Artificial Intelligence Transformers

Table of Contents

1 — NLP TRANSFORMERS

A. Encoder-Only Transformers

Model	What It Is	Why It Exists	Key Features	Use Cases
BERT	Bidirectional encoder	Understand full sentence context	Bidirectional attention	Classification, QA
RoBERTa	Improved BERT	Fixes BERT's training limits	No NSP, more data	Same as BERT
DistilBERT	Compressed BERT	Edge/mobile NLP	97% performance	On-device NLP
ALBERT	Memory-efficient BERT	Reduce parameters	Parameter sharing	Search, QA
XLNet	Permutation LM	Better context	Permutation attention	NLP understanding

---

B. Decoder-Only Transformers

Model	What It Is	Why It Exists	Key Features	Use Cases
GPT (1–4, 4o)	Autoregressive decoder	Text generation & reasoning	Massive datasets	ChatGPT, coding

---

C. Encoder–Decoder Transformers

Model	What It Is	Why It Exists	Key Features	Use Cases

T5	Text → Text transformer	Unified NLP interface	Text-to-text	Summarization
BART	BERT + GPT hybrid	High-quality generation	Denoising seq2seq	Rewriting, QA

---

2 — VISION TRANSFORMERS

Model	What It Is	Why It Exists	Key Features	Use Cases
ViT	Patch-based transformer	Replace CNNs	Patch embeddings	Classification
DeiT	Efficient ViT	Train with less data	Distillation	Lightweight CV
Swin	Hierarchical transformer	Scale to large images	Shifted windows	Detection
MAE	Self-supervised ViT	Reduce labeling cost	Masked autoencoding	Pretraining
BEiT	Visual BERT	Token-based vision	Discrete visual tokens	Recognition

---

3 — AUDIO & SPEECH TRANSFORMERS

Model	What It Is	Why It Exists	Key Features	Use Cases
Whisper	Transformer ASR	Noisy multilingual speech	Robust encoder–decoder	Transcription
Wav2Vec 2.0	Raw audio transformer	Learn from unlabeled audio	Self-supervised learning	ASR
HuBERT	Hidden-unit transformer	Better speech representations	Predict masked units	ASR
FastSpeech	Transformer TTS	Extremely fast speech generation	Non-autoregressive	TTS
Transformer-TTS	Transformer-based Tacotron	High-quality speech synthesis	Attention-based	Voice generation

---

4 — MULTIMODAL TRANSFORMERS (Text + Image + Audio + Video)

Model	What It Is	Why It Exists	Key Features	Use Cases
CLIP	Vision + text transformer	Image–text understanding	Contrastive learning	Search
BLIP	Vision–language transformer	Strong multimodal alignment	Multi-stage training	Captioning
Flamingo	Few-shot multimodal LM	Multimodal reasoning	Large context window	VQA
LLaVA	Vision + LLM	Add vision to LLMs	ViT + LLM fusion	Image chat
GPT-4o	Fully multimodal	Unified model for all modes	Native audio/video	Multimodal AI
Kosmos	Grounded multimodal transformer	Visual reasoning	Joint embeddings	VQA

---

5 — EFFICIENT / SPECIALIZED TRANSFORMERS

Model	What It Is	Why It Exists	Key Features	Use Cases
Longformer	Long-text transformer	Process long documents	Sliding-window attention	Legal, research
Reformer	Efficient memory transformer	Reduce computation	LSH attention	Long sequences
Performer	Linear transformer	Scale attention	Random feature kernels	Real-time AI
Linformer	Efficient transformer	Reduce attention cost	Low-rank projection	Edge NLP
RetNet	Transformer recurrence	Memory-efficient inference	Recurrent representations	Streaming LLMs
BigBird	Sparse-attention transformer	Very long sequence handling	Sparse tokens	Document AI
TinyBERT / MobileBERT / TinyLlama	Compressed transformers	Deploy on mobile	Distilled + quantized	On-device AI