How GPTs Have Evolved from 2017 – 2026: The 14 Improvements of Modern GPTs

Chai K Toh

Safe AI Foundation USA

20 JULY 2026

Natural language processing (NLP) is the ability to understand, interpret, and generate human language in text or speech form. It combines linguistic rules, statistical models, and machine learning (including deep learning) to bridge the gap between human communication and computer understanding.

Prior to the arrival of the transformer architecture, NLP uses statistical models like n-grams and Hidden Markov Models (HMMs), and later Gated Recurrent Unit (GRUs), and Long Short-Term Memory (LSTM) networks.

Traditional NLP methods — from early rule-based systems to older statistical models — often failed to handle the full complexity of human language because they were built on assumptions and techniques that no longer match the scale and diversity of modern language data.

╠ 1. Limited Linguistic Coverage and Diversity

Early NLP systems were designed for specific languages or domains, often ignoring the thousands of languages with unique grammars, vocabularies, and cultural nuances. This made them brittle when applied to new or under-represented languages.

╠ 2. Ambiguity and Context Sensitivity

The human language is inherently ambiguous — words can have multiple meanings (polysemy), sounds the same but differ in meaning (homophony), and meaning depends heavily on context. Traditional approaches relied on fixed dictionaries and syntactic rules, which could not adapt meaning dynamically. Even modern deep learning (DL) models can still misinterpret sentences without strong contextual cues.

╠ 3. Over-Reliance on Syntactic Rules

Older systems focused heavily on parsing sentence structure (syntax) and matching patterns, but human meaning (semantics) is deeply tied to context,

pragmatics, and world knowledge. Rule-based systems often missed subtle inferences or idiomatic expressions.

╠╡ 4. Scalability and Data Limitations

Pre-2010s NLP models were trained on small, curated datasets that could not generalize well to the vast, unstructured, and noisy data of the web. This led to poor performance in real-world applications like open-domain question answering or conversational systems.

╠╡ 5. Inability to Capture Semantic Richness

Traditional models struggled with rich semantic content — metaphors, sarcasm, cultural references — because they lacked the ability to learn latent representations of meaning from large-scale data.

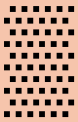
╠╡ 6. Domain and Task Rigidity

Many older systems were task-specific and required manual feature engineering for each application. This made them inflexible and hard to adapt to new domains or user needs.

The transformer architecture was introduced in 2017 by Vaswani et al. in the paper “Attention Is All You Need” as a direct response to the limitations of earlier sequence modeling approaches. The transformer architecture overcomes the sequential bottlenecks, training inefficiencies, and long-range dependency limitations of RNNs (Recurrent Neutral Networks) and LSTMs, while enabling parallel, scalable, and context-rich sequence modeling.

An Large Language Model (LLM) is any AI model trained to understand and generate human language. A GPT (Generative Pre-trained Transformer) is a specific model within the LLM category that uses the transformer architecture. This is the distinction between LLMs and GPTs. The term GPT is not reserved solely for Open AI. Germini, Llama, Meta, Claude can all be termed as a GPT.

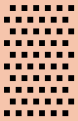
Evolution Timeline...



2017:

Transformer was born.

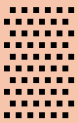
→ Better sequence modeling, parallelism



2018 – 2020:

GPT Scaling.

→ Bigger models, more data, emergent abilities



2020 – 2022:

Instruction Tuning & RLHF.

→ Models become assistants

Evolution Timeline...

Evolution Timeline...

2022 – 2024:

Tools, Multi-modality & Longer context.

→ Models become interactive systems

2024 – 2026:

Reasoning Systems, Agents & Automated Workflows.

→ Models increasingly perform multi-step tasks

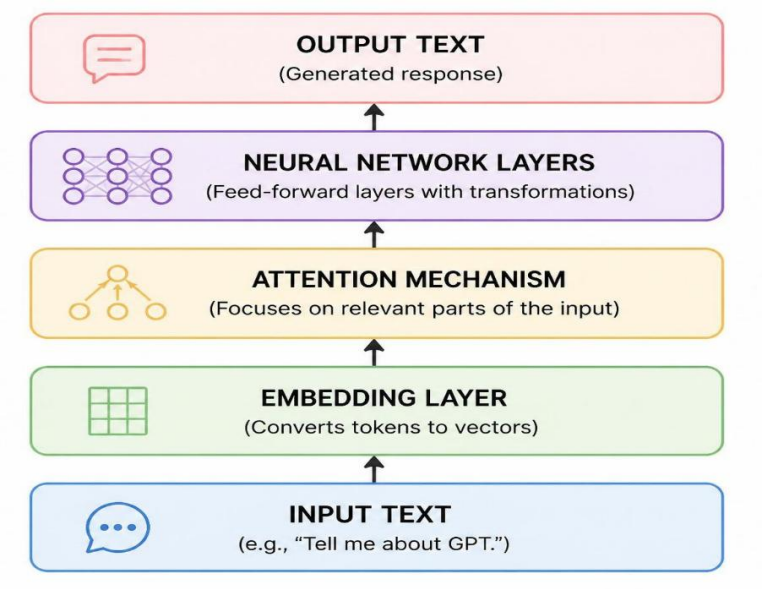


Fig 1: A Simplified Block Diagram of a GPT.

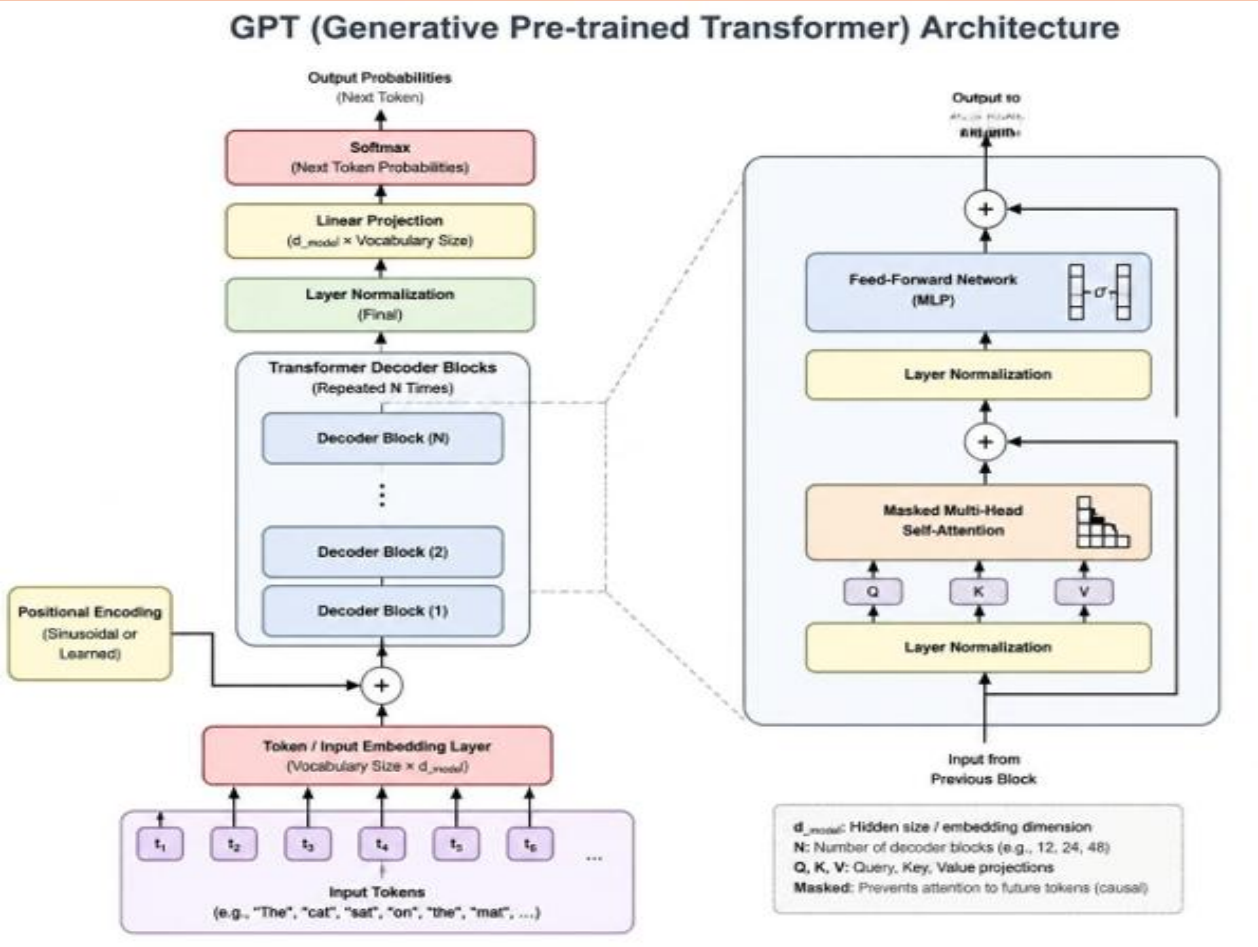


Fig 2: A GPT in detailed form, revealing internal blocks.

Below, we elaborate the 14 advances made to date that result in modern GPTs:

1. LARGE-SCALE PRE-TRAINING

The original Transformer was a general neural architecture. GPT (Generative Pre-trained Transformer) added the idea of training a Transformer decoder on enormous amounts of text using self-supervised next-token prediction.

Key additions:

- Massive datasets
- Hundreds of billions (or more) of training tokens
- Large parameter counts
- Distributed training across thousands of GPUs/accelerators

This turned the Transformer from a sequence model into a broad language-learning system.

2. SCALING LAWS

Researchers discovered that model performance tends to improve predictably with increases in:

- Parameters
- Data
- Compute

Parameters are numbers, specifically referring to weights in a neural network. This led to systematic scaling rather than simply making models bigger by trial and error. This also resulted in the greater demand for data centers.

3. INSTRUCTION TUNING

Early GPT models mostly predicted text. Instruction tuning changed the objective.

Instead of: "Continue this text."

The model learns: "Follow this human request."

This involves curated examples of:

- Questions and answers
- Tasks and solutions
- Explanations
- Formatting preferences

This explains why GPTs' answer tend to be longer now, with follow up options.

4. REINFORCED LEARNING from HUMAN FEEDBACK (RLHF)

A major shift was teaching LLM models human preferences or feedback. This requires actual action from the GPT user. A simplified version is:

1. Humans rank model responses
2. A reward model learns those preferences
3. The AI is optimized to produce higher-rated responses

This helped produce assistants that are more helpful, less likely to produce undesirable outputs, and better aligned with conversational expectations.

5. PREFERENCE OPTIMIZATION METHODS beyond RLHF

Researchers have developed alternatives and improvements, including:

- Direct Preference Optimization (DPO)
- Reinforcement learning variants
- Constitutional AI approaches
- AI feedback methods

Preference optimization methods are designed to answer a question that next-token prediction alone doesn't address. It captures what people like (in the answers) and work on producing those answers. These make alignment training more stable and scalable.

6. BETTER ATTENTION MECHANISMS and ARCHITECTURES

The original Transformer has evolved with many improvements. Examples are:

- More efficient attention mechanisms
- Grouped-query attention & Multi-query attention
- Rotary positional embeddings (RoPE)
- Mixture-of-Experts (MoE) architectures
- Improved normalization methods
- Better activation functions

These improve efficiency, context length, and reasoning ability.

7. LONGER CONTEXT WINDOW

Early Transformers could only handle relatively short sequences. Modern systems can process:

- Long documents and books
- Codebases
- Extended conversations

This required advances in:

- Positional encoding
- Attention efficiency
- Memory management
- Retrieval methods

8. RETRIEVAL-AUGMENTED GENERATION (RAG)

GPTs now have answers augmented by external documents and databases. Instead of relying only on stored model weights:

1. The system searches a database or documents.
2. Relevant information is retrieved.
3. The model uses it to answer.

This allows:

- More up-to-date information
- Company-specific knowledge
- Document analysis

9. TOOLS USAGE & AGENTIC SYSTEMS

In 2022, LangChain (an open-source framework) was created to integrate LLMs into applications. Modern GPT systems can now be connected to:

- Web search
- Code execution
- Databases
- APIs
- File analysis
- External applications

The LLM model becomes a reasoning layer that can act through software.

10. MULTI-MODAL LEARNING

GPT-like systems have expanded beyond text. Inputs can now include:

- Images
- Audio
- Video
- Code
- Other structured data

This requires:

- Vision encoders
- Audio models
- Cross-modal alignment techniques

11. SYNTHETIC DATA GENERATION

Modern training increasingly uses AI-generated data, such as:

- Problem sets and conversations
- Code examples and reasoning traces
- Critiques of other outputs

The challenge is to avoid feedback loops where models amplify their own errors.

12. BETTER REASONING TECHNIQUES

Recent systems use methods such as:

- Chain-of-thought (COT)-style internal reasoning
- Self-consistency methods
- Search over possible solutions
- Verification passes & critic models
- Planning loops

The key idea is moving from:

"Generate the first plausible answer"

toward:

"Generate, evaluate, and refine."

13. SAFETY & ALIGNMENT SYSTEMS

Modern GPTs have included additional layers, such as:

- Safety training and monitoring
- Red-team evaluation & risk assessments
- Refusal behavior tuning
- Policy-guided behavior

The safety aspects have to include acceptable ethical behavior, compliances to laws, input-model-output security, responsible behavior, etc. This area is currently lacking and demands further work and improvements.

▒ 14. INFERENCE-TIME IMPROVEMENTS

Not all progress comes from training. Systems can be improved through:

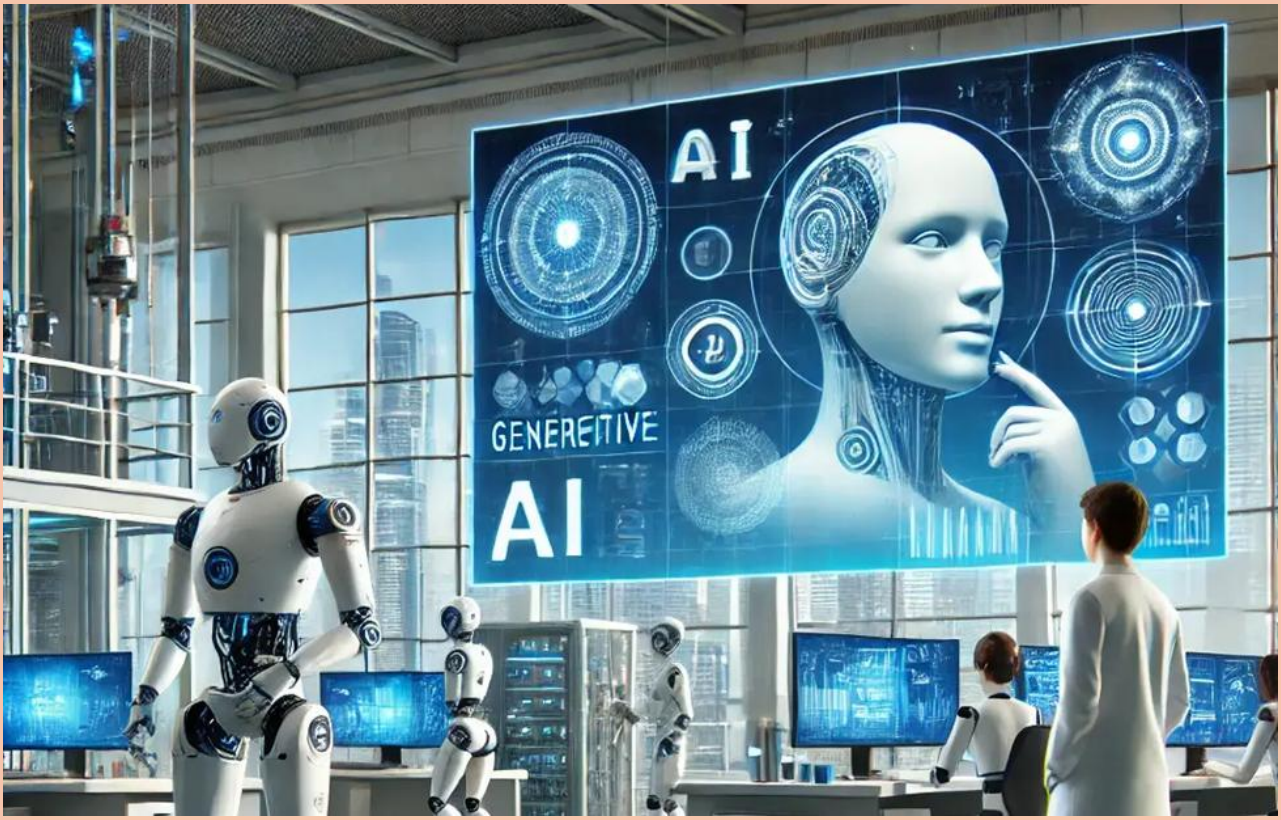
- Better prompting
- Dynamic computation
- Routing between models & tool selection
- Verification & memory systems

These 14 improvements have advanced GPTs to what they are today! Transformer was the engine, but Modern GPT Systems can be viewed as:

> **A Complete AI Stack** =
>
> Architecture + Training Curriculum + Safety + Retrieval + Tools + Reasoning Methods + Infrastructure.

The biggest conceptual shift has been moving from "a model that predicts text" to a system that can “reason, use information, and act”.

\\\ end ///



SAFE
AI
FOUNDATION

www.safeaifoundation.com

The SAFE AI Foundation was formed in 2025, in the heart of the Silicon Valley, California, USA. Its mission is to embrace the AI revolution and promote the use of good AI to improve on the quality of life and the economy. It will address the issues of bad AI and its impacts on society. The foundation will inform, educate, and mitigate the dangers of bad AI and help the society to embrace good AI. The foundation will interact with industries, universities and public sector to jointly advance the progress of Good AI. The foundation is a non-profit 501c(3) organization, registered in the State of California, USA.

Note: We have simplified our explanations to reach the general audiences. Engineers and recent graduates are encouraged to do further readings. A GPT is used to assist us in producing some parts of this document.