Generative AI: GPT, DALL-E, Codex & Stable Diffusion

---

1. Introduction

- Generative AI refers to machine learning models that **create** new content: text, images, code, sounds.
- Key recent models include **GPT** (text generation), **DALL-E** (text-to-image), **Codex** (code generation), **Stable Diffusion** (image generation via diffusion models).
- Under the hood many of these use **Transformers**, **Diffusion Models**, sometimes **GANs** (Generative Adversarial Networks).

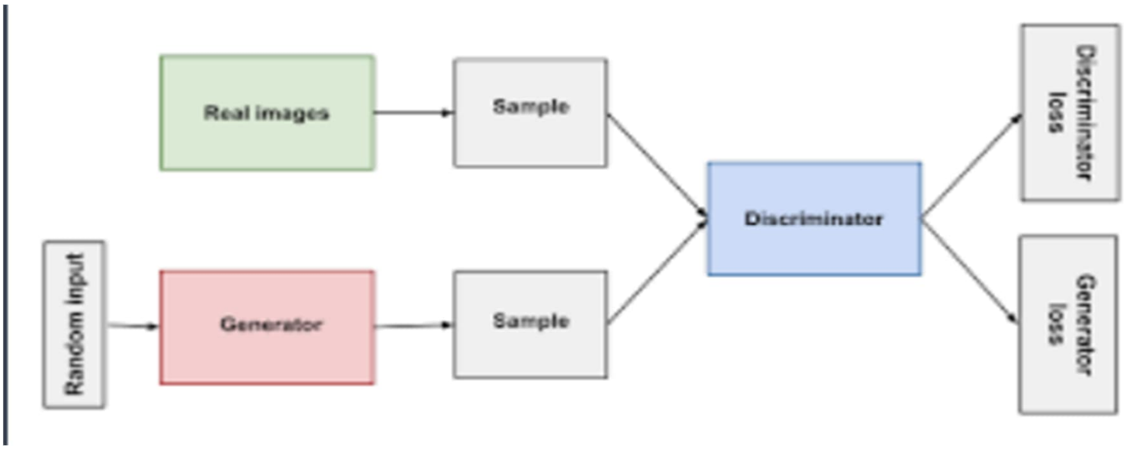
Purpose of this document:

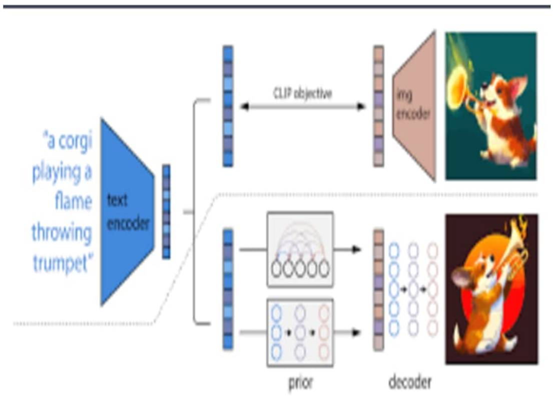
- Explain how each model works, → strengths/weaknesses → compare them → real-life uses → where things are heading.

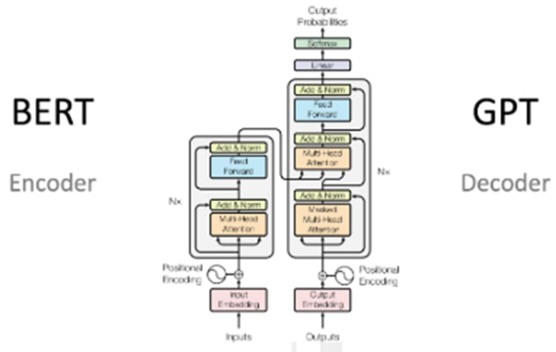
---

2. Core Architectures & Concepts

Architecture / Concept	What It Is	Key Ideas / Mechanisms
Transformers	Neural network architecture introduced in "Attention Is All You Need" (Vaswani et al., 2017). Bestarion+1	Self-attention, multi-head attention, feed-forward layers; processes sequences in parallel rather than step by step (as in RNNs) Medium+1
GANs (Generative Adversarial Networks)	Two networks: Generator + Discriminator contest each other.	Generator tries to create data to "fool" Discriminator; Discriminator distinguishes real vs fake. Good at sharp images but training unstable. Bestarion
Diffusion Models	Start from noise; gradually denoise to produce data matching desired distribution.	Forward process adds noise; reverse process learns to remove noise step-by-step. Stable and high quality output. Bestarion+1







3. Model Overviews

Here are summaries of the four models.

Model	Type / Task	Basic Mechanism	Inputs & Outputs	Strengths
GPT (e.g. GPT-3, GPT-4)	Large Language Model (text generation, understanding)	Transformer decoder / autoregressive: predict next token given prior tokens; large pretraining on text corpora.	Input: text prompt; Output: text continuation, answers, summaries, etc.	Great fluency; broad general knowledge; very strong at language tasks; versatile (can be used for chat, summarization,

				translation, etc.).
Codex	Code generation & understanding	GPT-style model (language modeling) but trained significantly on code + code-related data.	Input: prompt in (natural language + maybe code), output: code snippet, autocomplete etc.	Helps programmers; automates boilerplate; can generate working code; integrate with tools (e.g. GitHub Copilot).
DALL-E / DALL-E 2 / DALL-E 3	Text-to-Image generation	Earlier versions used discrete VAE + autoregressive sequence modelling; later versions use diffusion conditioned on text embeddings (often via CLIP or similar). Wikipedia+2inceptivetechnologies.com+2	Input: textual description prompt; Output: image fulfilling prompt.	Very good at creative image generation; high quality; increasingly good at following prompts semantically; usable for design, art, storyboarding etc.
Stable Diffusion	Text-to-Image (with open accessibility)	Diffusion model + conditioning on text embeddings; latent space diffusion (faster, more efficient) etc. inceptivetechnologies.com+1	Input: text prompt (and optionally image for image-to-image or inpainting etc.); Output:	Open source; more customizable; usable on consumer-class GPUs; large community; supports

			high resolution image.	image editing, variations etc.

---

4. Comparisons

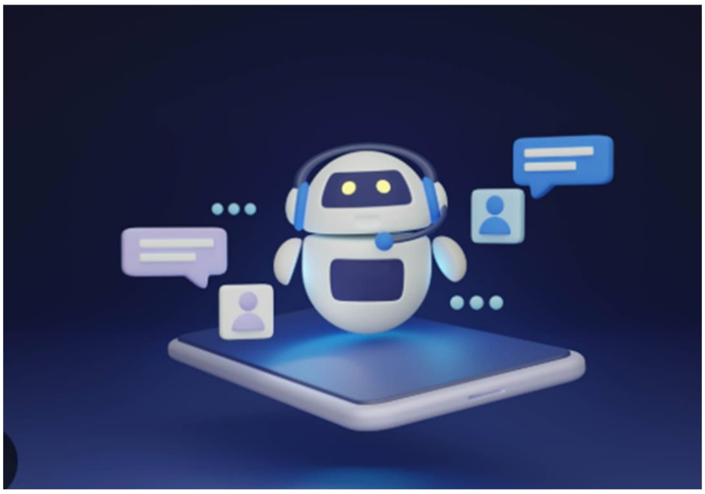
Here's a side-by-side comparison of the models in different dimensions.

Comparison Dimension	**GPT / Codex**	**DALL-E**	**Stable Diffusion**
Speed of generation	Text generation is fast for short outputs; code generation depends on size.	Typically fast for images once model is loaded; earlier versions slower.	Slightly slower because diffusion needs multiple denoising steps; but latent diffusion speeds things up.
Prompt adherence / Semantic fidelity	Very good (for text); sometimes hallucinates; needs good prompt engineering.	High; often good at visualizing what prompt describes; sometimes misses details.	Also high; improvements over time; good trade-offs for resolution vs fidelity.
Resource requirements / accessibility	Large compute for training; inference can be lighter.	Image generation models require GPU; inference cost non-trivial.	Because of optimizations and open source, more accessible to developers; can run locally with decent GPUs.
Flexibility / Customization	Highly flexible for many text tasks; can be fine-tuned, adapted.	Some versions allow style, variation, editing features; version dependency.	Very flexible; many community tools, models, fine-tuning; supports image-to-image etc.
Limitations / Weaknesses	Produces incorrect info ("hallucinations"); bias in training data;	Can misinterpret prompts; may produce	Slower sampling; sometimes struggles with specific details;

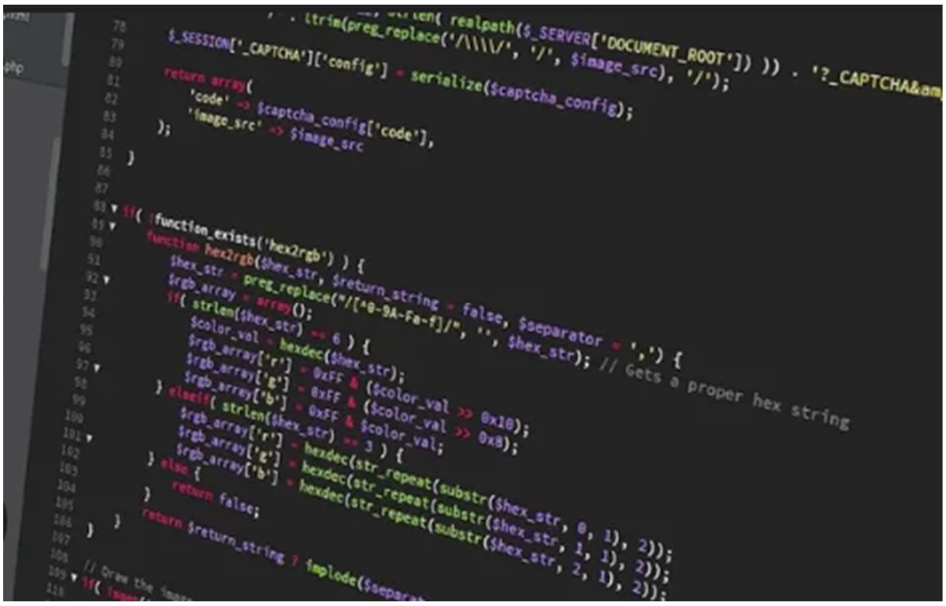
	very large models are expensive.	artifacts; sometimes lacks fine detailed control.	requires strong GPU for best performance; data bias issues.

---

5. Real-Life Applications & Use Cases

- **GPT / Language Models**
  • Chatbots & virtual assistants (customer support)
  • Content generation: articles, marketing copy, summarization, translation
  • Legal, medical, technical writing assistance
  

- **Codex / Code Generators**
  • Autocomplete / code suggestion tools (e.g. GitHub Copilot)
  • Learning aid: explain code, generate examples
  • Automating repetitive coding tasks



- **DALL-E / Stable Diffusion**
  • Art & illustration: concept art, storyboards, design mockups
  • Marketing & advertising visuals
  • Product visualisation (e.g. furniture, fashion)
  • Image editing: inpainting, image variation, style transfer

- **Hybrid / Multimodal Uses**
  • Generating images from text prompts within chat interfaces (e.g. using GPT + image models)
  • Tools for creators: combining code, text, and images

---

6. Why Diffusion > GANs in Many Modern Image Models

- GANs were historically very popular: high fidelity output, fast sampling; but problematic training (mode collapse, instability). Bestarion+1
- Diffusion models offer more stable training, better diversity, often better visual quality especially under prompt conditioning. Bestarion+2arXiv+2
- The trade-off is that diffusion often requires more computational work at inference (multiple iterative steps) vs GANs which can generate in a single forward pass.

---

7. Tables: Architectures & Trade-offs

Here are two comparative tables you can include:

Table A: Architecture & Training Differences

Feature	GANs	Diffusion Models	Transformers (Autoregressive e.g. GPT)
Training stability	Often unstable; issues like mode collapse	More stable, well-behaved training	Stable when scaled; pretraining + fine-tuning works well
Inference speed	Very fast (single pass)	Slower (many diffusion steps)	Fast for text generation; depends on model size for large outputs
Output diversity / safety	Risk of missing modes; sometimes less diverse	Good diversity; more robust	For text/code, risk of hallucination, bias but high capability
Conditional control (style, prompt adherence)	Possible but tricky	Strong in prompt conditioning; style control improving	Strong in prompts; control via prompt design / fine-tuning

Table B: Model Comparison — GPT vs Codex vs DALL-E vs Stable Diffusion

Model	Primary Domain	Approx. Model Size / Training Data	Best Use Case	Weakness / Caveat
GPT	Languag e (text)	Very large text corpora; billions of parameters	Writing, summarizati	May produce incorrect/fabric

			on, chat, knowledge tasks	ated info; bias; large resource usage
Codex	Code + Language	Code dataset + natural text; tuned to code style	Generate code, help developers, auto-complete	Sometimes produces incorrect or insecure code; not always context-aware
DALL-E	Text→Image	Large image-text paired datasets; uses CLIP and diffusion/autoregressive parts Wikipedia+2inceptivetechnologies.com+2	High quality image generation from text; creative art	Cost / compute; limitations in detail & control; prompt sensitivity
Stable Diffusion	Text→Image (open access)	Uses latent diffusion; somewhat more efficient; trained on diverse image-text pairs inceptivetechnologies.com+1	Customized art, image editing, community use, inpainting	Slower generation (many steps); sometimes artifacts; dependency on prompt/data quality

---

8. Limitations, Ethical & Practical Challenges

- Bias & fairness: Models reflect biases in their training data; can generate biased or inappropriate content.
- Hallucination (in text models): GPT may produce statements that are grammatically correct but factually false.
- Prompt sensitivity: Small changes in prompt can lead to very different outputs.
- Resource & energy cost: Very large models need massive compute and power for training and inference.
- Intellectual property & copyright issues: Training on large scraped datasets raises questions of ownership, copyright.

---

9. Future Directions

- Better efficiency (faster inference, fewer parameters) using techniques like distillation, quantization, sparse / efficient transformer variants.
- More controllability: controlling style, safety, details, reducing unwanted outputs.
- Multimodal models that combine text, images, video, audio more seamlessly.
- Democratization: open-source models, tools accessible to smaller groups (Stable Diffusion is an example).
- Ethical & regulatory frameworks for use, dataset sourcing, bias mitigation.

---

10. Conclusion

- GPT, Codex, DALL-E, Stable Diffusion represent a new age of generative AI capabilities. Each has its domain and strengths.
- Diffusion models are pushing image generation quality forward, while transformers remain central for language, text, code.
- In real use, often the best solutions come from combining models or using hybrid pipelines.
- As technology matures, concerns of cost, ethics, interpretability & controllability will be as important as raw capability.