上的提升。
accuracy	0.72->约0.81	上升	准确率提高。
整体表现	-	正面	模型成功收敛，性能良好。

微调模型调用结果展示：

```
[27]: question = "酒店服务人员在与客人沟通时应避免哪些不当行为？"
response = tuned_model.generate_content(question)
print(response.text)
```

在与客人沟通时，酒店服务人员应避免以下不当行为：

1. **不专业态度**：语气轻浮、态度傲慢或不友好，不尊重客人的需求和感受。
2. **缺乏倾听**：打断客人讲话，没有认真倾听他们的需求或投诉。
3. **不清晰的沟通**：使用含糊不清的语言，让客人感到困惑或误解。
4. **不耐烦的表现**：在客人提出问题或要求时表现出不耐烦或烦躁的情绪。
5. **不解决问题**：未能有效解决客人的问题或投诉，给客人留下不满意的印象。
6. **个人情绪化**：将个人情绪带入服务中，对客人的态度带有个人情感色彩。
7. **不遵守隐私**：在未经授权的情况下分享客人的个人信息，侵犯客人的隐私。
8. **语言障碍**：缺乏必要的语言技能，无法与国际客人有效沟通。
9. **缺乏同理心**：未能理解客人的情绪或需求，无法提供个性化的服务。
10. **不遵守公司政策**：在没有明确的理由下，拒绝按照酒店政策提供服务或解决方案。
11. **过度承诺**：给出无法实现的承诺，导致客人期望落空。
12. **不尊重文化差异**：未能尊重不同文化背景客人的习俗和习惯，导致误解或冲突。

通过避免这些不当行为，酒店服务人员可以提供更优质的服务，提升客人的满意度和忠诚度。

```
[22]: # Save the response from model in the Cloud Storage bucket
```

```
[28]: question = "如何利用数字技术增强酒店客房的市场推广效果？"
response = tuned_model.generate_content(question)
print(response.text)
```

利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些具体的方法：

1. **搜索引擎优化（SEO）**：通过优化酒店网站和在线列表，使其在搜索引擎结果中排名更高。使用相关的关键词、高质量的内容和用户友好的设计来吸引更多的有机流量。
2. **社交媒体营销**：利用社交媒体平台（如Facebook、Instagram、Twitter等）来推广酒店客房。发布吸引人的照片、视频和内容，与潜在客户互动，进行有针对性的广告投放。
3. **内容营销**：创建有价值的内容，如博客文章、视频和图片，展示酒店的独特之处和周边景点的吸引力。这有助于吸引潜在客户并建立品牌声誉。
4. **电子邮件营销**：建立客户数据库，通过电子邮件向潜在和现有客户发送个性化的优惠、促销信息和活动邀请。确保邮件内容吸引人，设计良好，并符合移动设备浏览。
5. **在线旅行社（OTA）优化**：确保酒店在主要的OTA平台（如Booking.com、Expedia等）上信息完整且最新。使用高质量的照片、吸引人的描述和竞争力的价格来吸引客户。
6. **用户生成内容（UGC）**：鼓励客户在社交媒体和评论平台上分享他们的体验，使用UGC来建立信任和真实性。可以通过竞赛、奖励或简单的感谢来激励用户分享。
7. **移动应用和网站优化**：确保酒店的网站和移动应用在各种设备上都能正常工作，并提供无缝的用户体验。使用移动广告来吸引移动用户。
8. **个性化推荐和优惠**：使用数据分析来了解客户偏好和预订历史，提供个性化的推荐和优惠，提高客户忠诚度和转化率。
9. **视频营销**：制作高质量的视频内容，展示酒店设施、客房环境和周边景点。视频可以在YouTube、Vimeo等平台上发布，也可以嵌入到酒店网站和社交媒体中。
10. **虚拟现实（VR）和增强现实（AR）**：提供虚拟现实体验让潜在客户在预订前体验酒店客房和设施。增强现实应用可以提供额外的信息或互动体验。

通过综合运用这些数字技术，酒店可以有效地增强客房的市场推广效果，吸引更多的客户并提高入住率。

资源创建流程及部署流程：

标准流程：

第一步 创建一个微调任务

The screenshot shows the Google Cloud Vertex AI Tuning interface. On the left, there's a sidebar with various tools like Dashboard, Model Garden, Pipelines, Notebooks, and Agent Builder. The main area is titled 'Tuning' and has a button '+ Create tuned model'. Below it, a section says 'In Vertex AI Studio, you can tune and distill foundation models to optimize them for specific tasks or knowledge domains. Learn more about tuning models'. It shows a table of tuning jobs, with one entry for 'gemini-2.0-flash-lite-hotel-qa' which is 'Failed' with a red status indicator. The table includes columns for Name, Base model, Method, Status, Created, Updated, and Notification.

进行base模型选择

This screenshot shows the 'Create a tuned model' interface. The sidebar is identical to the previous one. The main area starts with 'Model details' and then 'Tuning dataset' (which is checked). Below that is a 'Start tuning' button. The 'Base model' dropdown menu is open, showing several options: 'gemini-2.5-flash' (selected), 'gemini-2.0-flash-lite-001', 'gemini-2.0-flash-001', 'gemini-1.5-flash-002', 'gemini-1.5-pro-002', and 'translation-llm-002'. A tooltip for 'gemini-2.5-flash' indicates it's a good option for well-defined tasks with labeled data.

进行数据关联

Google Cloud CY-AISpeci-demo

Search (S) for resources, docs, products, and more

Search (S)

Vertex AI Create a tuned model

Tools

- Dashboard
- Model Garden
- Pipelines
- Notebooks
- Colab Enterprise
- Workbench
- Vertex AI Studio
- Overview
- Create prompt
- Media Studio
- Stream realtime
- Prompt gallery
- Prompt management
- Tuning
- Agent Builder
- Agent Garden
- Agent Engine
- RAG Engine
- Vertex AI Search
- Vector Search

Model details

Tuning dataset

Start tuning

Tuning dataset

The dataset is a JSON file where each line contains a single example. The number of recommended examples varies by task. View the dataset documentation to learn how to prepare one or download a sample dataset for Gemini models.

Upload file to Cloud Storage

Existing file on Cloud Storage

Select JSONL file * hotel_train_data.jsonl

The JSONL file containing the dataset

Dataset location * gs://peft-model-cy-aispeci-demo

Browse

Model validation

Generates validation metrics during tuning to help you measure model performance. View sample dataset

Enable model validation

Validation dataset

The validation dataset is a JSONL file containing model prompt and response examples (one per line). You can use 10 to 5000 examples.

Validation dataset * gs://peft-model-cy-aispeci-demo/hotel_validation_data.jsonl

Browse

查看微调数据的基本情况

Vertex AI Workbench Tuning details: Hotel assitant-Gemini Tuning Job

Monitor Dataset Details Test View model details Evaluate

Messages 923 Tokens 369078

Total Total

Dataset sample

Displaying the first 3 rows of the tuning dataset (samples may be truncated for preview)

Message

如何利用数字技术增强酒店客房的市场推广效果? ① 利用数据分析技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些策略和方法: 1. **虚拟现实 (VR) 和增强现实 (AR)**: - 使用VR技术提供虚拟参观, 让潜在客户在预订前体验客房和酒店设施。- 利用AR应用程序, 客户可以通过手机查看房间的3D模型或获取更多信息。 2. **个性化营销**: - 使用大数据分析客户的偏好和行为, 提供个性化的推荐和优惠。- 通过电子邮件或社交媒体推送个性化的内容和促销活动。 3. **社交媒体和影响者营销**: - 在Instagram, Facebook等平台上分享高质量的客房照片和视频。- 与旅游博主和影响者合作, 通过他们的渠道推广酒店。 4. **搜索引擎优化 (SEO) 和内容营销**: - 优化酒店网站的SEO, 提高在搜索引擎中的排名。- 创建有价值的内容, 如博客文章、视频和...

酒店服务员在与客人沟通时应避免以下不当行为? ② 酒店服务员在与客人沟通时应避免以下不当行为: 1. **态度冷漠**: 表现出不耐烦或漠不关心会让客人感到不受重视。 2. **使用不礼貌的语言**: 避免使用粗鲁或不专业的语言, 保持礼貌和尊重。 3. **打断客人**: 在客人表达需求或意见时, 不要打断他们, 耐心倾听。 4. **争论或辩解**: 即使客人有错误, 也应避免与客人争论, 尽量以解决问题为导向。 5. **忽视客人的请求**: 对客人的合理请求应及时回应, 不应推诿或忽视。 6. **过度承诺**: 不要向客人承诺无法实现的服务或解决方案, 以免引发不满。 7. **泄露隐私**: 严格保密客人的个人信息和隐私, 不在公开场合讨论。 8. **缺乏专业知识**: 对酒店的服务和设施不熟悉会影响服务质量, 应确保了解相关信息。 9. **情绪化反应**: 保持冷静和专业, 不因个人情绪影响与客人的沟通。 10. **忽视文化差异...

Data Datasets

Number of input token per example

Range	Min	Max
37-39	37	52
40-42	37	52
43-44	27	52
45-47	27	52
48-50	10	52
51-52	10	52

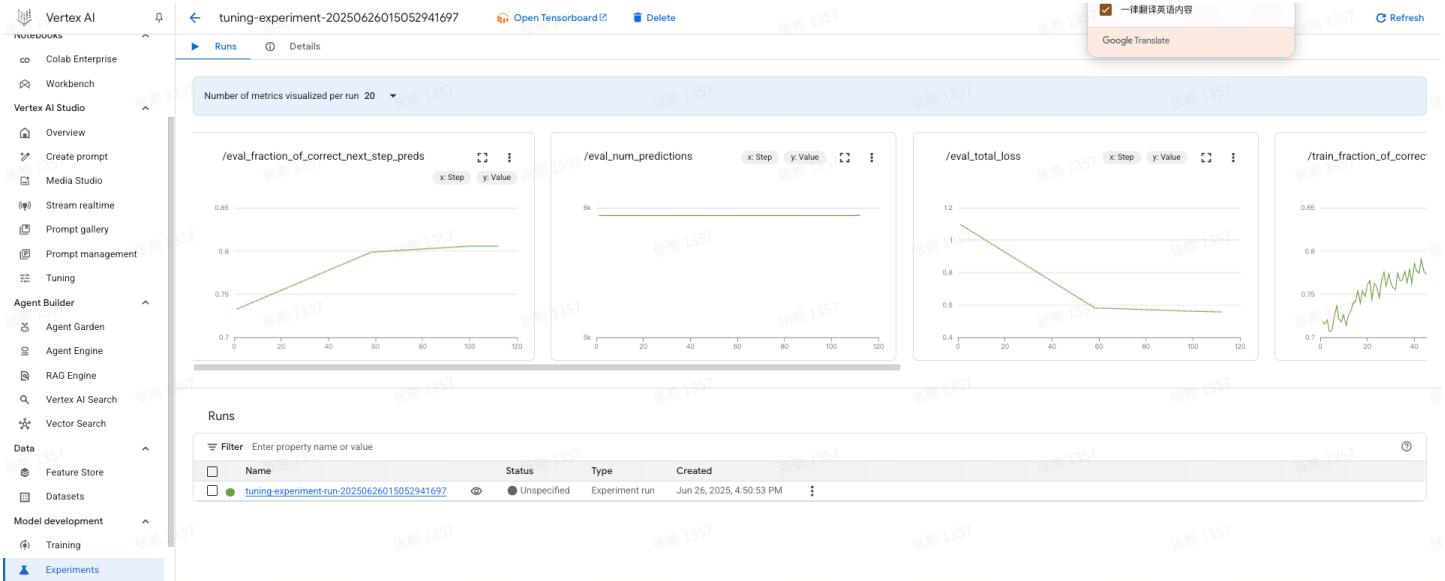
Mean 43 Median 43 p5 39 p95 48

Number of output token per example

Range	Min	Max
54-144	54	594
145-234	145	594
235-324	356	594
325-414	356	594
415-504	356	594
505-594	458	594

Min 54 Max 594 Mean 356 Median 353 p5 272 p95 458

查看训练情况:



查看部署情况：

Name	ID	Status	Models	Deployment resource pool	Region	Monitoring	Most recent alerts	Last updated	API	Labels
Hotel assistant-Gemini Tuning Job	4014462088545042432	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 5:02:57 PM	—	google-ver... 8669907845... tune-type: sft
Hotel assistant-Gemini Tuning Job	8378450127467053056	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 5:00:51 PM	—	google-ver... 8669907845... tune-type: sft
Hotel assistant-Gemini Tuning Job	4825110021471731712	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 4:46:39 PM	—	google-ver... 69989220759... tune-type: sft
hotel_assistant_job_0626	247423101598432800	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 4:46:26 PM	—	google-ver... 3832845307... tune-type: sft
Hotel assistant-Gemini Tuning Job	7473226602365583360	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 4:44:43 PM	—	google-ver... 69989220759... tune-type: sft
Hotel assistant-Gemini Tuning Job	1607288087715512320	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 4:34:42 PM	—	google-ver... 5674169667... tune-type: sft
Hotel assistant-Gemini Tuning Job	6200959707633418240	Active	1	—	us-west1	Disabled	—	Jun 26, 2025, 4:32:46 PM	—	google-ver... 5674169667... tune-type: sft
Hotel assistant-Gemini Tuning Job	1199712321438482432	Active	1	—	us-west1	Disabled	—	Jun 25, 2025, 7:26:33 PM	—	google-ver... 4340154200... tune-type: sft
Hotel assistant-Gemini Tuning Job	8684694902128246784	Active	1	—	us-west1	Disabled	—	Jun 25, 2025, 2:11:21 PM	—	google-ver... 8951840218... tune-type: sft
Hotel assistant-Gemini Tuning Job	1771106524161114112	Active	1	—	us-west1	Disabled	—	Jun 21, 2025, 3:49:01 PM	—	google-ver... 8146593086... tune-type: sft
Hotel assistant-Gemini Tuning Job	2507445063236190208	Active	1	—	us-west1	Disabled	—	Jun 21, 2025, 11:09:52 AM	—	google-ver... 4441256493... tune-type: sft
Hotel assistant-Gemini Tuning Job	3563539175854571520	Active	1	—	us-west1	Disabled	—	Jun 21, 2025, 11:06:53 AM	—	google-ver... 4441256493... tune-type: sft
Hotel assistant-Gemini Tuning Job	263917535233677216	Active	1	—	us-west1	Disabled	—	Jun 20, 2025, 8:04:55 PM	—	google-ver... 614218018... tune-type: sft
Hotel assistant-Gemini Tuning Job	5787785228147752	Active	1	—	us-west1	Disabled	—	Jun 20, 2025, 8:03:24 PM	—	google-ver... 614218018... tune-type: sft
Hotel assistant-Gemini Tuning Job	85276318656123700736	Active	1	—	us-west1	Disabled	—	Jun 20, 2025, 8:02:55 PM	—	google-ver... 614218018... tune-type: sft

Google Cloud | CY-AISpeci-demo | 张鹏 1357

Search (I) for resources, docs, products, and more | Search

Vertex AI | Hotel assitant-Gemini Tuning Job | Edit settings

Details

Endpoint ID	4014462088545042432
Region	us-west1
Logs	View Logs
Model Monitoring	Disabled

Deployed [Sort by]

Model	ID	Status	Deployment resource pool	Most recent alerts	Monitoring	Traffic split	Compute nodes	Type	Created
Hotel assitant-Gemini Tuning Job (Version 1, Checkpoint 2)	1294329695055118336	Ready	-	-	Disabled	100%	Auto (1 minimum, 1 maximum)	Large model	Jun 26, 2025, 5:02:55 PM

[Deploy another model](#)

Chart interval: 1 hour 6 hours 12 hours 1 day 2 days 4 days 7 days 14 days 30 days

[Performance](#) Resource usage

Predictions/second

No data is available for the selected time frame.

Prediction error percentage

UTC+8 10:10 AM 10:15 AM 10:20 AM 10:25 AM 10:30 AM 10:35 AM 10:40 AM 10:45 AM 10:50 AM 10:55 AM 11:00 AM 11:05 AM

自定义流程：

在workbecnh 构建实例

Google Cloud | CY-AISpeci-demo | 张鹏 1357

Search (I) for resources, docs, products, and more | Search

Vertex AI | Workbench | Instances

Workbench | Create New | Refresh | Learn

Instances | Executions | Schedules

View: Instances User-managed Notebooks Managed Notebooks

JupyterLab 4 is now available in Vertex AI Workbench. Dismiss

Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more

Filter

Instance name	Zone	Auto upgrade	Version	Machine Type	GPUs	Owner	Created	Labels
instance-gemini-finetuning-polymericclod-us	us-west1-a	M130	4 vCPUs, 15 GB RAM	NVIDIA T4 x1	70435298590-compute@developer.gserviceaccount.com	Jun 20, 2025, 7:32:32PM	consumer-p...:cy-aис...:...	

Google Cloud | qwiklabs-gcp-04-0380eb7c441f | AI Workbench

Vertex AI / Workbench

Tools

- Dashboard
- Model Garden
- Pipelines

Notebooks

- Colab Enterprise
- Workbench

Vertex AI Studio

- Overview
- Create prompt
- Media Studio
- Stream realtime
- Prompt gallery
- Prompt management
- Tuning

Agent Builder

- Agent Garden
- Agent Engine
- RAG Engine
- Vertex AI Search
- Vector Search

Data

Workbench

Instances Executions Schedules

View: Instances User-managed Notebooks Managed Notebooks

JupyterLab 4 is now available in Vertex AI Workbench.

Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. [Learn more](#)

Filter

Instance name	Zone	Auto upgrade	Version	Machine type

You don't have any instances in this project.

Create New

New instance

Name * student-workbench-instance

Region * us-west4 (Las Vegas) Zone * us-west4-a

Some regions are restricted due to a policy set by your organization. [Learn more](#)

Attach 1 NVIDIA T4 GPU

Enable Dataproc Serverless Interactive Sessions

Enable access to Dataproc Spark kernels

Network in this project

Shared network

Network default

Subnetwork * default(10.182.0.0/20)

Instance properties

Machine type e2-standard-4

Data disk 100 GB Balanced persistent disk

Permission Compute Engine default service account

Estimated cost \$162.00 monthly, \$0.22 hourly

Advanced options Cancel Create

Google Cloud | qwiklabs-gcp-04-0380eb7c441f | AI Workbench

Vertex AI / Workbench

Tools

- Dashboard
- Model Garden
- Pipelines

Notebooks

- Colab Enterprise
- Workbench

Vertex AI Studio

- Overview
- Create prompt
- Media Studio
- Stream realtime
- Prompt gallery
- Prompt management
- Tuning

Agent Builder

- Agent Garden
- Agent Engine
- RAG Engine
- Vertex AI Search
- Vector Search

Data

Provisioned Throughput

Workbench

Instances Executions Schedules

View: Instances User-managed Notebooks Managed Notebooks

JupyterLab 4 is now available in Vertex AI Workbench.

Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. [Learn more](#)

Filter

Instance name	Zone	Auto upgrade	Version	Machine type

You don't have any instances in this project.

Create New

New instance

Name * student-workbench-instance

Region * us-west4 (Las Vegas) Zone * us-west4-a

Some regions are restricted due to a policy set by your organization. [Learn more](#)

Attach 1 NVIDIA T4 GPU

Enable Dataproc Serverless Interactive Sessions

Enable access to Dataproc Spark kernels

Network in this project

Shared network

Network default

Subnetwork * default(10.182.0.0/20)

Instance properties

Machine type n1-standard-4

Data disk 100 GB Balanced persistent disk

Permission Compute Engine default service account

Estimated cost \$375.48 monthly, \$0.51 hourly

Advanced options Cancel Create

Google Cloud | qwiklabs-gcp-04-0380eb7c441f | AI Workbench / Instances / Create instance

Tools

- Dashboard
- Model Garden
- Pipelines

Notebooks

- Colab Enterprise
- Workbench

Vertex AI Studio

- Overview
- Create prompt
- Media Studio
- Stream realtime
- Prompt gallery
- Prompt management
- Tuning

Agent Builder

- Agent Garden
- Agent Engine
- RAG Engine
- Vertex AI Search
- Vector Search

Data

Provisioned Throughput

Tutorials

Create instance

Details

Name * student-workbench-instance

Environment

Machine type

Disk

Networking

IAM and security

System health

Labels

+ Add label

Pricing summary

\$375.48 monthly estimate

That's about \$0.514 hourly

Pay for what you use: No upfront costs and per second billing

Networking cost also applies. [Learn more](#)

Details

Network tags

Assign tags to your Workbench resource. [Learn more](#)

Workbench type

Instead of creating a Vertex AI Workbench instance, you can create a user-managed or managed notebook using the same configuration. These are older versions of Workbench and not recommended unless you have a specific need.

Type Instance

Continue

Recommended for you

Introduction to Vertex AI Workbench

Create a new Vertex AI Workbench instance

Query data in BigQuery from within JupyterLab

Use cases for Vertex AI

Terraform samples

Architecture guides for AI and machine learning

All Vertex AI documentation

[Vertex AI / Workbench / Instances / Create instance](#)

Create instance

[Tools](#) [Create instance](#) [Learn](#)

[Details](#)

[Environment](#)

- Machine type
- Disks
- Networking
- IAM and security
- System health

Environment

All environments use JupyterLab 3 by default and have the latest NVIDIA GPU and Intel libraries and drivers installed. You can specify a previous version instead. [Learn more](#)

JupyterLab Version

- JupyterLab 3.x
- JupyterLab 4.x [New](#)

Use custom container

Version

- Use the latest version
- Use a previous version

To learn more about specific versions, see the [Vertex AI Workbench release notes](#).

Post-startup script

Path to post-startup script [Browse](#)

Cloud Storage path to script that automatically runs after the instance boots up.

Metadata

Some metadata keys including `data-disk-uri`, `framework`, `notebooks-api`, `notebooks-api-version`, `nvidia-driver-gcs-path`, `proxy-uri`, `restriction`, `shutdown-script`, `title`, `version` are reserved for system use only. If you use these variable names below, they will be overwritten by system values.

[+ Add metadata](#)

[Back](#) [Continue](#)

Recommended for you

[Introduction to Vertex AI Workbench](#)

[Help document](#)

Vertex AI Workbench lets you perform your data science workflow in a JupyterLab notebook-based development environment.

[Create a new Vertex AI Workbench instance](#)

[Help document](#)

Create a new Vertex AI Workbench instance with the latest machine learning and data science libraries installed.

[Query data in BigQuery from within JupyterLab](#)

[Help document](#)

Access BigQuery data without leaving the JupyterLab interface.

[Use cases for Vertex AI](#)

[Help document](#)

Explore use cases, best practices, and industry solutions.

[Terraform samples](#)

[Help document](#)

See examples of using Terraform to create Vertex AI resources.

[Architecture guides for AI and machine learning](#)

[Help document](#)

Discover best practices and reference architectures for AI and machine learning.

[Google Cloud](#) [qwiklabs-gcp-04-0380eb7c441f](#)

[Vertex AI / Workbench / Instances / Create instance](#)

[AI Workbench](#) [Tutorial](#)

Create instance

[Tools](#) [Create instance](#) [Learn](#)

[Details](#)

[Environment](#)

[Machine type](#)

- Disks
- Networking
- IAM and security
- System health

Machine type

[General purpose](#) [GPUs](#)

Machine types for common workloads, optimized for cost and flexibility

Series	Description	vCPUs	Memory
E2	Low cost, day-to-day computing	2 - 32	4 - 128 GB
N2	Balanced price & performance	2 - 128	4 - 864 GB
N2D	Balanced price & performance	2 - 224	4 - 896 GB
<input checked="" type="radio"/> N1	Balanced price & performance	2 - 96	3.6 - 624 GB

Machine type: [n1-standard-4 \(4 vCPU, 2 core, 15 GB memory\)](#)

GPUs

The number of attached GPUs affects the VM's maximum number of memory and CPUs. [Learn More](#)

GPU type: [NVIDIA T4](#) Number of GPUs: [1](#)

CPU platform and GPU Reservations

Reservations: [Don't use](#)

Use an existing Compute Engine reservation when creating this Notebook.

Shielded VM

Turn on all settings for the most secure configuration. [Learn more](#)

Secure Boot

Created bucket qwiklabs-gcp-04-0380eb7c441f-model-dataset

Create instance

Disks

- Details
- Environment
- Machine type
- Disks
- Networking
- IAM and security
- System health

Pricing summary

\$375.48 monthly estimate
That's about \$0.514 hourly
Pay for what you use: No upfront costs and per second billing
Networking cost also applies. [Learn more](#)

Encryption

- Google-managed encryption key
Keys owned by Google
- Cloud KMS key
Keys owned by customers

Back Continue

Recommended for you

- [Introduction to Vertex AI Workbench](#)
- [Create a new Vertex AI Workbench instance](#)
- [Query data in BigQuery from within JupyterLab](#)
- [Use cases for Vertex AI](#)
- [Terraform samples](#)
- [Architecture guides for AI and machine learning](#)

Create instance

Networking

- Details
- Environment
- Machine type
- Disks
- Networking
- IAM and security
- System health

Pricing summary

\$375.48 monthly estimate
That's about \$0.514 hourly
Pay for what you use: No upfront costs and per second billing
Networking cost also applies. [Learn more](#)

Networking

The instance requires internet access to be used. Make sure one of the following is selected. [Learn more](#)

- Assign an external IP address
- Select a network that has internet access
- Turn on [Private Google Access](#)

Network in this project

Network in this project

Shared network

Back Continue

Recommended for you

- [Introduction to Vertex AI Workbench](#)
- [Create a new Vertex AI Workbench instance](#)
- [Query data in BigQuery from within JupyterLab](#)
- [Use cases for Vertex AI](#)
- [Terraform samples](#)
- [Architecture guides for AI and machine learning](#)

Create instance

IAM and security

- Details
- Environment
- Machine type
- Disks
- Networking
- IAM and security
- System health

Pricing summary

\$375.48 monthly estimate
That's about \$0.514 hourly
Pay for what you use: No upfront costs and per second billing
Networking cost also applies. [Learn more](#)

Service account

Anyone with the iam.serviceAccounts.actAs can access the instance account

Single user

Restricts access to one user

Use default Compute Engine service account

Security options

- Root access to the instance
- nbconvert
- File download
- Terminal access

Back Continue

Recommended for you

- [Introduction to Vertex AI Workbench](#)
- [Create a new Vertex AI Workbench instance](#)
- [Query data in BigQuery from within JupyterLab](#)
- [Use cases for Vertex AI](#)
- [Terraform samples](#)
- [Architecture guides for AI and machine learning](#)

System health

- Details
- Environment
- Machine type
- Disks
- Networking
- IAM and security
- System health

Pricing summary

\$375.48 monthly estimate
That's about \$0.514 hourly
Pay for what you use: No upfront costs and per second billing
Networking cost also applies. [Learn more](#)

Instance name	Zone	Auto upgrade	Version	Machine Type	GPUs	Owner	Created
student-workbench-instance	us-west4-a	-	M130	4 vCPUs, 15 GB RAM	NVIDIA T4 x 1	679098231003-compute@developer.gserviceaccount.com	Jun 19, 2025, 10:41:26 AM

第二步：打开open Jupyterlab

```
[2]: gsutil ls -l gs://qwiklabs-gcp-00-4e744fb7420-lab-data/gemini-supervised-tuning-qa-challenge.ipynb
22913 2025-06-19T02:56:10Z gs://qwiklabs-gcp-00-4e744fb7420-lab-data/gemini-supervised-tuning-qa-challenge.ipynb
TOTAL: 1 objects, 22913 bytes (22.38 KiB)
```

构建训练集和验证集

```
1 project_id_output = !gcloud config list --format 'value(core.project)'  
2 >/dev/null  
2 PROJECT_ID = project_id_output[0]  
3 REGION = !gcloud compute project-info describe --format="value[]"  
 (commonInstanceMetadata.items.google-compute-default-region)"  
4 LOCATION = "asia-east2"  
5  
6 BUCKET_NAME = f"[PROJECT_ID]-model-dataset"  
7 # BUCKET_URI = f"gs://{BUCKET_NAME}"  
8 BUCKET_URI = "gs://peft-model-cy-aispeci-demo"  
9  
10 import vertexai  
11 from vertexai.generative_models import (  
12     GenerativeModel,  
13     Part,  
14     HarmCategory,  
15     HarmBlockThreshold,  
16     GenerationConfig,  
17 )  
18 from vertexai.preview.tuning import sft  
19 from typing import Union  
20 import pandas as pd  
21 from google.cloud import bigquery  
22 from sklearn.model_selection import train_test_split  
23 import datetime  
24 import time  
25 local_train_file = "hotel_train_data.jsonl"  
26 local_validation_file = "hotel_validation_data.jsonl"  
27 vertexai.init(project=PROJECT_ID, location=LOCATION)  
28 print(f"✓ Vertex AI 初始化完成 - 项目: {PROJECT_ID}, 区域: {LOCATION}")  
29  
30 train_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_train_data.jsonl"  
31 validation_dataset_uri = f"gs://peft-model-cy-aispeci-  
demo/hotel_validation_data.jsonl"
```

```
张熊 1357
project_id_output = !gcloud config list --format 'value(core.project)' 2>/dev/null
PROJECT_ID = project_id_output[0]
REGION = !gcloud compute project-info describe --format="value[](commonInstanceMetadata.items.google-compute-default-region)"
LOCATION = "asia-east2"

BUCKET_NAME = f"{PROJECT_ID}-model-dataset"
# BUCKET_URI = f"gs://{BUCKET_NAME}"
BUCKET_URI = "gs://peft-model-cy-aispeci-demo"

import vertexai
from vertexai.generative_models import (
    GenerativeModel,
    Part,
    HarmCategory,
    HarmBlockThreshold,
    GenerationConfig,
)
from vertexai.preview.tuning import sft
from typing import Union
import pandas as pd
from google.cloud import bigquery
from sklearn.model_selection import train_test_split
import datetime
import time
local_train_file = "hotel_train_data.jsonl"
local_validation_file = "hotel_validation_data.jsonl"
vertexai.init(project=PROJECT_ID, location=LOCATION)
print(f"✅ Vertex AI 初始化完成 - 项目: {PROJECT_ID}, 区域: {LOCATION}")

train_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_train_data.jsonl"
validation_dataset_uri = f"gs://peft-model-cy-aispeci-demo/hotel_validation_data.jsonl"
```

Python

第三步：基于训练集进行模型微调训练

代码块

```
1 sft_tuning_job = sft.train(
2             source_model="gemini-1.5-pro-002",
3             train_dataset=train_dataset_uri,
4             validation_dataset=validation_dataset_uri,
5             epochs=3,
6             learning_rate_multiplier=1.0,
7             tuned_model_display_name="Hotel assiant-Gemini Tuning Job"
8         )
```

第四步：查看训练是否完成

```
[7]: sft_tuning_job_name = sft_tuning_job.resource_name
sft_tuning_job_name
[7]: 'projects/704352985590/locations/us-west1/tuningJobs/8669907845124194304'

[13]: while not sft_tuning_job.refresh().hasEnded:
    time.sleep(60)

[14]: tuned_model_name = sft_tuning_job.tuned_model_name
tuned_model_name
[14]: 'projects/704352985590/locations/us-west1/models/6857245004113379328@1'

[15]: sft_tuning_job.list()
[15]: [<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd53a020370>
resource name: projects/704352985590/locations/us-west1/tuningJobs/8669907845124194304,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215bee0>
resource name: projects/704352985590/locations/us-west1/tuningJobs/7530497139399458816,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215b850>
resource name: projects/704352985590/locations/us-west1/tuningJobs/6989220759184867328,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215ba0>
resource name: projects/704352985590/locations/us-west1/tuningJobs/383284530762481664,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215b820>
resource name: projects/704352985590/locations/us-west1/tuningJobs/5674169667992682496,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539fafb20>
resource name: projects/704352985590/locations/us-west1/tuningJobs/4340154200244617216,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539fafc10>
resource name: projects/704352985590/locations/us-west1/tuningJobs/8951840218672005120,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539fafd70>
resource name: projects/704352985590/locations/us-west1/tuningJobs/8146593086860951552,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539faeb60>
resource name: projects/704352985590/locations/us-west1/tuningJobs/4441256493441875968,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539faf370>
resource name: projects/704352985590/locations/us-west1/tuningJobs/6141218018216116224]

[16]: tuned_model_endpoint_name = sft_tuning_job.tuned_model_endpoint_name
tuned_model_endpoint_name
```

第五步：基于微调的模型进行远端部署

```
5]: sft_tuning_job.list()
[5]: [<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd53a020370>
resource name: projects/704352985590/locations/us-west1/tuningJobs/8669907845124194304,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215bee0>
resource name: projects/704352985590/locations/us-west1/tuningJobs/7530497139399458816,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215b850>
resource name: projects/704352985590/locations/us-west1/tuningJobs/6989220759184867328,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215ba0>
resource name: projects/704352985590/locations/us-west1/tuningJobs/383284530762481664,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd55215b820>
resource name: projects/704352985590/locations/us-west1/tuningJobs/5674169667992682496,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539fafb20>
resource name: projects/704352985590/locations/us-west1/tuningJobs/4340154200244617216,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539fafc10>
resource name: projects/704352985590/locations/us-west1/tuningJobs/8951840218672005120,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539fafd70>
resource name: projects/704352985590/locations/us-west1/tuningJobs/8146593086860951552,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539faeb60>
resource name: projects/704352985590/locations/us-west1/tuningJobs/4441256493441875968,
<vertexai.tuning._supervised_tuning.SupervisedTuningJob object at 0x7fd539faf370>
resource name: projects/704352985590/locations/us-west1/tuningJobs/6141218018216116224]

6]: tuned_model_endpoint_name = sft_tuning_job.tuned_model_endpoint_name
tuned_model_endpoint_name
[6]: 'projects/704352985590/locations/us-west1/endpoints/4014462088545042432'

7]: tuned_model = GenerativeModel(tuned_model_endpoint_name) x q
print(tuned_model)

<vertexai.generative_models.GenerativeModel object at 0x7fd53a020ee0>
/home/jupyter/.local/lib/python3.10/site-packages/vertexai/generative_models/_generative_models.py:433: UserWarning: This feature is deprecated as of June 24, 2025 and will be removed on June 24, 2026. For details, see https://cloud.google.com/vertex-ai/generative-ai/docs/deprecations/genai-vertexai-sdk.
  warning_logs.show_deprecation_warning()

8]: tuned_model_endpoint_name = sft_tuning_job.tuned_model_endpoint_name
tuned_model_endpoint_name
```

第六步：基于微调的模型进行调用

```
on June 24, 2026. For details, see https://cloud.google.com/vertex-ai/generative-ai/docs/deprecations/genai-vertexai-sdk.
warning_logs.show_deprecation_warning()

18]: tuned_model_endpoint_name = sft_tuning_job.tuned_model_endpoint_name
tuned_model_endpoint_name
[18]: 'projects/704352985590/locations/us-west1/endpoints/4014462088545042432'

28]: question = "如何利用数字技术增强酒店客房的市场推广效果？"
response = tuned_model.generate_content(question)

print(response.text)
利用数字技术增强酒店客房的市场推广效果可以从多个方面入手。以下是一些具体的方法：

1. **搜索引擎优化（SEO）**：通过优化酒店网站和在线列表，使其在搜索引擎结果中排名更高。使用相关的关键词、高质量的内容和用户友好的设计来吸引更多的有机流量。
2. **社交媒体营销**：利用社交媒体平台（如Facebook、Instagram、Twitter等）来推广酒店客房。发布吸引人的照片、视频和内容，与潜在客户互动，进行有针对性的广告投放。
3. **内容营销**：创建有价值的内容，如博客文章、视频和图片，展示酒店的独特之处和周边景点的吸引力。这有助于吸引潜在客户并建立品牌声誉。
4. **电子邮件营销**：建立客户数据库，通过电子邮件向潜在和现有客户发送个性化的优惠、促销信息和活动邀请。确保邮件内容吸引人，设计良好，并符合移动设备浏览。
5. **在线旅行社（OTA）优化**：确保酒店在主要的OTA平台（如Booking.com、Expedia等）上信息完整且最新。使用高质量的照片、吸引人的描述和竞争力的价格来吸引客户。
6. **用户生成内容（UGC）**：鼓励客户在社交媒体和评论平台上分享他们的体验，使用UGC来建立信任和真实性。可以通过竞赛、奖励或简单的感谢来激励用户分享。
7. **移动应用和网站优化**：确保酒店的网站和移动应用在各种设备上都能正常工作，并提供无缝的用户体验。使用移动广告来吸引移动用户。
8. **个性化推荐和优惠**：使用数据分析来了解客户偏好和预订历史，提供个性化的推荐和优惠，提高客户忠诚度和转化率。
9. **视频营销**：制作高质量的视频内容，展示酒店设施、客房环境和周边景点。视频可以在YouTube、Vimeo等平台上发布，也可以嵌入到酒店网站和社交媒体中。
10. **虚拟现实（VR）和增强现实（AR）**：提供虚拟现实体验让潜在客户在预订前体验酒店客房和设施。增强现实应用可以提供额外的信息或互动体验。
通过综合运用这些数字技术，酒店可以有效地增强客房的市场推广效果，吸引更多的客户并提高入住率。
70]: # Save the response from model in the Cloud Storage bucket
```

```
1 tuned_model_endpoint_name = sft_tuning_job.tuned_model_endpoint_name
2 tuned_model_endpoint_name
3
4 question = "如何利用数字技术增强酒店客房的市场推广效果?"
5 response = tuned_model.generate_content(question)
6
7 print(response.text)
```

模型部署证明：

Region						
us-west1 (Oregon)						
Filter Enter a property name						
<input type="checkbox"/> Name	Default version	Deployment status	Description	Type	Sourc	
<input type="checkbox"/> Hotel assiant-Gemini Tuning	1	✓ Deployed	—	◆ Large model	Verte	Studio
<input type="checkbox"/> Hotel assiant-Gemini Tuning	1	✓ Deployed	—	◆ Large model	Verte	Studio
<input type="checkbox"/> Hotel assiant-Gemini Tuning	1	✓ Deployed	—	◆ Large model	Verte	Studio

Online prediction

Endpoints Deployment resource pools

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

To create an endpoint, you need at least one machine learning model. [Learn more](#)

Region us-west1 (Oregon) ?

Endpoints + Create

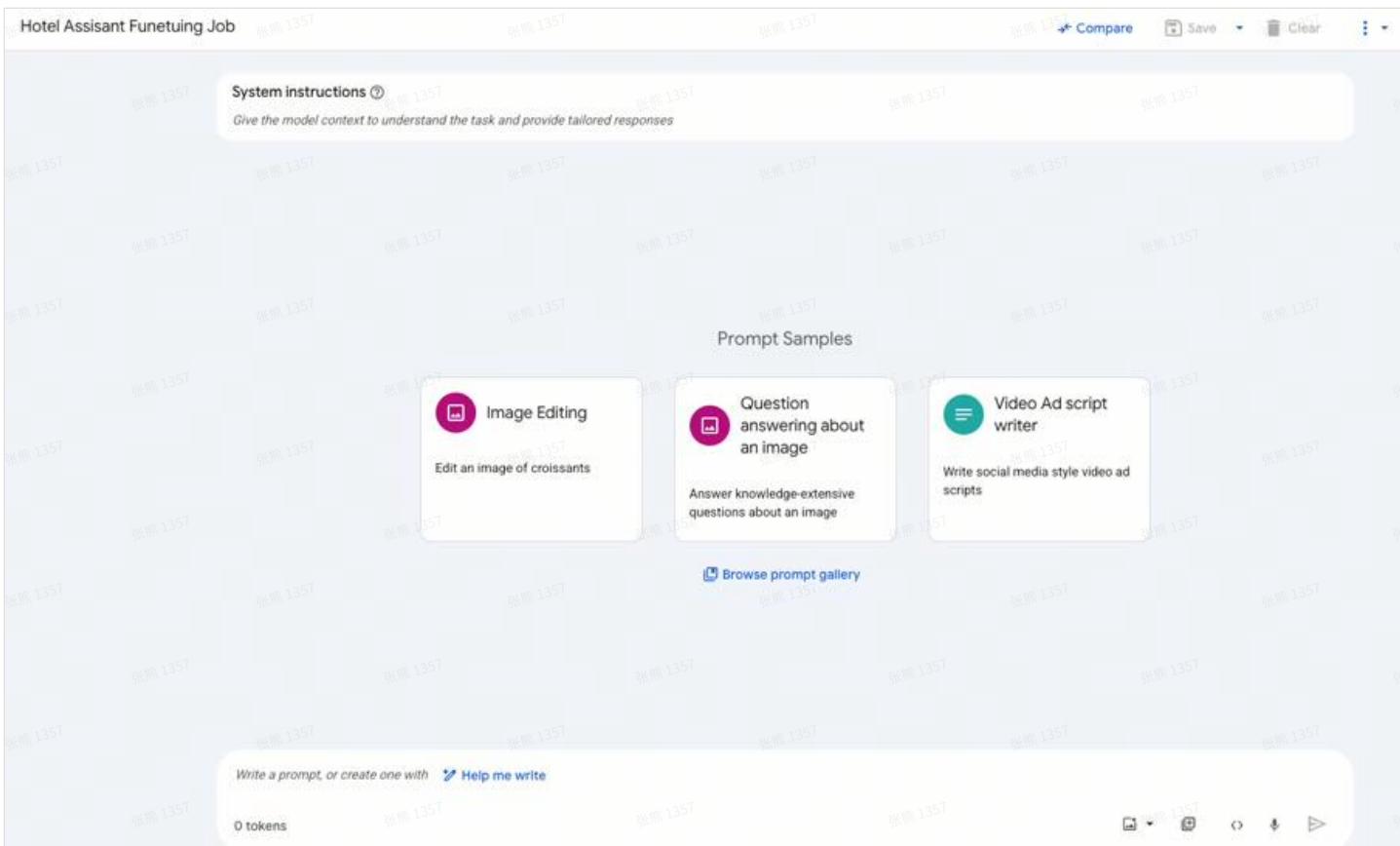
Filter Enter a property name

<input type="checkbox"/>	Name	ID	Status	Models	Deployment resource pool	Region	Monitoring	Most recent alerts	L
<input type="checkbox"/>	Hotel assiant-Gemini Tuning Job	1771106524161114112	Active	1	—	us-west1	Disabled	—	J 3
<input type="checkbox"/>	Hotel assiant-Gemini Tuning Job	2507445063236190208	Active	1	—	us-west1	Disabled	—	J 1
<input type="checkbox"/>	Hotel assiant-Gemini Tuning Job	3563539175854571520	Active	1	—	us-west1	Disabled	—	J 1
<input type="checkbox"/>	Hotel assiant-Gemini Tuning Job	2639175352336777216	Active	1	—	us-west1	Disabled	—	J 8
<input type="checkbox"/>	Hotel assiant-Gemini Tuning Job	578778522814775296	Active	1	—	us-west1	Disabled	—	J 8
<input type="checkbox"/>	Hotel assiant-Gemini Tuning Job	8527631865123700736	Active	1	—	us-west1	Disabled	—	J 8

案例case:

https://github.com/ChuanYang-AI/Demo3/blob/main/certification_materials/docs/hotel_ai_case_study.md

youtube视频demo录制:



FAQ: