Assignment 5: Emerging_Models_Research— Report

1. Introduction:

In this project I focused on two recent generative AI models that work in very different ways:

- **Qwen3 (Embedding model)** – a text model that turns sentences into dense vectors so we can measure meaning and similarity.

- **Stable Diffusion v1.5** – an image diffusion model that generates pictures from text prompts and can be tuned for a specific style.

I chose these two because together they cover both **language** and **vision**. Qwen3 helped me build a simple semantic search demo over a text dataset, while Stable Diffusion let me build a full image generation pipeline and experiment with prompts, parameters, and light fine-tuning. Stable Diffusion is the focus of the project because it required more engineering work (parameter tuning, optional LoRA fine-tune).

2. Comparison:

Qwen3 (Embedding):

In this project I used the **embedding version**, which does not generate long texts. Instead, it takes a sentence or paragraph and outputs a **vector.**

The pipeline I used:

1. Load Qwen3 tokenizer and model.

2. Encode a list of texts into embeddings (mean-pooling over hidden states, then L2-normalize).

3. For a query, encode it and compute cosine similarity with all document embeddings.

4. Sort by similarity and return top-k results.

3. Stable Diffusion v1.5

Stable Diffusion is a latent diffusion model for images:

- A text encoder encodes the prompt (e.g., "Astronaut in a jungle, cold color palette…")
- A UNet denoiser gradually removes noise from a random latent.
- A VAE converts the final latent back to an image.

Exposes a deploy() method with arguments initial_instance_count and instance_type to mirror the SageMaker call. Instead of provisioning instances, it simply returns a FakePredictor object and prints a message that deployment is being simulated.

In my project I used the StableDiffusionPipeline and then

a)Base generation:

Prompt: "Astronaut in a jungle, cold color palette, muted colors, detailed, 8k"

Parameters: guidance_scale ≈ 7.5, num_inference_steps ≈ 30.

This gave a solid but not perfect image.

b)Parameter "fine-tuning":

I experimented with:

Guidance scale (how strictly the image follows the prompt).

Number of denoising steps (20 vs 40 vs 70).

Negative prompts (e.g., "blurry, low quality, distorted, watermark").

c)LoRA-style fine-tune:

I added LoRA attention processors on top of the UNet and trained only those small layers on.

After a short training run, prompts like

"Futuristic city at night, neon lights, flying cars, cyberpunk, ultra detailed"

produced images with a more consistent futuristic neon style than the base model.

Stable Diffusion is heavier than Qwen3: it needs a GPU, more memory, and each image takes seconds rather than milliseconds. But visually the improvement LoRA is very clear.

4. Comparison:

Qwen3:

- It worked well for:
  - ➢ finding related profiles or documents,
  - ➢ grouping similar sentences together,
  - ➢ ranking answers by relevance.

- Weaknesses I noticed:

  - ➢ Very short are harder for the model; the results become noisy.

  - ➢ It only handles **text**, so any visual information has to be described in words first.

Overall, Qwen3 gave **stable and reliable text retrieval**, and it was easy to scale to many documents by caching embeddings.

5. Stable Diffusion:

Medium guidance (~7–9) → best balance between creativity and prompt accuracy.

Increasing denoising steps

> ➢ 20 steps → fast but softer and sometimes noisy.

> ➢ 40–50 steps → sharper, more detailed, better textures.

> ➢ 70+ steps → small quality gain but much slower; not worth it for a demo.

Using negative prompts ("blurry, low quality, distorted, extra limbs, watermark") helped remove common defects and gave cleaner images. With the small LoRA fine-tune (if you ran it), images from futuristic prompts became:

> ➢ more stylistically consistent across different prompts,

> ➢ more aligned with the look of the training images (e.g., neon colors, night scenes).

Compared to Qwen3, the Stable Diffusion results are much more visible: you can literally see the effect of each parameter change in the generated picture

6. Application:

Qwen3:

Semantic search inside large document collections (support tickets, FAQs, research papers).

Recommendation systems – match users with similar interests or match questions to existing answers.

RAG (retrieval-augmented generation) – use Qwen3 to pick the right context and then feed that context into a larger language model to generate answers.

Stable Diffusion:

Creative industries: concept art, storyboards, thumbnails, book covers, UI mockups.

Marketing and design: fast generation of variations for ads, posters, social media content

Education: quick illustrations for learning material or explanations that need pictures.

With LoRA or similar lightweight fine-tuning:

brands can train their own style, so all generated images follow a specific look. Compared with Qwen3, Stable Diffusion is more disruptive visually: people notice the change in how fast and cheaply images can be created. For this assignment, Stable Diffusion is clearly the **main star**:

- The astronaut-in-the-jungle image shows a strong mapping from text to visuals.

Query: Is ronaldo better than messi

#1 (score=0.8391) → Ronaldo scores goals grater than messi
#2 (score=0.7789) → Ronaldo Plays for portugual
#3 (score=0.7779) → Leonardo messi plays soccer
#4 (score=0.6203) → Aegentina won FIFA worldcup

- Parameter tuning (guidance, steps, negative prompts, and optional LoRA) gives **visible improvements** in style and detail.

- The demo makes it very easy to see the power of modern generative models.

Qwen3, on the other hand, plays a more **supporting role**:

- It does not create new content, but it **organizes and retrieves** existing text very well.

- In a real system, Qwen3 could be used to **understand a user's question** and fetch relevant facts, and then a model like Stable Diffusion could generate a picture that matches those facts.

7.Conclusion:

- **Qwen3** is strong at **understanding and comparing text**. It is stable, efficient, and ideal for search and recommendation tasks. In my experiments it handled synonyms and wording changes well and gave useful rankings of documents for a given query.

- **Stable Diffusion v1.5** is strong at **creating new visual content**. Through prompt design, parameter tuning, and optional LoRA fine-tuning, I could push the model toward very specific futuristic styles and clearly see the impact of each setting on the final images.

If I had to pick one as the main model for this assignment, it is **Stable Diffusion**:

- It required more complex configuration (text encoder, noise scheduler, UNet, VAE).

- It showed clear trade-offs between quality and speed when I changed guidance scale and inference steps.

At the same time, the two models **complement each other**:

- Qwen3 can be used to **understand user intent and retrieve relevant text**.

- Stable Diffusion can then **turn that intent into images**.

 In summary, the Qwen3 example shows **meaning-based retrieval**, while the Stable Diffusion example shows **high-quality visual generation**. Together they highlight how different types of generative models can be combined: one to understand and search, and the other to create.