SWOT Analysis of Generative AI Models

Strengths - internal attributes and resources that support a successful outcome.
Weaknesses - internal attributes and resources that may hinder performance or achievement.
Opportunities - External factors that the subject could leverage to its advantage.
Threats - External challenges that could cause trouble or jeopardize success.

GPT-5 by OpenAI (paid tool)	Gemini 1.5 Pro by Google DeepMind (free)	Claude 3 Opus by Anthropic (paid tool)	DeepSeek (free)	Co-Pilot by Microsoft (paid tool)
Architecture: GPT-4 (Generative Pre-Trained Transformer 4) is built on the Transformer architecture, significantly enhanced from its predecessors. Underlying code considered IP of OpenAI. Parameter event business mode with hyper-parameter tuning and neural network.	**Architecture:** Proprietary LLM Utilizes a Mixture of Experts (MoE) architecture, a departure from the traditional Transformer architecture used in many large language models.	**Architecture:** Clause API, Amazon Bedrock, and Google Vertex AI Model Garden. Modular. hyper-parameter tuning for model customizations and prompt engineering. Evolution of Anthropics's AI models, built on a robust Transformer architecture with a focus on safety and alignment	**Architecture:** Transformer with several innovations including MoE and Multi-head Attention (MLA). The hybrid approach allows for efficient resource allocation. multimodal	**Architecture:** Powered by OpenAI's Codex model. Open source technologies SignalR, Adaptive Cards, Markdown, and object-basin used.
Parameter Count: It boosts 175 billion parameters, making it one of the largest models developed by OpenAI	**Parameter Count:** 1 trillion parameters plus experts	**Parameter Count:** It boosts 137 - 175 billion parameters designed for balance, making it one of the largest models developed by OpenAI	**Parameter Count:** 236 billion parameters	**Parameter Count:** 3.3 billion

Trained Data: Use of weights and biases for encodes. Capable of processing contexts of 128,000 tokens in the latest version, significantly enhancing its ability to handle long-term content and complex dialogues. Speculatively trained on 1 trillion tokens.	**Trained Data:** Trained on massive datasets of multilingual and multimodal data, including publicly available sources and Google Cloud specific information. Accuracy on content, good explainability. Use of weights and biases to enable model predictions and generate text	**Trained Data:** Well trained on workloads and increased knowledge base. Trained on carefully curated dataset that emphasizes ethical considerations and diverse representations of text.	**Trained Data:** Massive amounts of code and natural language. V3 trained on 14.8 trillion tokens. Coder trained on 2 trillion tokens with 87% code and 13% linguistic data.	**Trained Data:** Leverages GPT-4o which has context window of up to 128k for certain use cases
Token Context Window: solid process to determine how much data to process	**Token Context Window**: Supports a token context window of up to 2 million tokens, the longest context window of any large-scale foundational model yet, enabling unprecendented context retention and long-form generation capabilities	**Token Context Window**: Handles a token context window of 200,000 tokens, ensuring coherent responses over extended conversations and detailed document analysis.	**Token Context Window:** High	**Token Context Window:** leverage GPT
Performance: More hallucinations compared to previous version. Improved language	**Performance:** Computationally expensive. Leading benchmark for Massive Multi -	**Performance:** Self-awareness, markdown formatting, model variants, fast, does not maximize.	**Performance:** Outstandingly performant, fast, high accuracy. Limited language	**Performance:** Improved File handling, Outlook availability, Loop integration, suggests

capabilities. Cloud adversely impacts performance. If incorrect prompts inputted - output is suboptimal or false. SOTA. Leading benchmark for Massive Multi - Task Language Understanding (MMLU)	Task Language Understanding (MMLU)	memory. Strong in NLP and operations tasks. Leading benchmark for Massive Multi - Task Language Understanding (MMLU)	support.	autocompletes or blocks of code, access third party services via plugins
Use Cases: Context, translation, Dall-E integration, reading documents, information retrieval, image generation, translation, summaries. General purpose and conversational tasks. Generate human like text input. Long-form content and complex dialogues. Real-world examples. Conversion optimization. Creating Posts.	**Use Cases:** Task versatility, accuracy. SMM Creating Posts Excels at multimodal applications Illustrative tools AI-assisted documentation. Templates for project planning.	**Use Cases:** Substantive responses to questions from diverse topics. Analyze complex queries. Content creation. Detailed document and extended conversations. Content Marketing, Lead Paragraph. SMM. Paid Promotion. Outreach. Landing pages. Headlines, Lead Paragraphs. Creating Posts. Lead capture and strategic webinar planning. Illustrative tools	**Use Cases:** MoE for specialized interests. Domain specific and privacy-conscious tasks. Task-specific precision, complex problem solving, reasoning expertise, math expertise, data analysis, context understanding, research. Developer and researcher empowerment. Demographic Generation - lead capture and strategic webinar planning. Creating Posts	**Use Cases:** GitHub integration for coding and debugging Coding assistance. Developer empowerment, answers complex questions, generates creative content, multilingual, summarizes information, integration with Excel, rewrite content, GitHub

Marketing Analytics. Illustrative tools				
Attention Mechanism ; Advanced and complex core to Transformer architecture - gives more weight to relevant words through larger context window.	**Attention Mechanism:** MoE - (Mixture of Elelments) Designation of experts within large neural network and training gateway network to activate best suited input	**Attention Mechanism:** Enhanced Transformer architecture, incorporating sparse attention to improve accuracy and performance trained on diverse data types to enable versatility in various tasks.	**Attention Mechanism:** MoE - only activates subset of parameters at a given time. Designation of experts within large neural network and training gateway network to activate best suited input	**Attention Mechanism:**
Fine-Tuning: Excels at filter outputs Improved accuracy and relevance, efficiency, reduced bias, personalized experience. Computationally expensive	**Fine-Tuning**: Improved performance on specific tasks. Domain expertise. Format customization, Ability to handle edge cases. Cons: requires Labeled data, computational extensive, non ideal for evolving, real-time, dynamic information, requires expertise.	**Fine-Tuning**: Not GA	**Fine-Tuning:** Domain-specific knowledge. Improved accuracy. Bias reduction. Task adaptation.	**Fine-Tuning:** Beneficial combined with MSFT 365: data security and privacy, enhanced customizations, improved accuracy and relevance, reduced dependency on prompt engineering
Ethics and Safety: Closed environment mitigate risks	**Ethics and Safety**: Closed source code	**Ethics and Safety:** Closed source model. Guardrails,	**Ethics and Safety:** SOTA data security. Models are	**Ethics and Safety:** Enhanced Data Protection. Core

and misuse		harvested, defends against hallucinations automatically.	open-source.	engine and application are closed source.

Earlier Versions of Language Models

N-Gram Model
2010s

Probabilistic model that requires tokenization input to assign probability distribution to predict next word in a sequence.

Tokenization Inputs:

Unigram N=1- prediction based on individual word probabilities
Bigram N=2 - prediction based on previous one word
Trigram N=3- prediction based on previous two words

Essential in NLP for token “n” prediction and machine translation. Rather than computing entire word history - discards information outside of token windows - produces approximations on just a few last words. Purely a statistical model; only counts based; no syntax, no word meaning, no word knowledge. Out of vocabulary is a common challenge. It is critical to have relevant training data for worthy output. Any ill-formed texts, falsehoods, nonsensical statements will render low probabilities.

Recurrent Neural Network (RNN)
2012-2015

Sequential neural network. Processes one word at a time. Ability to handle variable context length but prone to high latency and memory intensive. Performance is slow.