September 1, 2023

# Everything We Know About GPT-4

by Stephen M. Walker II, Co-Founder / CEO

**TOP TIP**

Q3 2023: This model card represents the combined publicly available GPT-4 model research – we stand on the shoulders of giants (and aggregate their data points) in this analysis.

## GPT-4 2023's State of the Art LLM

GPT-4 represents a major leap forward in large language model capabilities. Developed by OpenAI, it builds on the architecture and strengths of GPT-3 while achieving new levels of scale and performance.

With GPT-4, OpenAI's objective was to create a model that was over 10x larger than GPT-3. This requires not just greater compute for training, but entirely new approaches to model architecture and inference serving.

Some key facts about GPT-4:

- **Total parameters** — ~1.8 trillion (over 10x more than GPT-3)

- **Architecture** — Uses a **mixture of experts (MoE)** model to improve scalability
- **Training compute** — Trained on ~25,000 Nvidia A100 GPUs over 90-100 days
- **Training data** — Trained on a dataset of ~13 trillion tokens
- **Inference compute** — Runs on clusters of 128 A100 GPUs for efficient deployment
- **Context length** — Supports up to 32,000 tokens of context

**TOP TIP**

Review GPT-4 outputs carefully before use, as the model can generate harmful, biased, or factually incorrect text without proper oversight.

# GPT-4 Model Card

## Model Details

| Parameter | Detail |
|---|---|
| Organization | OpenAI |
| Model name | GPT-4 |
| Model type | Transformer with Mixture-of-Experts |
| Parameters | 1.8 trillion |
| Context Window | 8-32,000 tokens |
| Launch Date | March 2023 |

| Parameter | Detail |
| --- | --- |
| Current Version | 1.1 (Release 06.13) |
| Training dataset | 13 trillion tokens (web text, books, other) |

## Compute

| Compute | Detail |
| --- | --- |
| Training | 90 days on 25,000 Nvidia A100 GPUs |
| Inference | 128 A100 GPU clusters |

## Training Data

| Parameter | Detail |
| --- | --- |
| Data sources | CommonCrawl, WebText2, books, Wikipedia, Reddit, Amazon reviews |
| Data volume | ~13 trillion tokens |
| Data prep | Deduplication, cleaning, filtering |
| Potential biases | Language, gender, race representation |

## API and Data Format

- Chat Completion API
- Multi-turn message types
- System, Function, User, Assistant
- JSONL fine-tuning with message arrays

## Intended Use

- Text generation
- Question answering
- Classification
- Conversational agents

## Factors

- **Language** — English
- **Capabilities** — Text generation, question answering, text classification
- **Modalities** — Text
- **Ethical considerations** — Potential for bias, harmful outputs, misuse

## Metrics

- **Perplexity** — Unknown
- **F1** — Unknown
- **Accuracy** — Unknown

## Limitations

- Fine-tuning unavailable (GA release target October 2023)

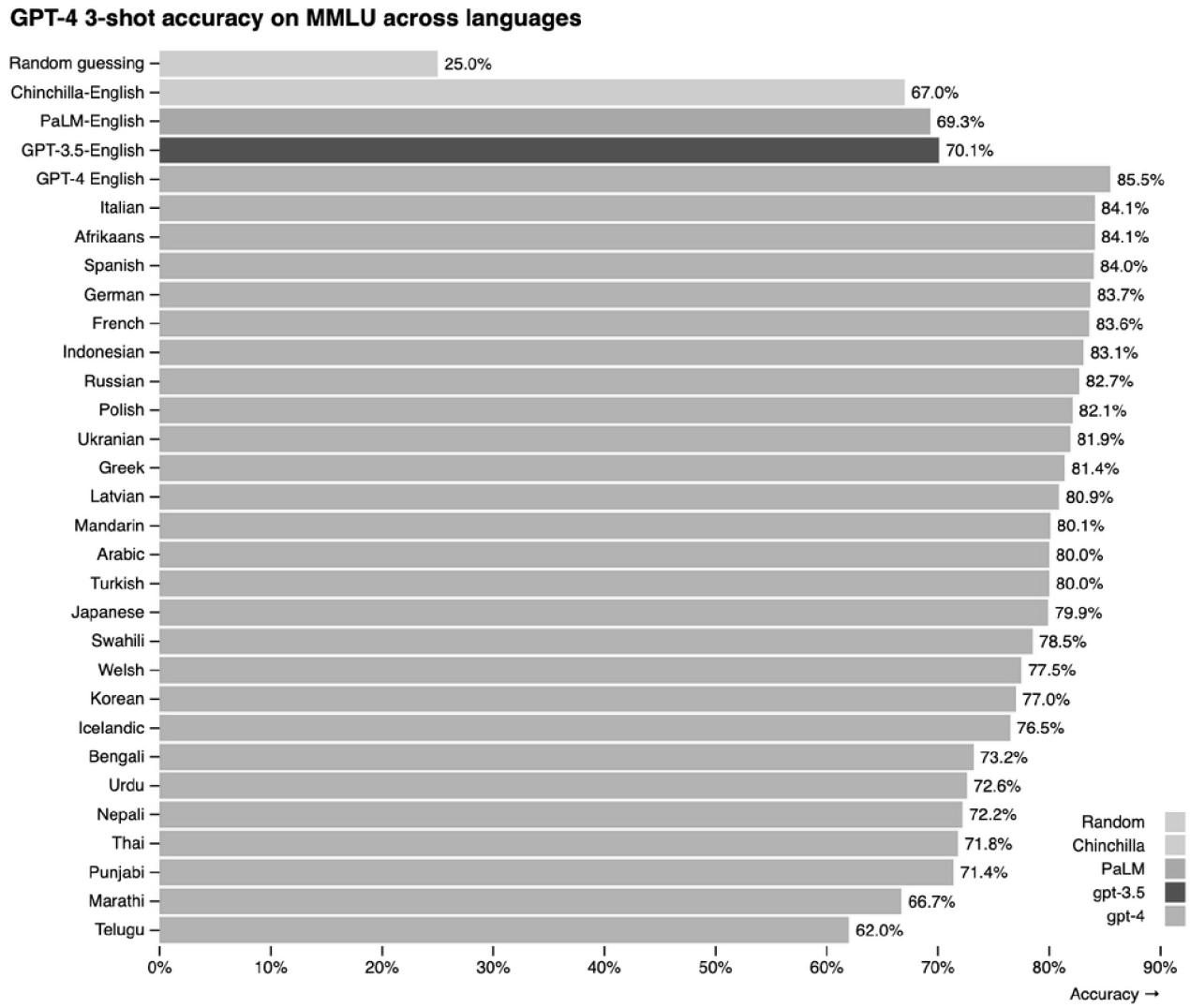
- Potential for harmful, biased outputs
- Lack of grounded reasoning
- Factually incorrect outputs
- Model mistakes as truth

## Performance Controls

- Temperature
- Top-k sampling
- Top-p sampling

## Language Supports

GPT-4 was tested on a translated version of the MMLU benchmark in 26 different languages. It outperformed GPT-3.5 and other LLMs in 24 out of the 26 languages tested, including low-resource languages like Latvian, Welsh, and Swahili. The Datacamp and MakeUseOf articles also note GPT-4's multilingual capabilities, with support for translation between English, French, German, Spanish, Chinese, Japanese, Korean and more. Translated Labs points out that GPT-4 has disparities in performance between English and other languages due to the predominance of English in its training data. Their T-LM product helps address this by translating prompts to enhance GPT-4's capabilities in 200 languages.




### Ethical Considerations

GPT-4 has potential for misuse and harmful societal impacts. Review outputs carefully before use. Do not treat as factual statements. For questions or concerns, contact **safety@openai.com**

## Model Architecture

The model architecture of GPT-4 moves away from a standard transformer approach. Instead it utilizes a mixture of experts (MoE) design.

In the MoE architecture, there are separate expert neural networks that specialize in certain tasks or data types. For each inference query, the appropriate expert models are selected to handle that specific input.

This provides two major advantages:

- The overall model can scale up in size significantly, while only routing inference through a small subset of expert parameters for any given query. This keeps inference costs practical.

2. The mixture of experts can develop specialized knowledge, improving overall capabilities.

Specifically, GPT-4 consists of:

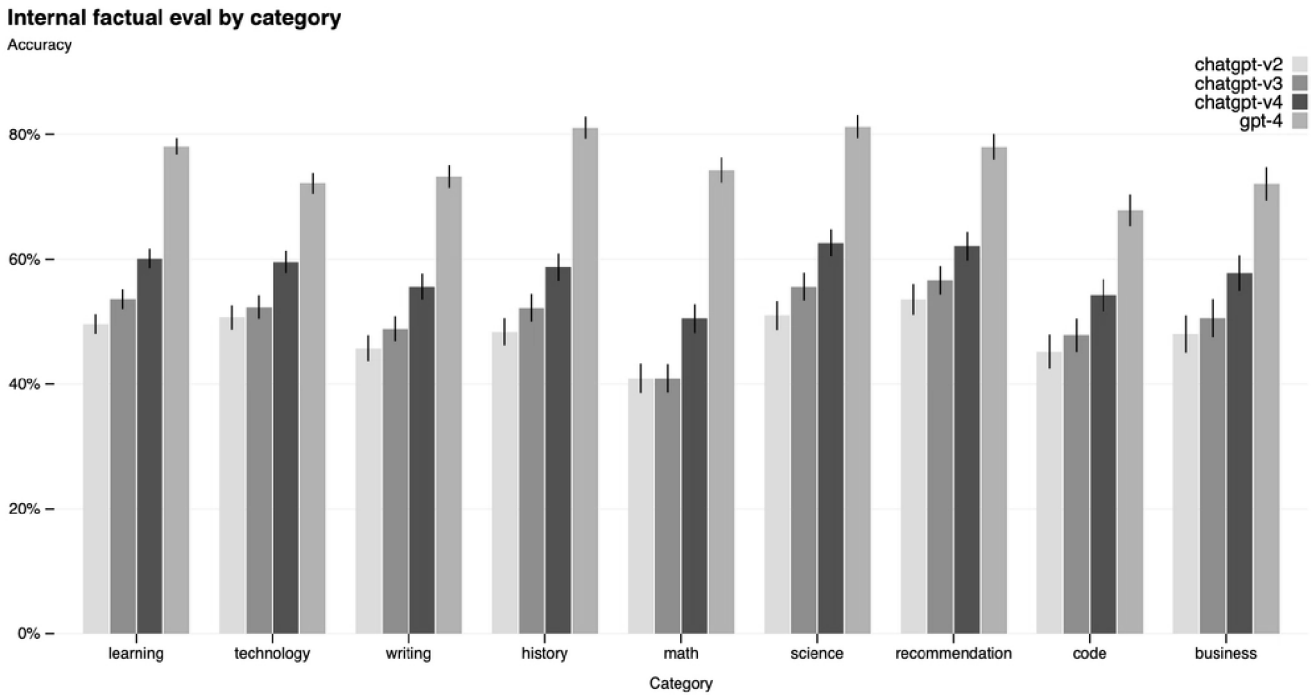
- 16 expert models, each with ~111B parameters
- 2 experts are activated per inference query
- 55B shared parameters for attention
- Results in ~280B parameters used per inference pass

> **TOP TIP**
>
> It is likely that this architecture prevents a true Temperature 0 setting resulting in inference variance caused by the sampling and routing to mixture of experts. Additionally the CUDA driver floating point operations are non-additive. This theory was confirmed in a 1:1 discussion with Stephen Wolfram in September 2023.

This architecture allows GPT-4 to reach over 1.8 trillion parameters in total, while only utilizing several hundred billion per query.

## Training

Training a model as large as GPT-4 requires extensive computational resources. It pushed the limits of existing infrastructure.




Key facts about the GPT-4 training process:

- Trained on ~25,000 Nvidia A100 GPUs simultaneously
- The batch size increased over time, eventually reaching 60 million tokens
- Trained for a total of 90-100 days continuously
- Required 2.15e25 **floating point operations (FLOPs)** in total
- Trained on a dataset of ~13 trillion tokens

To make this feasible, extensive parallelism techniques were used:

- 8-way tensor parallelism to distribute the model across GPUs
- 15-way pipeline parallelism to split batches into stages
- Various clustering topologies to maximize inter-GPU bandwidth

The result was one of the largest compute jobs ever for an AI model.

## Inference

Deploying GPT-4 for inference at scale is a significant challenge due to its size and mixture of experts architecture. Efficient inference directly impacts costs.

Key facts about GPT-4 inference:

- Runs on clusters of 128 A100 GPUs
- Leverages 8-way tensor parallelism and 16-way pipeline parallelism
- Carefully balances latency, throughput, and utilization
- May use speculative decoding to improve throughput by 2-3x
- Multi-query attention reduces memory needs for long contexts

Inference clusters are designed to maximize throughput and hardware utilization. This keeps costs lower per query.

There are still challenges around consistently batching queries for diverse expert models. But overall, the infrastructure can effectively deploy GPT-4 without pricing becoming prohibitive.

## Understanding Token Dropping in GPT-4

The mixture-of-experts (MoE) architecture used in GPT-4 relies on a token routing mechanism to determine which experts process each token. This can lead to certain tokens being "dropped" or unprocessed.

GPT-4 uses a simple top-2 token routing approach, where each token is sent to the 2 most likely experts according to the router. The experts themselves have a set capacity limit on how many tokens they can process per batch.

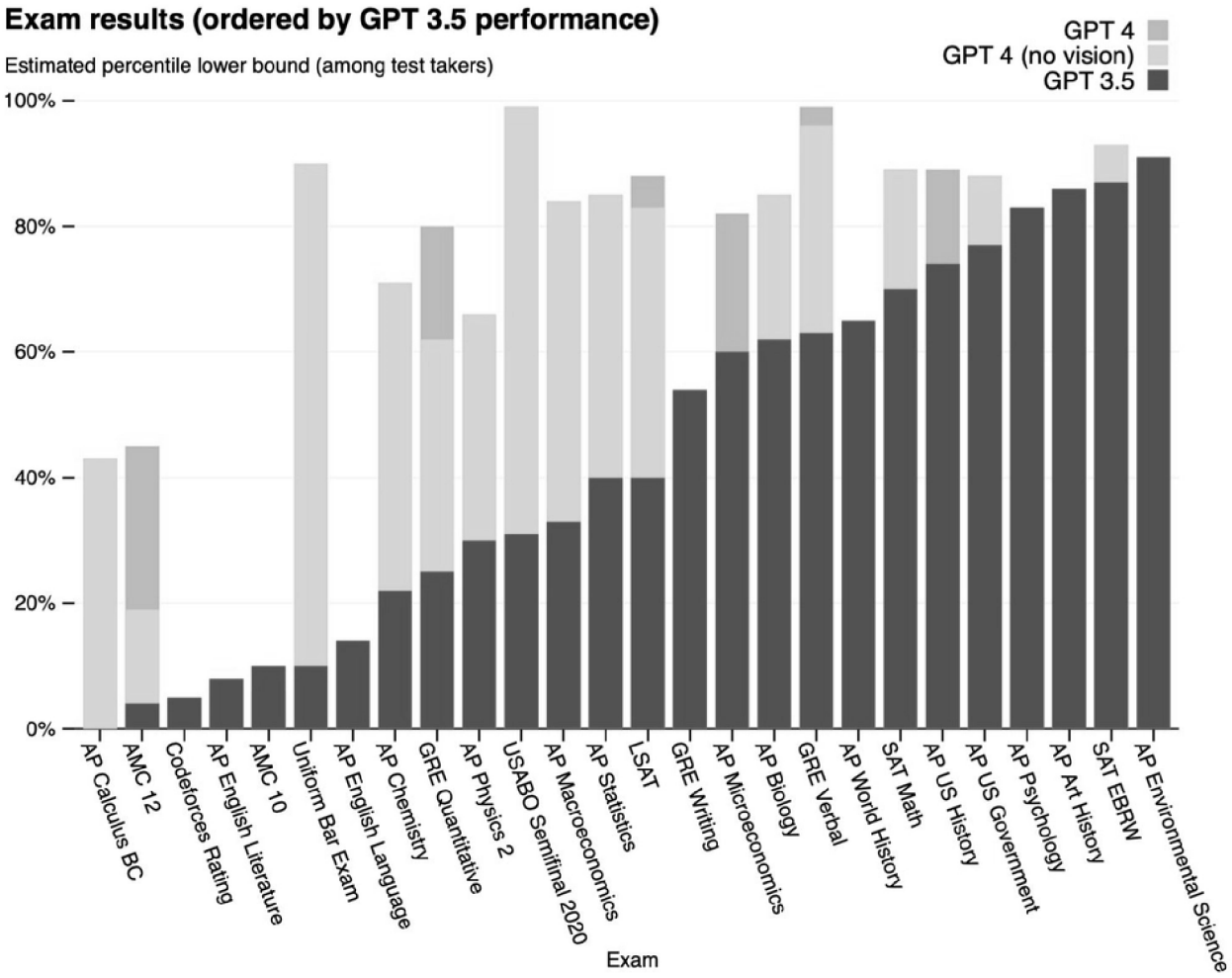
When aggregated across long input sequences and large batch sizes, the expert capacity is often exceeded, resulting in tokens being dropped. Counterintuitively, some level of dropping is actually beneficial for model performance and efficiency, as it prevents overloading experts.

The drops are non-deterministic - running the same prompt twice can lead to different drops each time. This is because the tokens are dropped differently across batches depending on capacity. The model itself remains deterministic.

While OpenAI could tweak expert capacity and reduce drops, this would substantially increase inference time and cost. The current tradeoff enables inexpensive deployment at scale. Dropping is inherent to sparse MoE designs.

Understanding how routing leads to drops provides insight into observations of randomness in GPT-4. The drops vary across usages, but the model logic itself does not.

## The Future




GPT-4 demonstrates impressive progress in language model foundations. However, future models will likely need to expand beyond a purely text-based approach.

Some areas of focus moving forward:

- Architectures that natively support vision, audio, speech, and text together
- Training models end-to-end across different data modalities

- Expanding beyond mixtures of experts for greater scalability
- Increasing training data diversity and size by orders of magnitude
- Advancing multi-modal capabilities for complex reasoning
- Optimizing model designs for real-world task performance

With each generation, OpenAI is pushing closer towards artificial general intelligence. While they are further along than any other LLM/AI research company, we are still far away from true general intelligence, lacking key attributes such as volition, decision making, memory, real-time knowledge synthesis, and other attributes.

GPT-4 shows they have the technical capabilities to make massive leaps forward with each iteration.

The future capabilities of these models remain incredibly exciting.

## GPT-4 System Card

The GPT-4 System Card comprehensively analyzes the model's safety challenges and the interventions OpenAI has implemented to mitigate potential harms. It reflects insights from over 50 experts and draws from the model card and system card concepts. GPT-4 introduces new risk surfaces with its advanced capabilities, necessitating robust safety measures.

GPT-4V, the vision-enabled extension of GPT-4, brings multimodal capabilities to the table, allowing the model to interpret image inputs and perform tasks beyond the scope of language-only systems. The safety properties of GPT-4V are detailed in a dedicated system card.

To enhance GPT-4's safety and alignment, OpenAI has refined the model to favor responses that raters deem high-quality and to avoid generating low-quality outputs. Despite these efforts, the Federation of American Scientists points out that the controls are not foolproof, and the mitigation strategies have limitations.

The System Card also outlines GPT-4's limitations, such as generating plausible yet inaccurate text, and its strengths, like improved illicit advice generation and dual-use capabilities. It documents OpenAI's comprehensive safety approach, encompassing

measurements, model adjustments, product and system interventions, and collaboration with external experts.

Supplementary information from Wikipedia notes that GPT-4 is a multimodal LLM that surpasses its predecessors in reliability, creativity, and nuanced instruction handling. It details two versions of GPT-4, each with different context window sizes, and mentions the introduction of GPT-4 Turbo with enhanced features.

A Reddit discussion illustrates the significance of safety measures by comparing examples of GPT-4's outputs with and without safety protocols, underscoring the necessity of these interventions.

## Research Sources

- **Knowing Enough About MOE**
- **Non-Determinism in GPT-4**
- **Continuous Batching LLM Inference**
- **GPT-4 Architecture Infrastructure**
- **From Sparse to Soft Mixtures of Experts**

---

## FAQs

### What are the potential risks associated with the current version of the GPT-4 AI language model, and how does OpenAI plan to address hate speech and harmful content in its system?

OpenAI has disclosed that GPT-4, its latest AI language model, has been rigorously tested through red teaming to uncover potential risks and failure modes. The accompanying technical report for the GPT-4 model card reveals that, although the model has been

enhanced with additional data and human feedback to decrease hate speech occurrences, ongoing efforts are essential to qualitatively evaluate and mitigate the potential for generating harmful content inadvertently.

The organization asserts that the system is programmed to reject prompts from hate groups and has undergone a meticulous post-training refinement to tailor the model's responses to sensitive inquiries. Despite these proactive measures, the extensive dataset employed for training could still harbor inaccuracies or biases, underscoring the necessity of persistent human oversight and reinforcement learning. The AI system's algorithm, known as **reinforcement learning from human feedback (RLHF)**, aims to refine the model by integrating users' intentions and more sophisticated instructions. For further details on AI systems, consult the GPT-4 system card.

Moreover, OpenAI has ventured into expanding **GPT-4's multimodal** capabilities, enabling it to process and analyze image inputs. However, they stress the importance of feedback from early access users to identify any issues, which allows AI researchers to conduct additional fine-tuning.

As a foundational model for large language models, the current iteration of GPT-4 excels at answering questions and predicting subsequent words in a sequence. Nonetheless, it necessitates further data and reinforcement learning to hone its responses and reduce the likelihood of endorsing self-harm or addressing other sensitive matters.

## How does OpenAI ensure that the GPT-4 AI system remains up-to-date, and what measures are taken to improve its performance and safety?

OpenAI's approach to maintaining the GPT-4 AI system involves a continuous cycle of red teaming, where AI researchers actively seek out and address potential risks and failure modes. The current model, which is the latest in the series of GPT models, has been developed based on a large dataset and has undergone a rigorous post-training process to enhance its performance.

A technical report released by OpenAI details the training dataset and the steps taken to fine-tune the AI language model using advanced techniques such as reinforcement learning and human review. The current version of GPT-4 has undergone further refinement to address issues related to hate speech and the system's ability to handle sensitive requests.

OpenAI claims that the model is designed to refuse prompts that may lead to the generation of harmful content, and it has been fine-tuned with additional data to better understand and

answer questions from users.

The AI system also benefits from early access feedback, which allows AI researchers to fine-tune the model's responses based on real-world use. To address the nuanced instructions of users and reduce the occurrence of factual errors, OpenAI employs a combination of language models, reinforcement learning, and human feedback.

The GPT4 system card also highlights the model's ability to analyze image inputs, showcasing its multimodal capabilities. As part of its commitment to safety, OpenAI emphasizes the importance of human feedback in the AI system's learning process, ensuring that the model aligns with the user's intent and reduces the risk of promoting self-harm or other sensitive content. The AI systems are based on the advanced model and multimodal model capabilities, using system messages for much more nuanced instructions in natural language.

# It's time to build

Collaborate with your team on reliable Generative AI features.
Want expert guidance? Book a 1:1 onboarding session from your dashboard.

Start for free →