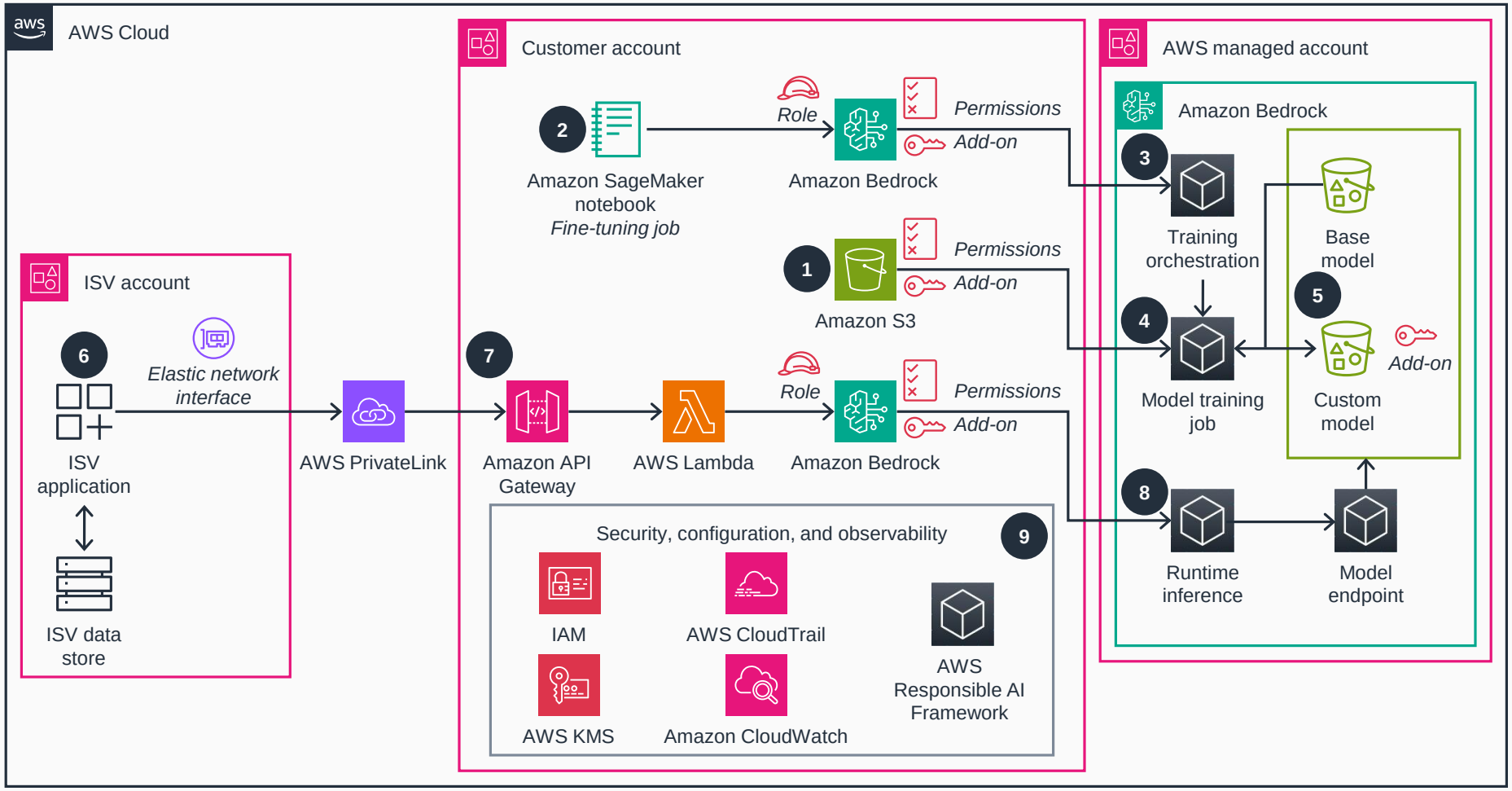
Guidance for Integrating a Custom Foundation Model with Advertising and Marketing ISVs on AWS

This architecture diagram shows how to securely import inferences into your ISV application from your customers' Amazon Bedrock FMs to centralize generative AI efforts and enrich your application. This slide details steps 1-7.



Model Customization

1. The customer moves any labeled examples that are needed for LLM customization to an **Amazon Simple Storage Service (Amazon S3)** bucket.
2. The customer uses **Amazon SageMaker** notebooks to write LLM-fine-tuning code. The customer then uses the **Amazon Bedrock** software development kit within the notebook to adjust model parameters and improve its performance on specific tasks or in certain domains. Alternatively, the customer can use the **Amazon Bedrock** console to run and monitor these tuning jobs.
3. The fine-tuning request will call the training orchestration component of **Amazon Bedrock**. This invokes a model training job.
4. The model training job uses a base model from an **S3** bucket managed by AWS and a labeled dataset from the customer's **S3** bucket. To improve the security posture, the customer can give **Amazon S3** access to the labeled datasets through virtual private cloud (VPC) configurations (such as subnets, security groups, or endpoints).
5. A new custom model is deployed in the fine-tuned model bucket, encrypted using **AWS Key Management Service (AWS KMS)** customer-managed keys. Only the customer can access the customized models, and no customer data is used to further train any **Amazon Bedrock** models.

Model Inference

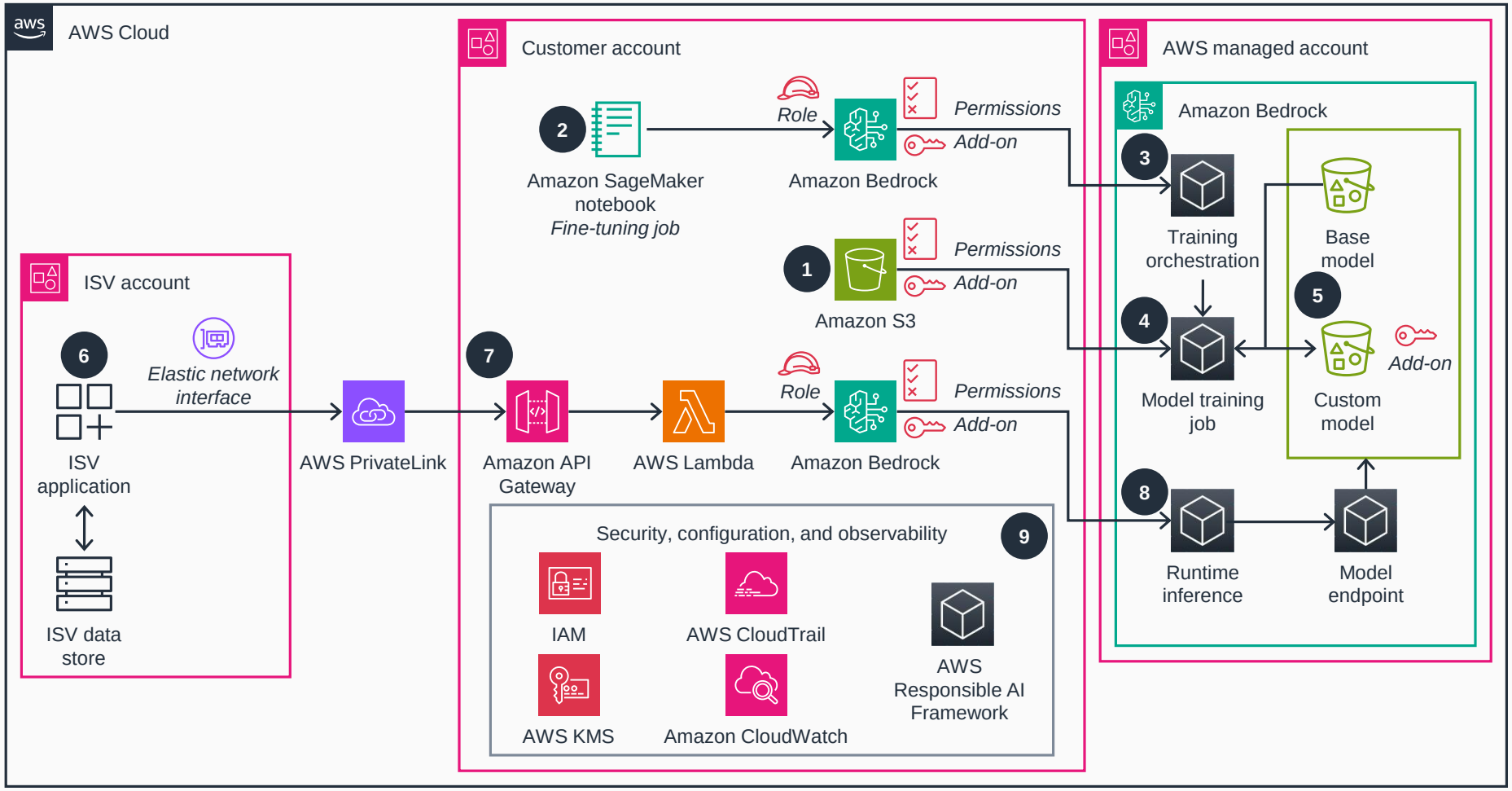
6. As the ISV, you can invoke the **Amazon Bedrock** API from your application to run inferences on available models. Use the inference response and store it in your application's data store.
7. Your customer creates a REST API on **Amazon API Gateway** as the entry point for you to access the fine-tuned LLM inference endpoint. An **AWS Lambda** function brokers the connection between **API Gateway** and the **Amazon Bedrock** inference endpoint. You can then access the **API Gateway** endpoint through **AWS PrivateLink** without exposing the traffic to internet.

Reviewed for technical accuracy April 15, 2024

AWS Reference Architecture

Guidance for Integrating a Custom Foundation Model with Advertising and Marketing ISVs on AWS

This architecture diagram shows how to securely import inferences into your ISV application from your customers' Amazon Bedrock FMs to centralize generative AI efforts and enrich your application. This slide details steps 8-9.



8 Your inference request is passed to the runtime inference component of **Amazon Bedrock**, based on customer-specified permissions. This sends a call to the correct provisioned capacity compute environment, where the model processes the request and returns the results. The same API can be used to derive inferences from base and custom models in a secure manner by configuring VPC settings.

Security, Configuration, and Observability

9 Use the following AWS services and resources to implement security and access control:

- **AWS Identity and Access Management (IAM)**: Control access to your customized FMs, allowing or denying access to specific FMs, as well as controlling which services can receive inferences and who can log into the **Amazon Bedrock** management console.
- **AWS KMS**: Your customized FMs are encrypted using **AWS KMS** keys and are encrypted in storage.
- **Amazon CloudWatch**: This service monitors logs and metrics across all the services used in this Guidance. You can also track input and output tokens using these metrics.
- **AWS CloudTrail**: This service monitors API activity and troubleshoots issues as you integrate other systems.
- Responsible AI: Reach out to your AWS team to learn more about building responsible generative AI applications.

Reviewed for technical accuracy April 15, 2024

AWS Reference Architecture