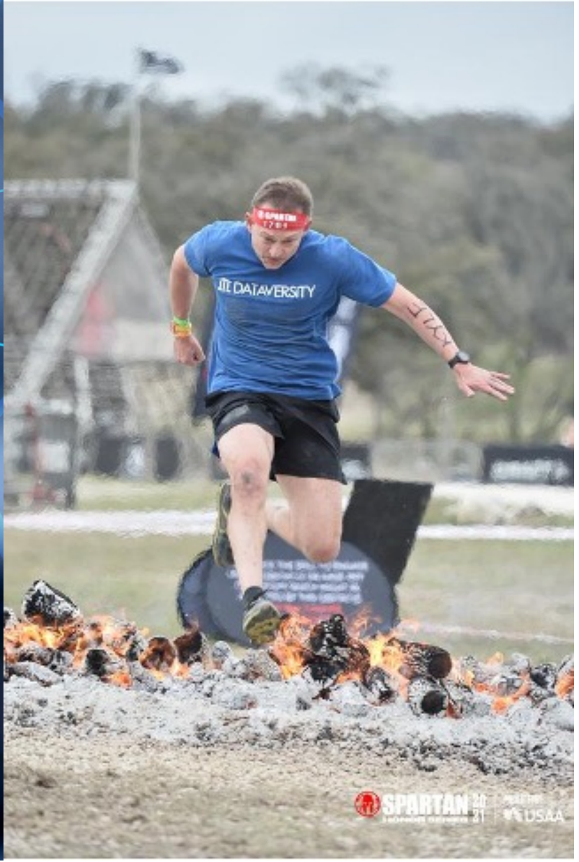
MCKNIGHT
CONSULTING GROUP

AI Language Models for Enterprises



Presented by: William McKnight

“#1 Global Influencer in Big Data” Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444

McKnight Consulting Group Partial Technology Implementation Expertise

Big/Analytic/Vector/Mixed Data Management

databricks

Pinecone

Google BigQuery

snowflake

MySQL HeatWave

Azure Synapse

teradata.

amazon REDSHIFT

CLOUDERA

SAP Datasphere

Exasol

VERTICA

Starburst

ACTIAN™ activate your data

Azure HDInsight

SingleStore

Apache Impala

amazon EMR

Yellowbrick

Cloud Dataproc

ORACLE AUTONOMOUS DATABASE

ORACLE

Data Movement and APIs

Airbyte

Informatica

Qlik Replicate®

Fivetran

SYNADIA

CONFLUENT

talend

apigee

MATILLION

Stitch

Kong

Astera

Data Management

MC MONTE CARLO

Anomalo

SNOWPLOW

SAP Datasphere

alteryx DESIGNER CLOUD

Profisee

splunk>

NGINX Part of F5

Collibra

IMMUTA™

elasticsearch

Azure MLOps

Operational/Transactional Data Management

DIAMANTI

Cockroach

Couchbase

DataStax

SQLite

MongoDB.

PostgreSQL

yugabyteDB

neo4j

ORACLE

Google Cloud Spanner

redis

Azure Cosmos DB

Microsoft® SQL Server®

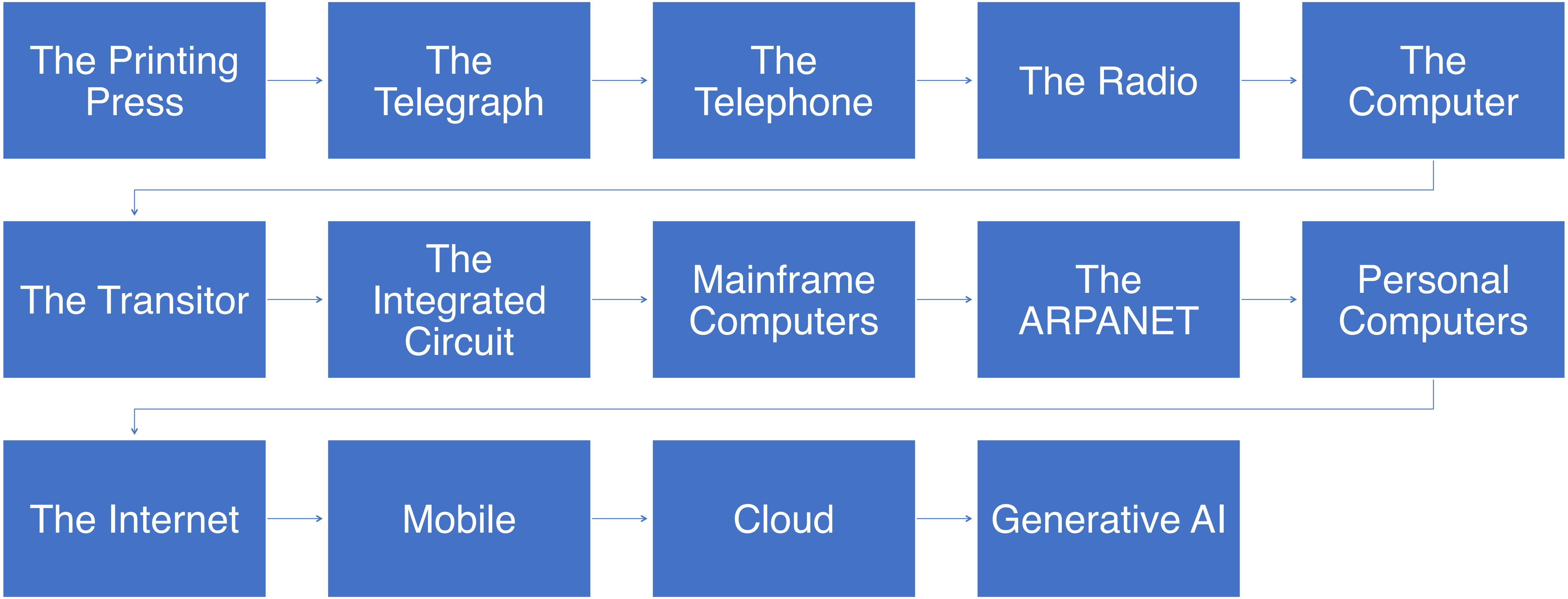
Amazon Aurora

MCKNIGHT CONSULTING GROUP

Dataversity Advanced Analytics with William McKnight 2024

1. 2024 Trends in Advanced Analytics
2. Data Integration – Newsflash: We Still Just Move Data!
3. Choosing Your Provider for Implementing a Data Fabric
4. Architecting a Modern Data Platform
5. What The? Another Database Model – Vector Databases Explained
6. Every Database is Multi-Modal; What Does this Mean to an Enterprise?
7. The ROI of Master Data Management is (usually) there – Let’s Run the Numbers
8. Promising AI Use Cases for the Enterprise in 2024
9. AI Language Models for Enterprises
10. Remembering Data Quality when the Data Spicket is Turned Way Up
11. What Does Information Management Maturity Look Like in 2024
12. How to Become an AI Ready Organization

The Big Technology Waves 1440-Present



The Printing Press

The Telegraph

The Telephone

The Radio

The Computer

The Transitor

The Integrated Circuit

Mainframe Computers

The ARPANET

Personal Computers

The Internet

Mobile

Cloud

Generative AI

History of LLMs

- Evolution of neural networks
- RNNs predicted next word in sentence in early 2000s
- 2017 Google DeepMind Team paper on Transformers
- 2018 Open AI developed GPT-1
- Traditional programming is instruction-based
- LLMs is teaching not ‘how’ but giving examples and asking it to learn

Large Language Models

A type of artificial intelligence (AI) model designed to process and understand human language

  - Trained on vast amounts of text data
  - Learns patterns and relationships in language

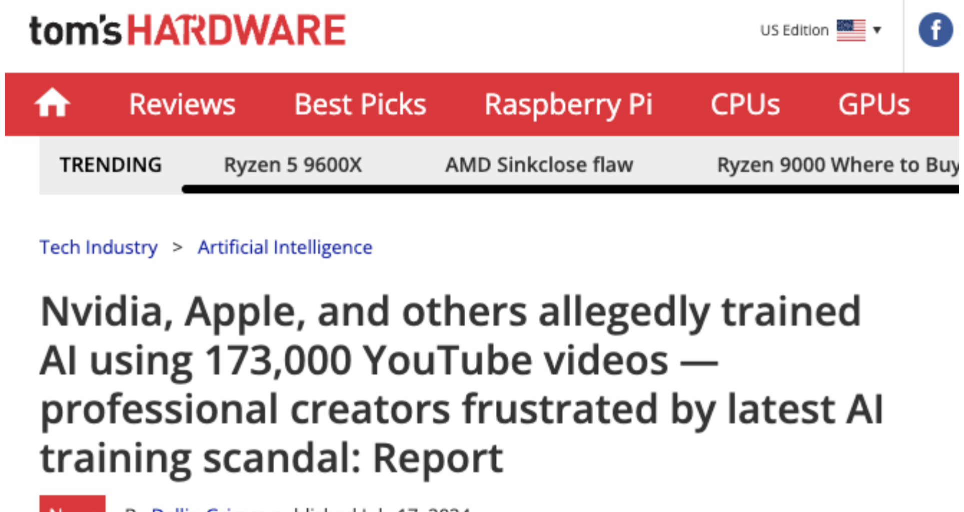
- **Architecture**:
  - Deep learning architectures
  - Transformers
  - Recurrent neural networks (RNNs)
  - Convolutional neural networks (CNNs)
- **Capabilities**:
  - Language understanding
  - Text generation
  - Translation
  - Summarization
  - Sentiment analysis
  - Question answering

LLMs are Trained

Large Language Models (LLMs) are trained on vast amounts of text data to learn patterns and relationships in language. This training process enables LLMs to understand and generate human-like language.

Training Data:

1. **Web pages**
2. **Books and articles**
3. **User-generated content** (e.g., social media, forums)
4. **Product reviews**
5. **Wikipedia**

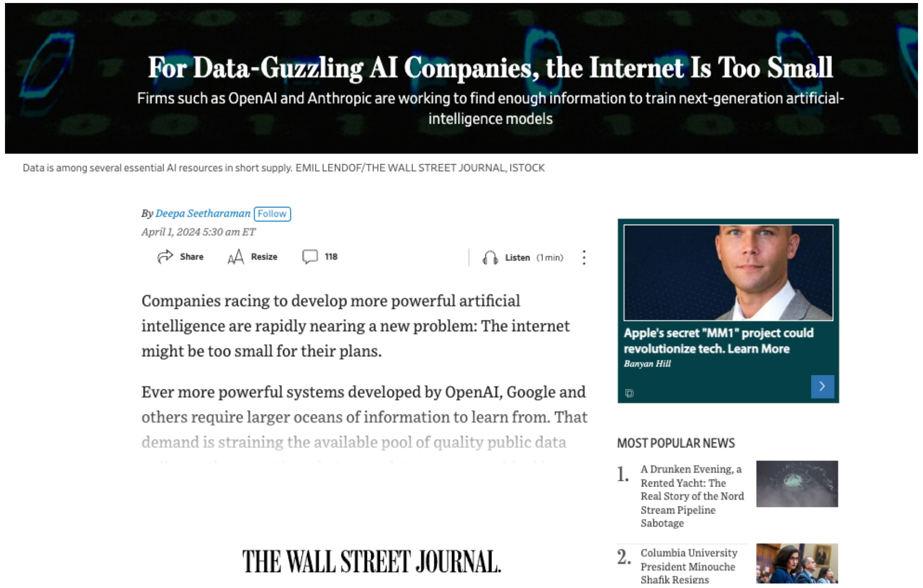


Tech Industry > Artificial Intelligence

Nvidia, Apple, and others allegedly trained AI using 173,000 YouTube videos — professional creators frustrated by latest AI training scandal: Report

News By Dallin Grimm published July 17, 2024

No one



For Data-Guzzling AI Companies, the Internet Is Too Small

Firms such as OpenAI and Anthropic are working to find enough information to train next-generation artificial-intelligence models

Data is among several essential AI resources in short supply. EMIL LENDOF/THE WALL STREET JOURNAL, ISTOCK

By Deepa Seetharaman Follow
April 1, 2024 5:30 am ET

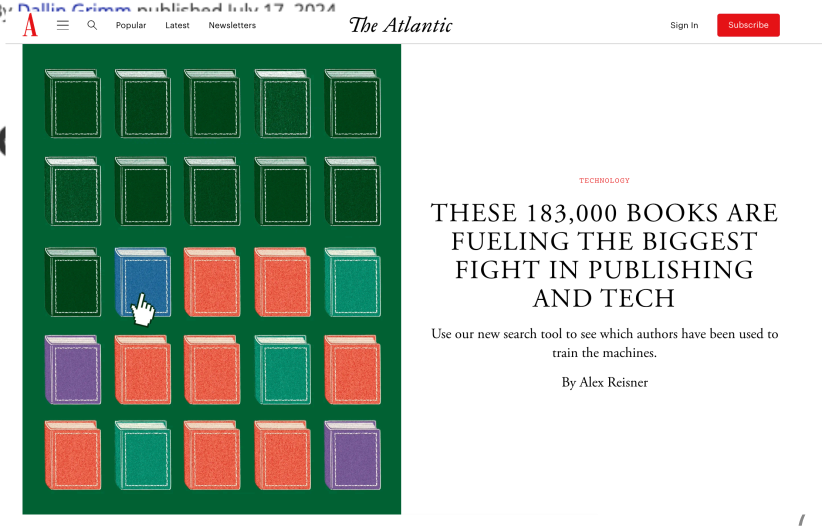
Companies racing to develop more powerful artificial intelligence are rapidly nearing a new problem: The internet might be too small for their plans.

Ever more powerful systems developed by OpenAI, Google and others require larger oceans of information to learn from. That demand is straining the available pool of quality public data

MOST POPULAR NEWS

1. A Drunken Evening, a Rented Yacht: The Real Story of the Nord Stream Pipeline Sabotage
2. Columbia University President Minouche Shafik Resigns

THE WALL STREET JOURNAL.



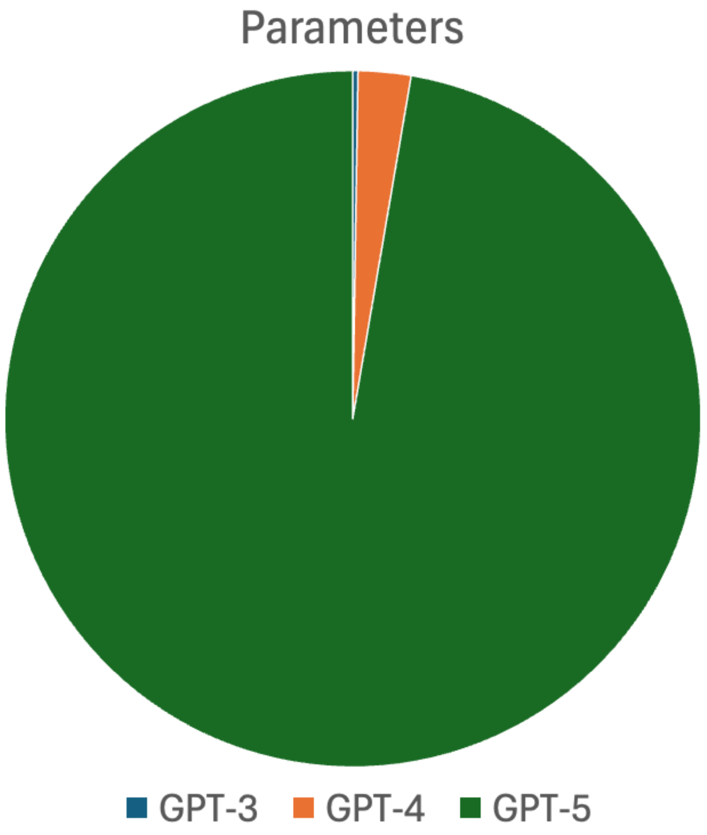
TECHNOLOGY

THESE 183,000 BOOKS ARE FUELING THE BIGGEST FIGHT IN PUBLISHING AND TECH

Use our new search tool to see which authors have been used to train the machines.

By Alex Reisner

....On Synthetic Data



- **What is Synthetic Data?**
  - Artificially generated information that mimics real-world data
  - Used to train AI models when human-generated datasets are exhausted or unavailable
  - Used to increase quantity and quality of data
- **Synthetic Data in GPT-5**
  - Makes up around 70% of the GPT-5 dataset
  - Enables unprecedented data quality and quantity
  - Contributes to a significant leap forward in model performance

Buyer Beware

- Easier for the app to do it all and back off
- Organizational culture
- Give error rather than BS
- LLM Guard
- Labels?
- User Education

EU AI Act

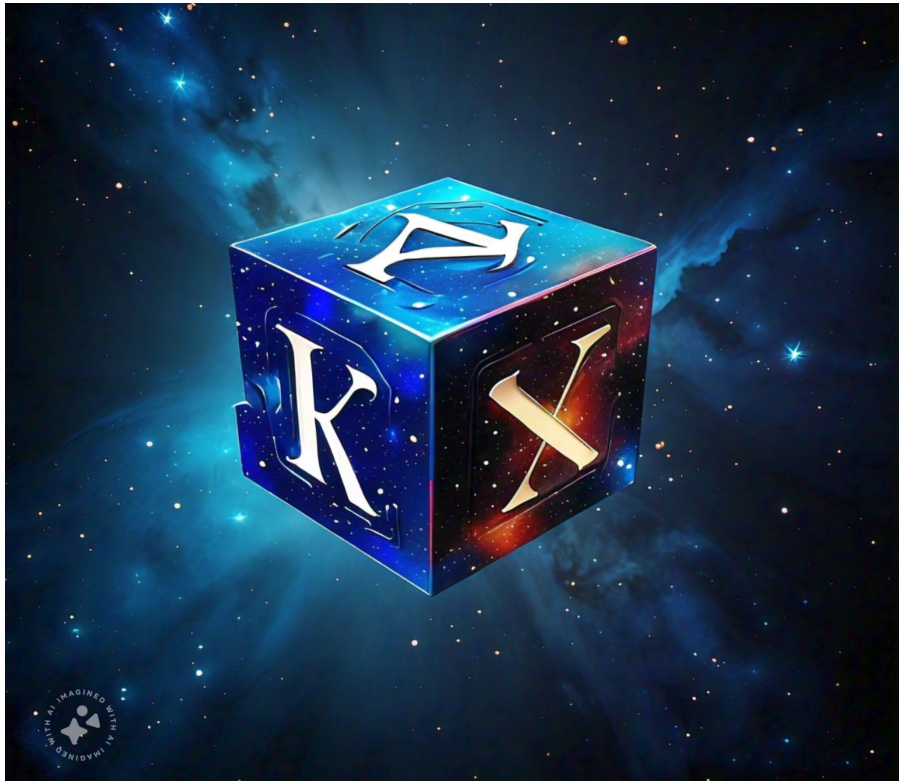
- Regulated AI
- Common risk structure for AI
- Specific requirements for GenAI providers and users
  - Providers include customers passing along access to LLMs
- Fines
- Began 8/1/24
- it's about change management
- Companies in US that operate in EU need to comply

LLM Steps

1. Tokenization
2. Embeddings (vectors)
3. Transformers

How LLMs are Trained

1. **Data Collection**: Gathering a massive dataset of text from various sources (e.g., books, articles, websites).
2. **Preprocessing**: Cleaning and formatting the data (e.g., tokenization, stopword removal).
3. **Model Architecture**: Designing the LLM's architecture (e.g., transformer, recurrent neural network).
4. **Training**: Feeding the preprocessed data into the model, adjusting parameters to minimize errors.



To Build an LLM

Building an LLM requires significant expertise, resources, and effort.

1. Data

2. Compute Resources

3. Model Architecture

4. Training: Algorithm, Objective, Batching

5. Software and Tools: TensorFlow, PyTorch, or JAX and libraries

6. Expertise

Challenges

- **Data quality and availability**
- **Computational resources and scalability**
- **Model complexity and training instability**
- **Evaluation and fine-tuning**

Who Has Built LLMs

RAG and LLM

Retrieval-Augmented Generation combines the strengths of LLMs and external knowledge sources

Enhances accuracy, relevance, and diversity of generated content

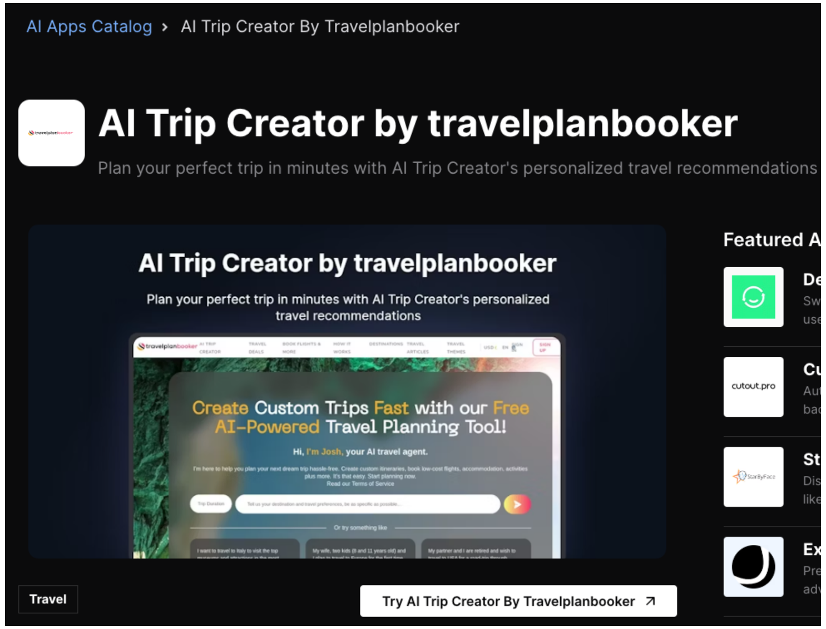
- **LLMs and RAG: A Powerful Duo**:
  - LLMs provide context and language understanding
  - RAG leverages external knowledge to augment LLM capabilities
- **Key Benefits**:
  - Improved accuracy and relevance
  - Increased diversity and novelty
  - Enhanced ability to handle nuanced and complex topics
- **Applications**:
  - Content generation (text, chatbots, etc.)
  - Question answering and information retrieval
  - Summarization and knowledge graph construction

LLM Use Cases – Customer Service

- **Brink's Home**
  - Leveraged AI to optimize service call scheduling and cross-sell recommendations
  - Boosted Average DTC Package Size
  - Increasing customer acquisition cost (CAC) and competitive pressure.
  - Significantly increased average DTC package size and revenue
- **Compliance Aspekte**
  - Replaced rule-based algorithms with AI-powered features
  - Developed a new "co-pilot" chatbot with advanced capabilities
  - Automatically associates documents with compliance requirements
  - Provides relevant insights and instructions

Automated Customer Service Use Cases

- **TravelPlanBooker**
  - Rule-based system struggled with complex user inputs.
  - Prone to conflicts and scalability issues.
  - Implemented generative and conversational AI for:
    - Handling abstract user queries.
    - Offering nuanced travel advice.
    - Users adjusted plans on the go.



AI Apps Catalog › AI Trip Creator By Travelplanbooker

AI Trip Creator by travelplanbooker

Plan your perfect trip in minutes with AI Trip Creator's personalized travel recommendations

AI Trip Creator by travelplanbooker

Plan your perfect trip in minutes with AI Trip Creator's personalized travel recommendations

Create Custom Trips Fast with our Free AI–Powered Travel Planning Tool!

Hi, I'm Josh, your AI travel agent.

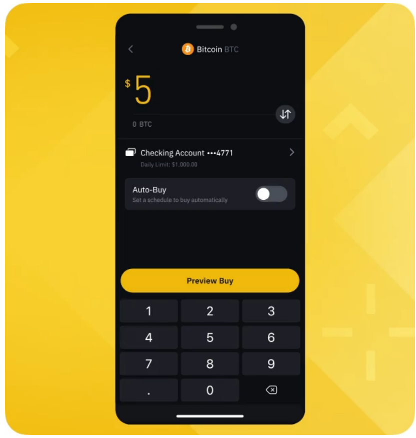
I'm here to help you plan your next dream trip hassle-free. Create custom itineraries, book low-cost flights, accommodation, activities plus more. It's that easy. Start planning now.
Read our Terms of Service

Automated Customer Service Use Cases

- **Estée Lauder**
  - Voice-enabled makeup assistant for visually impaired users.
- **Lufthansa Group**
  - Its AI takes in crew availability and locations, passenger demand, aircraft maintenance status, weather, and many other variables.
    - It will then send suggested scenarios – for example, a particular aircraft for a specific flight – to human operations controllers to support their decision-making.
  - Uses AI to manage high volumes of customer queries about canceled and rescheduled flights, which helped improve the overall customer experience.

LLM Use Case - AI-Driven Financial Advisory Services

- **Enova**
  - AI platform offers tailored financial analysis and advice, demonstrating the impact of AI in personalizing financial services
  - Enova utilizes AI to evaluate creditworthiness more accurately, making credit accessible to those traditionally underserved by traditional financial institutions.
  - AI-powered systems analyze vast amounts of data to detect fraudulent activities, protecting both Enova and its customers.
  - AI helps identify distinct customer segments based on behavior, demographics, and financial profiles, enabling tailored product offerings.
  - AI algorithms determine optimal pricing strategies for different customer segments, maximizing revenue while maintaining competitiveness.
  - AI-driven chatbots and virtual assistants provide 24/7 customer support and personalized interactions.
- **Binance**
  - AI algorithms analyze vast amounts of transaction data to identify suspicious activities and prevent fraudulent accounts.
  - AI helps assess market risks, protect user assets, and ensure regulatory compliance.
  - Trading Bot Development: Binance offers AI-powered trading bots that automate trading strategies based on market data analysis.
  - AI-driven chatbots provide efficient and round-the-clock customer support.
  - AI tools analyze market trends and sentiment to provide valuable insights to traders.
  - AI is used to strengthen security measures, such as identifying potential vulnerabilities and threats.


Bitcoin BTC

$5

0 BTC

Checking Account •••4771

Daily Limit: $1,000.00

Auto-Buy

Set a schedule to buy automatically

Other LLM Use Cases

- Programming Assistants
- Summarization
- Language Translation
- Essay Writing
- Summarization
- Image Captioning

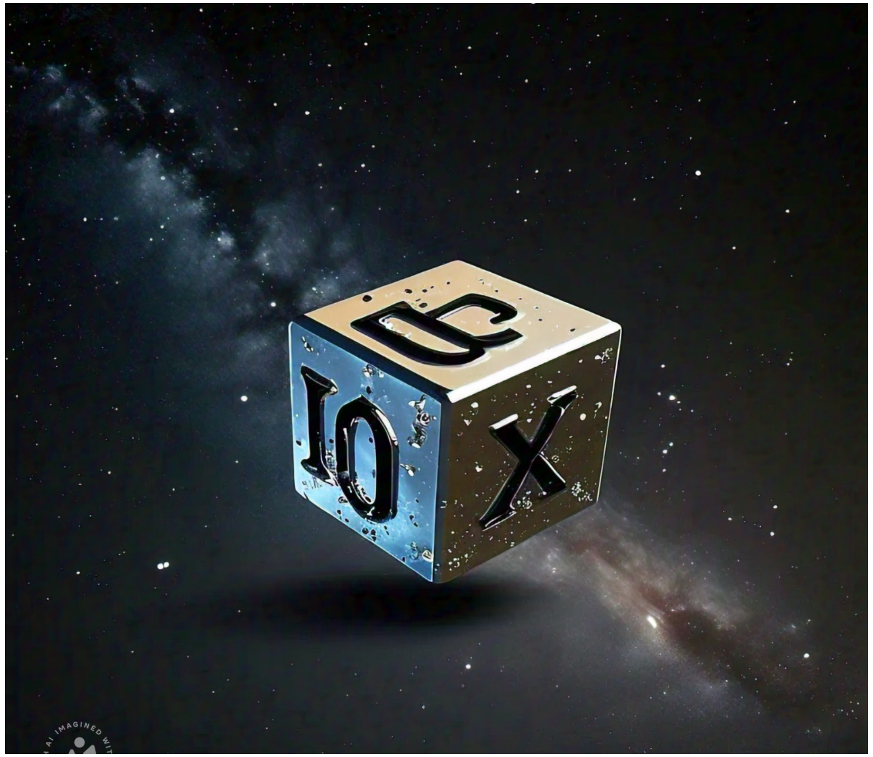
How to use LLMs

- Cloud
- License
- API
    - Ie, Meta AI’s official API for LLaMA
    - Register for an API Key
    - Configure API Settings
    - Use Python/other to send API requests to LLaMA
    - Receive output data
- Pipelines

```
import requests

api_key = "YOUR_API_KEY"
model_name = "llama"
input_text = "Your input text here"

url = f"https://api.huggingface.co/models/{model_name}"
headers = {"Authorization": f"Bearer {api_key}"}
data = {"inputs": input_text, "parameters": {"temperature": 0.7}}

response = requests.post(url, headers=headers, json=data)
output = response.json()["generated_text"]

print(output)
```

Temperature

- **Low temperature (e.g., 0.1-0.5)**: More deterministic, predictable output. The model is more likely to repeat familiar patterns and phrases.
- **Medium temperature (e.g., 0.5-1.0)**: Balanced output, suitable for most applications. The model generates coherent text with some variation.
- **High temperature (e.g., 1.0-2.0)**: More random, creative output. The model is more likely to produce novel phrases, but may also generate less coherent or relevant text.



Top LLMs

GPT (Generative Pre-trained Transformer)

- Developed by OpenAI
- Strengths: Exceptional accuracy, versatility, and API documentation
- Prices: Varying subscription plans, $0.5/1M tokens to $5/1M tokens
- Image models available, $0.016/image to $0.040/image

Gemini

- Developed by Google
- Strengths: Efficiency, scalability, and context window
- Prices: API pricing based on usage, $0.35/million tokens; personal use, $19/user/month

LLaMA (Large Language Model Meta AI)

- Developed by Meta AI
- Strengths: Resource-efficient, and scenario-specific models
- Prices: Free for personal use, API services available

Claude

- Developed by Anthropic
- Strengths: Speed, affordability, and conversational AI capabilities
- Prices: Personal use, free to $20 (pro) or $30 (teams); API, $3/M tokens (input) and $15 (output)

Other LLMs

Mistral

- Open-source LLMS with a focus on flexibility and customization
- Designed for large-scale deployments and complex use cases
- Offers advanced features like multi-tenancy, SSO, and custom branding
- Strong community support and extensible architecture

Gemma

- Modern, cloud-based LLMS with a user-friendly interface
- Emphasizes ease of use, intuitive design, and mobile responsiveness
- Supports gamification, social learning, and personalized learning paths
- Integrates with popular tools like Zoom, Google Drive, and Office 365

Code

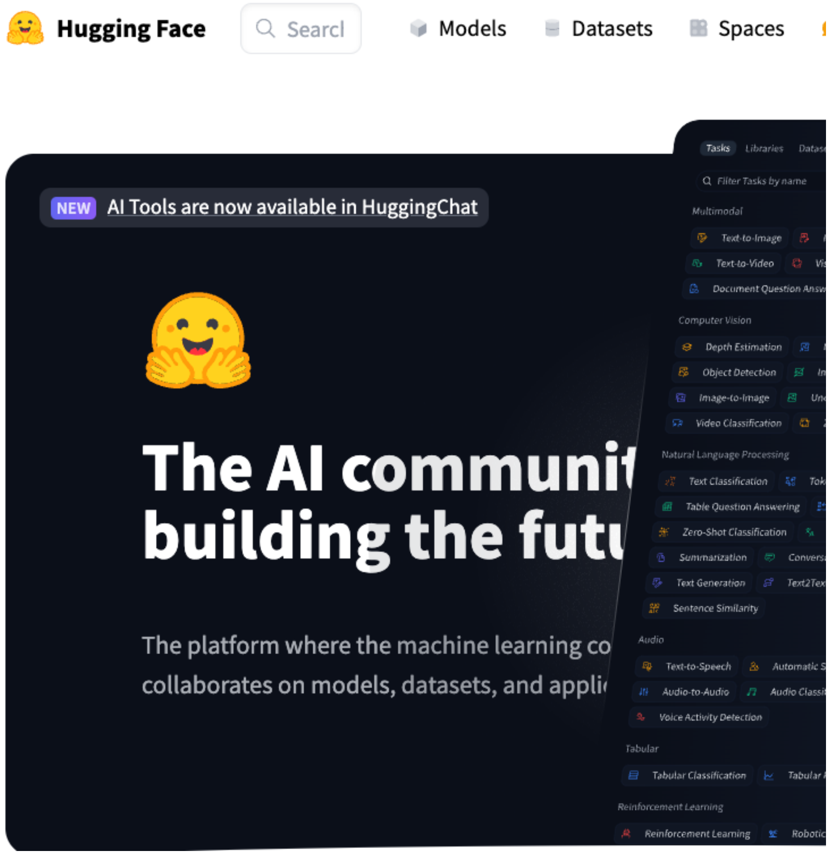
- Specifically designed for coding and software development training
- Offers features like code review, version control, and real-time feedback
- Supports a wide range of programming languages and frameworks
- Integrates with popular development tools like GitHub and Bitbucket

Reka

- AI-powered LLMs with a focus on adaptive learning and skill development
- Uses machine learning to personalize learning paths and recommend content
- Offers features like natural language processing, sentiment analysis, and predictive analytics
- Designed for enterprise-level deployments and large-scale skill development initiatives

Hugging Face

- Hugging Face is best known for its open-source Transformers library, which provides a wide range of pre-trained models and a simple interface for using them.
- Hugging Face offers a vast collection of pre-trained models for various NLP tasks, such as language translation, question answering, and text generation.
- Hugging Face has a large and active community of developers, researchers, and practitioners who contribute to the library and share their knowledge.
- Hugging Face's library provides an interface for integrating pre-trained models into applications.



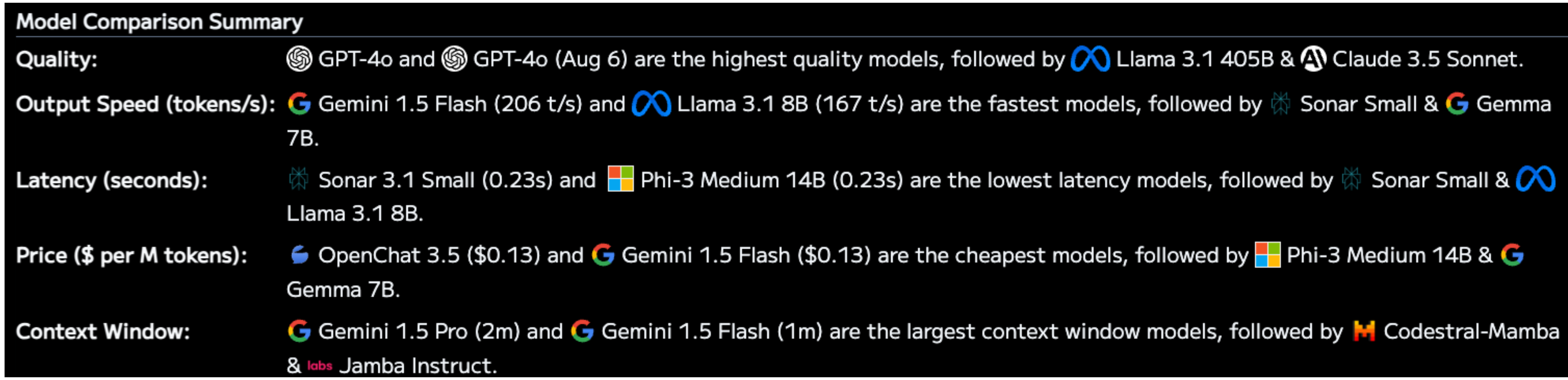
NEW AI Tools are now available in HuggingChat

The AI communit
building the futu

The platform where the machine learning co
collaborates on models, datasets, and applic

How to Evaluate LLMs

- Quality
- Speed
- Price
- Quality and Output Speed
- Quality and Ability (General Ability [Chatbot], Reasoning&Knowledge [MMLU], Coding [Human Evaluation])
- Quality and Context Window, Input Token Price
- Quality and Price
- Latency and Output Speed
- Output Speed by Input Token Count
- Output Speed Variance
- Latency
- Latency by Input Token Length
- Latency Variance
- Total Response Time
- Total Response Time by Input Token (Context) Length
- Total Response Time Variance



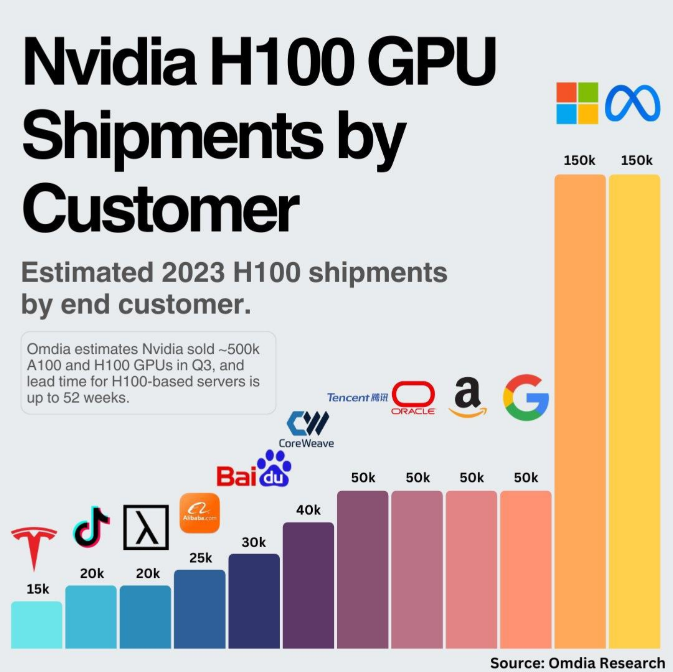
Model Comparison Summary

Quality:	GPT-4o and GPT-4o (Aug 6) are the highest quality models, followed by Llama 3.1 405B & Claude 3.5 Sonnet.
Output Speed (tokens/s):	Gemini 1.5 Flash (206 t/s) and Llama 3.1 8B (167 t/s) are the fastest models, followed by Sonar Small & Gemma 7B.
Latency (seconds):	Sonar 3.1 Small (0.23s) and Phi-3 Medium 14B (0.23s) are the lowest latency models, followed by Sonar Small & Llama 3.1 8B.
Price ($ per M tokens):	OpenChat 3.5 ($0.13) and Gemini 1.5 Flash ($0.13) are the cheapest models, followed by Phi-3 Medium 14B & Gemma 7B.
Context Window:	Gemini 1.5 Pro (2m) and Gemini 1.5 Flash (1m) are the largest context window models, followed by Codestral-Mamba & Jamba Instruct.

Artificialanalysis.ai

LLM Chips

- NVIDIA
- Intel
- AMD
- SambaNova

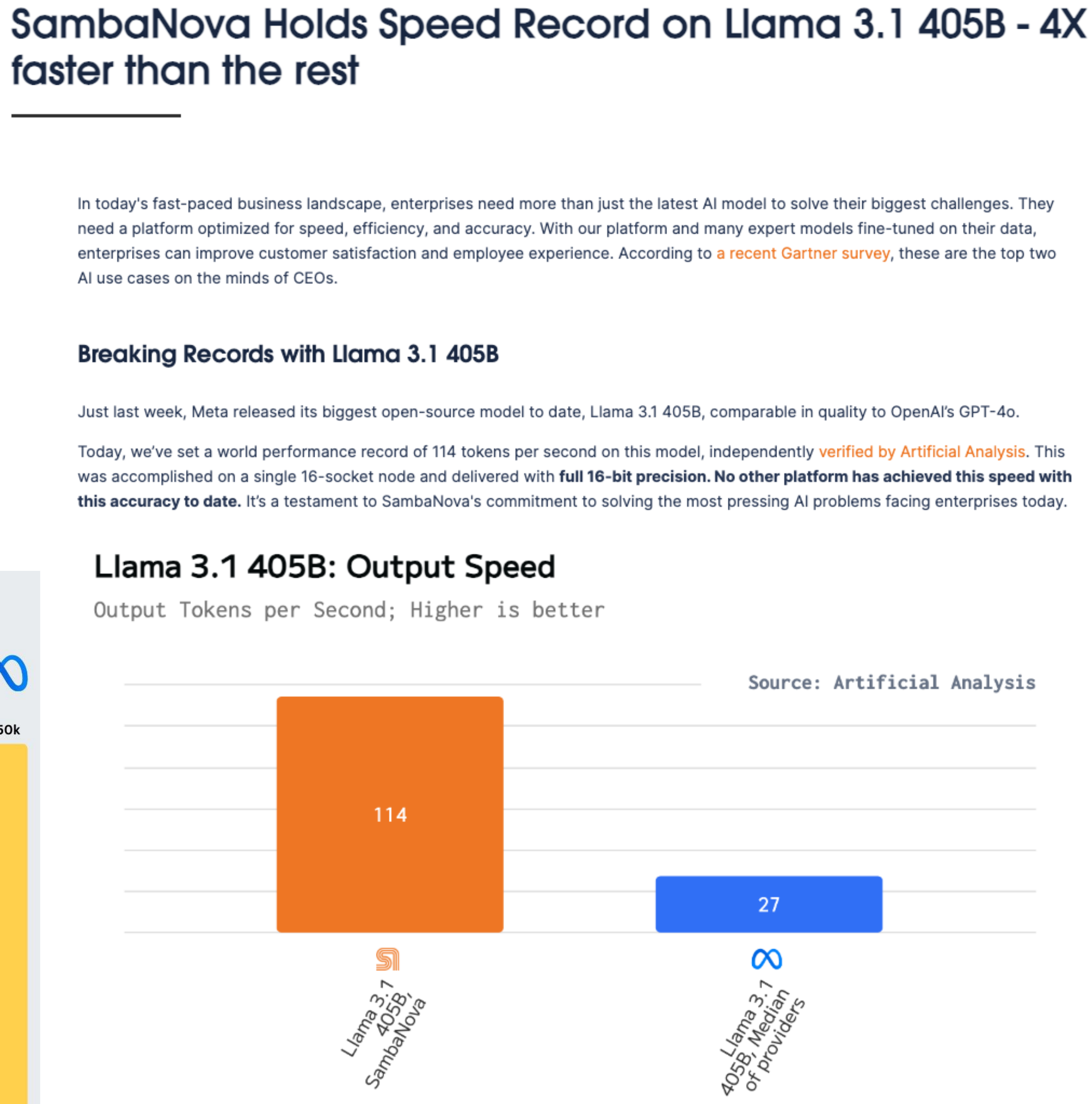


Nvidia H100 GPU Shipments by Customer

Estimated 2023 H100 shipments by end customer.

Omdia estimates Nvidia sold ~500k A100 and H100 GPUs in Q3, and lead time for H100-based servers is up to 52 weeks.

Source: Omdia Research



SambaNova Holds Speed Record on Llama 3.1 405B - 4X faster than the rest

In today's fast-paced business landscape, enterprises need more than just the latest AI model to solve their biggest challenges. They need a platform optimized for speed, efficiency, and accuracy. With our platform and many expert models fine-tuned on their data, enterprises can improve customer satisfaction and employee experience. According to a recent Gartner survey, these are the top two AI use cases on the minds of CEOs.

Breaking Records with Llama 3.1 405B

Just last week, Meta released its biggest open-source model to date, Llama 3.1 405B, comparable in quality to OpenAI's GPT-4o.

Today, we've set a world performance record of 114 tokens per second on this model, independently verified by Artificial Analysis. This was accomplished on a single 16-socket node and delivered with **full 16-bit precision. No other platform has achieved this speed with this accuracy to date.** It's a testament to SambaNova's commitment to solving the most pressing AI problems facing enterprises today.

Llama 3.1 405B: Output Speed

Output Tokens per Second; Higher is better

Source: Artificial Analysis

114
27
Llama 3.1 405B, SambaNova
Llama 3.1 405B, Median of providers

Hardware Acceleration and Multi-Worker

Hardware Acceleration:

- Bud Labs leverages specialized hardware to accelerate AI computations, such as:
    1. **GPUs (Graphics Processing Units)**: Utilizing NVIDIA GPUs for parallel processing and matrix operations.
    2. **TPUs (Tensor Processing Units)**: Employing Google TPUs for optimized machine learning computations.
    3. **ASICs (Application-Specific Integrated Circuits)**: Designing custom ASICs for specific AI workloads.

Multi-Worker:

- Bud Labs' Multi-Worker approach involves distributing AI workloads across multiple computing resources, including:
    1. **Multi-GPU**: Splitting workloads across multiple GPUs for parallel processing.
    2. **Multi-Node**: Distributing workloads across multiple machines or nodes for scalable processing.
    3. **Multi-Tasking**: Running multiple AI tasks concurrently to maximize resource utilization.

Coreweave and PyTorch

CoreWeave

Flexible GPU instances: Offering a range of GPU instances, from single GPUs to multi-GPU configurations.

Bare-metal infrastructure: Providing direct access to underlying hardware for maximum performance.

Container-optimized: Designed for containerized workloads, with support for Kubernetes and Docker.

High-performance storage: Offering fast storage options, including NVMe and SSDs.

Scalability: Enabling easy scaling of resources to meet changing workload demands.

Pytorch

Dynamic Computation Graph: PyTorch's computation graph is built on the fly, allowing for more flexible and rapid prototyping.

Automatic Differentiation: PyTorch automatically computes gradients, making it easier to implement backpropagation and optimize models.

Pythonic API: PyTorch's API is designed to be intuitive and Pythonic, making it easy to learn and use.

GPU Acceleration: PyTorch supports GPU acceleration, enabling fast computation and training of large models.

Modular Architecture: PyTorch's modular design allows for easy extension and customization of models and algorithms.

Distributed Training: PyTorch supports distributed training, enabling scalable training of large models.

Limitations

- Math, Logic, Reasoning, Bias (i.e., Human Opinions), Safety
- Data is Static
- Hallucinations (With Confidence)
- Hardware Intensive
- Ethics
  - Scamming
  - Misinformation
  - Fake Images, text, opinions

The Future May Be Small Language Models

1. **Efficiency**: Small language models require less computational resources and energy, making them more efficient.
2. **Faster deployment**: Small models can be deployed faster, enabling quicker response times and real-time processing.
3. **Lower costs**: Training and maintaining small models is less expensive than large models.
4. **Specialized knowledge**: Small models can be fine-tuned for specific domains or tasks, capturing unique knowledge.
5. **Easier interpretability**: Small models are often more interpretable, allowing for better understanding of decision-making processes.
6. **Reduced data requirements**: Small models can achieve good performance with less data, reducing data collection and storage needs.
7. **Improved safety**: Small models may be less prone to biases and ethical concerns due to their smaller size and focused training.
8. **Edge AI**: Small models are well-suited for edge AI applications, where resources are limited.
9. **Federated learning**: Small models can be used in federated learning, enabling collaborative training without sharing sensitive data.
10. **Practical applications**: Small models can be used in practical applications like chatbots, virtual assistants, and language translation, where large models may be overkill.

Future of LLMs

- Fact-Checking Itself
- Multi-Modality
- Improve Reasoning Ability
- Bigger Context Windows

Summary

- A type of artificial intelligence (AI) model designed to process and understand human language
- Large Language Models (LLMs) are trained on vast amounts of text data to learn patterns and relationships in language
- LLM use cases are abundant
- Top LLMs include GPT, Gemni, Llama, Claude
- Quality, Speed and Price are drivers of LLM selection
- The future may be Small Language Models

MCKNIGHT
CONSULTING GROUP

AI Language Models for Enterprises

Presented by: William McKnight

"#1 Global Influencer in Big Data" Thinkers360

President, McKnight Consulting Group

3 X Inc 5000

/in/wmcknight

www.mcknightcg.com
(214) 514-1444