Como fazer Fine-Tuning em modelos do Amazon Nova com Sagemaker AI

*Este é um guia baseado na sessão **AIM402 - Advanced model customization with Amazon Nova**, atendida por Bruno Vilardi e Alan Patrick da Dati no AWS Re:Invent 2025 no dia 01/12/2025, em Las Vegas - US.*

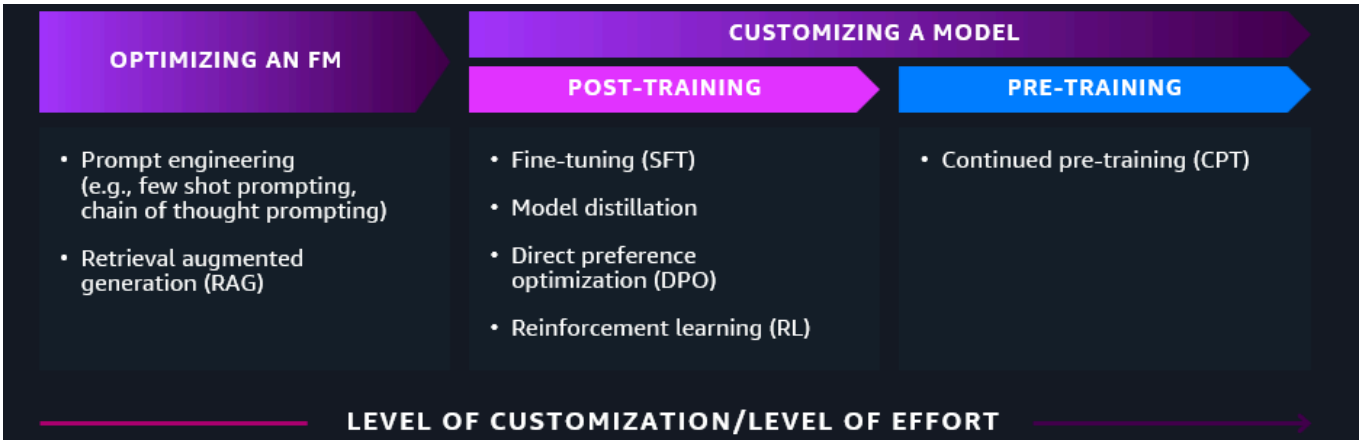
Resumo

O grande valor da IA Generativa só é desbloqueado com o uso dos dados da empresa e, para isso, existem diversas técnicas diferentes que estão sendo utilizadas por corporações ao redor do Globo. Neste artigo apresentaremos uma delas: o Fine-Tuning a partir de uma solução pronta do Sagemaker AI

1. Introdução

A personalização de modelos fundacionais (FMs) com os dados proprietários tornou-se uma etapa fundamental para empresas que buscam utilizar inteligência artificial generativa. Para isso, existem algumas técnicas, como:

- Engenharia de Prompt: Processo de definir e ajustar instruções claras e estruturadas para guiar o comportamento do modelo.
- RAG: Combina busca de dados em bases externas com a geração de texto, permitindo que o modelo responda usando informações atualizadas e específicas.
- Fine-tuning: Muda os pesos internos do modelo fundacional com dados rotulados (pares de pergunta e resposta esperada) para ajustar seu comportamento e aumentar precisão em tarefas específicas.
- Model Distilation: Utiliza um modelo maior ("mais inteligente") para ensinar um modelo menor a ser especialista em uma determinada área.
- Reinforcement Learning: Ajusta o modelo com base em feedback humano ou métricas de recompensa para melhorar decisões e alinhar o comportamento.
- Direct Preference Optimization (DPO): Técnica de RL (com Human In The Loop) que treina o modelo diretamente com preferências humanas definidas em pares, sem uso de funções de recompensa complexas.
- Continuous pre-training (CPT): Continuidade do treinamento do modelo, utilizando dados não rotulados de um domínio específico.



Atualmente, é muito falado sobre as técnicas de prompt engineering e RAG, principalmente para projetos iniciais, pois elas permitam um bom desempenho com baixo custo. Porém, a experiência da Arquiteta de Soluções da AWS em diversos de seus clientes demonstrou que esses métodos têm algumas limitações, como:

- Alucinações
- Dependência de ferramenta externa para buscas
- Latência
- Falta de adequação a vocabulário específico da indústria

Para superar essas limitações, o workshop apresentou outra técnicas de customização: o Fine-Tuning, aplicada ao Amazon Nova, o modelo generativo da AWS.

1.1 O workshop da AWS

O workshop foi conduzido em um pequeno grupo de 6 pessoas, com pessoas de diferentes países e empresas. Nesse grupo, uma Arquiteta de Soluções da AWS conduziu, dando explicação teórica e ajudando no passo a passo do laboratório.

Esse laboratório é basicamente uma conta AWS que acessamos e seguimos um guia de passo a passo para realizar a atividade proposta. Nesse caso, acessamos um Notebook Python a partir do Sagemaker AI para realizar todos os processos que serão descritos nesse artigo.

2. Preparando os dados

O processo de Fine-Tuning do Amazon Nova precisa que os dados estejam no formato JSON. Para esse workshop, utilizamos o dataset público de dados financeiros ibm-research/finqa. Transformamos esse dataset para o formato que o Amazon Nova espera:

```
{
    "system": [{"text": Content of the System prompt}],
    "messages": [
        {
            "role": "user",
            "content": ["text": Content of the user prompt]
        },
        {
            "role": "assistant",
            "content": ["text": Content of the answer]
        },
        ...
    ]
}
```

Essa transformação de dados é realizada pelo próprio notebook Python

O dataset já vem com separações dos dados em Treinamento, Validação e Teste, então podemos manter como veio.

3. Realizando o Fine-Tuning

Agora com os dados prontos, vamos começar o trabalho de Fine-Tuning do modelo. Para isso, utilizaremos o Estimator no Sagemaker AI (sim, muito parecido com a forma que realizamos treinamento de ML tradicional).

Para isso:

- Definimos o tipo e quantidade de instâncias
  - Utilizamos 1 ml.g5.12xlarge
- Definimos a imagem do container que será usado para o treinamento
  - Utilizamos a "708977205387.dkr.ecr.us-east-1.amazonaws.com/nova-fine-tune-repo:SM-TJ-SFT-ai-league", que contem todas as dependências necessárias
- Definimos qual modelo será usado como base para o Fine-Tuning
  - Utilizamos o Amazon Nova Lite
- Definimos a Receita*
  - Utilizamos a "fine-tuning/nova/nova_lite_g5_g6_12x_gpu_lora_sft"

*Aqui está a cereja do bolo, a AWS criou essas receitas padrão, com uma pré-configuração de todos os parâmetros necessários para realizar o Fine-Tuning do Amazon Nova. Esses artigos podem ser encontrados em link.

Com isso, definimos o Estimator:

```
estimator = PyTorch(
output_path=output_path,
base_job_name=job_name,
role=role,
disable_profiler=True,
debugger_hook_config=False,
instance_count=instance_count,
instance_type=instance_type,
training_recipe=recipe,
recipe_overrides=recipe_overrides,
max_run=432000,
sagemaker_session=sess,
image_uri=image_uri
)
```

E simplesmente executamos o comando

```
estimator.fit(inputs={"train": train_input, "validation": val_input})
```

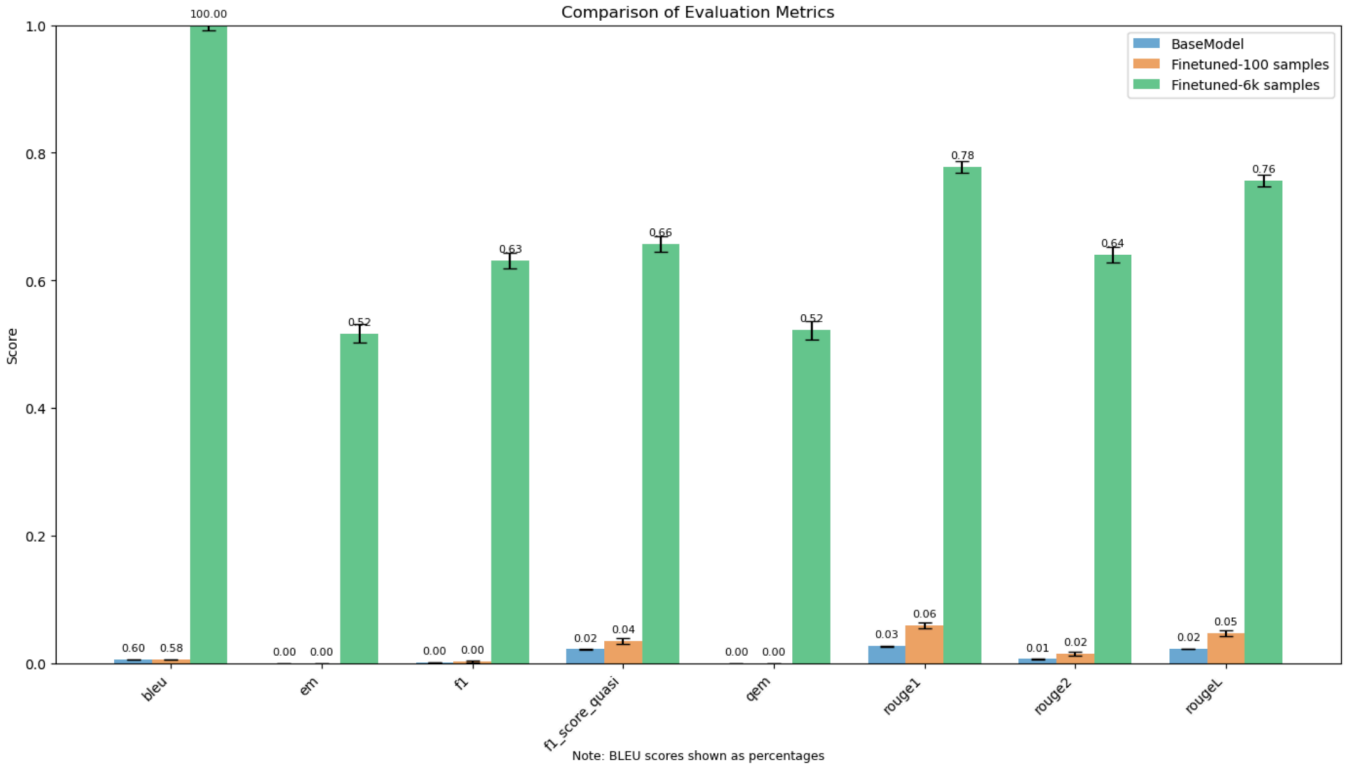
Com isso, a instância e o container são provisionados automaticamente, dando início ao processo de Fine-Tuning. Agora é só esperar o treinamento finalizar.

O Dataset original tem alguns milhares de registros, e por isso levaria mais de 6 horas para fazer o treinamento, então, no workshop reduzimos para apenas 100 registros, o que reduz para uma média de 35 minutos (o meu levou 32.5 minutos).

A recomendação da Arquiteta de Soluções conduzindo o workshop foi que para cenários reais, precisamos utilizar pelo menos 10 mil registros para um bom Fine-Tuning.

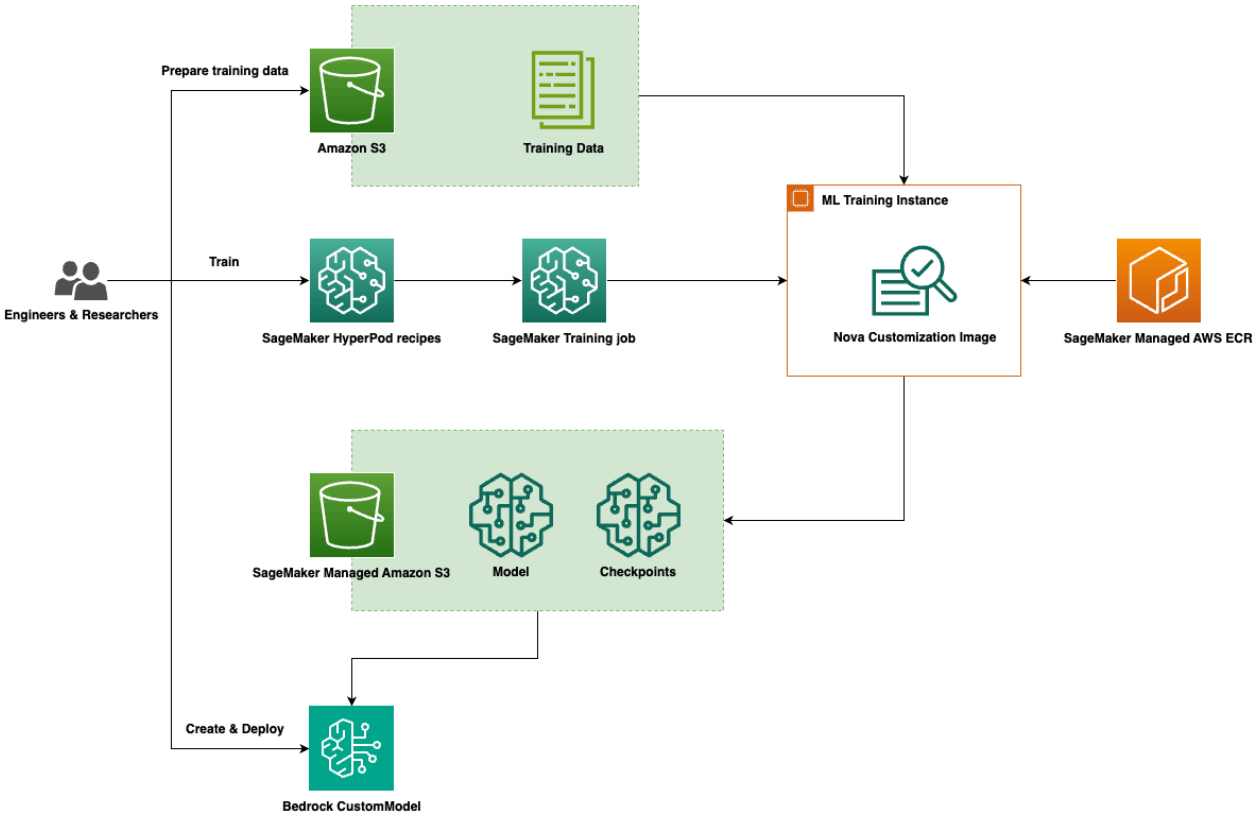
5. Resultados

Avaliamos o comportamento de 3 cenários: modelo Amazon Nova Lite original, modelo após Fine-Tuning com 100 amostras (que nós fizemos) e modelo após Fine-Tuning com 6k amostras (feito previamente pela AWS) e foi gerado o seguinte resultado:



Fica evidente que o Fine-Tuning aumenta na acuracidade e, quanto mais dados, melhor ficam as respostas. Esse resultado empírico deste laboratório foi fortemente reforçado pela Arquiteta de Soluções da AWS que conduziu, dizendo que seus clientes estão usando essa técnica amplamente. O principal caso de uso compartilhado por ela foi o de geração de código, em que o modelo é customizado para utilizar os padrões de código definidos e utilizados neste cliente.

5. Arquitetura final do workshop



Basicamente, utilizamos dados externos para atualizar os pesos do modelo do Amazon Nova Lite, conforme esse novo contexto. Interessante ressaltar que os pesos dos modelos da família Amazon Nova não são públicos, ou seja, todo o processo de treinamento é realizado em uma conta de serviço AWS e o deploy dele só pode ser feito dentro da AWS (com Bedrock CustomModel Import ou SageMaker AI).

6. Conclusão e principais Insights

- O RAG e Prompt Engineering são técnicas muito interessantes para iniciar na utilização corporativa de IA Generativa, porém, em maiores escalas podem começar a haver problemas de baixa acurácia devido a alucinação
- Quanto mais dados, melhor o resultado do Fine-Tuning
  - Mas lembrem-se da qualidade do dado! Não adianta colocar muitos dados que não possuam qualidade
- CPT (continuous pré-training) é uma técnica muito poderosa, porém com um custo inicial muito alto também. Poucos casos de uso fazem sentido utilizar.
- Fine-Tuning é uma abordagem muito interessante para customizar os modelos com seus dados, uma vez que permite:
  - Maior longevidade: os modelos especializados performam melhor do que os originais, por isso, empresas que tem modelos com Fine-Tuning podem se preocupar menos

com ter que ficar mudando o modelo a cada novo modelo que é lançado no mercado, já que o seu modelo vai provavelmente performar melhor.
    - Atualmente é possível realizar a inferência dos modelos com Fine-Tuning on demand, quebrando a barreira inicial de ter que fazer provisionado. Para mais informações, link
- O workshop foi especificamente para o Amazon Nova, porém podemos utilizar o mesmo processo para qualquer modelo com pesos públicos, basta fazermos os ajustes da receita