key concepts in modern AI:

---

1. LLM (Large Language Model)

- **Definition**: A type of artificial intelligence model trained on massive amounts of text data to understand and generate human-like language.
- **Example**: GPT-4 (what you're interacting with now), Claude, Gemini.
- **Uses**: Chatbots, text generation, summarization, translation.

---

2. NLP (Natural Language Processing)

- **Definition**: A field of AI focused on enabling computers to understand, interpret, and generate human language.
- **Includes tasks like**:
  - Sentiment analysis
  - Named entity recognition
  - Machine translation
  - Question answering

---

3. RAG (Retrieval-Augmented Generation)

- **Definition**: A technique that combines LLMs with external knowledge retrieval systems.
- **How it works**: Instead of relying only on its training, the model fetches relevant documents (e.g., from a database or search engine) to answer a question more accurately.
- **Example**: ChatGPT with browsing, or search-augmented chatbots.

---

4. AI Agent

- **Definition**: An AI system that can autonomously perform tasks, often over multiple steps, to achieve a goal.
- **Key traits**:
  - Decision-making
  - Task planning
  - Memory
  - Autonomy
- **Examples**: AutoGPT, babyAGI, AI assistants that can schedule meetings, send emails, or code.

---

5. Workflow (in AI context)

- **Definition**: A structured sequence of tasks or processes, often involving multiple AI tools or models.
- **Use in AI**:
  - Chaining models together (e.g., image → caption → summary)
  - Managing task flow in AI agents
  - Automation pipelines in business or data science

---

6. Fine-Tuning

- **Definition**: The process of training a pre-trained model further on a specific dataset to specialize it for a particular task or domain.
- **Why it matters**: Makes general models perform better on domain-specific problems (e.g., legal or medical texts).

---

7. Diffusion (Models)

- **Definition**: A type of generative model used to create images, videos, or audio from noise by reversing a diffusion process.
- **How it works**: Starts with random noise, gradually "denoises" to create coherent outputs.
- **Example**: DALL·E, Stable Diffusion, Midjourney.

---

8. Transformer

- **Definition**: A neural network architecture that's foundational to most modern LLMs and generative models.
- **Key innovation**: Uses "attention" to weigh parts of input data differently, enabling better understanding of context.
- **Introduced in**: The 2017 paper *"Attention Is All You Need"*.

Here are some **examples of Transformer models**, grouped by task and generation:

□ Foundational Transformer Models

These are general-purpose models that use the Transformer architecture:

1. **BERT (Bidirectional Encoder Representations from Transformers)**
   - **Use**: Text classification, question answering.
   - **Key trait**: Reads text bidirectionally.
   - **Company**: Google (2018)
2. **GPT series (Generative Pretrained Transformer)**
   - **Use**: Text generation, dialogue, code generation.
   - **Key trait**: Unidirectional (left to right), autoregressive.
   - **Company**: OpenAI
3. **T5 (Text-To-Text Transfer Transformer)**
   - **Use**: Converts all NLP tasks into a text-to-text format.
   - **Company**: Google
4. **RoBERTa (Robustly Optimized BERT Approach)**
   - **Use**: Improved version of BERT with better training.
   - **Company**: Facebook AI

□ Multimodal Transformers

Handle **multiple types of input** like text, images, or audio.

1. **DALL·E / DALL·E 2 / DALL·E 3**
   - **Use**: Text-to-image generation
   - **Company**: OpenAI
2. **CLIP**
   - **Use**: Connects vision and language (e.g., match images to captions)
   - **Company**: OpenAI
3. **Flamingo / Gemini / GPT-4o**
   - **Use**: Visual + language understanding in a single model
   - **Companies**: DeepMind (Flamingo, Gemini), OpenAI (GPT-4o)

□ Code Transformers

Specialized for programming languages:

1. **Codex**
   - **Use**: Code generation (used in GitHub Copilot)
   - **Company**: OpenAI
2. **CodeBERT / CodeT5**
   - **Use**: Code understanding and generation

- **Companies**: Microsoft Research, Salesforce Research

□ Audio / Speech Transformers

1. **Whisper**
   - **Use**: Speech-to-text (automatic speech recognition)
   - **Company**: OpenAI
2. **Wav2Vec 2.0**
   - **Use**: Speech recognition, audio feature learning
   - **Company**: Facebook AI

□ Transformer in Vision (ViT)

1. **ViT (Vision Transformer)**
   - **Use**: Image classification, object detection
   - **Company**: Google Research
   - **How**: Splits images into patches and processes them like words in text.