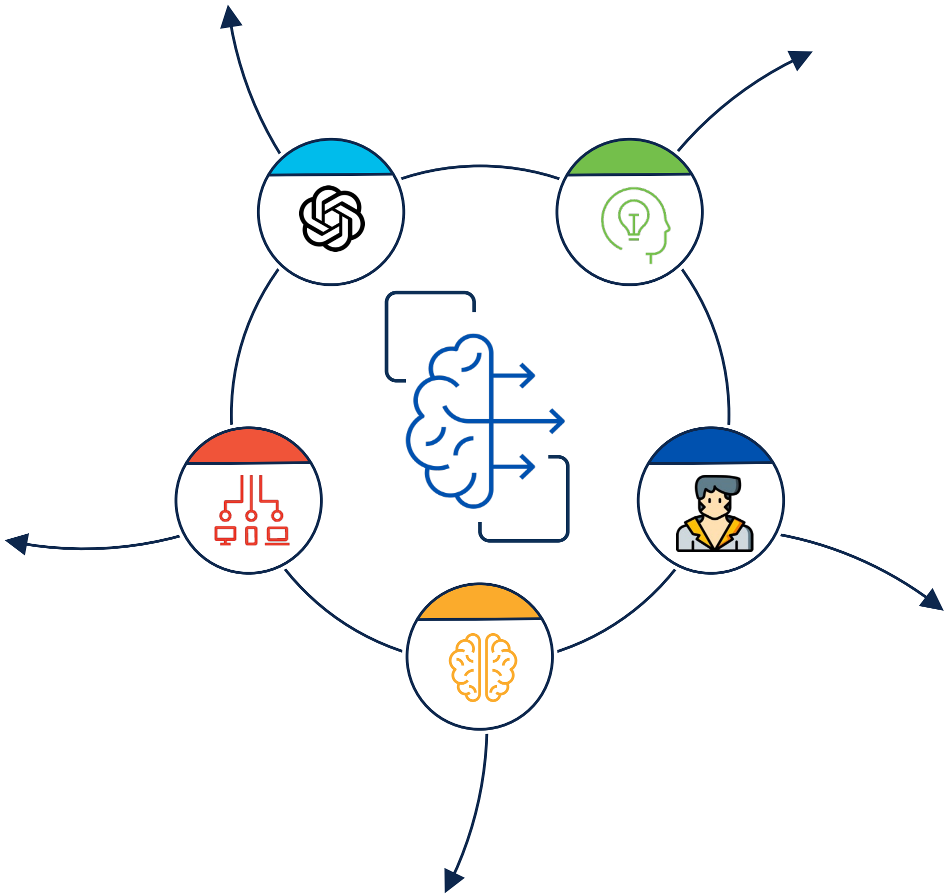
cisco

OpenNetworkIng Conference Japan 2023

OUTLINE PROGRAM SPONSORS ACCESS Facebook

# A 12-year journey developing breakthrough AI products for Networking

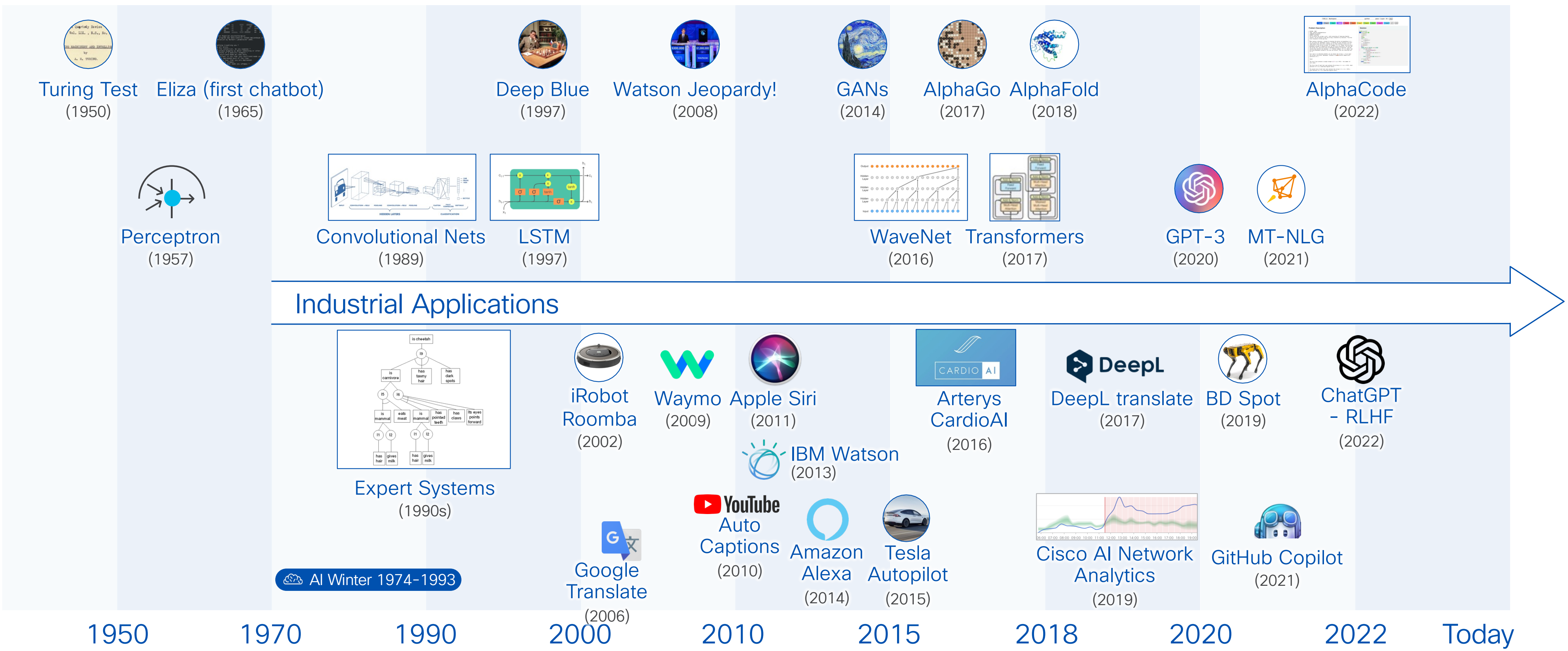
## ネットワーキングのための画期的なAI製品を開発する12年間の旅路



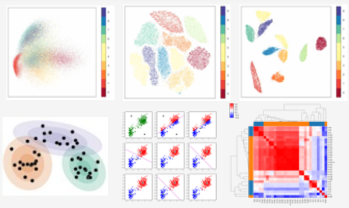
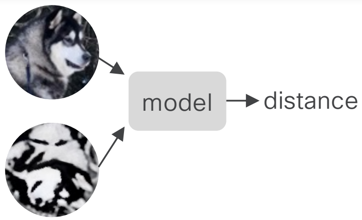
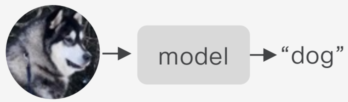
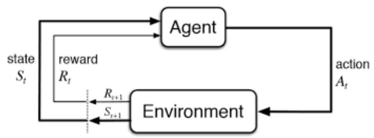
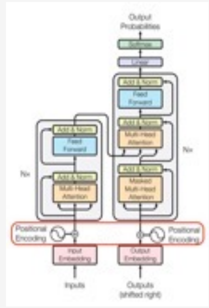

JP Vasseur, PhD - jpv@cisco.com
Cisco Fellow, ML/AI


October 2023

# A brief history of AI/ML and its applications



# Learning Strategies and Key Challenges

| Unsupervised Learning | Self-Supervised Learning | Supervised Learning | Reinforcement Learning | Large Language modes |
|---|---|---|---|---|
|  |  |  |  |  |
| Learn from **patterns** | Learn from **structure** | Learn from **experts** | Learn from **experience** | |

The Dark Matter of AI Yann Lecun

$10^8$ — $10^7$ — $10^6$ — $10^5$ — $10^4$ — $10^3$ — $10^2$ — $10^1$ — 1 bit →

## Number of related topics and approaches

Bits of information per sample

### Explainable AI
How to make the model explain its predictions to a user?

### Adversarial AI
How to trick or exploit models for malicious purposes?

### Active Learning
How to collect training samples optimally?

### Ethical AI
How to design algorithms that recognize societal biases in their training data?

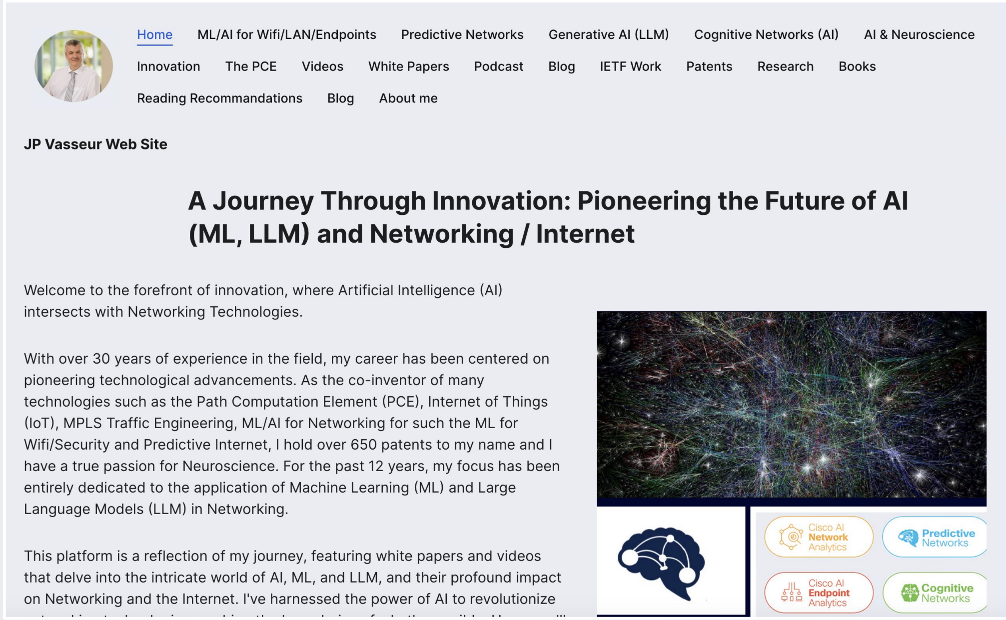
### Transfer Learning
How to re-use a model trained on task X on another task Y?

### Private AI
How to create models that never disclose private information from their training data?

### Multi-task Learning
How to train a single model that accomplishes several tasks?

### Federated Learning
Use of decentralized device to collaboratively learn a shared model without sharing local data

# Cisco AI/ML journey


Home ML/AI for Wifi/LAN/Endpoints Predictive Networks Generative AI (LLM) Cognitive Networks (AI) AI & Neuroscience
Innovation The PCE Videos White Papers Podcast Blog IETF Work Patents Research Books
Reading Recommandations Blog About me

**JP Vasseur Web Site**

**A Journey Through Innovation: Pioneering the Future of AI (ML, LLM) and Networking / Internet**

Welcome to the forefront of innovation, where Artificial Intelligence (AI) intersects with Networking Technologies.

With over 30 years of experience in the field, my career has been centered on pioneering technological advancements. As the co-inventor of many technologies such as the Path Computation Element (PCE), Internet of Things (IoT), MPLS Traffic Engineering, ML/AI for Networking for such the ML for Wifi/Security and Predictive Internet, I hold over 650 patents to my name and I have a true passion for Neuroscience. For the past 12 years, my focus has been entirely dedicated to the application of Machine Learning (ML) and Large Language Models (LLM) in Networking.

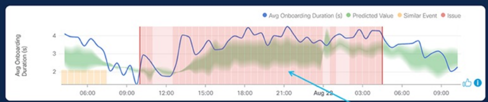
This platform is a reflection of my journey, featuring white papers and videos that delve into the intricate world of AI, ML, and LLM, and their profound impact on Networking and the Internet. I've harnessed the power of AI to revolutionize



[www.jpvasseur.me](http://www.jpvasseur.me)

# Our ML/AI Journey since 2012 ...

## Building a ML Model from Telemetry to detect anomalies
2016-2019



Raw Telemetry was used to build a model predicting the Average On-boarding time

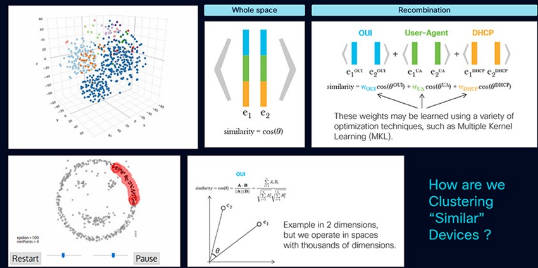
The green band is computed using Gradient Boosted Tree used to compute the lower/upper bounds for the Average On-Boarding time considering a number of networking parameters … (not just what we usually see in terms of average on-boarding time at a given time).

This is the value of Data … turning Telemetry into a Model so as to detect Anomalies

Input parameters for the model were: # radios, # sequences, # onboardings, # clients, proportion of clients with .1X, PSK or open authentication

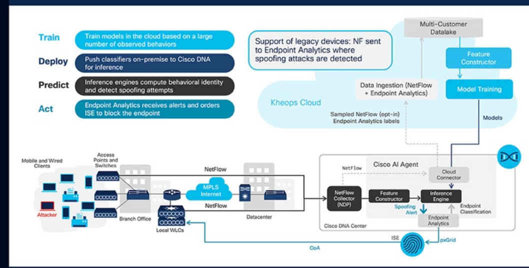
RAW Telemetry would have shown what were the usual average on-boarding time considering specific circumstances

## Cisco Endpoint Analytics
2020



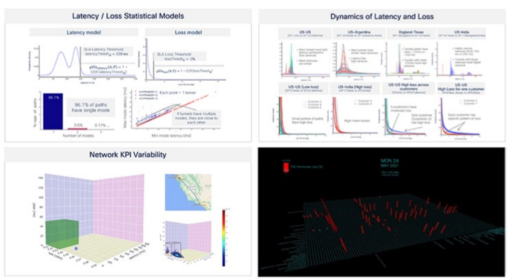

How are we Clustering "Similar" Devices ?

## Cisco AI Endpoint Analytics Detecting Spoofing Attacks with ML
2021




## Deep Study of Internet Dynamics
2019-…

Very deep study: millions of paths, thousands of SPs, multiple access types, …




## First Predictive Engine for the Internet
Mid 2020-Now

Imagine a world (only) reacting with no learning?

**The Internet**

**The Internet has been Reactive for 35 years…**

- Routing/QoS inherently static
- Multiple Recovery mechanisms using Protection and Restoration
  - Relies of fast detection of failure, followed by rerouting
  - Optical, Fast IGP (OSPF, IS-IS), IP FRR BGP, MPLS FRR
- Few Adaptive strategies based on recent events (backoff, … ) or approximate future

**No learning …**

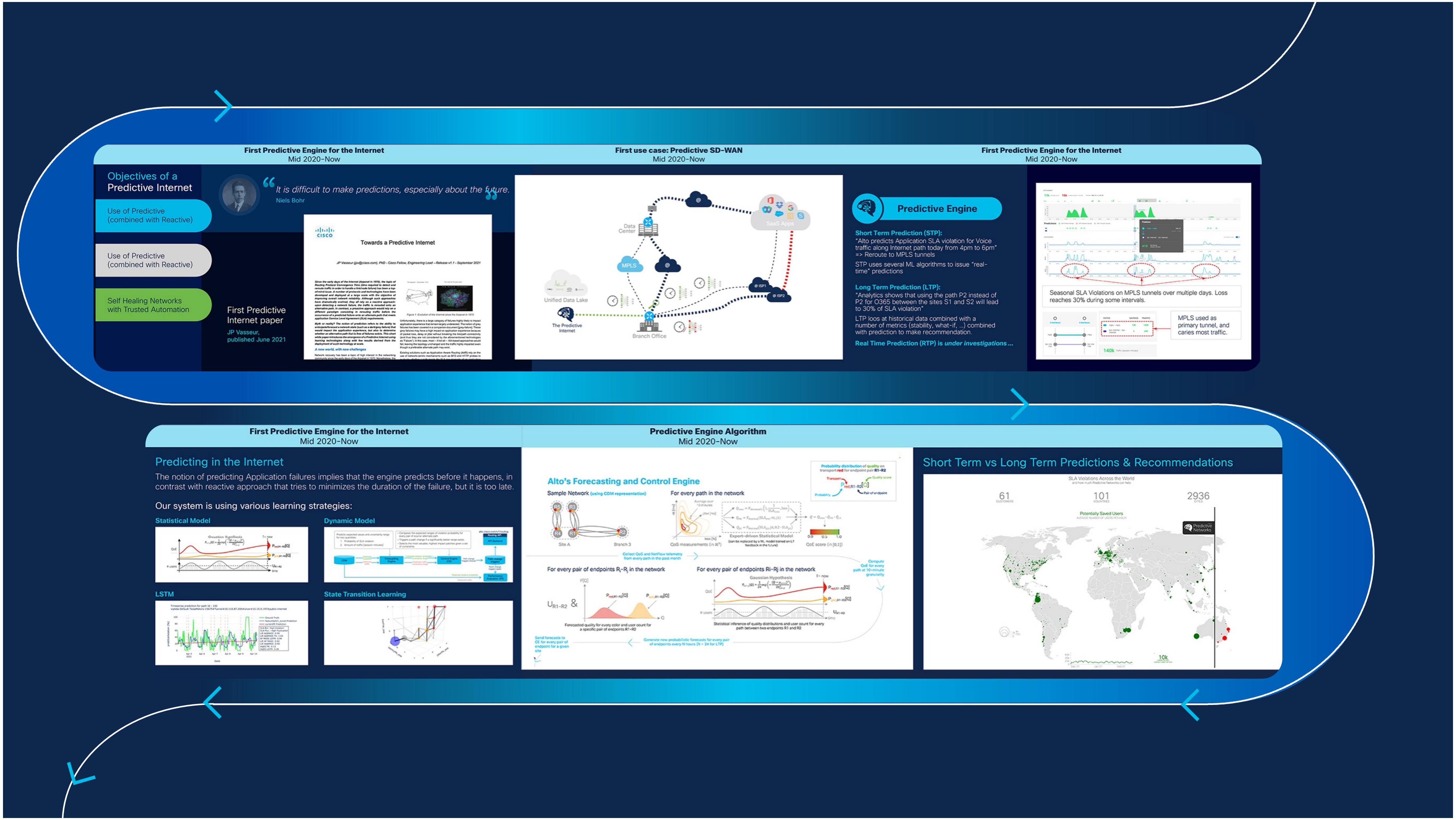
**The Human Brain**

- **Learns** process not entirely known: nature versus nurture, build a model of the world (observation), ability to predict seems central, experience *(Plasticity)*
- **Predicts** (e.g predictive coding) – Various theories
- **Plans** (higher executive functions)

**Predictive Networks (Networking "Brain")**

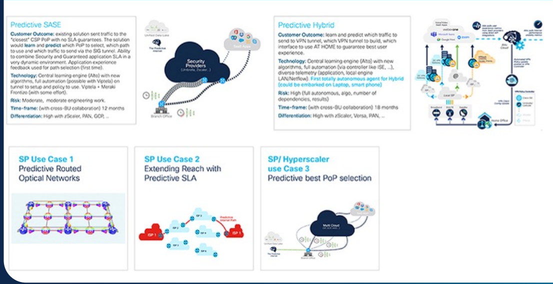
**Predictive Networks:**

- **Build** (ML/Statistical) models of the world (Internet & Application)
- **Predict** potential issues (application experience)
- **Learn** and keeps improving (Telemetry)
- **Plan** with Automation

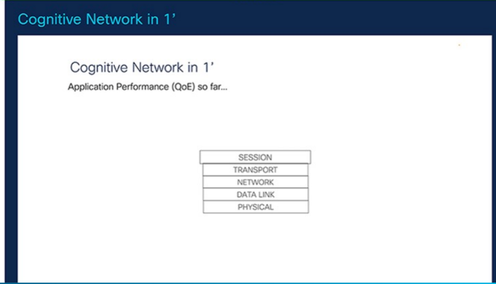


First Predictive Engine for the Internet
Mid 2020-Now
Objectives of a Predictive Internet
Use of Predictive (combined with Reactive)
Use of Predictive (combined with Reactive)
Self Healing Networks with Trusted Automation
"It is difficult to make predictions, especially about the future."
Niels Bohr
Towards a Predictive Internet
First Predictive Internet paper
JP Vasseur, published June 2021
First use case: Predictive SD-WAN
Mid 2020-Now
Data Center
SaaS Apps
MPLS
Unified Data Lake
The Predictive Internet
Branch Office
First Predictive Engine for the Internet
Mid 2020-Now
Predictive Engine
Short Term Prediction (STP):
"Alto predicts Application SLA violation for Voice traffic along Internet path today from 4pm to 6pm" => Reroute to MPLS tunnels
STP uses several ML algorithms to issue "real-time" predictions
Long Term Prediction (LTP):
"Analytics shows that using the path P2 instead of P2 for O365 between the sites S1 and S2 will lead to 30% of SLA violation"
LTP loos at historical data combined with a number of metrics (stability, what-if, ...) combined with prediction to make recommendation.
Real Time Prediction (RTP) is under investigations ...
Seasonal SLA Violations on MPLS tunnels over multiple days. Loss reaches 30% during some intervals.
MPLS used as primary tunnel, and caries most traffic.
First Predictive Engine for the Internet
Mid 2020-Now
Predicting in the Internet
The notion of predicting Application failures implies that the engine predicts before it happens, in contrast with reactive approach that tries to minimizes the duration of the failure, but it is too late.
Our system is using various learning strategies:
Statistical Model
Dynamic Model
LSTM
State Transition Learning
Predictive Engine Algorithm
Mid 2020-Now
Alto's Forecasting and Control Engine
Sample Network (using CDM representation)
For every path in the network
Site A
Branch 3
QoS measurements
Expert-driven Statistical Model
QoE score
Collect QoS and NetFlow telemetry from every path in the past month
For every pair of endpoints Ri-Rj in the network
For every pair of endpoints Ri-Rj in the network
Gaussian Hypothesis
Forecasted quality for every color and user count for a specific pair of endpoints R1-R2
Statistical inference of quality distributions and user count for every path between two endpoints R1 and R2
Generate new probabilistic forecasts for every pair of endpoints every N hours (N = 24 for LTP)
Short Term vs Long Term Predictions & Recommendations
SLA Violations Across the World
61
101
2936
Potentially Saved Users
Predictive Networks
10k

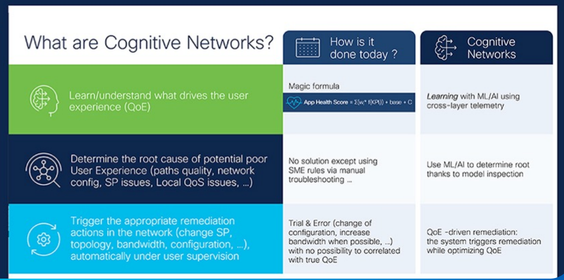
**Future of Predictive Networks**




**Cognitive Networks**
2022-...

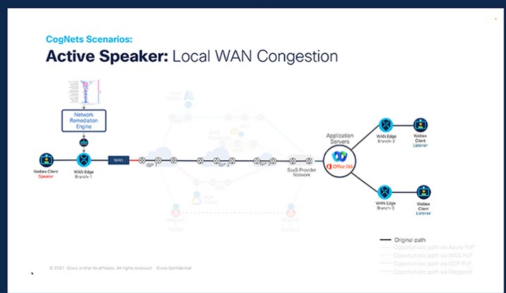



**Cognitive Networks**
2022-...




**First QoE Model for the Networking industry**
2022-...

**Automated actions to improve the QoE**
2022-...




**Cognitive Networks**
2022-...

Future

# Level of Interest for ML/AI

# Why two camps ?

OUTDATED

OUTDATED

OUTDATED

**Pro ML/AI** ... who believe
that ML/AI is the *only*
approach to build (intelli
useful systems

L/AI ... who are highly
L/AI is a pure fantasy
and does not work)
e technology is evil
will replace humanity

(I h are wr

- Be Pragm g a specific issue
(need DEEP cannot do)
- Do not build wron nothing to do with
(human level) intelligence useful for a broad range of problems)

OUTDATED

OUTDATED

OUTDATED

# Why being skeptical a
# ML/AI?

- A bit of fatigue a
- Over promise, Ov

developing ML
pro go

decade of
nt

approaches
any worked)

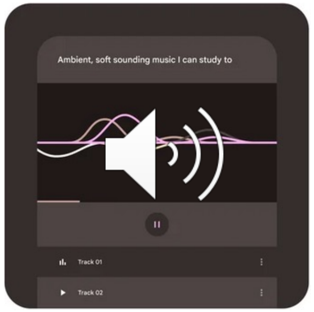
- Our have been deployed
a

e there and AI/ML for
king moving to the next phase ....

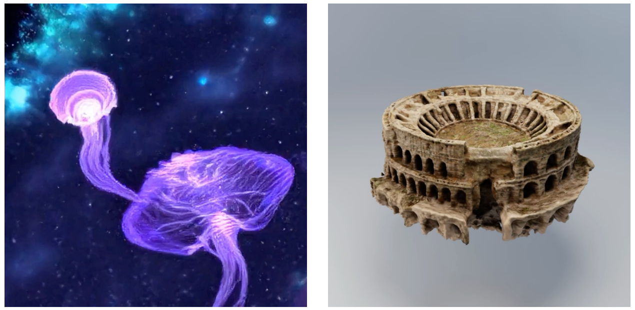
# What is Generative AI

Generative AI refers to a type of artificial intelligence that is capable of generating new and original data, such as images, music, text, or even entire videos, that are similar in style or structure to the data it has been trained on. Unlike other types of AI that are designed to recognize patterns or make predictions based on existing data, generative AI models are designed to create new data that is similar to the input data it has been trained on. …. Definition from a Generative AI ☺

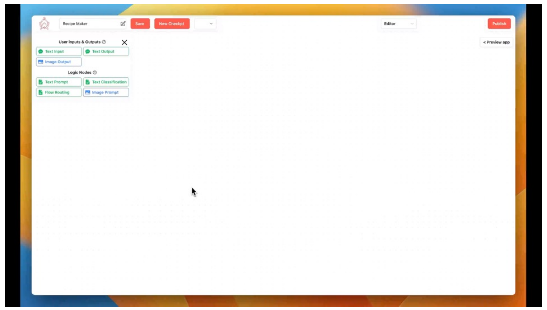
Image Generation




Music Generation
**(Source MusicLM)**



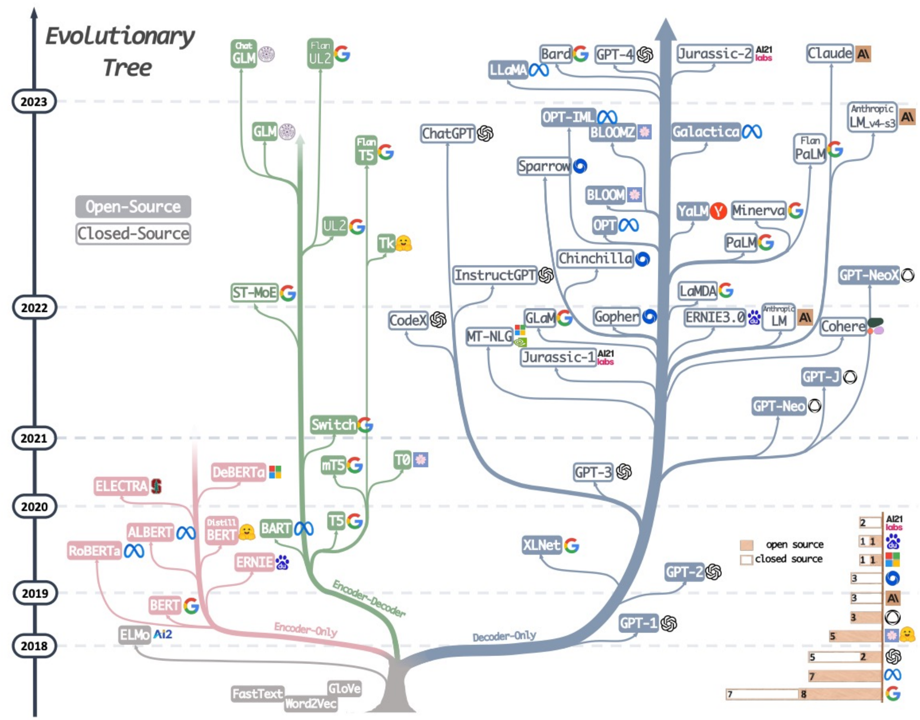
Text to 3d, text to Video
**(Source NVIDIA Picasso)**



Software/ Code Generation
**(Source ForgeAI)**

# “Current” state of LLMs (thousands of new models / week)

| | Model | Release Time | Size (B) | Base Model | Adaptation IT | Adaptation RLHF | Pre-train Data Scale | Latest Data Timestamp | Hardware (GPUs / TPUs) | Training Time | Evaluation ICL | Evaluation CoT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T5 [71] | Oct-2019 | 11 | - | - | - | 1T tokens | Apr-2019 | 1024 TPU v3 | - | ✓ | - |
| | mT5 [72] | Mar-2021 | 13 | - | - | - | 1T tokens | Apr-2019 | - | - | ✓ | - |
| | PanGu-$\alpha$ [73] | May-2021 | 13* | - | - | - | 1.1TB | - | 2048 Ascend 910 | - | ✓ | - |
| | CPM-2 [74] | May-2021 | 198 | - | - | - | 2.6TB | - | - | - | - | - |
| | T0 [28] | Oct-2021 | 11 | T5 | ✓ | - | - | - | 512 TPU v3 | 27 h | ✓ | - |
| | GPT-NeoX-20B [75] | Feb-2022 | 20 | - | - | - | 825GB | Dec-2022 | 96 40G A100 | - | ✓ | - |
| | CodeGen [76] | Mar-2022 | 16 | - | - | - | 577B tokens | - | - | - | ✓ | - |
| | Tk-Instruct [77] | Apr-2022 | 11 | T5 | ✓ | - | - | - | 256 TPU v3 | 4 h | ✓ | - |
| | UL2 [78] | Apr-2022 | 20 | - | ✓ | - | 1T tokens | Apr-2019 | 512 TPU v4 | - | ✓ | ✓ |
| | OPT [79] | May-2022 | 175 | - | - | - | 180B tokens | - | 992 80G A100 | - | ✓ | - |
| Publicly | NLLB [80] | Jul-2022 | 55 | - | - | - | - | - | - | 51968 h | ✓ | - |
| Available | BLOOM [66] | Jul-2022 | 176 | - | - | - | 366B | - | 384 80G A100 | 105 d | ✓ | - |
| | GLM [81] | Aug-2022 | 130 | - | - | - | 400B tokens | - | 768 40G A100 | 60 d | ✓ | - |
| | Flan-T5 [82] | Oct-2022 | 11 | T5 | ✓ | - | - | - | - | - | ✓ | ✓ |
| | mT0 [83] | Nov-2022 | 13 | mT5 | ✓ | - | - | - | - | - | ✓ | - |
| | Galactica [35] | Nov-2022 | 120 | - | - | - | 106B tokens | - | - | - | ✓ | ✓ |
| | BLOOMZ [83] | Nov-2022 | 176 | BLOOM | ✓ | - | - | - | - | - | ✓ | - |
| | OPT-IML [84] | Dec-2022 | 175 | OPT | ✓ | - | - | - | 128 40G A100 | - | ✓ | ✓ |
| | Pythia [85] | Jan-2023 | 12 | - | - | - | 300B tokens | - | 256 40G A100 | 72300 h | ✓ | - |
| | LLaMA [57] | Feb-2023 | 65 | - | - | - | 1.4T tokens | - | 2048 80G A100 | 21 d | ✓ | - |
| | GShard [86] | Jan-2020 | 600 | - | - | - | 1T tokens | - | 2048 TPU v3 | 4 d | - | - |
| | GPT-3 [55] | May-2020 | 175 | - | - | - | 300B tokens | - | - | - | ✓ | - |
| | LaMDA [87] | May-2021 | 137 | - | - | - | 2.81T tokens | - | 1024 TPU v3 | 57.7 d | - | - |
| | HyperCLOVA [88] | Jun-2021 | 82 | - | - | - | 300B tokens | - | 1024 A100 | 13.4 d | ✓ | - |
| | Codex [89] | Jul-2021 | 12 | GPT-3 | - | - | 100B tokens | May-2020 | - | - | ✓ | - |
| | ERNIE 3.0 [90] | Jul-2021 | 10 | - | - | - | 375B tokens | - | 384 V100 | - | ✓ | - |
| | Jurassic-1 [91] | Aug-2021 | 178 | - | - | - | 300B tokens | - | 800 GPU | - | ✓ | - |
| | FLAN [62] | Oct-2021 | 137 | LaMDA | ✓ | - | - | - | 128 TPU v3 | 60 h | ✓ | - |
| | MT-NLG [92] | Oct-2021 | 530 | - | - | - | 270B tokens | - | 4480 80G A100 | - | ✓ | - |
| | Yuan 1.0 [93] | Oct-2021 | 245 | - | - | - | 180B tokens | - | 2128 GPU | - | ✓ | - |
| | Anthropic [94] | Dec-2021 | 52 | - | - | - | 400B tokens | - | - | - | ✓ | - |
| | WebGPT [70] | Dec-2021 | 175 | GPT-3 | - | ✓ | - | - | - | - | ✓ | - |
| | Gopher [59] | Dec-2021 | 280 | - | - | - | 300B tokens | - | 4096 TPU v3 | 920 h | ✓ | - |
| | ERNIE 3.0 Titan [95] | Dec-2021 | 260 | - | - | - | 300B tokens | - | 2048 V100 | 28 d | ✓ | - |
| Closed Source | GLaM [96] | Dec-2021 | 1200 | - | - | - | 280B tokens | - | 1024 TPU v4 | 574 h | ✓ | - |
| | InstructGPT [61] | Jan-2022 | 175 | GPT-3 | ✓ | ✓ | - | - | - | - | ✓ | - |
| | AlphaCode [97] | Feb-2022 | 41 | - | - | - | 967B tokens | Jul-2021 | - | - | - | - |
| | Chinchilla [34] | Mar-2022 | 70 | - | - | - | 1.4T tokens | - | - | - | ✓ | - |
| | PaLM [56] | Apr-2022 | 540 | - | - | - | 780B tokens | - | 6144 TPU v4 | - | ✓ | ✓ |
| | AlexaTM [98] | Aug-2022 | 20 | - | - | - | 1.3T tokens | - | 128 A100 | 120 d | ✓ | ✓ |
| | Sparrow [99] | Sep-2022 | 70 | - | - | ✓ | - | - | 64 TPU v3 | - | ✓ | - |
| | WeLM [100] | Sep-2022 | 10 | - | - | - | 300B tokens | - | 128 A100 40G | 24 d | ✓ | - |
| | U-PaLM [101] | Oct-2022 | 540 | PaLM | - | - | - | - | 512 TPU v4 | 5 d | ✓ | ✓ |
| | Flan-PaLM [82] | Oct-2022 | 540 | PaLM | ✓ | - | - | - | 512 TPU v4 | 37 h | ✓ | ✓ |
| | Flan-U-PaLM [82] | Oct-2022 | 540 | U-PaLM | ✓ | - | - | - | - | - | ✓ | ✓ |
| | GPT-4 [46] | Mar-2023 | - | - | ✓ | ✓ | - | - | - | - | ✓ | ✓ |
| | PanGu-$\Sigma$ [102] | Mar-2023 | 1085 | PanGu-$\alpha$ | - | - | 329B tokens | - | 512 Ascend 910 | 100 d | ✓ | - |




**WARNING:** open-source does **not** mean commercially available.



Source: A Survey of Large Language Models and Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

# Examples of LLM Use Cases For Networking

## UI/CLI Replacement

- Interact with various devices and controllers via a ChatBot as opposed to the classic CLI or UI interface.

*Out of scope for now.*

## Troubleshooting

- Suggest potential root causes based on user prompt and proposes a troubleshooting strategy.
- Uses *tools* to interact with network domains and execute troubleshooting steps, interprets outputs and received telemetry to identify issues.
- Proposes remediation steps based on best practices.

## Performance Monitoring

- Analyse large amounts of data and highlight top/worst performers for key network metrics.
- Corelates metrics from different dashboards, tools or controllers (SD-WAN, Thousand Eyes, DNAC, etc) and builds new visualizations.

## Configuration Assistance

- Guidance for accomplishing various configuration tasks (steps, commands etc).
- Reviews existing configuration deployed against best practices. Makes improvement recommendations.
- Builds automation (scripts, playbooks) for common configuration tasks.

# Summary - Generative AI

(L)LM have been in the works for a long time ('48), long list of recent cutting technologies (transformers ('18), RLHF, ...) - first commercial BREAKTHROUGHT implementation recently available (Chat-GPT) on Nov '22

Works "surprisingly well" for several key tasks (e.g., text summarization, translation, code generation) with emergent properties (can/cannot do)

Number of use cases: Networking (conversational, troubleshooting with RCA, analytics, config management), Security & Collboration, Applications.

Architecture & Technologies: prompt tuning (tools, ICL, Thought reasoning, RAG, ...), model tuning (training strategies), generic large vs specialized open-source, knowledge DB with semantic search, agents, ... and overall architecture

Technical challenges: Reliability (determinism, hallucinations), Information Sourcing, Privacy, Security (prompt injection, ...), ...

Are LLMs the long-awaited Bing-Bang?

- Emerging properties keeps arising (general pattern matching engines, used for complex reasoning, anomaly detection)
- Never-seen before: combination of open innovation and major companies solving issues at unprecedented pace

# Lots of exciting AI topics

What have LLMs learned ?

Do LLM understands the world (probing classifier, ...) ?

LLM as general patterns matching

Interpretability (mechanistic, ...)

Tracing factual knowledge, Watermarking

LLM generalization and Grokking

Accessing trillion tokens

LLM and RL

LLM & Security

The bridge to possible