



# LaunchPad

## User Guide

### Authors:

Michael Yang      myang@logicworks.net  
Daan Grashoff    dgrashoff@logicworks.net  
Elmer Real        ereal@logicworks.net

### AIML Team

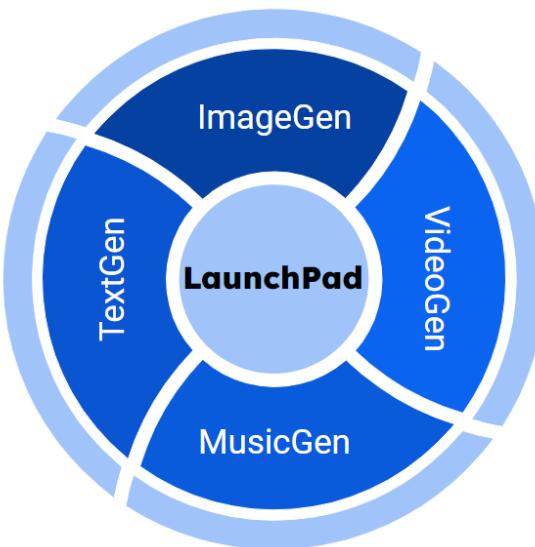
Date: Feb 2, 2024

# Table of Contents

<b>Table of Contents.....</b>	<b>1</b>
<b>About Launchpad.....</b>	<b>2</b>
<b>LaunchPad Solution Overview.....</b>	<b>2</b>
Security.....	5
Pre-requisites.....	6
Observations.....	7
CloudFormation Templates.....	8
Deploy Cloudformation template.....	8
VPC Template (Optional).....	13
LaunchPad Template.....	15
Increase service quota to run G5 on-demand instances.....	17
Access Endpoints.....	18
LaunchPad Pricing.....	19
Instances pricing.....	20
Monitoring.....	21
Cleanup Process.....	23
<b>TextGen.....</b>	<b>24</b>
Select Model.....	24
Download Model.....	24
Load Model.....	25
Setup Prompt Templates.....	26
Playground.....	27
Fine-tuning.....	28
Using the Fine-tuned Model.....	29
Evaluating the Fine-tuned Model:.....	30
Logicworks Extensions.....	31
S3 Data.....	31
Retrieval Augmented Generation.....	32
Deploy to Sagemaker Endpoint.....	34
Chat using Sagemaker Endpoint.....	35
<b>Technical Troubleshooting: Common Errors &amp; Tips.....</b>	<b>36</b>

## About LaunchPad

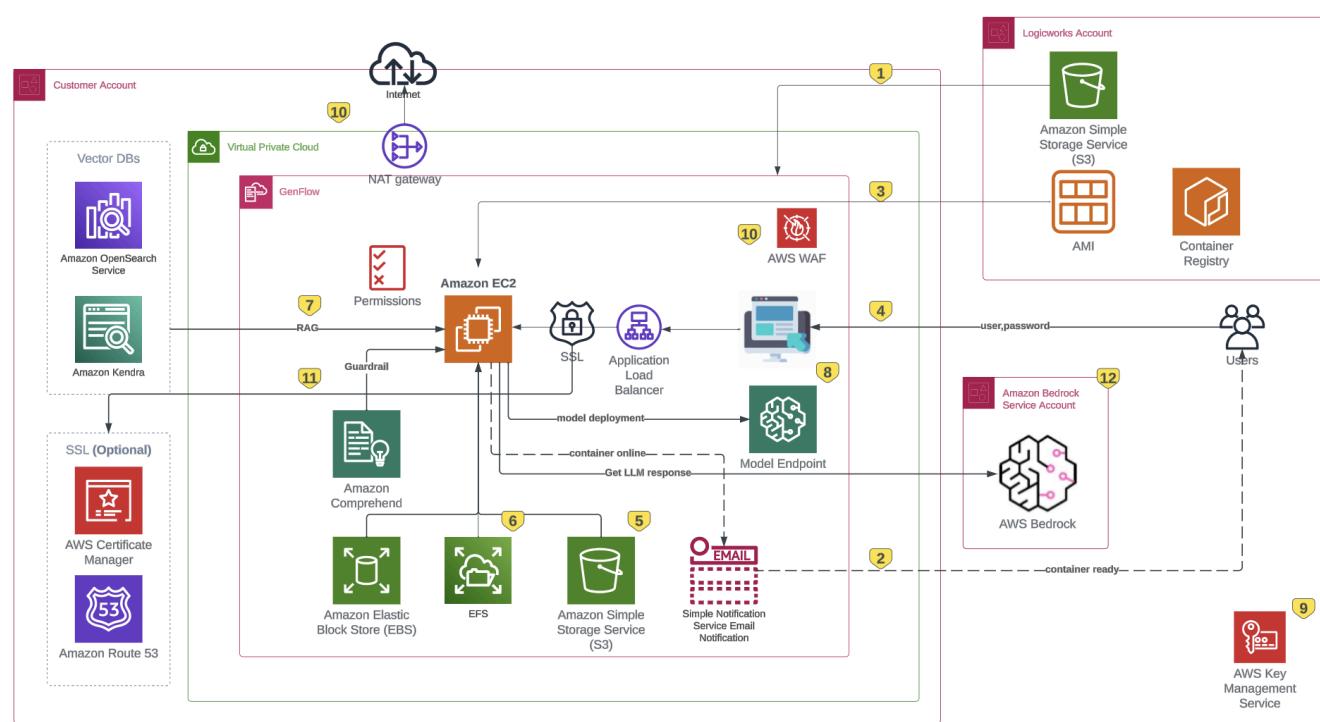
LaunchPad is a no-code tool designed to democratize access to generative AI models, specifically tailored for non-technical domain experts. This platform prioritizes data privacy and security, ensuring that sensitive information remains protected. LaunchPad encompasses a suite of applications including TextGen, ImageGen, VideoGen, and MusicGen, each delivered through AMI images curated by Logicworks. The solution provides a web UI interface playground, allowing users to explore the cutting-edge capabilities of state-of-the-art (SOTA) models in each respective modality.



Compared to AWS's Bedrock and Jumpstart, LaunchPad distinguishes itself by providing a seamless no-code environment. This environment grants users access to the entire array of models available in HuggingFace, all within their Virtual Private Cloud (VPC), without any data leaving their account. Moreover, LaunchPad offers the flexibility of customizable extensions, enabling users to integrate additional functionalities such as LORA fine-tuning, Kendra RAG integration, endpoint recommendation, and much more. This empowers users to tailor their AI experience to their specific needs and preferences.

# LaunchPad Solution Overview

The following image shows LaunchPad architecture and the different AWS services involved in the solution.



No	AWS Service	Description
1	Amazon IAM	AWS Identity and Access Management (IAM) is a web service that helps you securely control access to AWS resources. With IAM, you can centrally manage permissions that control which AWS resources users can access. The CloudFormation stack has been developed following the least privileged principle by enforcing permission to the EC2 instance through the proper IAM policies and roles.
2	Amazon EC2	Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. This computer service will host TextGen and ImageFlow.
3	Application Load Balancer	For security reasons, the EC2 instance will be deployed in a private subnet only accessible within the same VPC. The application load balancer provides a secure way to access the instance without exposing the instance itself to the internet.

4	AWS WAF	AWS WAF helps protect against common web exploits and bots that can affect availability, compromise security, or consume excessive resources. It will be associated with the Application Load Balancer to enhance LaunchPad security.
5	Amazon Sagemaker	Amazon SageMaker is a machine-learning platform that enables developers to create, train, and deploy machine-learning models in the cloud. SageMaker offers the ability to deploy machine learning models as endpoints, which can be invoked to make predictions on new data.
6	Amazon EBS	Amazon Elastic Block Store (Amazon EBS) is an easy-to-use, scalable, high-performance block-storage service designed for Amazon Elastic Compute Cloud (Amazon EC2). This volume will act as the principal file system of the EC2 instance.
7	Amazon EFS	Amazon Elastic File System (EFS) is designed to provide serverless, fully elastic file storage that lets you share file data without provisioning or managing storage capacity and performance. All data and models of TextGen/ImageFlow will be stored in an EFS that grows as needed with time.
8	Amazon S3	Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. You can provide the ARN of any existing S3 bucket with data to grant the proper permissions to download data from the instance.
9	Amazon SNS	Amazon Simple Notification Service (Amazon SNS) is a managed service that provides message delivery from publishers to subscribers. It will be used to notify users when the instance is ready.
10	AWS KMS	AWS Key Management Service (KMS) gives you centralized control over the cryptographic keys used to protect your data. The service is integrated with other AWS services, making it easier to encrypt data you store in these services and control access to the keys that decrypt it. This service will be used to enforce security at rest on the different storage services (EBS, EFS and S3 bucket). Only some AWS services will have the permissions to access it.
11	Amazon AMI	An Amazon Machine Image (AMI) is a template that contains a software configuration (for example, an operating system, an application server, and applications). From an AMI, you launch an instance, which is a copy of the AMI running as a virtual server in the cloud.

12	Elastic Container Registry	Amazon Elastic Container Registry (Amazon ECR) is an AWS managed container image registry service that is secure, scalable, and reliable. Amazon ECR supports private repositories with resource-based permissions using AWS IAM.
13	NAT Gateway	NAT Gateway is a highly available AWS managed service that makes it easy to connect to the Internet from instances within a private subnet in an Amazon Virtual Private Cloud (Amazon VPC). It allows instances to reach the internet without being reachable from the internet.
14	Amazon Kendra	Amazon Kendra is an intelligent enterprise search service that helps you search across different content repositories with built-in connectors.
15	Amazon OpenSearch Service	Amazon OpenSearch Service is an AWS-managed service that lets you run and scale OpenSearch clusters without having to worry about managing, monitoring, and maintaining your infrastructure, or having to build in-depth expertise in operating OpenSearch clusters.
16	AWS Bedrock	Amazon Bedrock is a fully managed service that makes high-performing foundation models (FMs) from leading AI startups and Amazon available for your use through a unified API.
17	Amazon Comprehend	Amazon Comprehend is a natural language processing (NLP) service that uses ML to extract meaning and insights in text.

## Security

Generative AI models available on the market have been trained with a large amount of data that makes them useful for general purposes. If you need to improve a generative model performance for specific use cases, it is required to re-train the model using LoRAs or use a retrieval-augmented generation (RAG) approach to consult an external knowledge base on the spot. Both alternatives involve the use of company proprietary data that may contain sensitive information about business processes or even personal information; therefore, security is a critical aspect to consider. LaunchPad enhances security in the following ways:

- Use of KMS keys to encrypt storage services like EBS, EFS, and S3. Key rotation has been enabled to automatically rotate KMS keys periodically.
- VPC is designed to allow internet access securely through the use of a NAT Gateway that allows the instance to download models from the internet, without being reachable directly.
- LaunchPad is accessed through an Application Load Balancer with WAF enabled. WAF is an AWS web application firewall that will protect LaunchPad web applications from various application layer attacks such as cross-site scripting (XSS), SQL injection, and cookie poisoning, among others. The WAF rules enabled are:
  - Admin protection managed rule group
    - The Admin protection rule group contains rules that allow you to block external access to exposed administrative pages. This might be useful if you run third-party software or want to reduce the risk of a malicious actor gaining administrative access to your application.
  - Known bad inputs managed rule group
    - The Known bad inputs rule group contains rules to block request patterns that are known to be invalid and are associated with exploitation or discovery of vulnerabilities. This can help reduce the risk of a malicious actor discovering a vulnerable application.
  - Linux operating system managed rule group
    - The Linux operating system rule group contains rules that block request patterns associated with the exploitation of vulnerabilities specific to Linux, including Linux-specific Local File Inclusion (LFI) attacks. This can help prevent attacks that expose file contents or run code for which the attacker should not have had access. You should evaluate this rule group if any part of your application runs on Linux. You should use this rule group in conjunction with the POSIX operating system rule group.
  - POSIX operating system managed rule group

- The POSIX operating system rule group contains rules that block request patterns associated with the exploitation of vulnerabilities specific to POSIX and POSIX-like operating systems, including Local File Inclusion (LFI) attacks. This can help prevent attacks that expose file contents or run code for which the attacker should not have had access. You should evaluate this rule group if any part of your application runs on a POSIX or POSIX-like operating system, including Linux, AIX, HP-UX, macOS, Solaris, FreeBSD, and OpenBSD.
- Amazon IP reputation list managed rule group
  - The Amazon IP reputation list rule group contains rules that are based on Amazon internal threat intelligence. This is useful if you would like to block IP addresses typically associated with bots or other threats. Blocking these IP addresses can help mitigate bots and reduce the risk of a malicious actor discovering a vulnerable application.
- Anonymous IP list managed rule group
  - The Anonymous IP list rule group contains rules to block requests from services that permit the obfuscation of viewer identity. These include requests from VPNs, proxies, Tor nodes, and web hosting providers. This rule group is useful if you want to filter out viewers that might be trying to hide their identity from your application. Blocking the IP addresses of these services can help mitigate bots and evasion of geographic restrictions.
- AWS WAF Bot Control rule group
  - The Bot Control managed rule group provides rules that manage requests from bots. Bots can consume excess resources, skew business metrics, cause downtime, and perform malicious activities.

You can get additional information about these WAF rules at the following link:

<https://docs.aws.amazon.com/waf/latest/developerguide/aws-managed-rule-groups-list.html>

## Pre-requisites

- AWS account with service quota available to run **G5 on-demand instances**. If your account is brand new, or you haven't used G5 instances before, please refer to this section.
- VPC and at least three subnets with internet access. If you don't have an existing VPC, you can deploy a simple VPC with public subnets.
- Provide your AWS Account to Logicworks team to request permission to access the AMI image and ECR Container.
- Currently, the AMI is only available in the region: us-east-2
- **(Optional)** Existing S3 bucket to upload/download models.
- **(Optional)** Existing EFS and Security group associated with its mount targets.

## Observations

- Sometimes the selected AZ **runs out of capacity to create GPU instances**. If that is the case, **delete** the CloudFormation stack and **try again in another Availability Zone**.
- If you did not provide a S3 bucket name, a new one will be created. If you are re-creating the stack, make sure to provide the bucket name to import it, if not, it will fail trying to create another bucket with the same name.
- The EFS volume and S3 bucket, created by the CloudFormation stack, will not be deleted. Make sure to delete them to avoid being charged.
- **IMPORTANT:** The CloudFormation stack name should be all lowercase.

## CloudFormation Templates

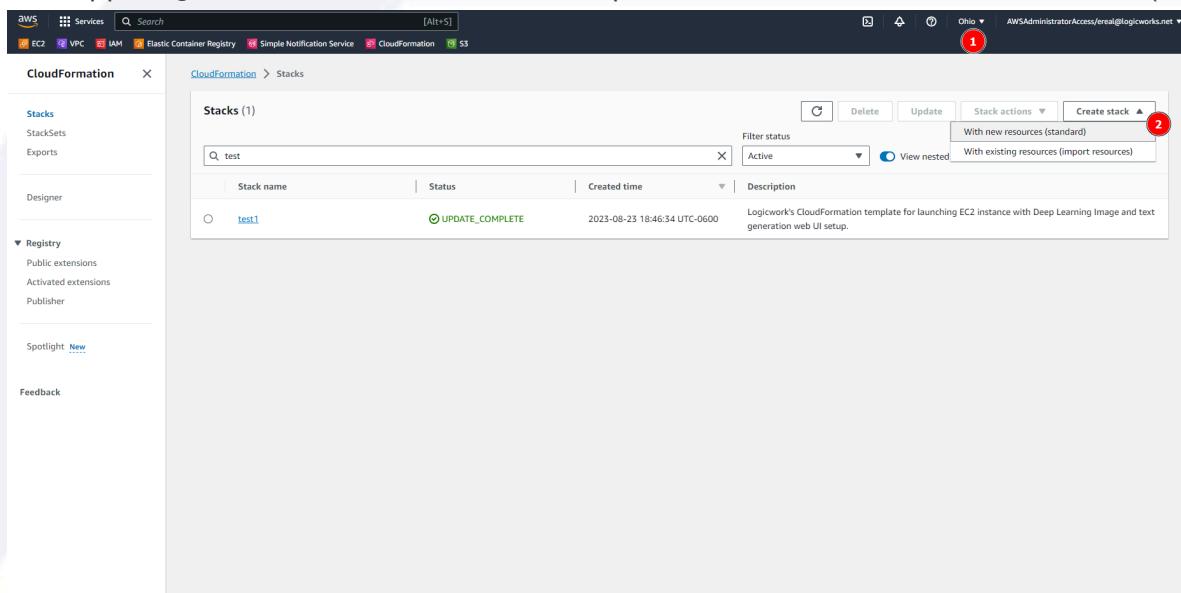
LaunchPad is deployed using best practices and enforcing security through infrastructure as code, to learn how to deploy a CloudFormation stack you can check this section: Deploy CloudFormation template.

As a prerequisite for deploying LaunchPad, it is required to have a VPC in place with at least 3 private subnets and 3 public subnets, where the resources can be deployed and internet access. If you don't have one, you can use this [VPC Template](#) to deploy one in your AWS account. Then you can use the [LaunchPad Template](#) to deploy the compute resources and unleash the Generative AI capabilities for your organization.

## Deploy Cloudformation template

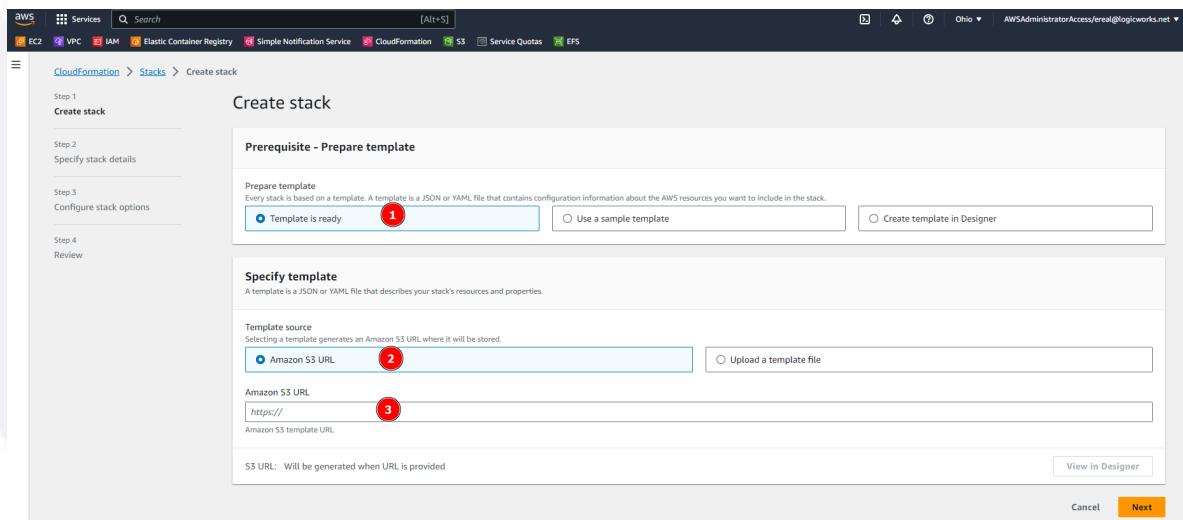
### THROUGH AWS CONSOLE

- Go to AWS Console
- Go to CloudFormation page <https://console.aws.amazon.com/cloudformation/>
- From the navigation bar, select the us-east-2 Region.
- On the upper right, click on the Create Stack dropdown and Click on With new resources (standard).



- Select template is ready option.
- Choose Amazon S3 URL option.
- In the Amazon S3 URL field, paste the template URL provided in the next sections.

# LaunchPad - User Guide



Prerequisite - Prepare template

Prepare template  
Every stack is based on a template. A template is a JSON or YAML file that contains configuration information about the AWS resources you want to include in the stack.

Template is ready 1

Use a sample template

Create template in Designer

Specify template  
A template is a JSON or YAML file that describes your stack's resources and properties.

Template source  
Selecting a template generates an Amazon S3 URL where it will be stored.

Amazon S3 URL 2

Upload a template file

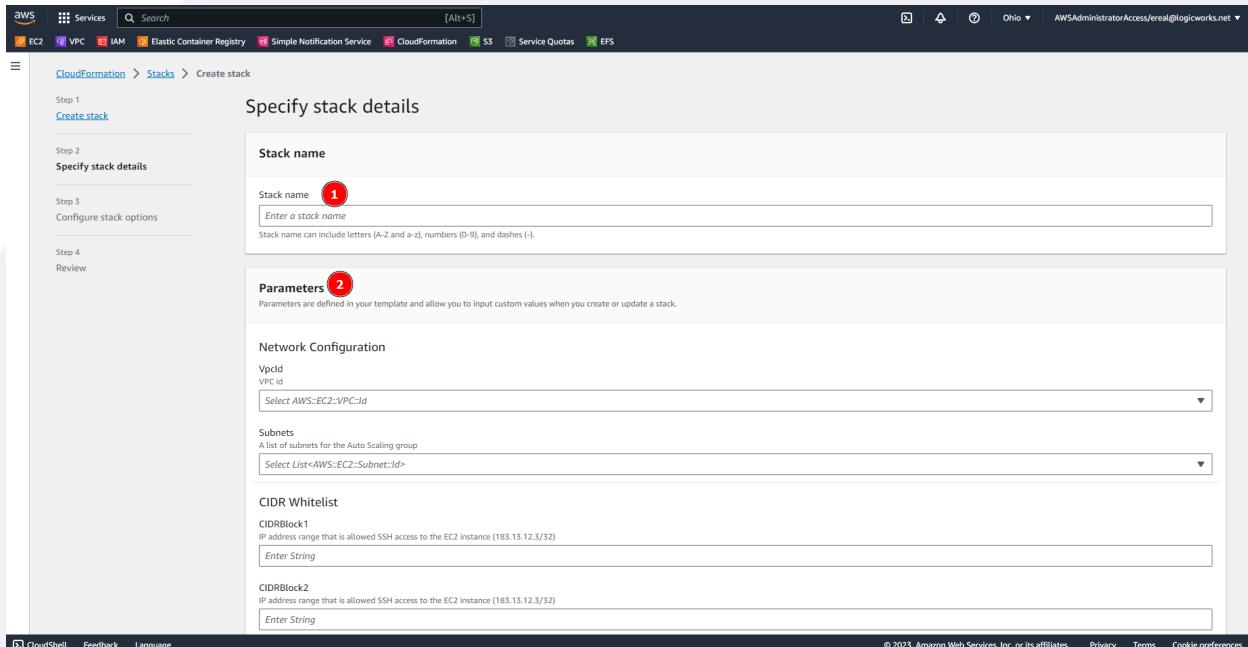
Amazon S3 URL  
3  
Amazon S3 template URL

S3 URL: Will be generated when URL is provided

[View in Designer](#)

[Cancel](#) Next

- Give a name to the CloudFormation stack to deploy. **It must be unique and lowercase.**
- Set the parameters values of the CloudFormation Stack. Check the Parameters Table of the respective CloudFormation stack.



Specify stack details

Stack name 1  
  
Stack name can include letters (A-Z and a-z), numbers (0-9), and dashes (-).

Parameters 2  
Parameters are defined in your template and allow you to input custom values when you create or update a stack.

Network Configuration  
VpcId  
VPC Id

Subnets  
A list of subnets for the Auto Scaling group

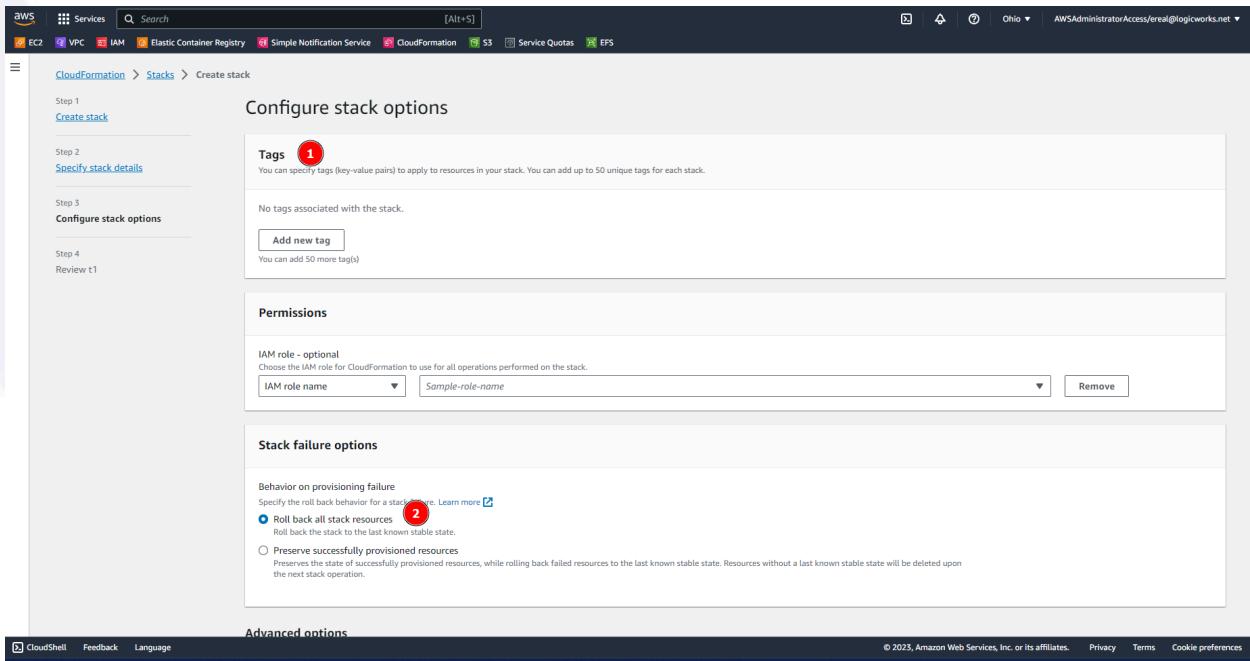
CIDR Whitelist  
CIDRBlock1  
IP address range that is allowed SSH access to the EC2 instance (183.13.12.3/32)  
  
CIDRBlock2  
IP address range that is allowed SSH access to the EC2 instance (183.13.12.3/32)

Roll back all stack resources

[Next Step](#)

- Add tags to add to the AWS Resources created by the CloudFormation stack.
- Select the option Roll back all stack resources.

# LaunchPad - User Guide



**Configure stack options**

**Tags** 1  
 You can specify tags (key-value pairs) to apply to resources in your stack. You can add up to 50 unique tags for each stack.  
 No tags associated with the stack.  
  
 You can add 50 more tag(s)

**Permissions**

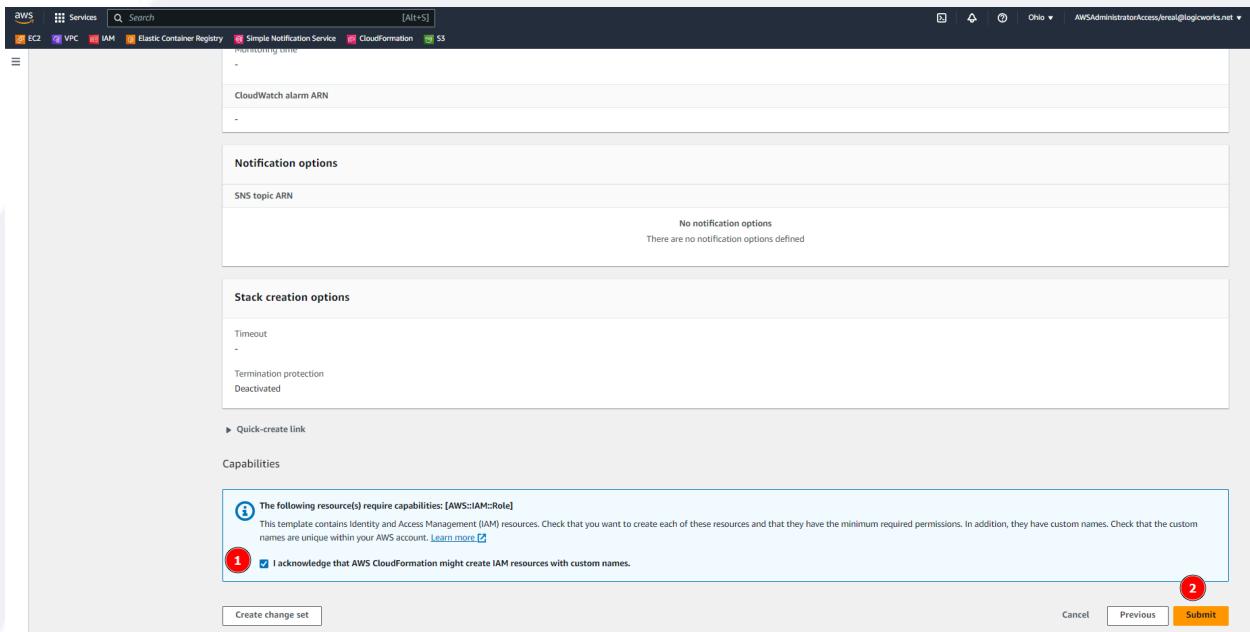
IAM role - optional  
 Choose the IAM role for CloudFormation to use for all operations performed on the stack.  
 IAM role name: Sample-role-name ▼ Remove

**Stack failure options**

Behavior on provisioning failure  
 Specify the roll back behavior for a stack. [Learn more](#) 2  
 Roll back all stack resources  
 Roll back the stack to the last known stable state.  
 Preserve successfully provisioned resources  
 Preserves the state of successfully provisioned resources, while rolling back failed resources to the last known stable state. Resources without a last known stable state will be deleted upon the next stack operation.

**Advanced options**

- Check the box to acknowledge the creation of some IAM roles and policies.
- Click the Submit button.



CloudWatch alarm ARN

**Notification options**  
 SNS topic ARN  
 No notification options  
 There are no notification options defined

**Stack creation options**

Timeout  
 -  
 Termination protection  
 Deactivated

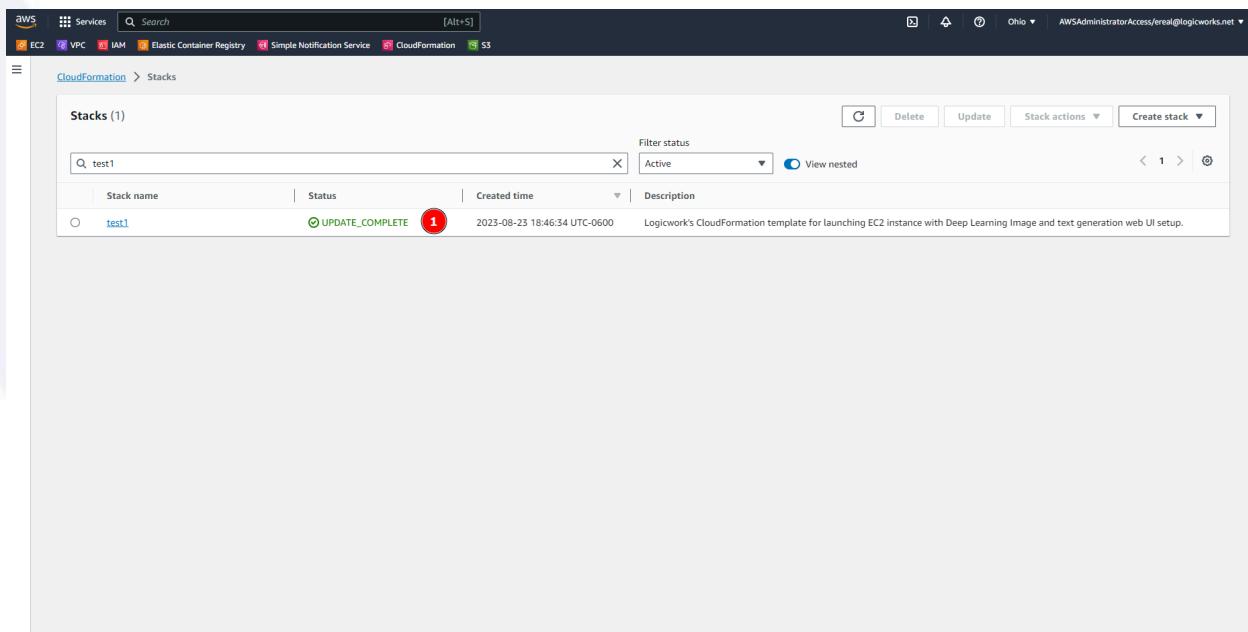
► Quick-create link

**Capabilities**

1  Acknowledge that AWS CloudFormation might create IAM resources with custom names. 2

Create change set Cancel Previous Submit

- Wait until the CloudFormation stack gets to the Create\_Complete or Update\_complete status.



## USING AWS CLI

You will need a terminal with access to AWS Services (Through IAM roles or Access Keys)

```
#!/bin/bash

# (Optional) Configure your access keys in the terminal
aws configure

# Set your stack's parameters
stack_name=<StackName>
template_url=<TemplateUrl>
region=<AWSRegion>

# Create the cloudformation stack
aws cloudformation create-stack \
--stack-name $stack_name \
--template-url $template_url \
--capabilities CAPABILITY_IAM \
--region $region

# Update the cloudformation stack
aws cloudformation update-stack \
--stack-name $stack_name \
--template-url $template_url \
--capabilities CAPABILITY_IAM \
--region $region
```

```
# Wait until update process is done...
aws cloudformation wait stack-update-complete \
--stack-name $stack_name --region $region
echo ">>>> Done..."
```

```
# Destroy the stack
aws cloudformation delete-stack --stack-name $stack_name --region $region
```

## OBSERVATIONS:

- If the CloudFormation stack fails the first time, you would need to delete the stack and start over the creation process.
- If you get the error "**Bucket name should not contain uppercase characters**", delete the stack and create a new one using only lowercase for the stack name.
- If you get the error: "**You have requested more vCPU capacity than your current vCPU limit of 0 allows for the instance bucket that the specified instance type belongs to. Please visit <http://aws.amazon.com/contact-us/ec2-request> to request an adjustment to this limit**". You would need to enable vCPU capacity in that AWS region using [these instructions](#).

## VPC Template (Optional)

### TEMPLATE URL

[https://genflow-artifacts.s3.us-east-2.amazonaws.com/templates/network\\_private\\_main.yml](https://genflow-artifacts.s3.us-east-2.amazonaws.com/templates/network_private_main.yml)

### DIAGRAM



### PARAMETERS

Name	Type	Description
Prefix	String	Keyword that identifies each resource created from different stacks that uses the same template.
Cidr	String	IP block for the VPC. Default: 10.0.0.0/16

## OUTPUTS

Name	Value
VpcID	ID of the VPC.
PublicASubnetID	Public subnet on Availability Zone A
PublicBSubnetID	Public subnet on Availability Zone B
PublicCSubnetID	Public subnet on Availability Zone C
PrivateASubnetID	Private subnet on Availability Zone A
PrivateBSubnetID	Private subnet on Availability Zone B
PrivateCSubnetID	Private subnet on Availability Zone C

## LaunchPad Template

### TEMPLATE URL

[https://genflow-artifacts.s3.us-east-2.amazonaws.com/templates/genflow\\_production\\_genflow\\_production\\_release-v1.yml](https://genflow-artifacts.s3.us-east-2.amazonaws.com/templates/genflow_production_genflow_production_release-v1.yml)

### PARAMETERS

Name	Type	Description
VpcId	AWS::EC2::VPC::Id	ID of the VPC where the instance will be deployed.
PublicSubnetA	AWS::EC2::Subnet::Id	ID of the Subnet to deploy the Application Load Balancer. Make sure that the subnet belongs to the VPC selected.
PublicSubnetB	AWS::EC2::Subnet::Id	ID of the Subnet to deploy the Application Load Balancer. Make sure that the subnet belongs to the VPC selected.
PublicSubnetC	AWS::EC2::Subnet::Id	ID of the Subnet to deploy the Application Load Balancer. Make sure that the subnet belongs to the VPC selected.
PrivateSubnetA	AWS::EC2::Subnet::Id	ID of the Subnet to deploy the instance. It is required to have Internet access. (Internet gateway or through a NAT Gateway/Instance). Make sure that the subnet belongs to the VPC selected.
PrivateSubnetB	AWS::EC2::Subnet::Id	ID of the Subnet to deploy the instance. It is required to have Internet access. (Internet gateway or through a NAT Gateway/Instance). Make sure that the subnet belongs to the VPC selected.
PrivateSubnetC	AWS::EC2::Subnet::Id	ID of the Subnet to deploy the instance. It is required to have Internet access. (Internet gateway or through a NAT Gateway/Instance). Make sure that the subnet belongs to the VPC selected.
PublicIP1	String	For security reasons, we recommend that you pass your public IP in this parameter to whitelist it and provide access only to you (i.e. 183.13.12.3). If you want to provide access to everyone on the internet,

# LaunchPad - User Guide

		you can write "PUBLIC". For production use, we do not recommend making it PUBLIC.
PublicIP2	String	For security reasons, we recommend that you pass your public IP in this parameter to whitelist it and provide access only to you (i.e. 183.13.12.3). If you want to provide access to everyone on the internet, you can write "PUBLIC". For production use, we do not recommend making it PUBLIC.
PublicIP3	String	For security reasons, we recommend that you pass your public IP in this parameter to whitelist it and provide access only to you (i.e. 183.13.12.3). If you want to provide access to everyone on the internet, you can write "PUBLIC". For production use, we do not recommend making it PUBLIC.
Email1ToNotify	String	Email that will receive an email notification once the instance is ready to be used.
Email2ToNotify	String	Email that will receive an email notification once the instance is ready to be used.
Email3ToNotify	String	Email that will receive an email notification once the instance is ready to be used.
AppType	String	Application to deploy on the EC2 instance.
AuthCreds	String	Authentication credentials for TextGen. Value should be like username:password; or comma-delimit multiple like u1:p1,u2:p2,u3:p3
GrafanaPassword	String	Authentication credentials for Grafana..
TextGenerationModels	String	Comma separated list of hugging face models to download. Default: NousResearch/Llama-2-7b-chat-hf
StableDiffusionModels	String	Comma separated list of hugging face models to download. Default: stabilityai/stable-diffusion-xl-base-1.0
InstanceType	String	Instance Size, it is required to have GPU available, G5 family is recommended.

InstanceAZ	String	AZ where the EC2 instance will be deployed.
StorageCapacity	Number	Storage capacity for the EBS
EfsId	String	(Optional) ID of existing EFS storage volume.
EfsSg	String	(Optional) Security group to allow access to the ec2 instance.
ExistingS3BucketName	String	Bucket name to allow EC2 to download and upload files.
LogicworksAccount	String	ID of the AWS Account where the Custom AMI and the Custom Docker Image were published.
LogicworksSupportPublicIP	String	For security reasons, we recommend that you pass your public IP in this parameter to whitelist it and provide access only to Logicworks support engineer (i.e. 183.13.12.3). If you want to provide access to everyone on the internet, you can write "PUBLIC". For production use, we do not recommend making it PUBLIC.

## OBSERVATIONS

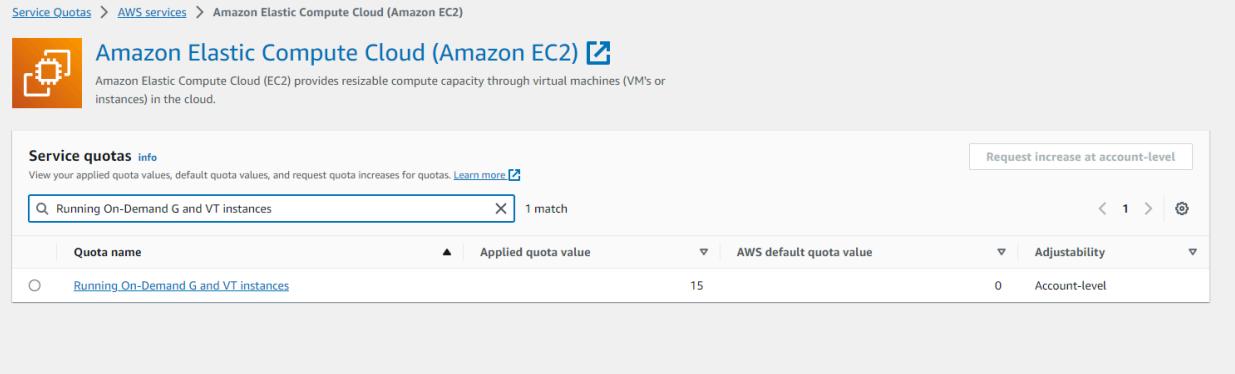
- If you have an existing EFS, you can reuse it with the current CloudFormation stack. You would need to have a Security Group associated with the mount targets and provide the **EfsID** and the security group ID (**EfsSg**) when deploying the stack.
- If you have an existing S3 bucket with models and LoRas previously trained, you can reuse it with the CloudFormation stack, providing the name as the **ExistingS3BucketName** parameter.
- You will need to provide the KMS key used when creating the EFS and S3 Bucket.

## Increase service quota to run G5 on-demand instances

- Open the Amazon EC2 console at  
<https://console.aws.amazon.com/servicequotas/home/services/ec2/quotas>
- From the navigation bar, select the Region in which to launch your resources. You can select any Region that's available to you, regardless of your location.
- Search for Running On-Demand G and VT instances
- Submit a quota increase request.
  - The quota depends on the vCPU number of the G5 instance size that you want to deploy. We recommend requesting a quota increase of 15, but you can always come back and request another quota increase if needed.

Instance Size	GPU	GPU Memory (GiB)	vCPUs
g5.xlarge	1	24	4
g5.2xlarge	1	24	8
g5.4xlarge	1	24	16
g5.8xlarge	1	24	32
g5.16xlarge	1	24	64

- This quota is regional. If you face a capacity limit error when trying to deploy the CloudFormation Stack or starting again a stopped instance. That is because sometimes availability zones run out of capacity to deploy certain EC2 instance sizes, and you would need to wait until there is available capacity, or change the instance size.



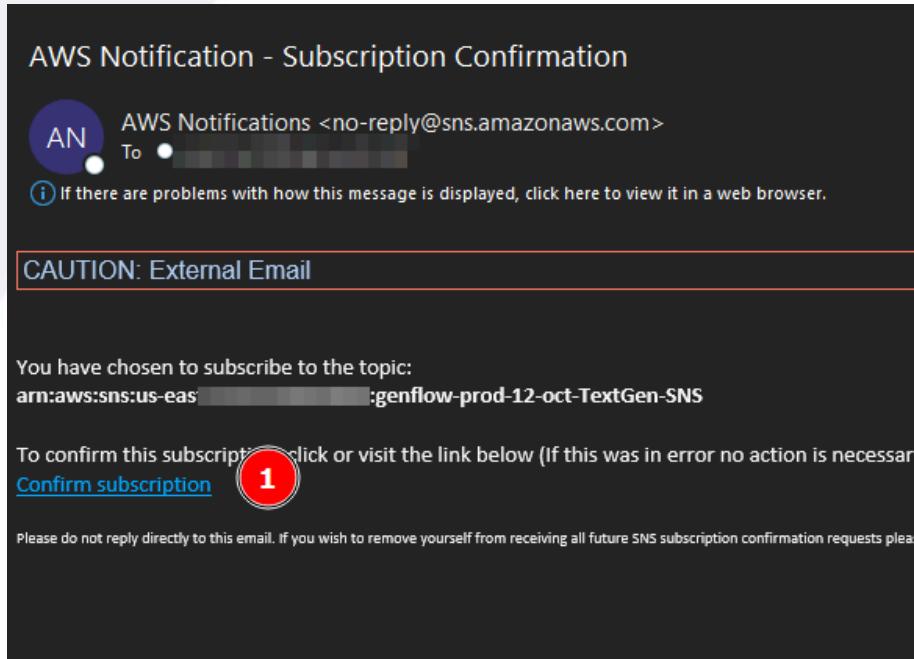
The screenshot shows the AWS Service Quotas console. The URL is [Service Quotas > AWS services > Amazon Elastic Compute Cloud \(Amazon EC2\)](#). The page title is "Amazon Elastic Compute Cloud (Amazon EC2)". A search bar at the top has "Running On-Demand G and VT instances" entered, with "1 match". Below the search is a table with one row:

Quota name	Applied quota value	AWS default quota value	Adjustability
Running On-Demand G and VT instances	15	0	Account-level

There is a "Request increase at account-level" button in the top right corner.

## Access Endpoints

Once you have deployed the LaunchPad CloudFormation template, it will create a CloudFormation stack with all the resources required. Each email registered will receive a confirmation email for the SNS subscription. Please click the Confirm subscription link to be able to receive your credentials and endpoints via email.



Once your instance is ready, you will receive an email like this one:

# LaunchPad - User Guide

AN AWS Notifications <no-reply@sns.amazonaws.com>  
To: [REDACTED]

[This message originated from an external sender outside your organization]

Dear User,

Your Rapidscale GenFlow instance has been configured with TextGen, and it is ready for you to unleash the power of GenerativeAI in your organization.

Please connect to:

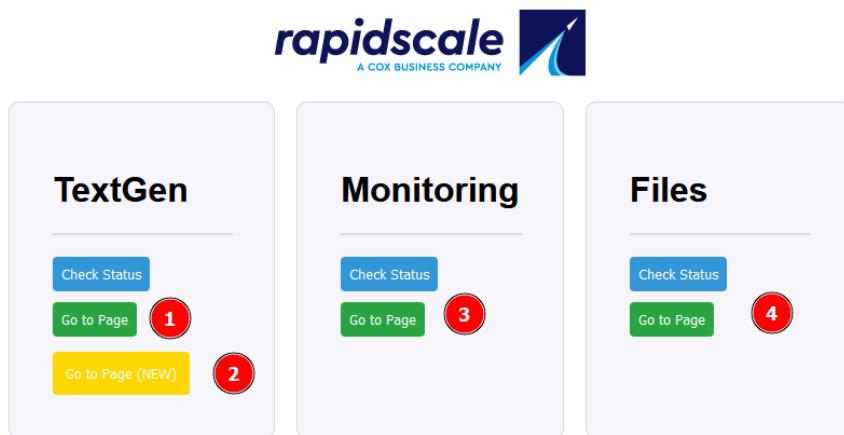
[REDACTED] and login with your credentials. If you don't know your credentials, please contact your GenFlow Administrator.

Thanks,  
Best Regards.

--  
If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:  
[REDACTED]

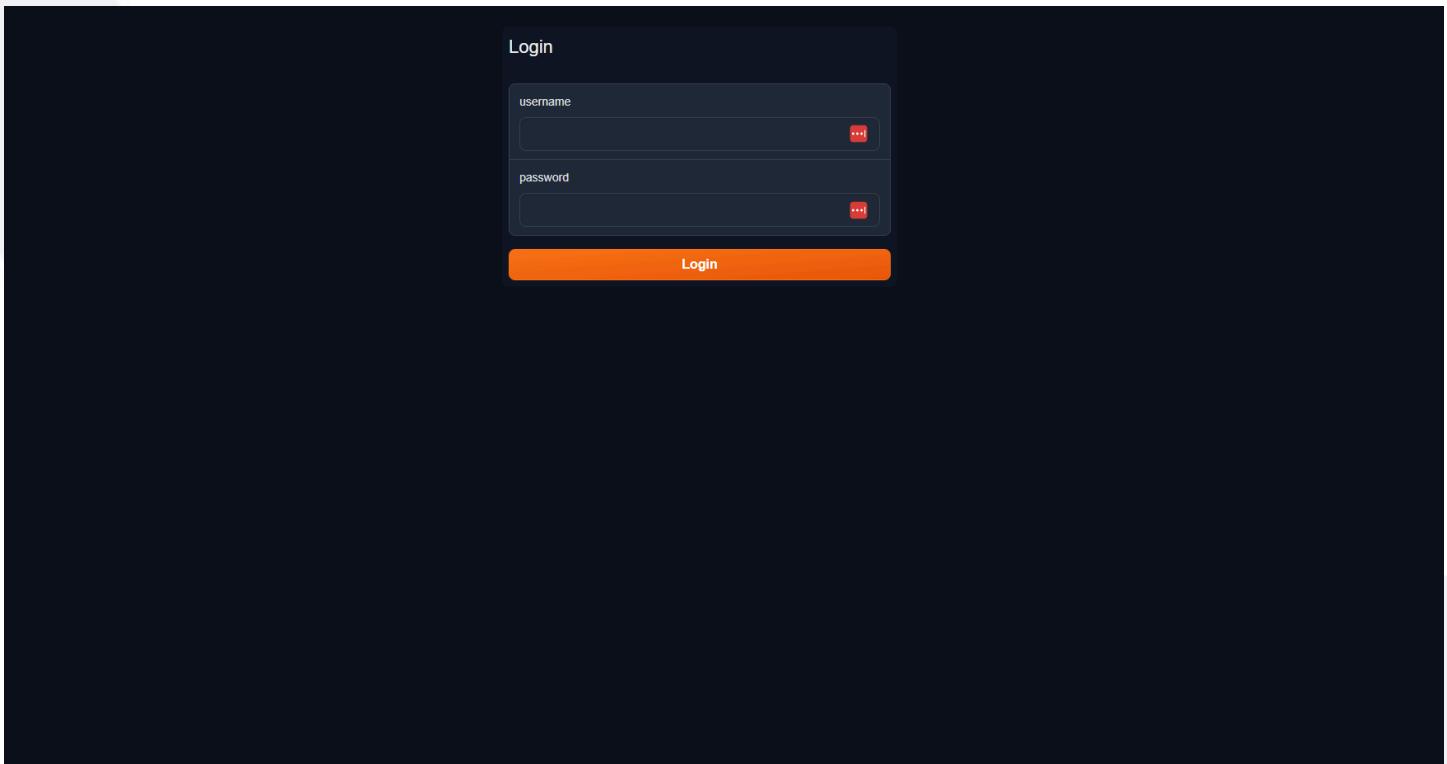
Please do not reply directly to this email. If you have any questions or comments regarding this email, please contact us at <https://aws.amazon.com/support>

The link that you have received will take you to the LaunchPad landing page

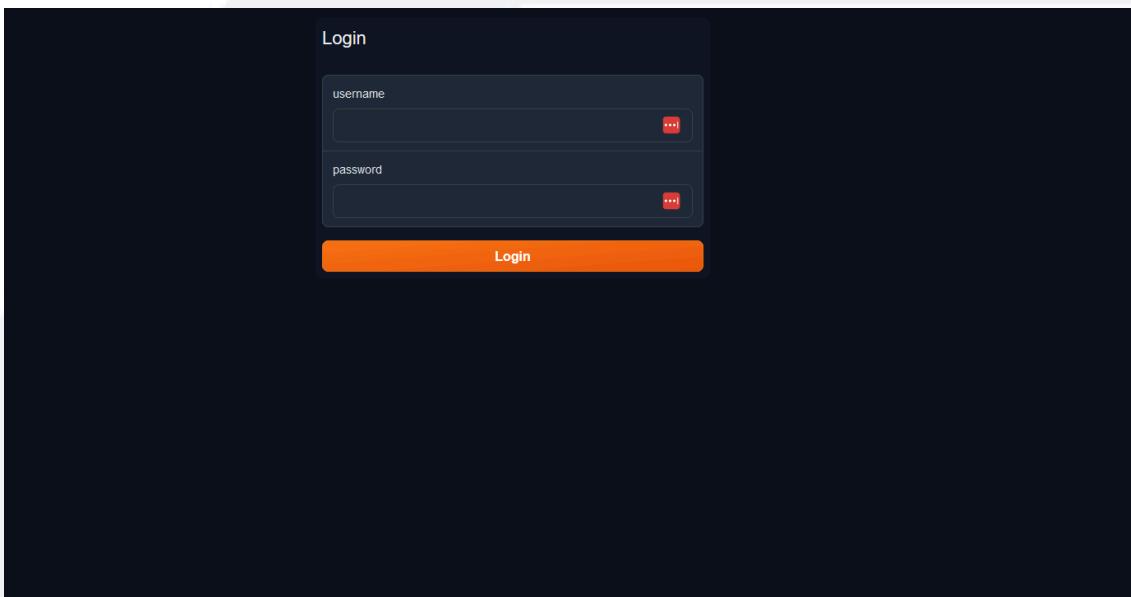


- Here you will be able to access to the different tools available to you on LaunchPad
  1. Go to TextGen page

# LaunchPad - User Guide

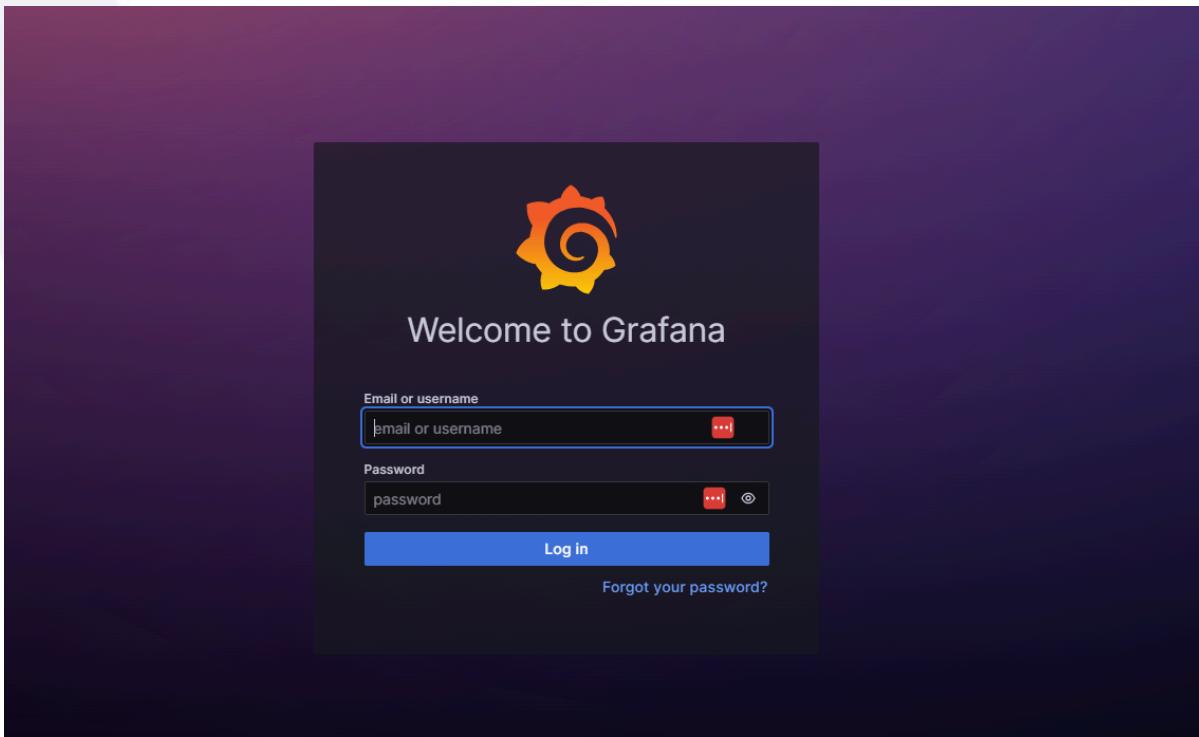


2. Go to the new TextGen page that comes with a web file system manager

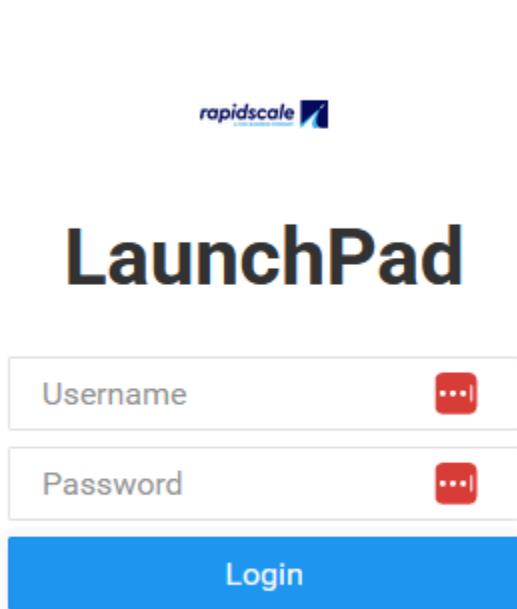


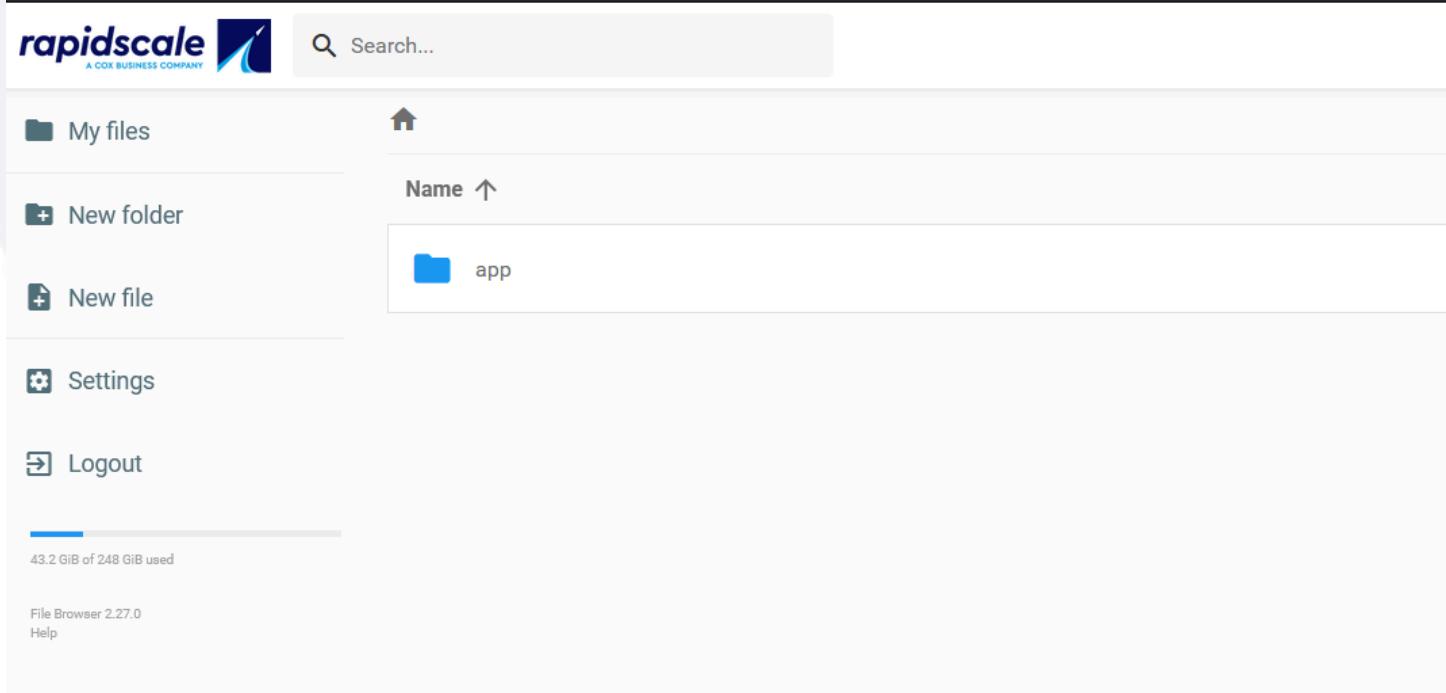
The image shows the same dark-themed login interface as the previous screenshot. A watermark for "rapidscale" is visible in the bottom right corner. To the right of the login form, the text "LaunchPad" is displayed above a smaller screenshot of the login interface, which has a blue "Login" button.

3. Go to Grafana to see the metrics of the server



4. Got directly to the web file system manager

A screenshot of the LaunchPad login page. The page has a white background with a "rapidscale" logo at the top left. The main title "LaunchPad" is centered in large, bold, black font. Below the title are two input fields: "Username" and "Password", each with a red "..." button to the right. A large blue "Login" button is centered below the fields.



The screenshot shows the LaunchPad file browser interface. On the left, there is a sidebar with the following options:

- My files
- New folder
- New file
- Settings
- Logout

Below the sidebar, there is a progress bar indicating "43.2 GiB of 248 GiB used". At the bottom, it says "File Browser 2.27.0" and "Help".

The main area displays a file tree with a single folder named "app".

- To access your instance through SSH use your instance-ip and the port 22. You will be able to connect through ssh only if you set a private key during the cloudformation deployment.
- Every email that is included in the SNS topic requires confirming the subscription to receive email notifications. If you miss the email, you can get the endpoints from the CloudFormation stack outputs.

## LaunchPad Pricing

Group	AWS Service	Description	Size	Unit	Hourly cost per Unit	Monthly Cost	TCO
Storage	EBS	GP3, No snapshots	250	GB	\$0.000	\$20.00	\$240.00
	EFS	100% data is frequently accessed.	500	GB	\$0.000	\$150.00	\$1,800.00
	S3	500 GB standard, 1000 GET requests, 1000 PUT requests	500	GB	\$0.000	\$11.50	\$138.00
Compute	EC2	g5.4xlarge - Single GPU - 1 GPU, 24 GPU Mem (GiB), 16 VCPUs 64 Memory (GiB)	1	instance	\$1.624	\$1,185.52	\$14,226.24
WAF	WAF ACL		1	ACL	\$0.007	\$5.00	\$60.00
	WAF Rule		8	Rule	\$0.001	\$8.00	\$96.00
	WAF Request		5	Million requests	\$0.001	\$3.00	\$36.00
Network	ALB		1	ALB	\$0.023	\$16.43	\$197.10
	ALB Data processed		1.4	LCU	\$0.008	\$8.18	\$98.11
	NAT Gateway		1	NAT	\$0.045	\$32.85	\$394.20
	NAT Gateway Data processed		500	GB	\$0.0001	\$22.50	\$270.00
Total					\$1.709	\$1,440.471	\$17,285.652

## Instances pricing

Category	Instance Size	GPU	GPU Memory (GiB)	vCPUs	Memory (GiB)	On Demand Price/hr*	On Demand Price/Monthly*
Single GPU	g5.xlarge	1.00	24.00	4.00	16.00	\$1.01	\$734.38
Single GPU	g5.2xlarge	1.00	24.00	8.00	32.00	\$1.21	\$884.76
Single GPU	g5.4xlarge	1.00	24.00	16.00	64.00	\$1.62	\$1,185.52
Single GPU	g5.8xlarge	1.00	24.00	32.00	128.00	\$2.45	\$1,787.04
Single GPU	g5.16xlarge	1.00	24.00	64.00	256.00	\$4.10	\$2,990.08

Multi GPU	g5.12xlarge	4.00	96.00	48.00	192.00	\$5.67	\$4,140.56
Multi GPU	g5.24xlarge	4.00	96.00	96.00	384.00	\$8.14	\$5,945.12
Multi GPU	g5.48xlarge	8.00	192.00	192.00	768.00	\$16.29	\$11,890.24

## Monitoring

TextGen uses Grafana as its monitoring tool. There is a Grafana dashboard built by the Logicworks team, to give you insights about the physician resources usage of your TextGen instance. Some of the metrics that this dashboard shows you, are the following:

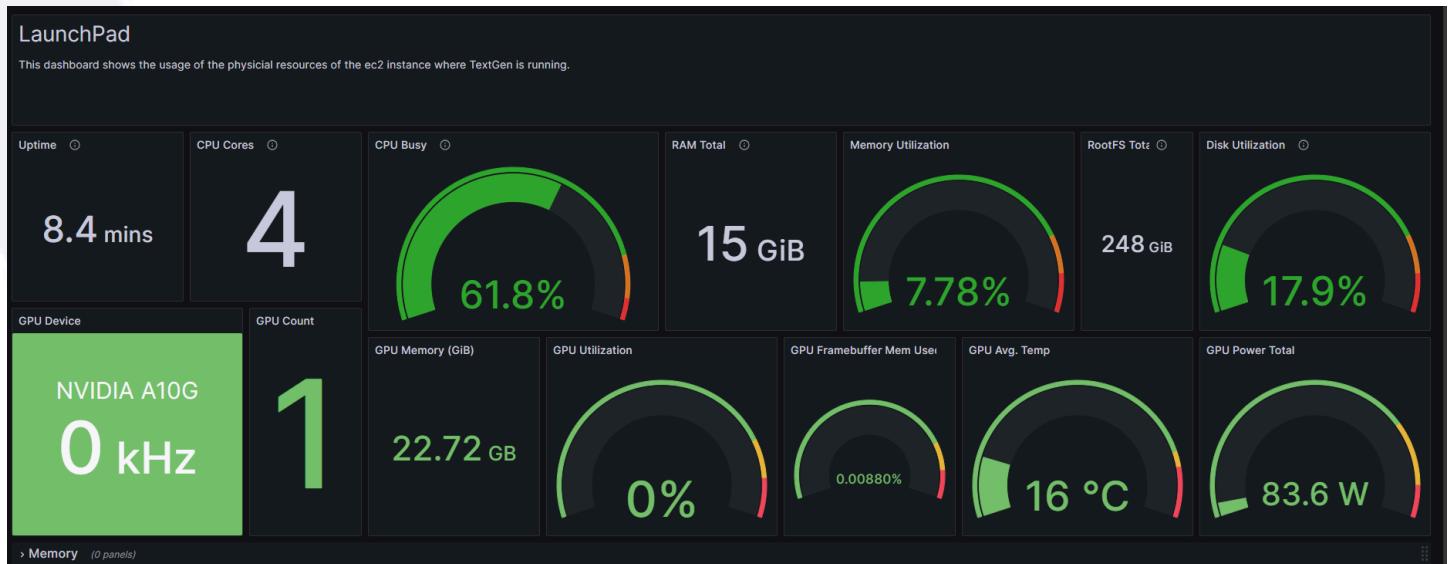
- **Uptime:** Time since last boot.
- **CPU Cores:** Number of CPU cores on the instance
- **CPU Busy:** Percentage of the CPU usage between all the cores available in the instance.
- **RAM Total:** Memory available in the instance.
- **Memory Utilization:** Percentage of the memory usage in the instance.
- **ROOT Fs:** Total local storage size in the instance.
- **Disk Utilization:** Percentage of the disk utilization in the instance.
- **CPU Device:** Name of the GPU device that the instance uses. When using it will show you the clock speed for the GPU.
- **GPU Count:** Number of GPUs available in the instance.
- **GPU Utilization:** Percentage of the GPU usage in the instance.
- **GPU AVG Temp:** Percentage of the GPU temperature.
- **GPU Power Total:** Percentage of the electricity consumed by the GPU.

To access the dashboard, you will need to use the endpoint sent to you via email notification.

To check out the dashboard, you would need to follow these instructions:

1. Sign in to Grafana.
  - a. The username is admin and the password is the one set by your administrator when deploying LaunchPad.
2. Click Dashboards in the left-side menu.
3. Expand the folder “General”.
4. Open the dashboard with the name “LaunchPad”.

# LaunchPad - User Guide



## Cleanup Process

When deleting LaunchPad CloudFormation stack, some AWS services are left in case later on you want to re-use them when creating a new stack. Those services are:

- EFS
- EC2 Security Groups
- S3 Bucket
- KMS Keys

If you are sure that you will not need those resources anymore, you can proceed to delete them following these instructions:

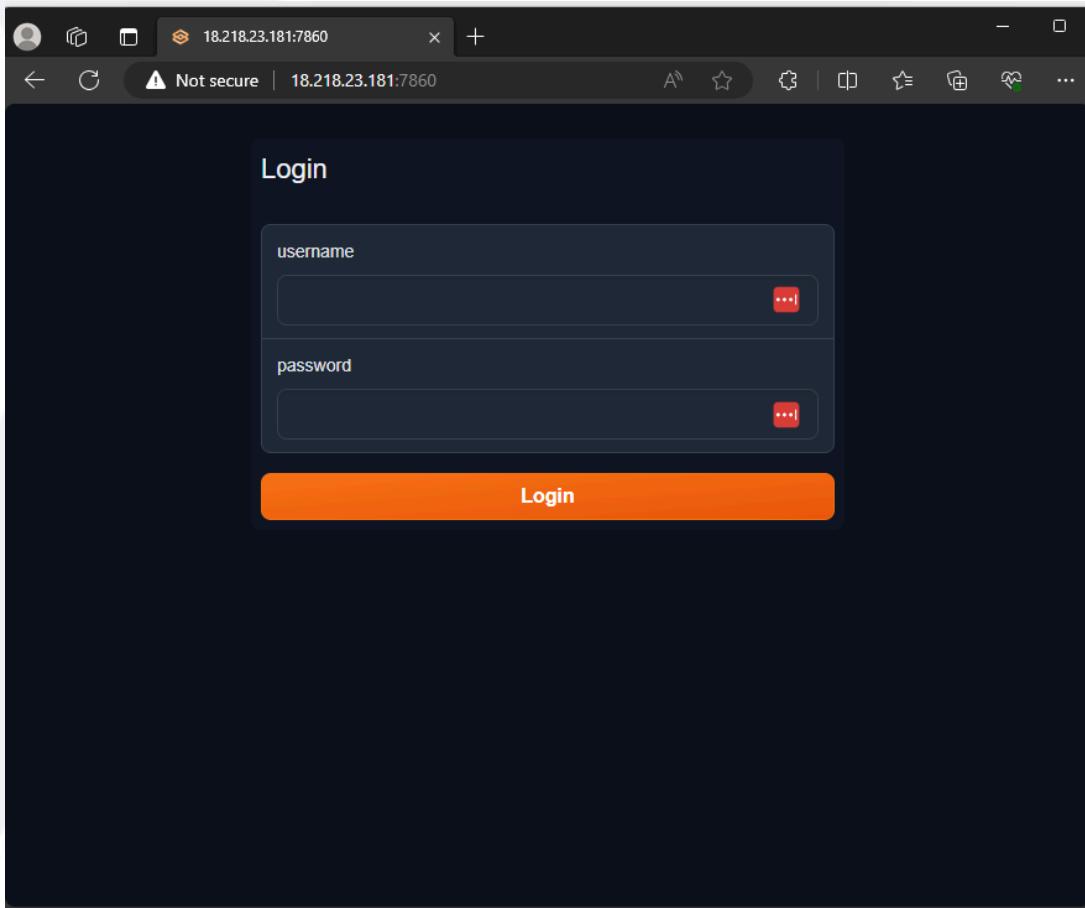
- Go to AWS Console
- Go to EFS page <https://console.aws.amazon.com/efs/>
- From the navigation bar, select the Region in which you launched your resources.
- Select the EFS that contains the name of the deleted CloudFormation stack and click on the top right the delete button.
- Go to EC2 page <https://console.aws.amazon.com/ec2/>
- From the right menu, select security groups.
- Select the EC2 Security group that contains the name of the deleted CloudFormation stack and click on the top right the actions>Delete security groups.
- Go to S3 page <https://console.aws.amazon.com/s3/>
- Select the S3 Bucket that contains the name of the deleted CloudFormation stack and click on the top right the empty button (the bucket should be empty to be able to delete it).
- Select the S3 Bucket that contains the name of the deleted CloudFormation stack and click on the top right the delete button
- Go to KMS page <https://console.aws.amazon.com/kms/>
- Select individually the KMS keys (for EFS and S3) that contain the name of the deleted CloudFormation stack and click on the top right key actions>schedule key deletion. Make sure to schedule the deletion by 7 days, so you don't have to wait too much time to create another stack with the same name.

## TextGen

Welcome to LaunchPad TextGen! Kick-start your journey with Generative AI by downloading any Language model from Hugging Face. As of now, we recommend the LLama-2 models, which are versatile and can be fine-tuned with Lora for optimal performance on a single g5.xlarge instance. Grab the model here. Once downloaded and loaded, you can instantly chat with the language model to test its capabilities.

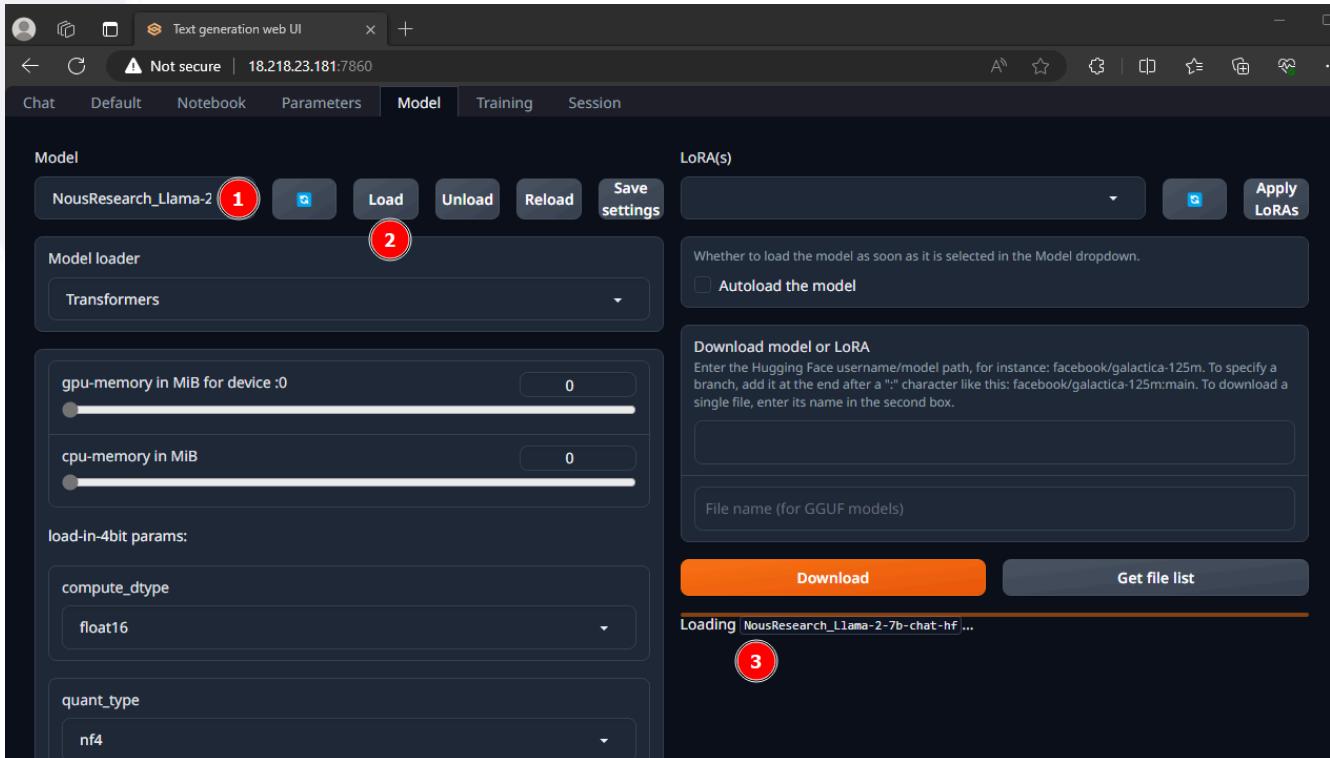
## Select a Model

Go to the TextGen endpoint that you have received via email and log in with your credentials.



# LaunchPad - User Guide

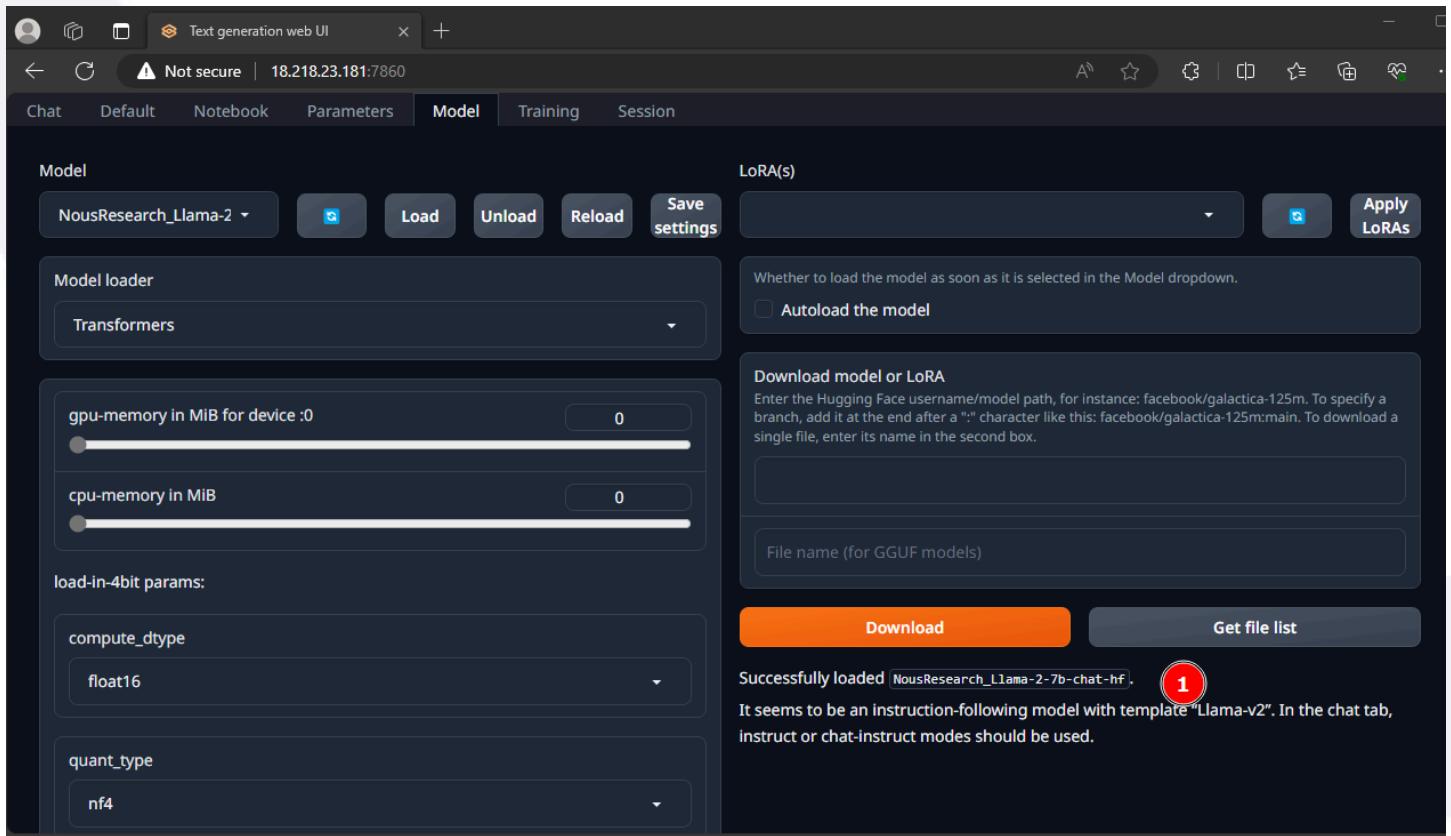
No model is loaded by default. Go to model, select the default model (1), click on the load button (2) and wait until the message indicates that the model has been loaded successfully (3).



If you go to grafana, you would see that when loading the model, the GPU memory usage increases. This is a good way to determine if the instance size that you have chosen is enough to use the model you want to use.



# LaunchPad - User Guide



The screenshot shows the 'Model' tab of the LaunchPad web UI. It includes sections for selecting a model (NousResearch\_Llama-2), managing LoRA settings, and configuring memory usage (gpu-memory in MiB for device 0 and cpu-memory in MiB). There are also dropdowns for Model loader (Transformers) and load-in-4bit parameters (compute\_dtype: float16, quant\_type: nf4). A message at the bottom indicates the model has been successfully loaded.

Model

NousResearch\_Llama-2

LoRA(s)

Model loader

Transformers

gpu-memory in MiB for device :0

cpu-memory in MiB

load-in-4bit params:

compute\_dtype

float16

quant\_type

nf4

Autoload the model

Download model or LoRA

File name (for GGU models)

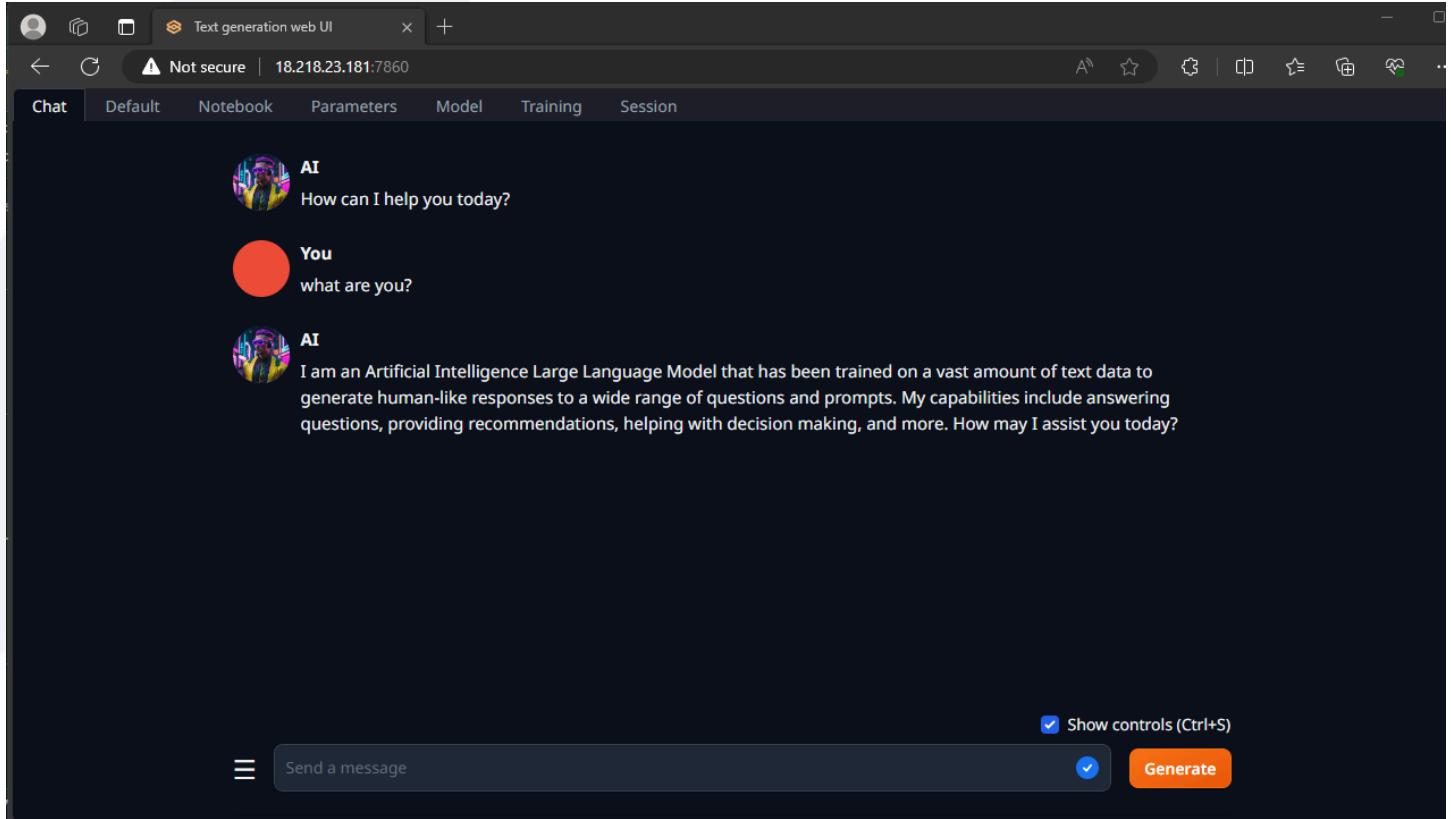
Download

Get file list

Successfully loaded NousResearch\_Llama-2-7b-chat-hf. 1

It seems to be an instruction-following model with template "Llama-v2". In the chat tab, instruct or chat-instruct modes should be used.

Then you can go back to the Chat page and start interacting with the model through a chat interface.



The screenshot shows the 'Chat' tab of the LaunchPad web UI. It displays a conversation between an AI (represented by a purple icon) and a user (represented by a red circle). The AI asks, "How can I help you today?", the user responds with "what are you?", and the AI provides a detailed explanation of its capabilities. At the bottom, there is a message input field with placeholder "Send a message" and a "Generate" button.

AI

How can I help you today?

You

what are you?

AI

I am an Artificial Intelligence Large Language Model that has been trained on a vast amount of text data to generate human-like responses to a wide range of questions and prompts. My capabilities include answering questions, providing recommendations, helping with decision making, and more. How may I assist you today?

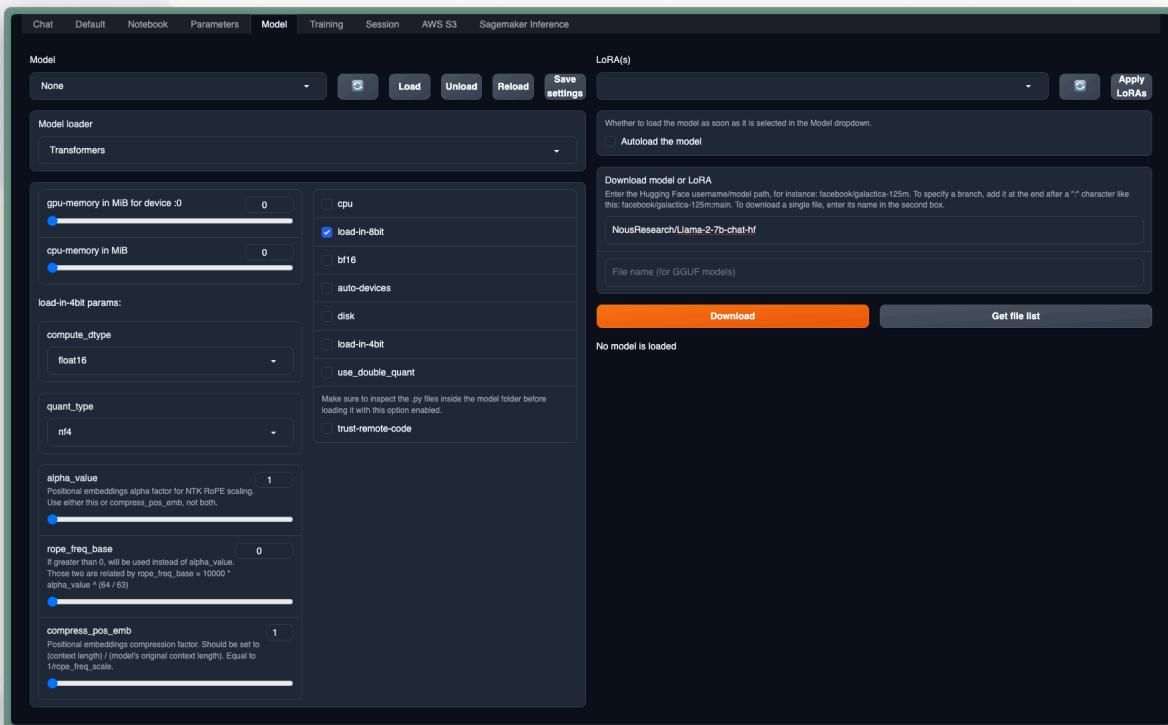
Show controls (Ctrl+S)

Send a message

Generate

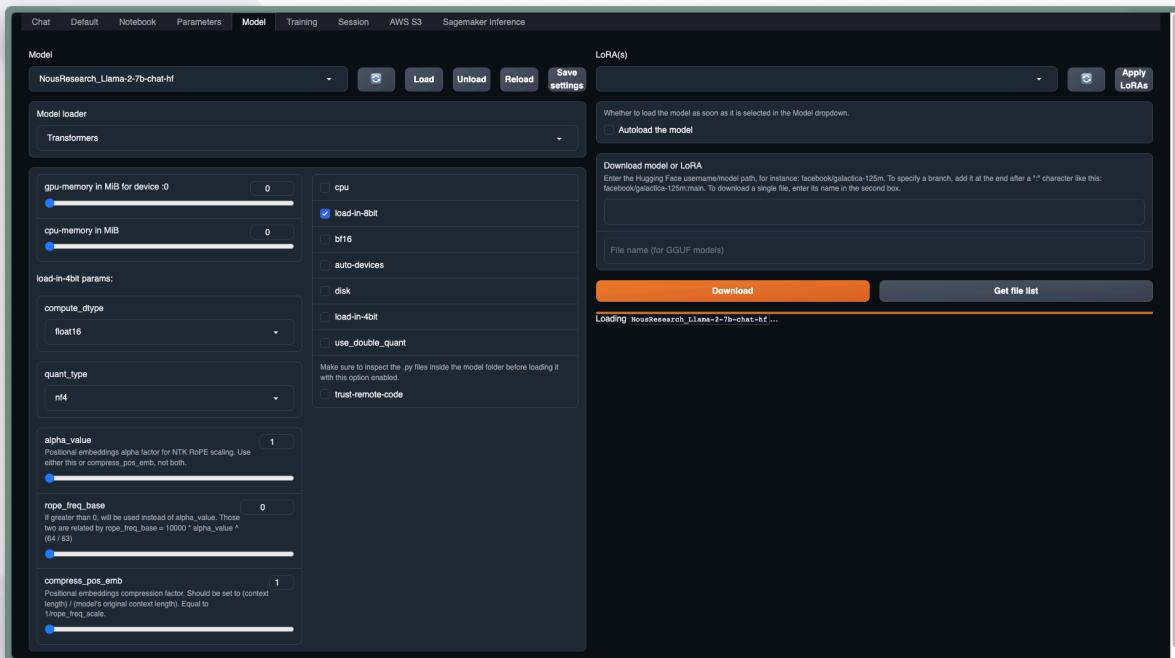
## Download Model

- Navigate to [Hugging Face](#) and identify the model of your choice.
- Copy the model ID, e.g., NousResearch/Llama-2-7b-chat-hf.
- In TextGen, paste the model ID into the 'Model' tab and press download to initiate the download.



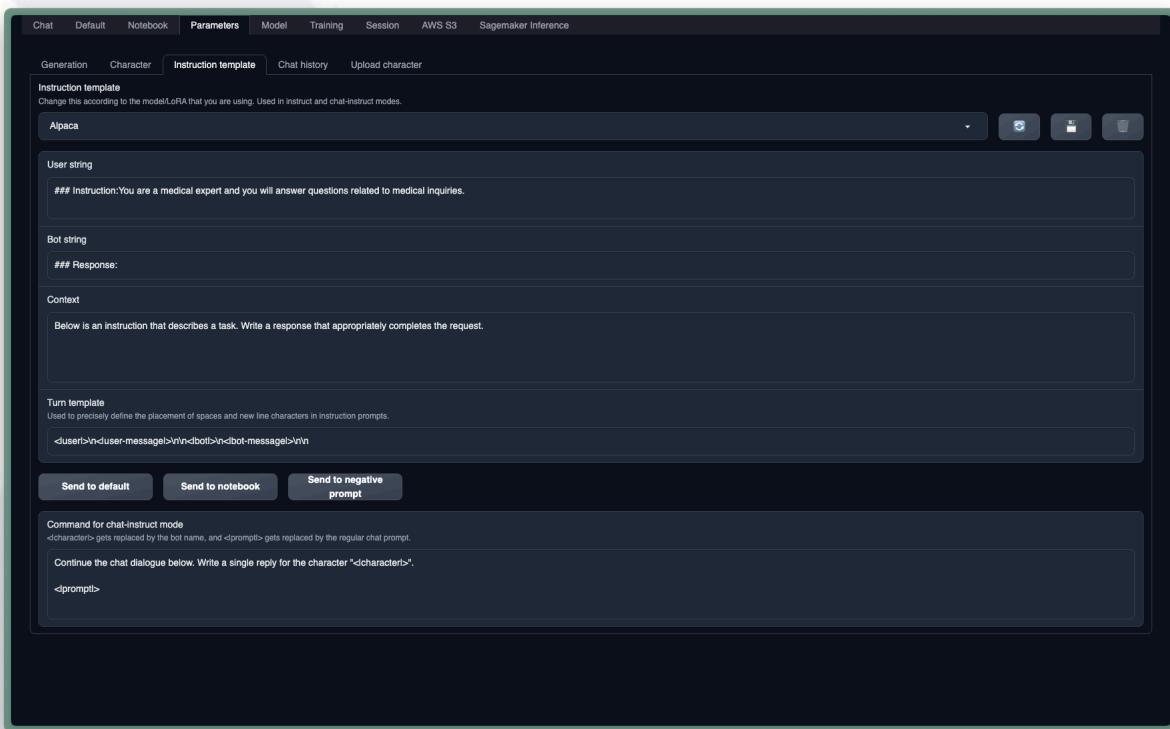
## Load Model

1. Once downloaded, refresh the model dropdown menu.
2. Choose your desired model from the dropdown.
3. Before loading, adjust parameters as needed. Here are two common scenarios:
  - a. For an HF model like LLama-2, opt for load-in-8bit to fit it into a g5.xlarge instance.
  - b. For GGML/GTPQ models, use llama.cpp to offload memory to the CPU, ensuring compatibility with smaller instance types.



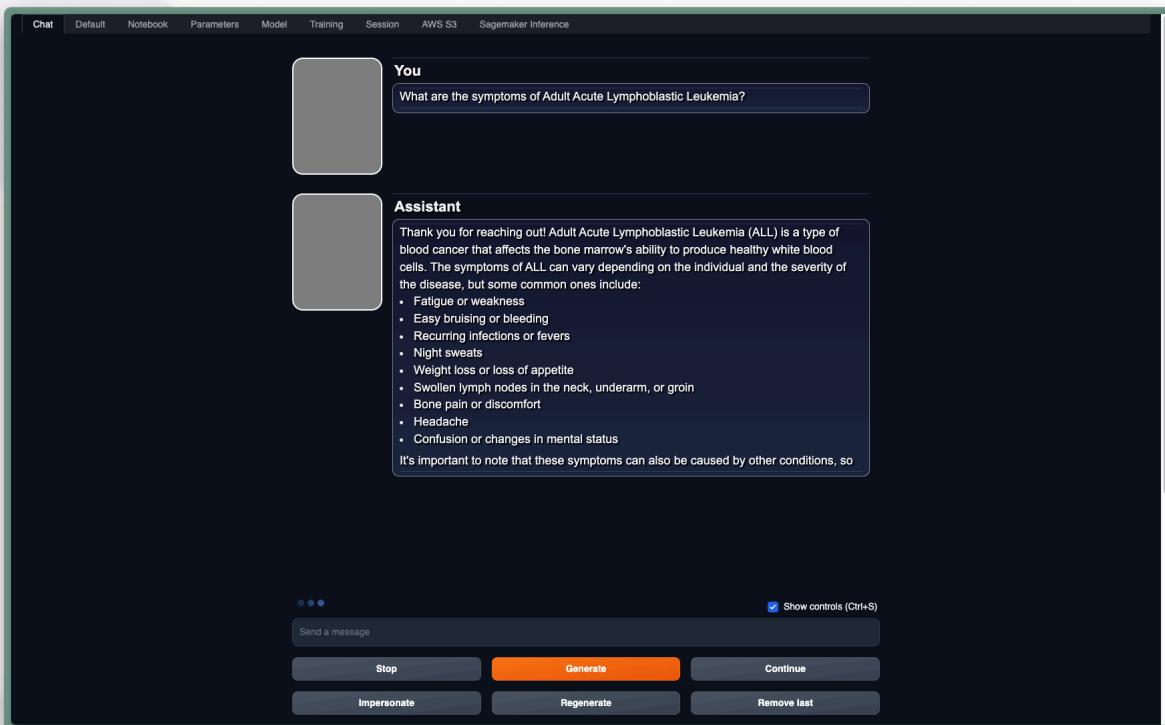
## Setup Prompt Templates

1. With the model loaded, head to the 'Parameters' tab adjacent to the 'Models' tab.
2. Navigate to 'Instruction Template' and choose a predefined template. For our purposes, options like llama-v2 or alpaca are ideal.
3. Fill in the fields to tailor your model's behavior. In this guide, we're setting it up as a medical expert.
4. We've opted for the alpaca instruction template, aligning with our intention to use a specific dataset later in this guide that pairs well with the alpaca instruction format.
5. In the 'User String' section, set the initial prompt to: "You are a medical expert and you will answer questions related to medical inquiries."



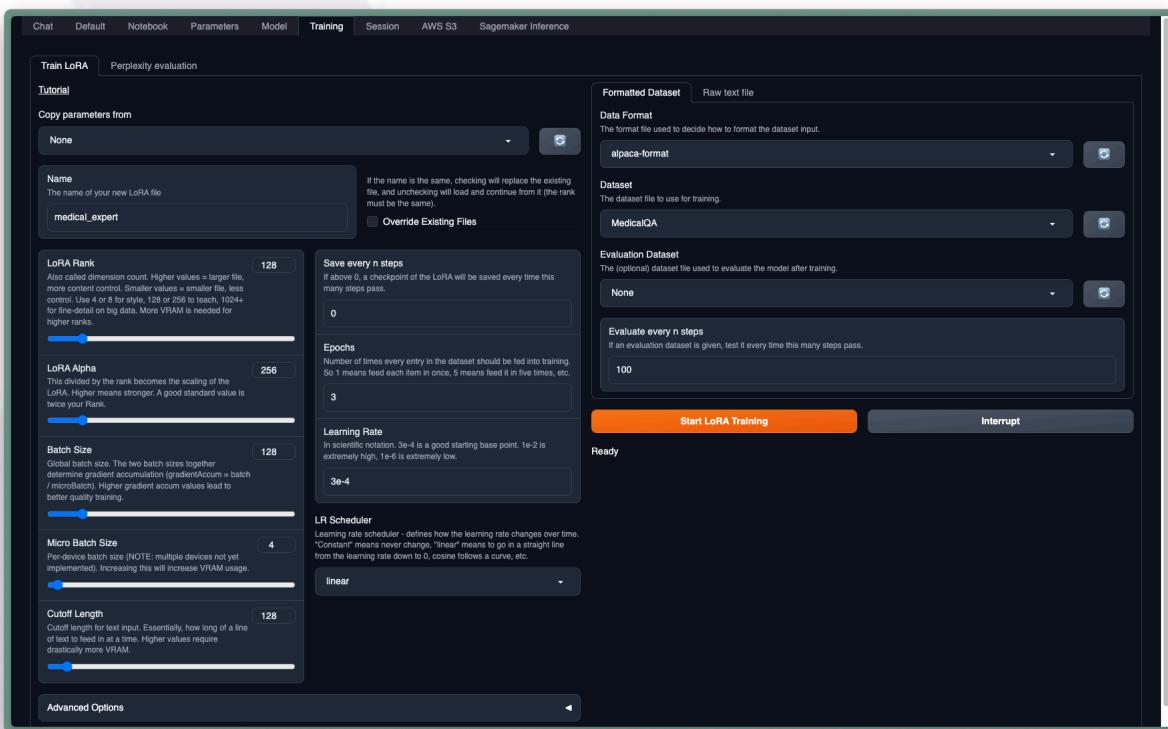
## Playground

1. Navigate to the 'Chat' tab to interact with your model.
2. To utilize the previously set template, scroll down and select 'chat-instruct' from the 'Mode' section.
3. Clear chat history anytime by clicking 'Clear History'.



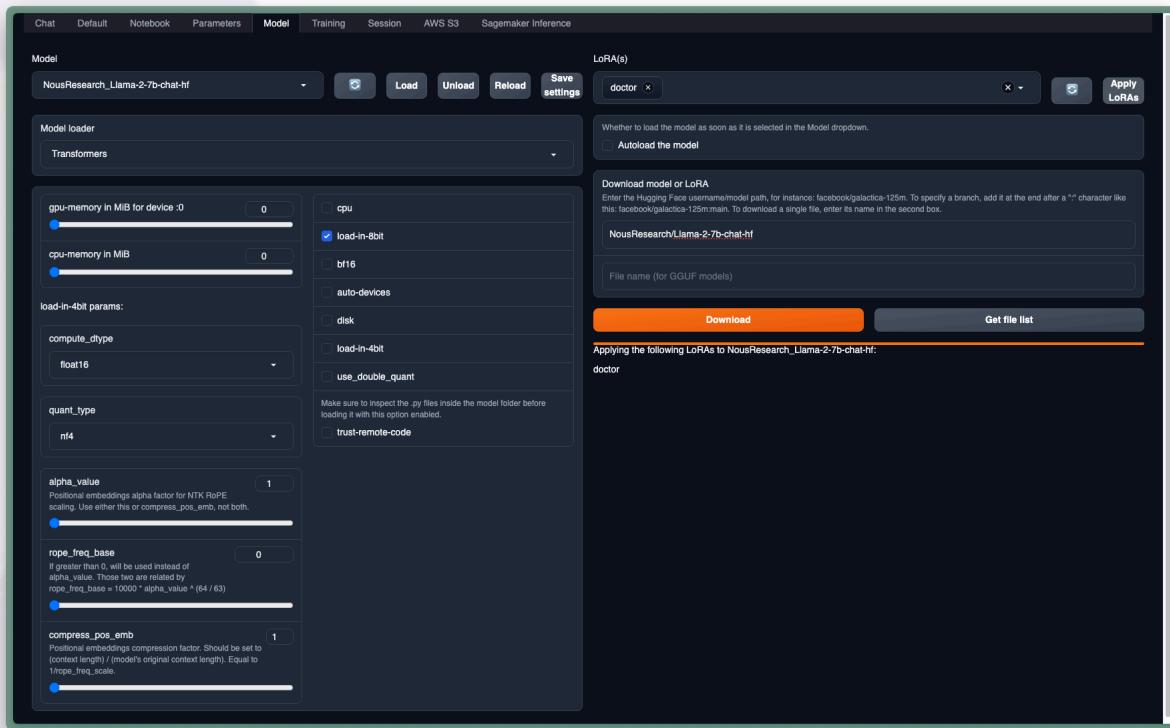
## Fine-tuning

1. Navigate to the 'Training' tab.
2. Name your LoRA.
3. On the right, set data format to 'alpaca-format' and choose the 'medicalQA' dataset. If not visible, hit 'Refresh'.
4. Adjust parameters:
  - a. Lora Rank (128)
  - b. Lora Alpha (256)
  - c. Cutoff Length (128)
  - d. Feel free to experiment with other settings.
5. Click 'Start LoRA Training'.



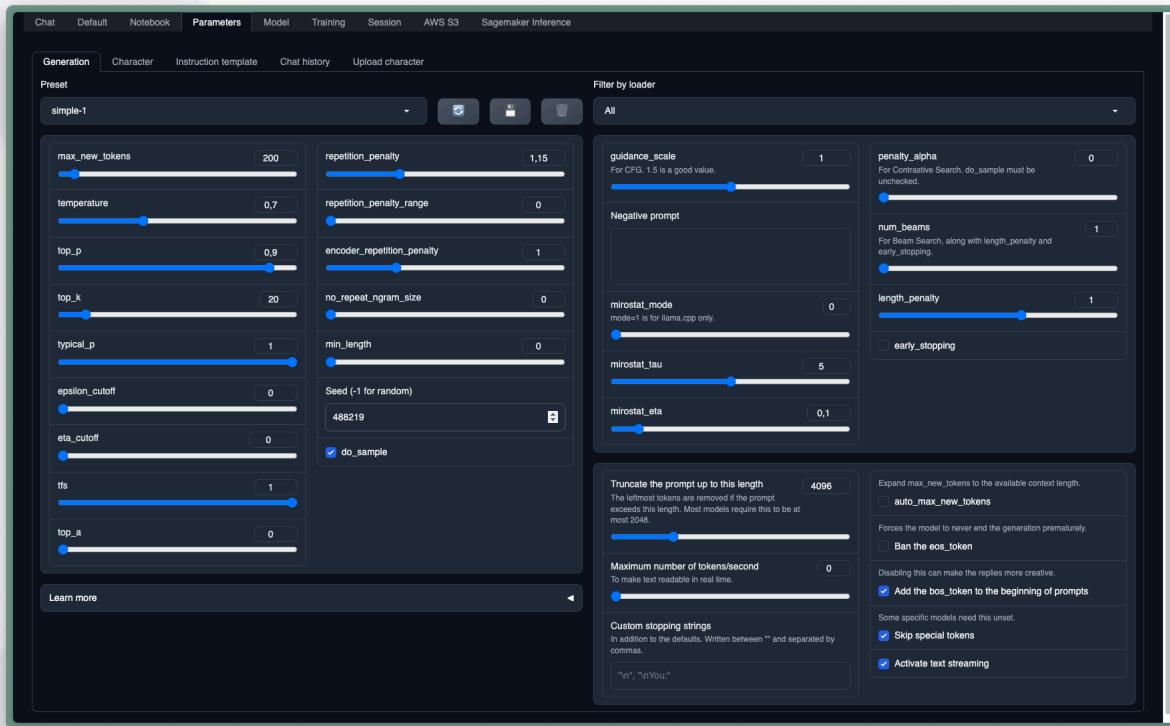
## Using the Fine-tuned Model

1. Head to the 'Model' tab and click 'Reload Model'.
2. From the dropdown, select your LoRA. If not visible, refresh the list.
3. Click 'Apply LoRA' to load it.
4. Engage with your enhanced model in the 'Chat' tab.



## Evaluating the Fine-tuned Model:

1. In the 'Parameters->Generation' tab, input a random seed to ensure consistent model outputs.
2. Toggle between loading and unloading the Lora to compare output differences.



## Logicworks Extensions

### S3 Data

#### S3 File Structure:

For the LaunchPad-TextGen app to seamlessly interact with your S3 bucket, it's crucial to maintain a specific file structure. This ensures that datasets, models, and Lora configurations are easily identifiable and accessible. Here's the recommended structure:

- Datasets: Store your training datasets in the following path: `dataset/train/*.json`. The app will look for `.json` files under this directory when you choose to download training data.
- Loras: All Lora configurations should be placed directly under the ``loras/`` directory.
- Artifacts: If you have model artifacts, they should be stored as: ``artifacts/*/model.tar.gz``. This structure allows the app to recognize and download model artifacts.
- Models: All other model files should be placed directly under the ``models/`` directory.

#### Selecting an S3 Bucket:

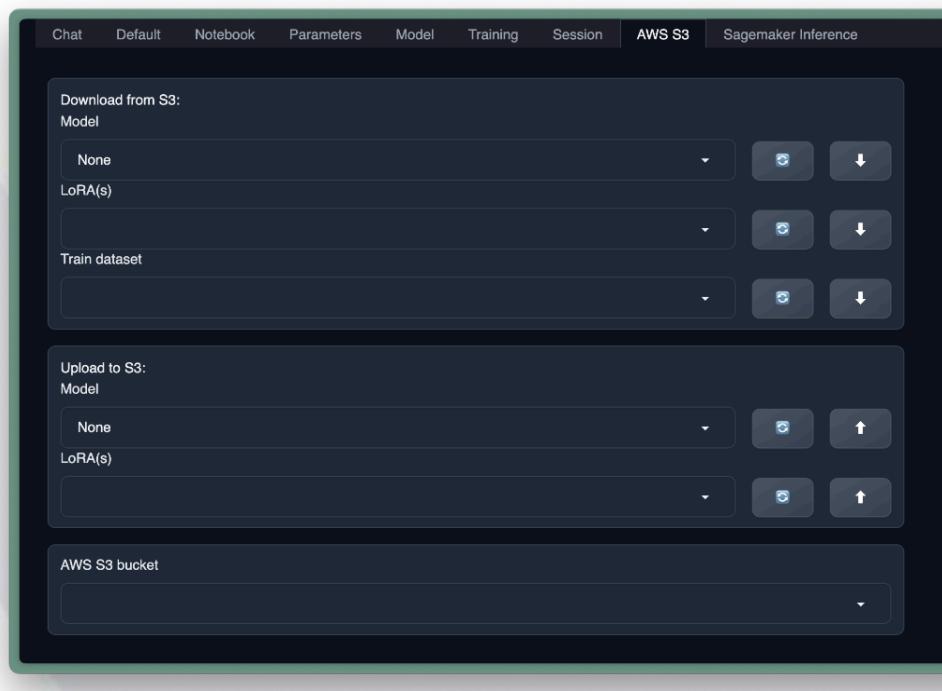
1. At the bottom of the 'AWS S3' tab, there's a dropdown menu where you can choose your preferred S3 bucket for download/upload operations.

#### Downloading from S3:

1. If you've stored models, Lora configurations, or training datasets on S3, they can be easily imported into the LaunchPad-TextGen app.
2. Navigate to the 'AWS S3' tab.
3. In the first section labeled 'Download Modules', you'll find three dropdown menus: 'Model', 'Lora's', and 'Train Dataset'.
4. Each dropdown is paired with a 'Refresh' button. Use this if you've recently added items to S3 and need to update the list of downloadable items.
5. After selecting the desired files from the dropdowns, click the 'Download' button to transfer them to the app.

#### Uploading to S3:

1. The LaunchPad-TextGen app also provides the capability to upload models and Lora configurations directly to your S3 bucket.
2. To initiate the upload:
  - a. Choose the model or Lora configuration you wish to upload.
  - b. Click on the 'Upload' button located adjacent to your selection.

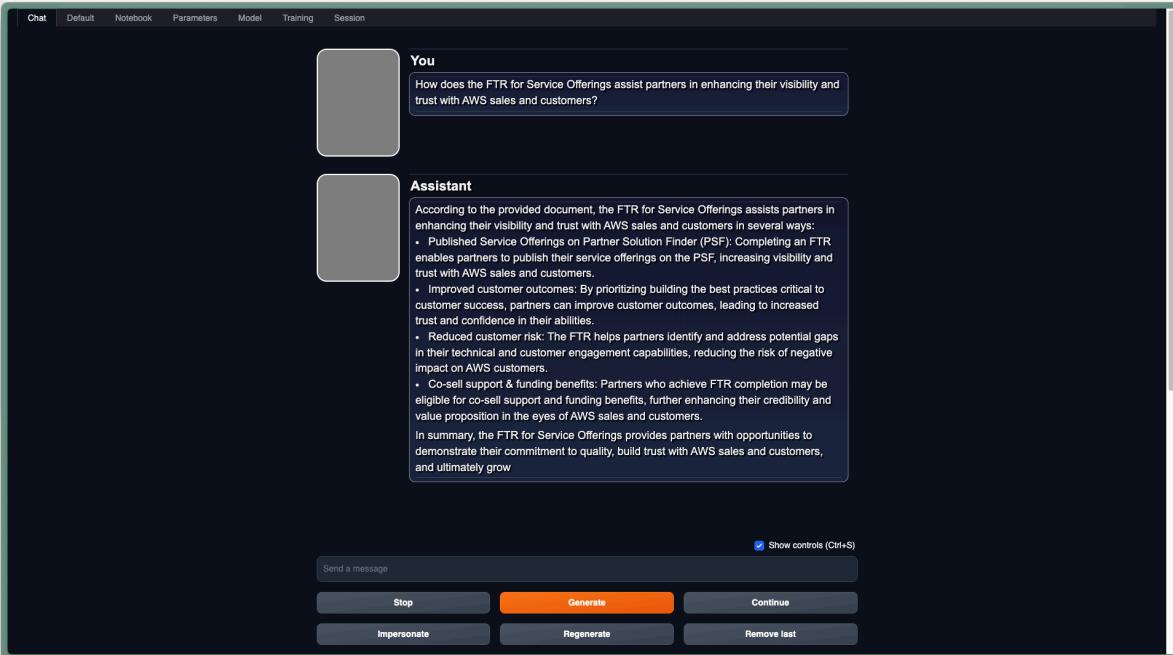


## Retrieval Augmented Generation

If you're looking to enhance your chat experience with the Kendra RAG extension, follow these straightforward steps:

1. Ensure API Extension is Active:
  - a. Before proceeding, confirm that the API extension is up and running, as it's essential for the Kendra RAG extension.
2. Navigate to the Chat Interface:
  - a. Open LaunchPad and head to the "Chat" tab, which is the first on the list.
3. Access the RAG Configuration:
  - a. Scroll until you find the dedicated RAG section.
4. Input the Kendra Index ID:
  - a. Visit your AWS console to locate and copy the Kendra Index ID.
  - b. Return to LaunchPad and paste this ID into the provided field in the RAG section.
5. Activate AWS Kendra Integration:
  - a. Check the box labeled "AWS Kendra" to enable the feature.
6. Apply Your Settings:
  - a. Click the "Submit" button.
  - b. LaunchPad will now channel your chat through the LangChain Kendra Conversational Retrieval Chain. This mechanism fetches documents from Kendra and formulates responses based on their content.

# LaunchPad - User Guide



The screenshot shows a dark-themed user interface for a chat application. At the top, there's a navigation bar with tabs: Chat, Default, Notebook, Parameters, Model, Training, and Session. Below the navigation bar, there are two main sections: 'You' and 'Assistant'. The 'You' section contains a message input field and a 'Send a message' button. The 'Assistant' section displays a response with several bullet points. At the bottom, there's a row of control buttons: Stop, Generate (highlighted in orange), Continue, Impersonate, Regenerate, and Remove last. A 'Show controls (Ctrl+S)' checkbox is also present.

**You**  
How does the FTR for Service Offerings assist partners in enhancing their visibility and trust with AWS sales and customers?

**Assistant**  
According to the provided document, the FTR for Service Offerings assists partners in enhancing their visibility and trust with AWS sales and customers in several ways:

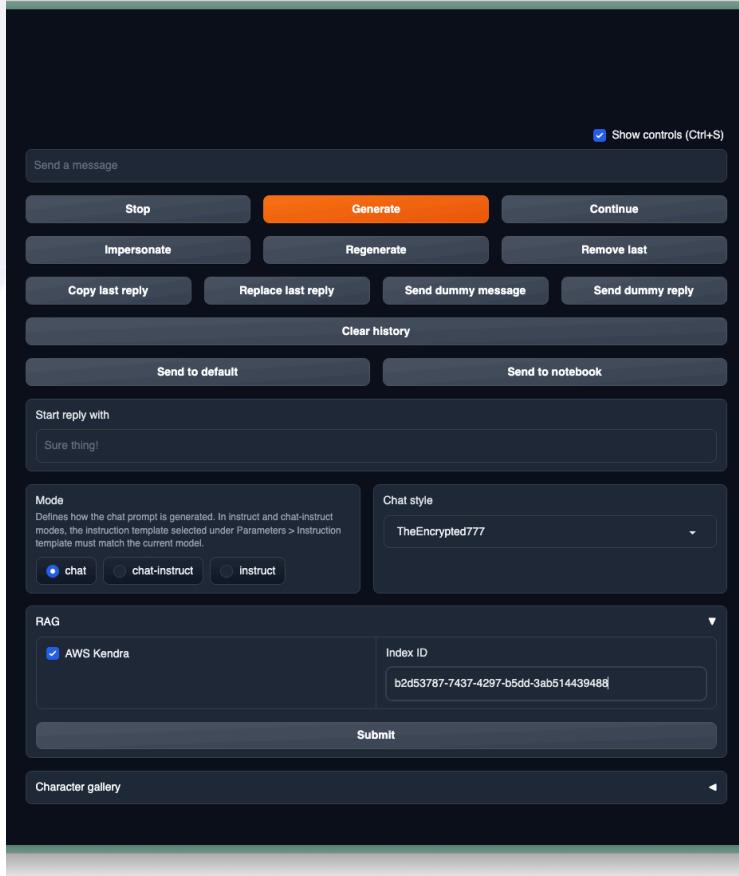
- Published Service Offerings on Partner Solution Finder (PSF): Completing an FTR enables partners to publish their service offerings on the PSF, increasing visibility and trust with AWS sales and customers.
- Improved customer outcomes: By prioritizing building the best practices critical to customer success, partners can improve customer outcomes, leading to increased trust and confidence in their abilities.
- Reduced customer risk: The FTR helps partners identify and address potential gaps in their technical and customer engagement capabilities, reducing the risk of negative impact on AWS customers.
- Co-sell support & funding benefits: Partners who achieve FTR completion may be eligible for co-sell support and funding benefits, further enhancing their credibility and value proposition in the eyes of AWS sales and customers.

In summary, the FTR for Service Offerings provides partners with opportunities to demonstrate their commitment to quality, build trust with AWS sales and customers, and ultimately grow.

Show controls (Ctrl+S)

Send a message

Stop Generate Continue  
Impersonate Regenerate Remove last



This screenshot shows the same dark-themed interface with additional configuration options visible. At the top, there are 'Stop', 'Generate' (orange), and 'Continue' buttons. Below them are 'Impersonate', 'Regenerate', and 'Remove last' buttons. Further down are 'Copy last reply', 'Replace last reply', 'Send dummy message', and 'Send dummy reply' buttons. A 'Clear history' button is also present. At the bottom, there are 'Send to default' and 'Send to notebook' buttons. On the left, there's a 'Start reply with' field containing 'Sure thing!'. On the right, there are 'Mode' and 'Chat style' settings. The 'Mode' section has radio buttons for 'chat' (selected), 'chat-instruct', and 'instruct'. The 'Chat style' dropdown is set to 'TheEncrypted777'. Below these are 'RAG' settings, which include a checked 'AWS Kendra' checkbox and an 'Index ID' field containing 'b2d53787-7437-4297-b5dd-3ab514439488'. A 'Submit' button is located at the bottom of the RAG section. A 'Character gallery' button is also visible.

Show controls (Ctrl+S)

Send a message

Stop Generate Continue  
Impersonate Regenerate Remove last

Copy last reply Replace last reply Send dummy message Send dummy reply  
Clear history

Send to default Send to notebook

Start reply with  
Sure thing!

Mode  
Defines how the chat prompt is generated. In instruct and chat-instruct modes, the instruction template selected under Parameters > Instruction template must match the current model.  
 chat  chat-instruct  instruct

Chat style  
TheEncrypted777

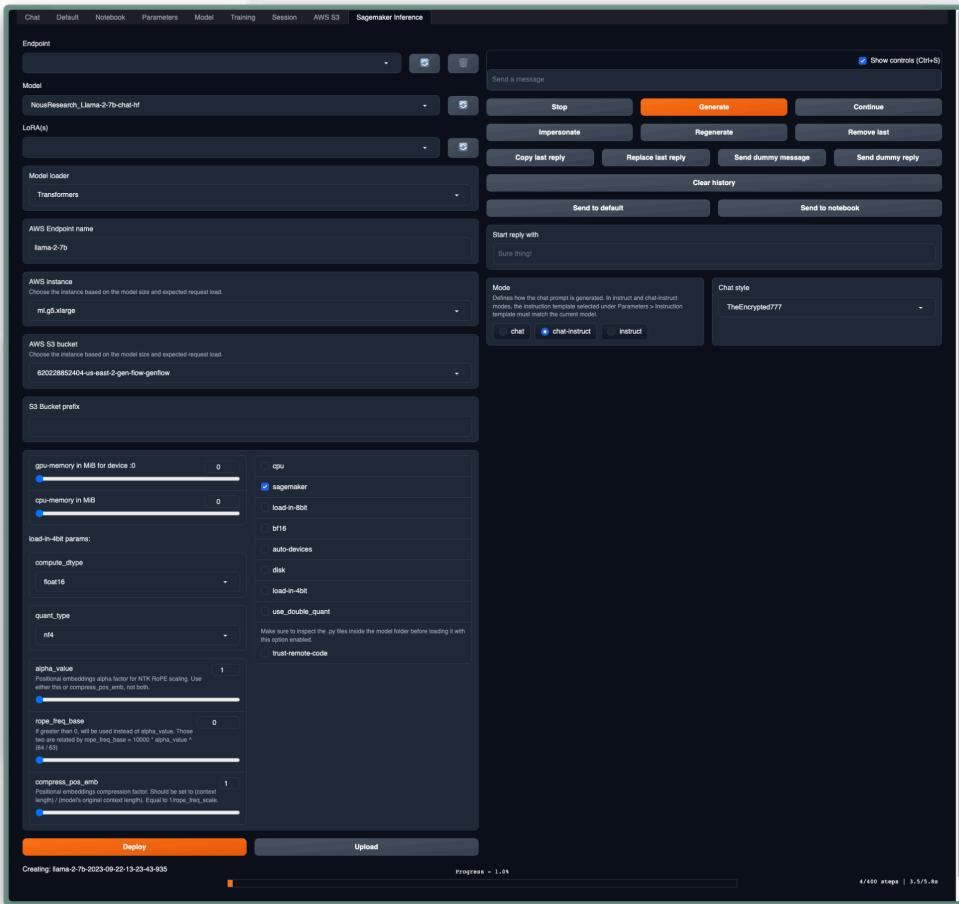
RAG  
 AWS Kendra Index ID  
b2d53787-7437-4297-b5dd-3ab514439488  
Submit

Character gallery

## Deploy to Sagemaker Endpoint

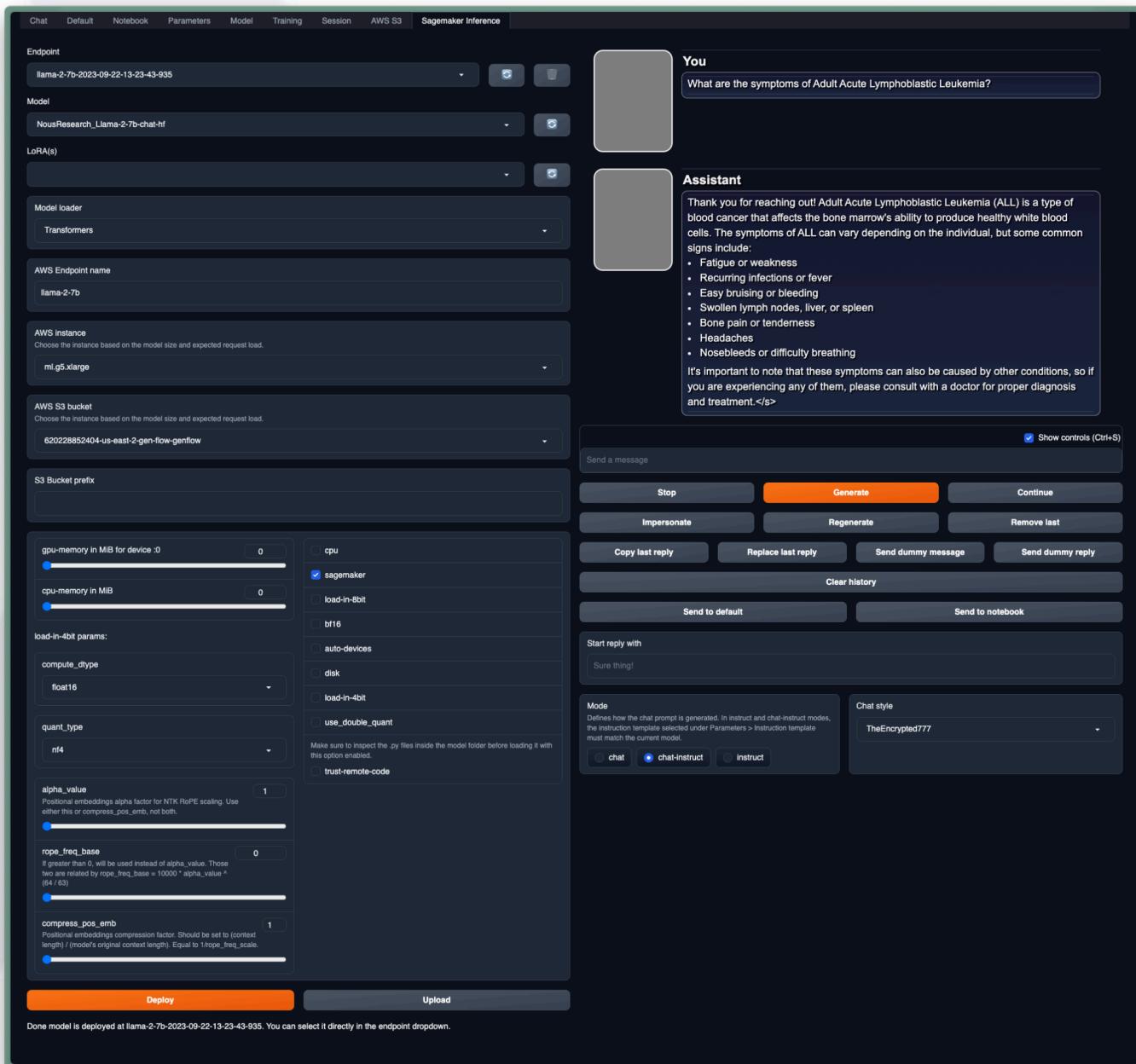
1. If you've downloaded a model from Hugging Face, you have the option to launch it directly to a Sagemaker endpoint.
2. Proceed to the 'Sagemaker Inference' section.
3. From the model dropdown, select your model.
4. Specify the AWS endpoint name.
5. Choose a suitable AWS instance size, ensuring it's compatible with your model.
6. From the dropdown menu, opt for an S3 bucket.
7. Hit the 'Deploy' button.
  - a. The system will compress your model using Gzip
  - b. Upload it to your chosen S3 bucket
  - c. Initiate deployment
8. The entire process is expected to take around 6 minutes, and you can monitor its status with the on-screen progress bar.

Note: Once you're done using the endpoint, you can easily delete it. Just click on the trash bin icon located to the right of the endpoint dropdown. This action will remove the Sagemaker endpoint.



## Chat using Sagemaker Endpoint

1. Stay within the 'Sagemaker Inference' tab and look to the right to find the chat window.
2. On the left, check the box labeled 'Sagemaker'. This informs TextGen that you intend to utilize the Sagemaker endpoint.
3. From the “endpoint” dropdown, select the endpoint you recently set up.
4. Engage with the chat as you typically would, using the 'Generate' button to interact.
5. Alternatively, you can also navigate to the 'Chat' tab to access and use the chat window in the same manner.



**You**

What are the symptoms of Adult Acute Lymphoblastic Leukemia?

**Assistant**

Thank you for reaching out! Adult Acute Lymphoblastic Leukemia (ALL) is a type of blood cancer that affects the bone marrow's ability to produce healthy white blood cells. The symptoms of ALL can vary depending on the individual, but some common signs include:

- Fatigue or weakness
- Recurring infections or fever
- Easy bruising or bleeding
- Swollen lymph nodes, liver, or spleen
- Bone pain or tenderness
- Headaches
- Nosebleeds or difficulty breathing

It's important to note that these symptoms can also be caused by other conditions, so if you are experiencing any of them, please consult with a doctor for proper diagnosis and treatment.

Show controls (Ctrl+S)

Send a message

Stop    **Generate**    Continue

Impersonate    Regenerate    Remove last

Copy last reply    Replace last reply    Send dummy message    Send dummy reply

Clear history

Send to default    Send to notebook

Start reply with

Sure thing!

Mode

Defines how the chat prompt is generated. In instruct and chat-instruct modes, the instruction template selected under Parameters > Instruction template must match the current model.

Chat style

TheEncrypted777

chat    **chat-instruct**    instruct

Deploy    Upload

Done model is deployed at llama-2-7b-2023-09-22-13-23-43-935. You can select it directly in the endpoint dropdown.

## Technical Troubleshooting: Common Errors & Tips

1. If your stack fails the first time because of a misconfiguration, the update button will not be enabled. You would need to destroy the stack and configure the parameters again.
2. If you received a
3. When deploying the **networking template**, the CIDR block should not overlap existing VPC CIDR blocks, and should be a valid CIDR range less than 28. If not, you can get an error like the following one.

Events (12)			
<input type="text"/> Search events			
Timestamp	Logical ID	Status	Status reason
2023-10-23 16:47:36 UTC+0200	GenFlow-network	⌚ DELETE_COMPLETE	-
2023-10-23 16:47:36 UTC+0200	VPC	⌚ DELETE_COMPLETE	-
2023-10-23 16:45:40 UTC+0200	GenFlow-network	ⓘ DELETE_IN_PROGRESS	User Initiated
2023-10-23 16:45:14 UTC+0200	InternetGateway	⌚ DELETE_COMPLETE	-
2023-10-23 16:45:12 UTC+0200	InternetGateway	ⓘ DELETE_IN_PROGRESS	-
2023-10-23 16:45:10 UTC+0200	GenFlow-network	✖ ROLLBACK_IN_PROGRESS	The following resource(s) failed to create: [InternetGateway, VPC]. Rollback requested by user.
2023-10-23 16:45:10 UTC+0200	InternetGateway	✖ CREATE_FAILED	Resource creation cancelled
2023-10-23 16:45:09 UTC+0200	VPC	✖ CREATE_FAILED	Resource handler returned message: "The CIDR '10.0.0.0/32' is invalid. (Service: Ec2, Status Code: 400, Request ID: 00b45ee0-5f57-4251-81b7-8442aeafb617)" (RequestToken: a0489c60-82be-5818-942d-c99ecde03782, HandlerErrorCode: GeneralServiceException)
2023-10-23 16:45:09 UTC+0200	InternetGateway	ⓘ CREATE_IN_PROGRESS	Resource creation Initiated
2023-10-23 16:45:08 UTC+0200	VPC	ⓘ CREATE_IN_PROGRESS	-
2023-10-23 16:45:08 UTC+0200	InternetGateway	ⓘ CREATE_IN_PROGRESS	-
2023-10-23 16:45:05 UTC+0200	GenFlow-network	ⓘ CREATE_IN_PROGRESS	User Initiated

4. When deploying LaunchPad template, make sure that the name is completely lowercase. Some AWS resources uses the stack name in their names and can only use lowercase letters.

2023-10-23 16:51:35 UTC+0200	KmsAliasS3	ⓘ CREATE_IN_PROGRESS	Resource creation Initiated
2023-10-23 16:51:35 UTC+0200	S3Bucket	✖ CREATE_FAILED	Bucket name should not contain uppercase characters

5. If you get the message “*you have requested more vCPU than your current vCPU limit of X allows for the instance ...*”. That means that you have run out of vCPU that you can deploy in that specific region. Follow these instructions: [Increase service quota to run G5 on-demand instances](#) to fix it.

=

Events (100+)			
<input type="text"/> Search events			
Timestamp	Logical ID	Status	Status reason
2023-10-16 11:12:04 UTC-0500	ArtifactsS3Policy	ⓘ DELETE_IN_PROGRESS	-
2023-10-16 11:12:04 UTC-0500	GenflowListenerJupyter	ⓘ DELETE_IN_PROGRESS	-
2023-10-16 11:11:57 UTC-0500	genflow2	☒ ROLLBACK_IN_PROGRESS	The following resource(s) failed to create: [GenAiEFS1MountTargetA, KmsAliasS3, KmsAliasEbs, KmsAliasEFS, GenFlowInstance1, S3Policy]. Rollback requested by user.
2023-10-16 11:11:56 UTC-0500	KmsAliasEFS	☒ CREATE_FAILED	Resource creation cancelled
2023-10-16 11:11:56 UTC-0500	KmsAliasEbs	☒ CREATE_FAILED	Resource creation cancelled
2023-10-16 11:11:56 UTC-0500	S3Policy	☒ CREATE_FAILED	Resource creation cancelled
2023-10-16 11:11:56 UTC-0500	GenAiEFS1MountTargetA	☒ CREATE_FAILED	Resource creation cancelled
2023-10-16 11:11:56 UTC-0500	KmsAliasS3	☒ CREATE_FAILED	Resource creation cancelled
2023-10-16 11:11:56 UTC-0500	GenFlowInstance1	☒ CREATE_FAILED	You have requested more vCPU capacity than your current vCPU limit of 0 allows for the instance bucket that the specified instance type belongs to. Please visit <a href="http://aws.amazon.com/contact-us/ec2-request">http://aws.amazon.com/contact-us/ec2-request</a> to request an adjustment to this limit. (Service: AmazonEC2; Status Code: 400; Error Code: VcpuLimitExceeded; Request ID: d48ce691-d3d8-4d2e-aa73-ce7375982e57; Proxy: null)
2023-10-16 11:11:55 UTC-0500	KmsPolicy	ⓘ CREATE_COMPLETE	-
2023-10-16 11:11:55 UTC-0500	GenFlowInstance1	ⓘ CREATE_IN_PROGRESS	-