GenAI Efficiency is About More than Models

Zorg Allport, Zhen Dong, Lutfi Erdogan, Amir Gholami, Sid Jha, *Kurt Keutzer*, Sehoon Kim, Nicholas Lee, Xiuyu Li, Monish Maheshwaran, Karttikeya Mangalam, Hiva Mohammadzadeh, Suhong Moon, Sebastian Nehrdich, Sheng Shen, Ryan Tabrizi, Chenfeng Xu, Wenchao Zhao, Banghua Zhu,
Fellow Faculty: Jiantiao Jiao, Sophia Shao

Keynote Address at the Young Professionals Symposium at the MLSYS Conference 2024

1 Hour of My Time Growing Up

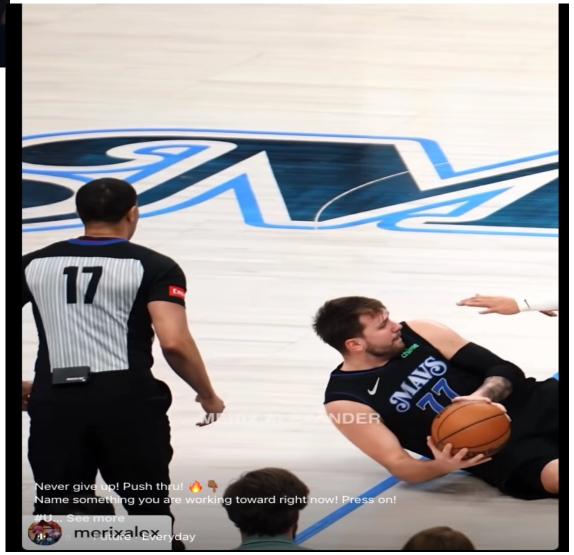
BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

1/5 of a Double Feature
2 movies was the norm

1/3 of a Baseball Game

1 Hour of Your Time Today!!!
A HUGE Responsibility

40 Instagram Reels



100 Avg Facebook Reels

The Inside Story of ChatGPT's Astonishing Potential | Greg Brockman | TED

4 Popular AI Ted Talks!

Three Aspects of My Talk

- My best and most heartfelt advice to young professionals in this field
- My enthusiasm for Compound GenAI Systems
- Research problems in Compound GenAI Systems and their role in MLSys and elsewhere

My (GenAI) Talk in One Slide

Machine Learning/Deep Learning have rapidly evolved through a number of eras:

- ML Era 1: Orchestration of statistics gave us **Machine Learning**
- ML Era 2: Orchestration of Machine Learning algorithms gave us **Neural Nets**
- ML Era 3: Orchestration of Neural Net model functions/components gave us the **Transformer**
- ML Era 4: Orchestration of Transformers gave us **Large Language Models**
- ML Era 5: Orchestration of Large Language Models gives us **Compound GenAI Systems**

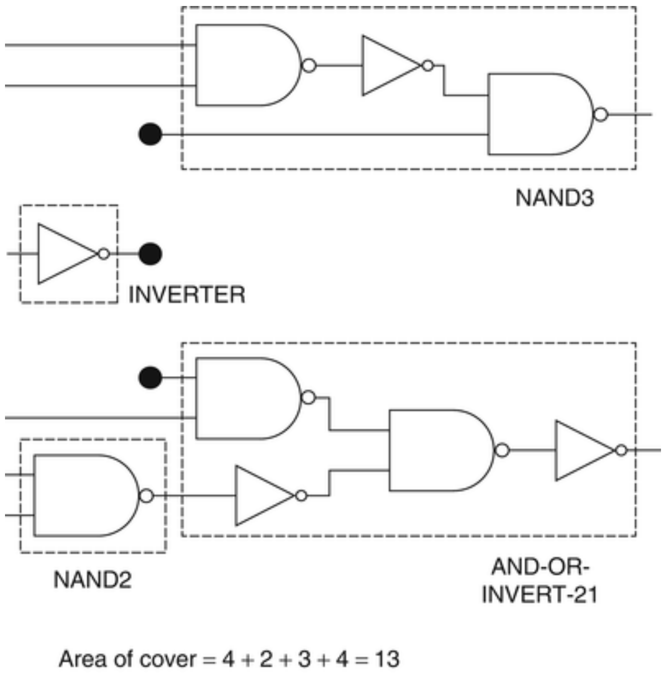
Compound GenAI Systems give us the next generation of key problems for ML Systems

For the Young Professionals:
Quote #1: A Quote that Has Shaped My Whole Career

“The right perspective, context, or point of view is worth 80 IQ points.” Alan Kay

Examples I Know Well

- DAGON
  - Every graduate student taking an advanced compiler course was familiar with the Aho-Corasick algorithm for string matching (1975) and the Sethi-Ullman algorithm for code generation in compilers (1970).
  - But, in 1987, researchers attacking the problem of matching Boolean equations to standard logical cells did not have that perspective.



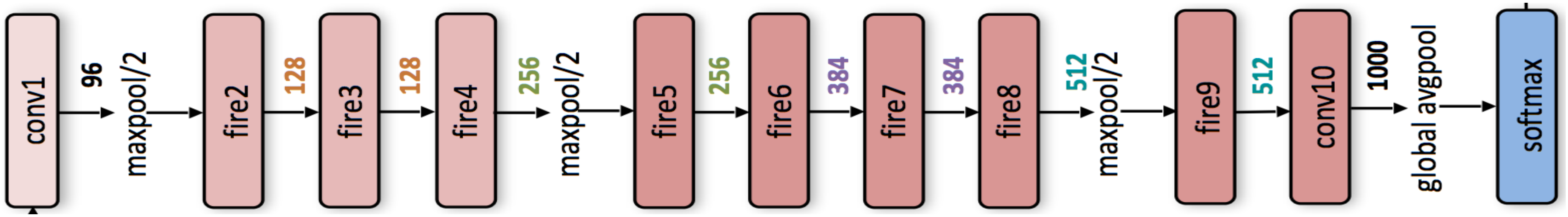
Sethi, Ravi, and Jeffrey D. Ullman. "The generation of optimal code for arithmetic expressions." *Journal of the ACM (JACM)* 17, no. 4 (1970): 715-728.

Aho, Alfred V., and Margaret J. Corasick. "Efficient string matching: an aid to bibliographic search." *Communications of the ACM* 18, no. 6 (1975): 333-340.

Keutzer, Kurt. "DAGON: Technology binding and local optimization by DAG matching." In *Proceedings of the 24th ACM/IEEE Design Automation Conference*, pp. 341-347. 1987.

Examples I Know Well

- SqueezeNet
  - Mainstream computer vision in 2015 was entirely focused on improving ImageNet accuracy using GPUs
  - The fields also had a good palette of Convolutional Neural Net model building elements
  - Our (Forrest Iandola and I) perspective from embedded systems taught us that there was always a “trickle down” of technology to the edge and soon NN’s would be everywhere.



Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." CVPR, pp. 1-9. 2015.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

A Quote that Has Shaped My Whole Career

"The right perspective, context, or point of view is worth 80 IQ points." Alan Kay

A different perspective can enable us to:

- Bring a fresh approach to an established problem
- Apply a well-established playbook to a new area
- Identify new research directions before others do
  - The first person to the beach picks up the most diamonds

My goal of this talk is to give you a new perspective

For the Young Professionals: Quote #2: The Value of History

The easiest way to predict the future is to study history.

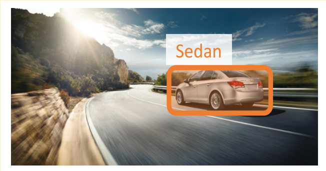
Corollary of:

- History repeats itself.
- There's nothing new under the sun.
- Etc.

History Often Gives a Useful Perspective
Our Applications 2007-2013

Computer Vision

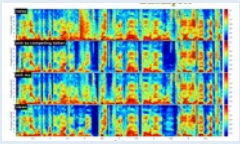
Image Classification



Object Detection

Image Segmentation

Audio Analysis And ASR



Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

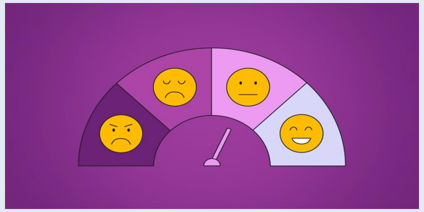
Multimedia and Rec Systems

Sentiment Analysis

Music Recommendation

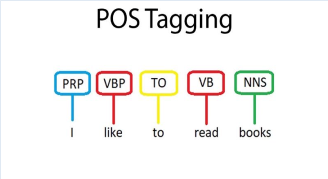
Ad Recommendation

Natural Language Processing



Sentiment Analysis

Spam Detection



Part of Speech Tagging

How Were We Solving Them?
ML Approaches in 2007 – 2013

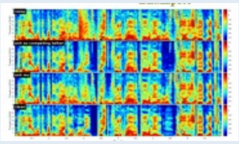
Image Classification

Object Detection

Image Segmentation

Computer Vision

Convolutions
Histograms
K-means
Support VMs …



Audio Enhancement

Call-center Sentiment Analysis

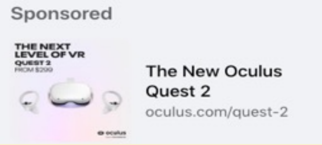
Speech Recognition

Audio Analysis And ASR

Gaussian Mixture Models
HMMs
Support VMs …

Sentiment Analysis

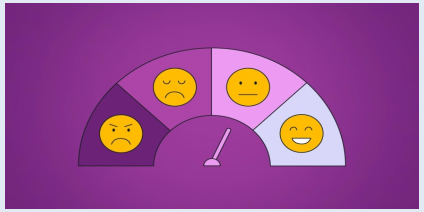
Music Recommendation



Ad Recommendation

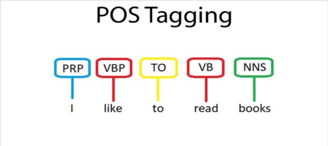
Multimedia and Rec Systems

Gaussian Mixture Models
HMMs
Support VMs …



Sentiment Analysis

Spam Detection

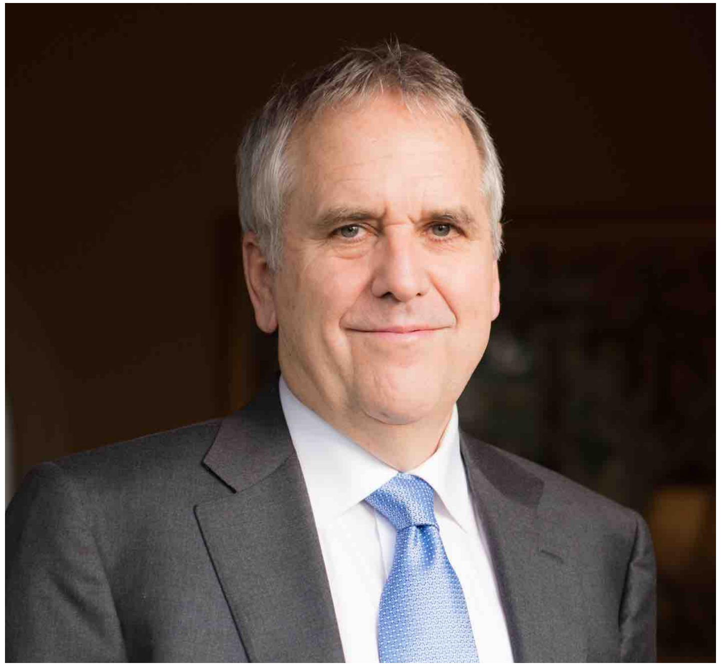


Part of Speech Tagging

Natural Language Processing

Bag-of-words,
Latent Dirichlet Allocation,
Hidden Markov Models

ML Era 1: (Traditional) Machine Learning



Michael Jordan's definition of Machine Learning:

"algorithms and supporting theory for making predictions and decisions **under uncertainty** based on **observed data.**"

Personal communication

Invited talk: SysML: Perspectives and Challenges,
Michael I. Jordan, SysML (MLSys) 2018

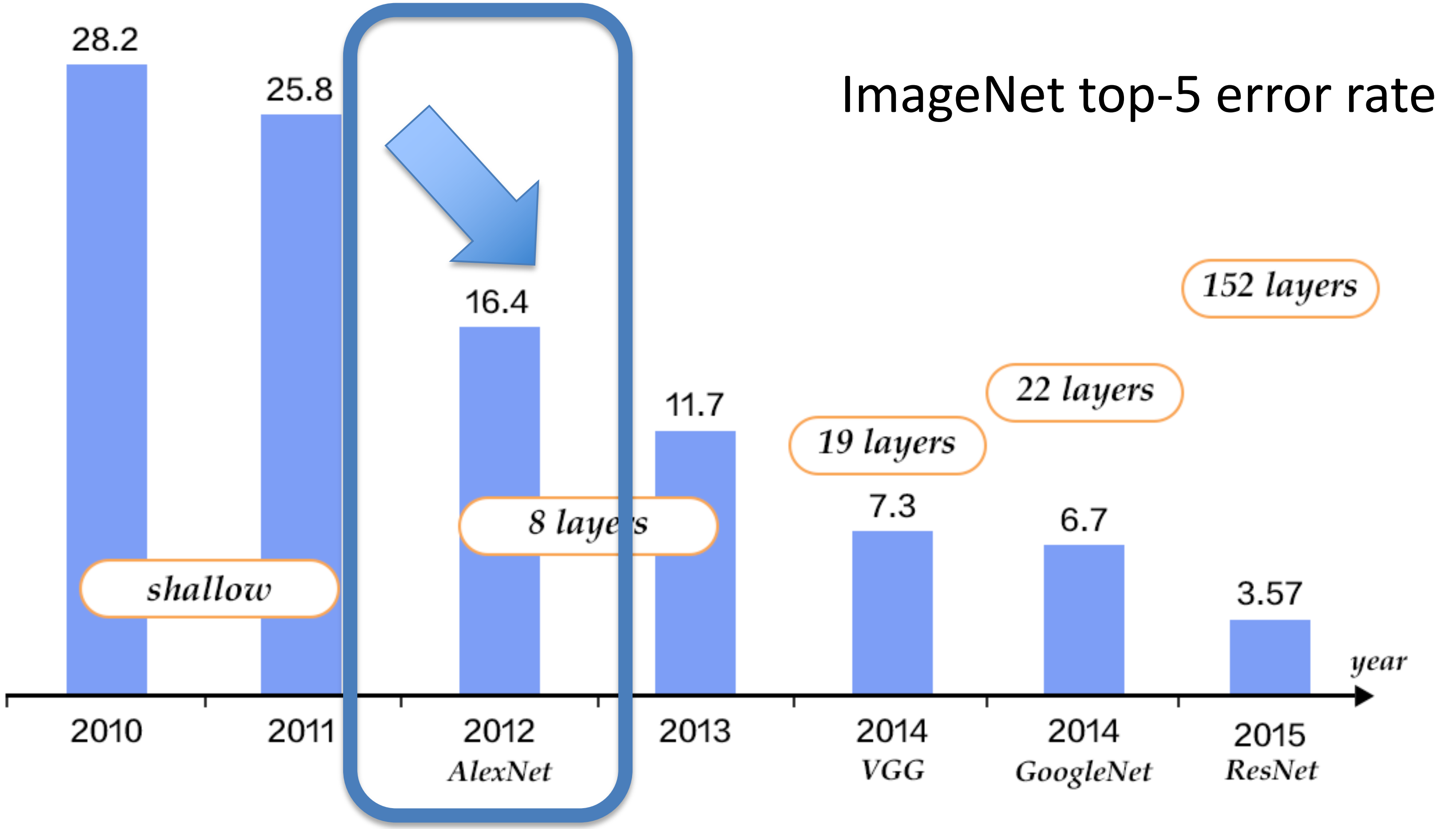
20 Selected *Machine Learning* Algorithms We Employed 2007-2013

- Computer vision
  - Convolution
  - K-means
  - Mean shift
  - Agglomerative algorithms
  - Vector distance
  - Histogram accumulation
  - Hough transform
  - Eigen decomposition
  - Feature matching
  - Support Vector machines

- Speech recognition and audio analysis
  - Convolution
  - K-means
  - Agglomerative hierarchical modeling
  - Orthogonal transformations
  - Gaussian Mixture models
  - Weighted-finite state transducers
  - Hidden-Markov-models
  - Dynamic Bayesian networks
  - Expectation maximization

Led by Bryan Catanzaro, we prided ourselves on making these traditional machine learning algorithms faster, particularly on GPUs and orchestrating them to solve real problems

Big Event #1
Accuracy Improvement after AlexNet (a DNN) 2012



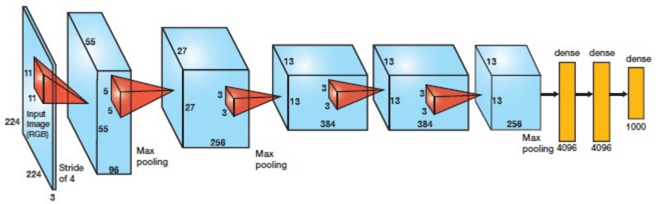
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *NeurIPS* 25 (2012).

Source: http://paddlepaddle.org/docs/develop/book/03.image_classification/index.html

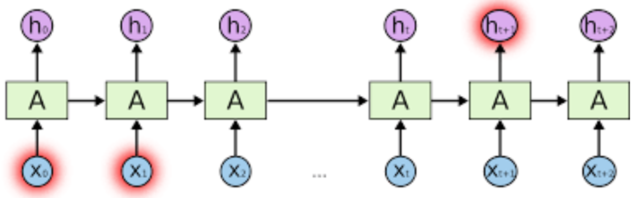
Soon Diverse Machine Learning Algorithms were Replaced by *a Single DNN*!

- Computer vision
  - Convolution
  - K-means
  - Mean shift
  - Agglomerative algorithms
  - Vector distance
  - Histogram accumulation
  - Hough transform
  - Eigen decomposition
  - Feature matching
  - Support Vector machines

- Speech recognition and audio analysis
  - Convolution
  - K-means
  - Agglomerative hierarchical modeling
  - Orthogonal transformations
  - Gaussian Mixture models
  - Hidden-markov models
  - Dynamic Bayesian network
  - Expectation maximization



Convolutional Neural Nets



Recurrent Neural Nets

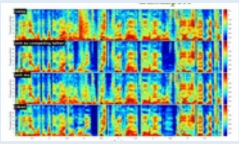
What Problems Were We Solving in 2013?

Image Classification

Object Detection

Image Segmentation

Computer Vision



Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

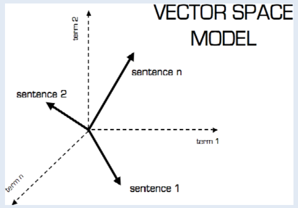
Audio Analysis And ASR

Sentiment Analysis

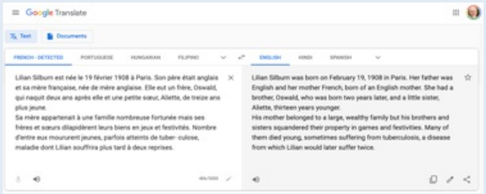
Music Recommendation

Ad Recommendation

Multimedia and Rec Systems



Semantic Similarity



Machine Translation

Named Entity Recognition

Natural Language Processing

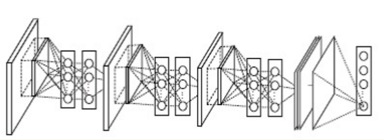
How? Approaches in 2013 – 2020

Image Classification

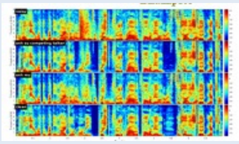
Object Detection

Image Segmentation

Computer Vision



Convolutional NN

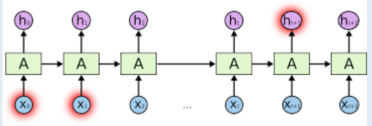


Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

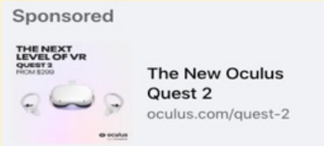
Audio Analysis And ASR



Recurrent NN

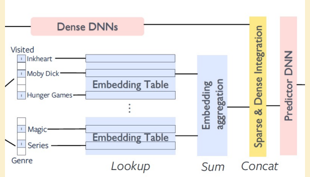
Sentiment Analysis

Music Recommendation

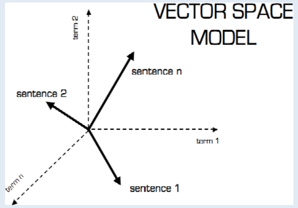


Ad Recommendation

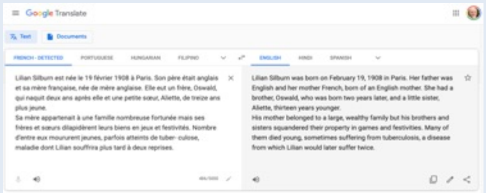
Multimedia and Rec Systems



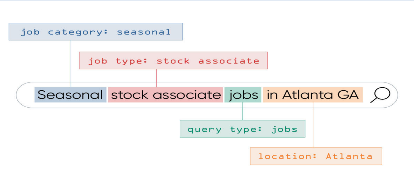
DLRM



Semantic Similarity

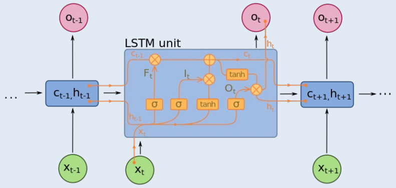


Machine Translation



Named Entity Recognition

Natural Language Processing



LSTM

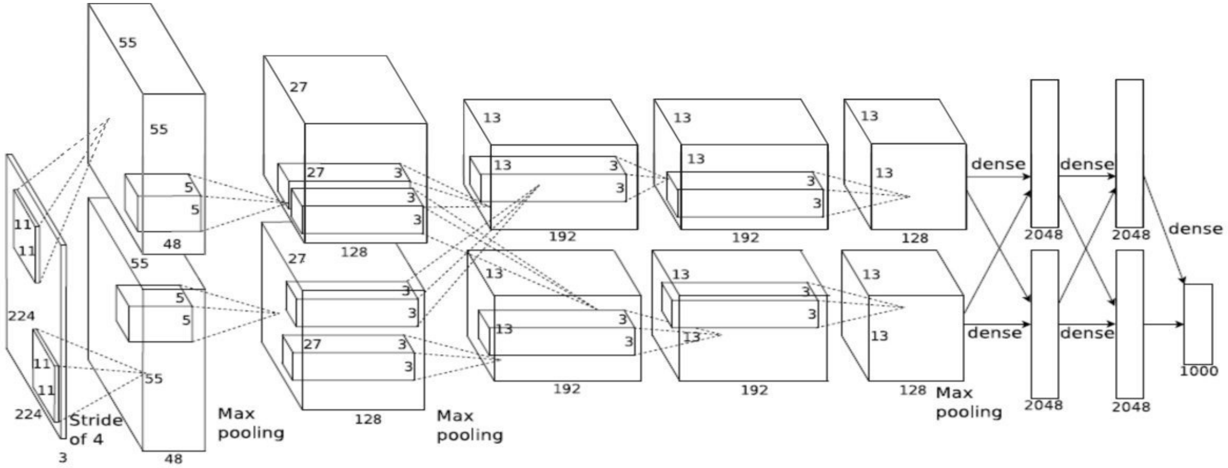
ML Era 2: Deep Learning Approach
Neural Nets

Yann LeCun

December 24, 2019 ·

Some folks still seem confused about what deep learning is. Here is a definition:

DL is constructing networks of parameterized functional modules & training them from examples using gradient-based optimization. That's it.



- LeCunn continues "This definition is orthogonal to the learning paradigm: reinforcement, supervised, or self-supervised."

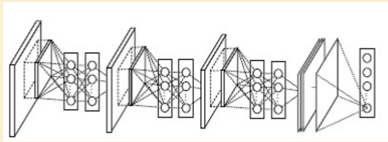
Big Surprise #1:No Algorithmic Approach Had Ever Had Such Broad Application

Image Classification

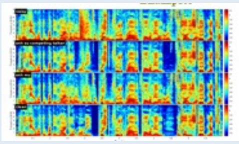
Object Detection

Image Segmentation

Computer Vision



Convolutional NN

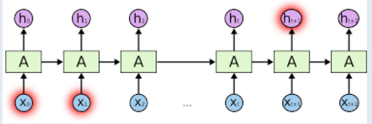


Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

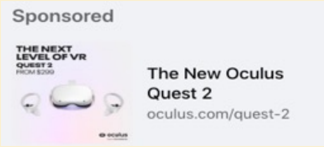
Audio Analysis And ASR



Recurrent NN

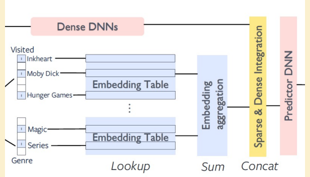
Sentiment Analysis

Music Recommendation

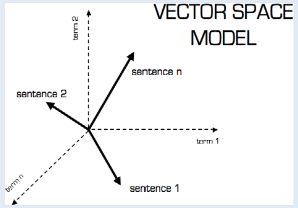


Ad Recommendation

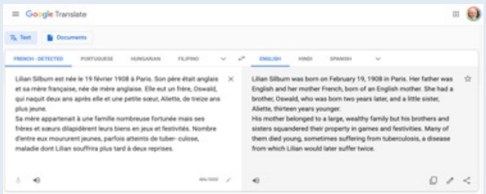
Multimedia and Rec Systems



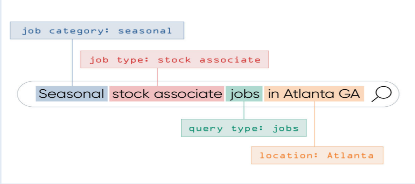
DLRM



Semantic Similarity

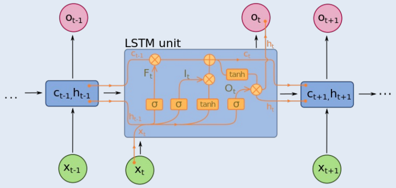


Machine Translation



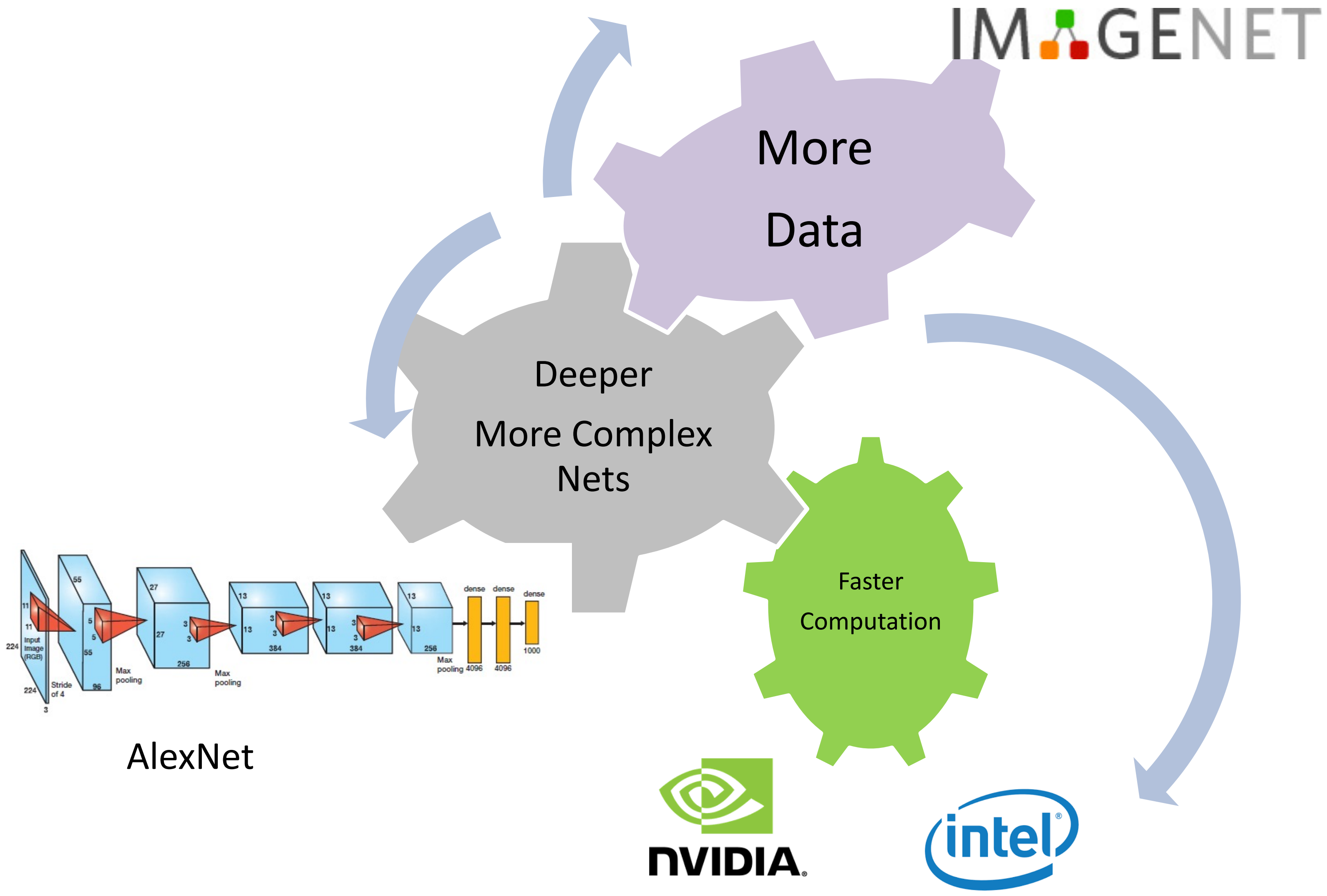
Named Entity Recognition

Natural Language Processing

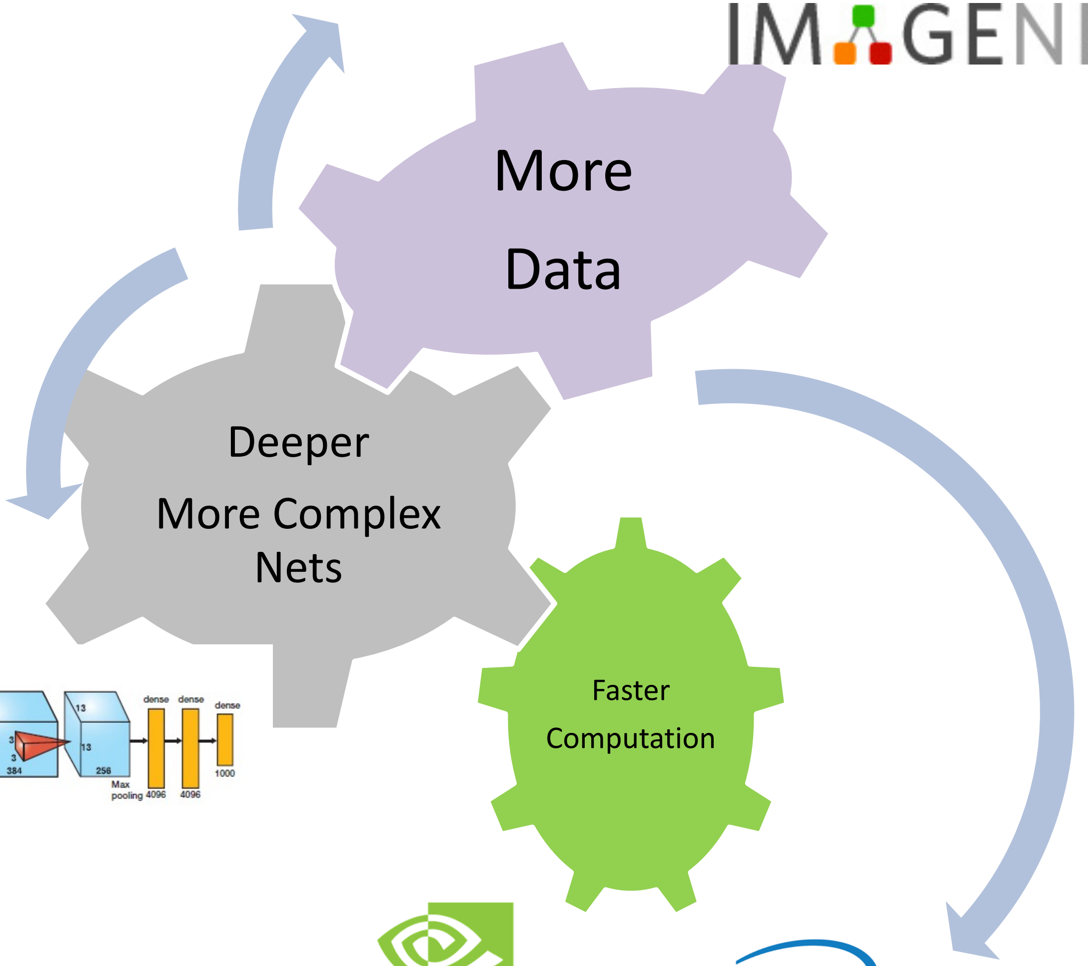


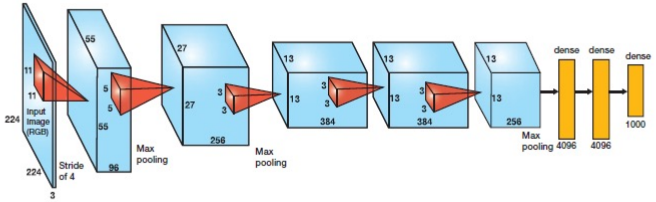
LSTM

Why Was Deep Learning So Successful?
Common View 2018: *Deep* Neural Nets!



We Thought NN Model Architectures Were the Key





AlexNet

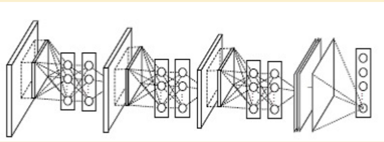
DNN Model Architecture Diversity

Image Classification

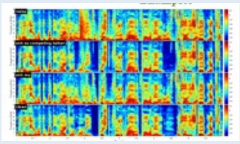
Object Detection

Image Segmentation

Computer Vision



Convolutional NN

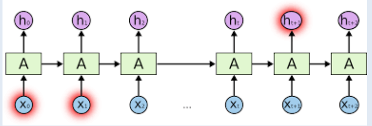


Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

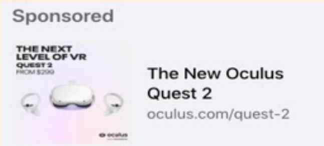
Audio Analysis And ASR



Recurrent NN

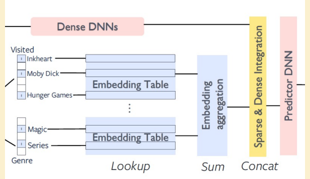
Sentiment Analysis

Music Recommendation

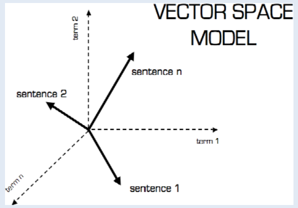


Ad Recommendation

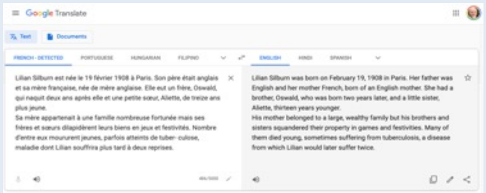
Multimedia and Rec Systems



DLRM



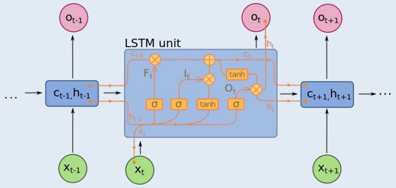
Semantic Similarity



Machine Translation

Named Entity Recognition

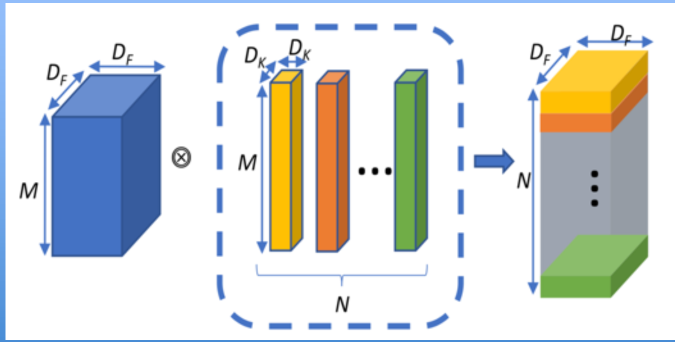
Natural Language Processing



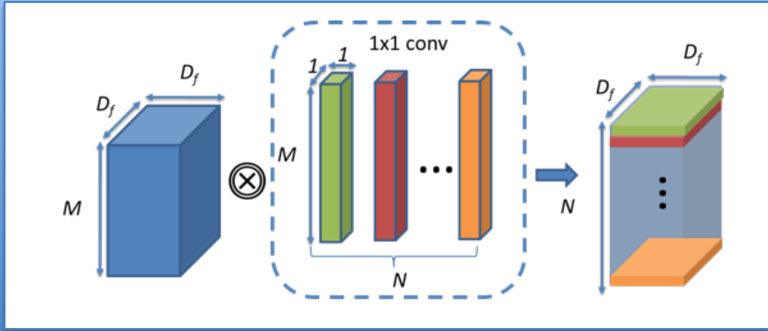
LSTM

Palette of a Computer VisionNeural Net Model Designer (~2020) Orchestation of NN Model Architectures Was the Key Skill

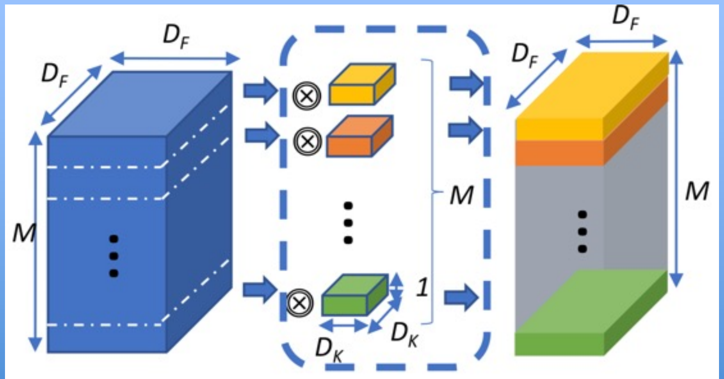
Spatial Convolution e.g. 3x3



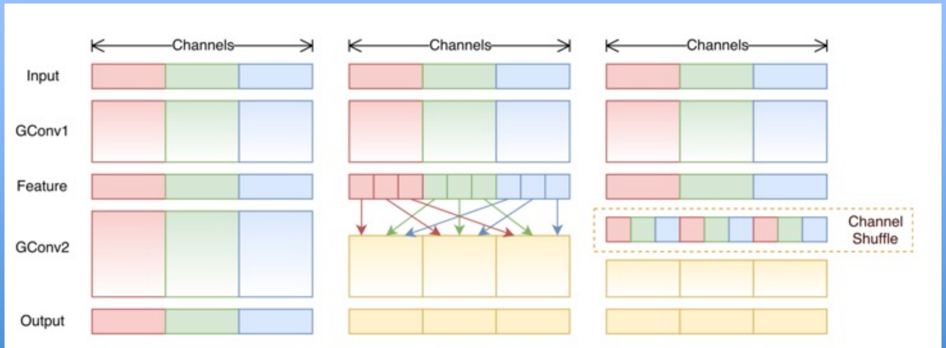
Pointwise Convolution 1x1



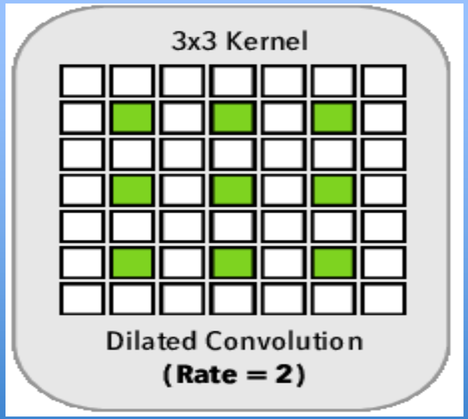
Depthwise Convolution



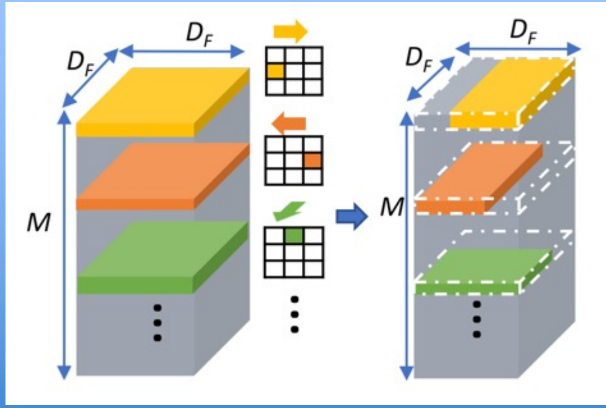
Channel Shuffle



Dilated Convolution



Shift

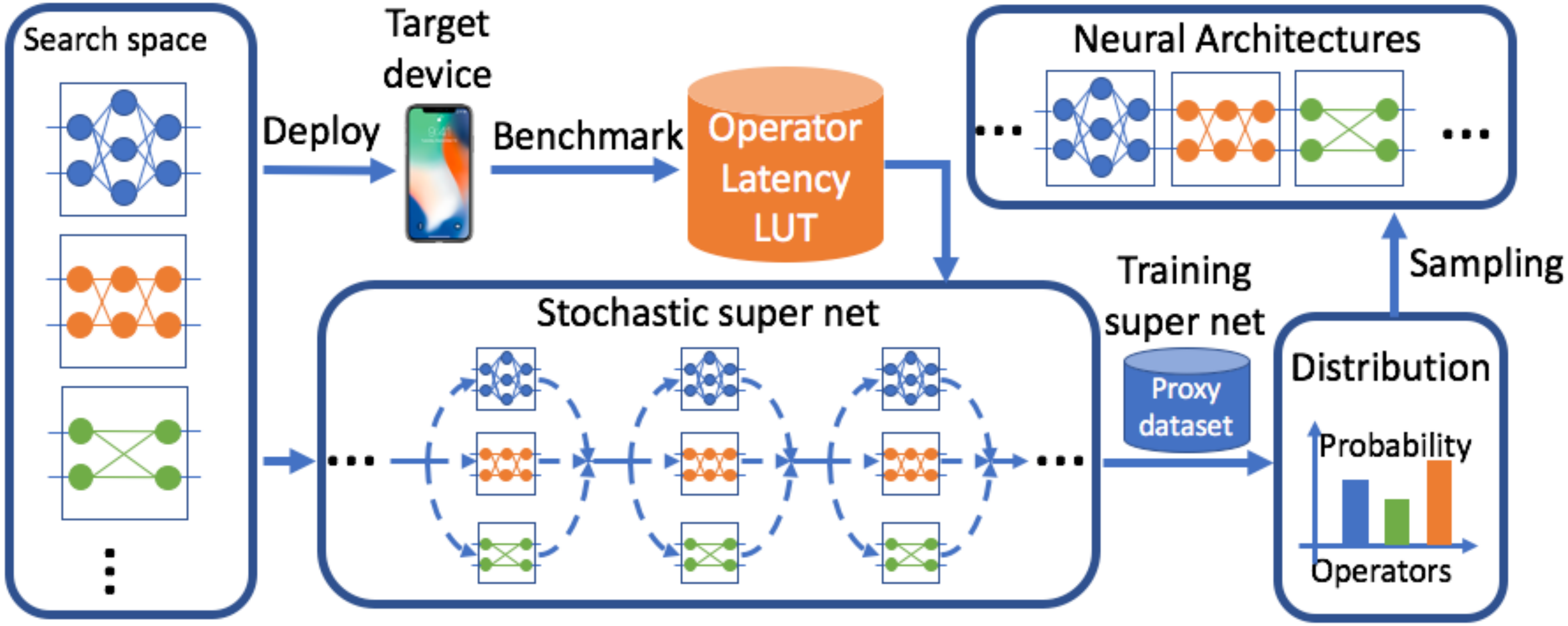


Design Space was so Large and Diverse that Searching It Required Significant Automation: Neural Architecture Search

CNN parameters to be explored:

- # Layers
- Type of convolution: spatial, group, dilated, shift
- Expansion factor
- ....

FBNet Family



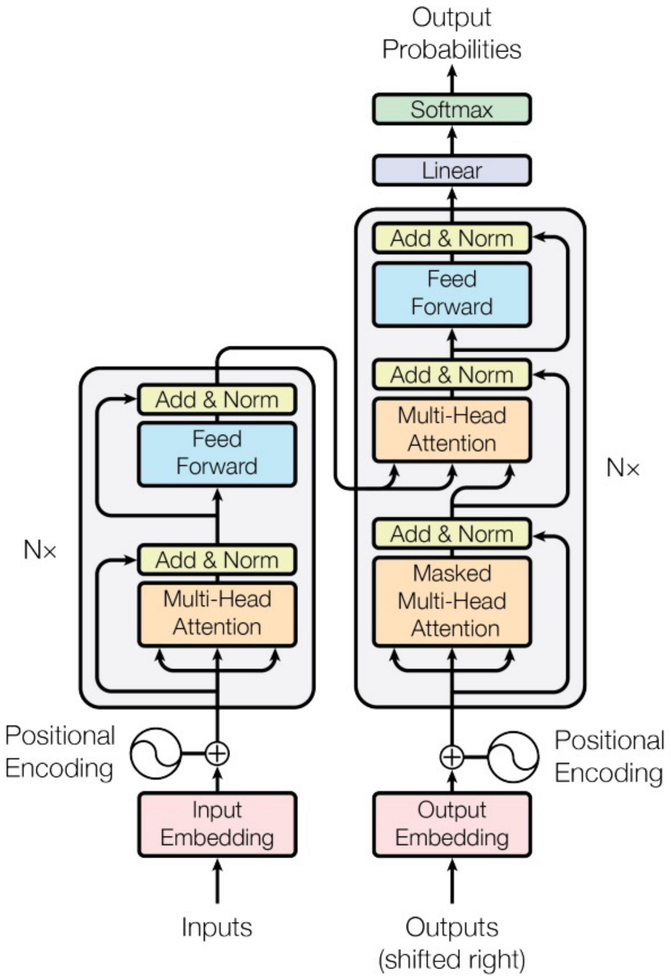
H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055, 2018.

Xie, S., Zheng, H., Liu, C., and Lin, L., SNAS: stochastic neural architecture search, ICLR, 2018.

Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y. and Keutzer, K., 2019. FBnet: Hardware-aware efficient convnet design via differentiable neural architecture search.

Guo, Z., Zhang, X., Mu, H., Heng, W., Liu, Z., Wei, Y., and Sun, J. Single path one-shot neural architecture search with uniform sampling, ECCV: 544-560, 2020.

ML Era #3.0
Then Came the Transformer



Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *NeurIPS* 30 (2017).

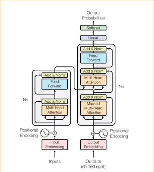
Within 3 Years Transformers Were Everywhere

Image Classification

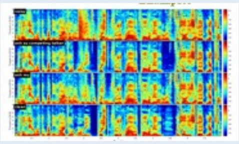
Object Detection

Image Segmentation

Computer Vision



Vision Transformer 2020

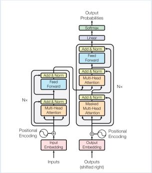


Audio Enhancement

Call-center Sentiment Analysis

Speech Recognition

Audio Analysis And ASR



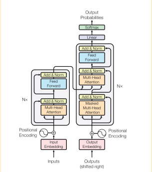
Transformer

Sentiment Analysis

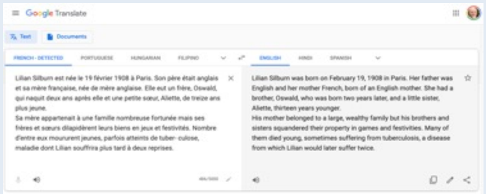
Music Recommendation

Ad Recommendation

Multimedia and Rec Systems



Transformer

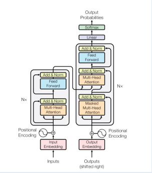


Translation

Question answering

Document Understanding

Natural Language Processing



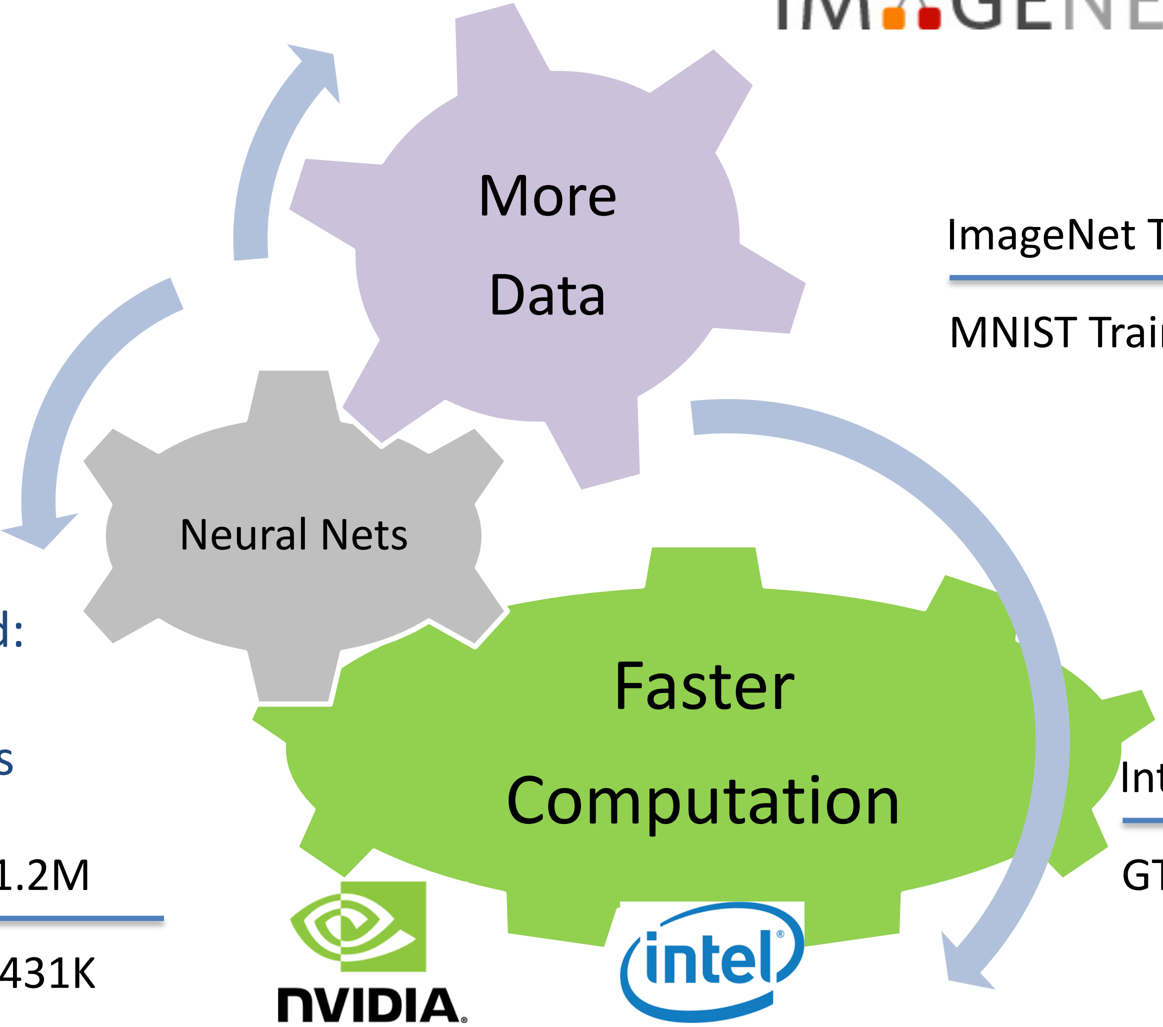
LLMs 2017

Big Surprise #2:
Model Architectures Are Converging

- Given the increasing diversity of applications, it would be natural to expect that the model architectures used in Deep Learning would be becoming diverse as well … but, the opposite is happening
- Broad convergence on transformer-based architectures

Data and Compute Capability Dominate Model Selection
SqueezeNet (2016) (~AlexNet Top-5) vs LeCunn's LeNet (1998)

Training harder on more data seems more impactful than NN model architecture



IMAGENET

ImageNet had 20X more training images than MNIST

ImageNet Training Images	1.2M
MNIST Training Images	60K

Deep Learning Wasn't Really About Deep Nets

- SqueezeNet only had:
- + 2 more layers
- 3x model parameters

SqueezeNet Parameters	1.2M
LeNet Parameters	431K

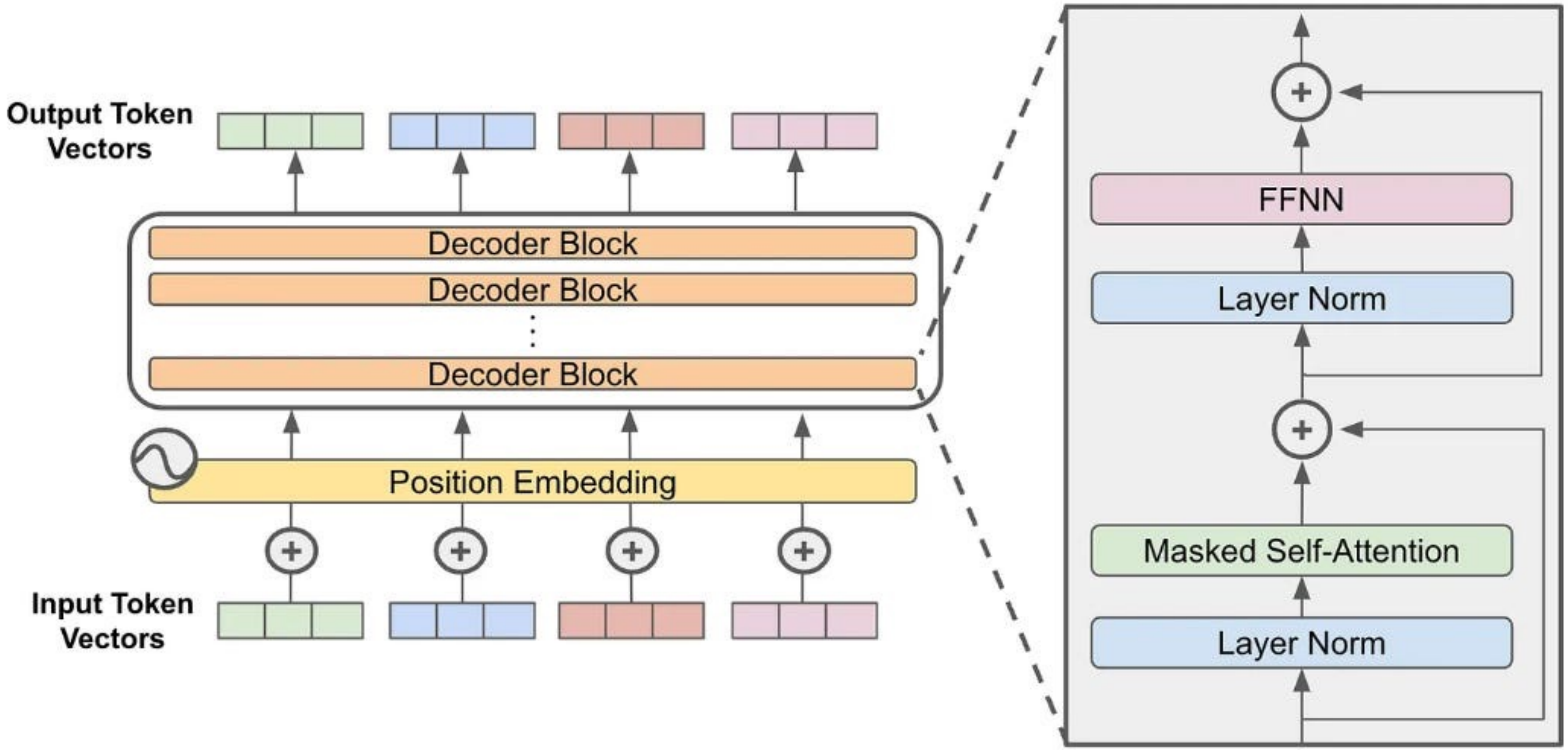
GTX 580 had 7,500 more Flops Than LeCunn's Pentium II

Intel Pentium II (1998)	1.5 TFLops
GTX 580 (2010)	.0002 TFlops

For Reflection #1

- Perhaps ***Neural Nets*** outperform ***traditional machine learning*** *algorithms* because they are better able to put very fine-grain parallelism to use through scaling …
- And, perhaps ***transformers*** outperform ***other Neural Net*** models for the same reason:
  - they are better able to put very fine-grain parallelism to use through scaling than other model architectures …

ML Era 4.0
Transformers orchestrated to create Large Language Models



ML Era 4.0: Large Language Models

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Prompt

Large Language Model(s)
E.g. GPT4

Generated/Predicted Output

KU You

Question/prompt

ChatGPT

Answer.

What Was the Point of All That?
Observe the Patterns

Machine Learning/Deep Learning have rapidly evolved through a number of eras:

- ML Era 1: Orchestration of statistics gave us **Machine Learning**
- ML Era 2: Orchestration of Machine Learning algorithms gave us **Neural Nets**
- ML Era 3: Orchestration of Neural Net model functions/components gave us the **Transformer**
- ML Era 4: Orchestration of Transformers gave us **Large Language Models**

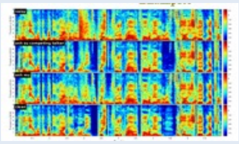
What Has Changed Since 2020?

Image Classification

Object Detection

Image Segmentation

Computer Vision


Audio Enhancement


Call-center Sentiment Analysis


Speech Recognition

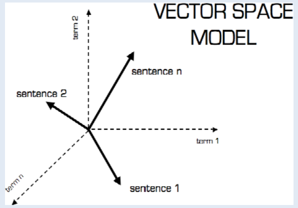
Audio Analysis And ASR


Sentiment Analysis

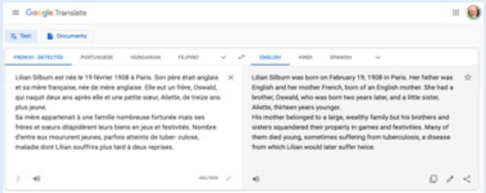

Music Recommendation

Ad Recommendation

Multimedia and Rec Systems



Semantic Similarity


Machine Translation

Named Entity Recognition

Natural Language Processing

Early eras of classical Machine Learning and Deep Learning were more:

- Single modality
- Analytical

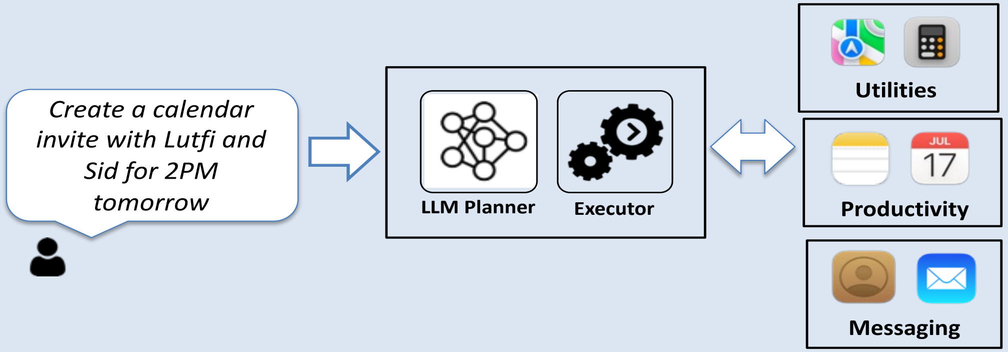
These techniques helped us find what we already knew was there

Examples:

- Objects In images
- Words in speech
- The best ad
- A Named Entity in text

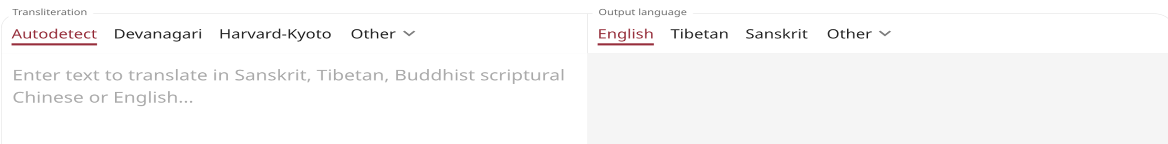
Synthetic , Multi-modal and Surprising!
Let’s Look at the System Challenges

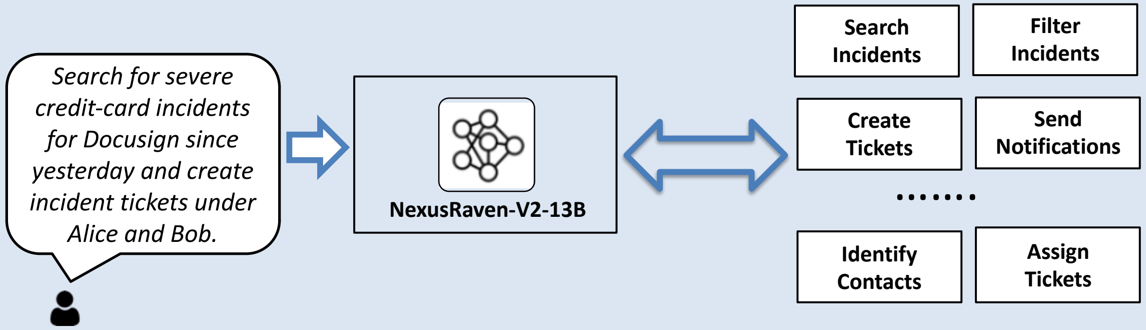
Image/video generation using **diffusion models**



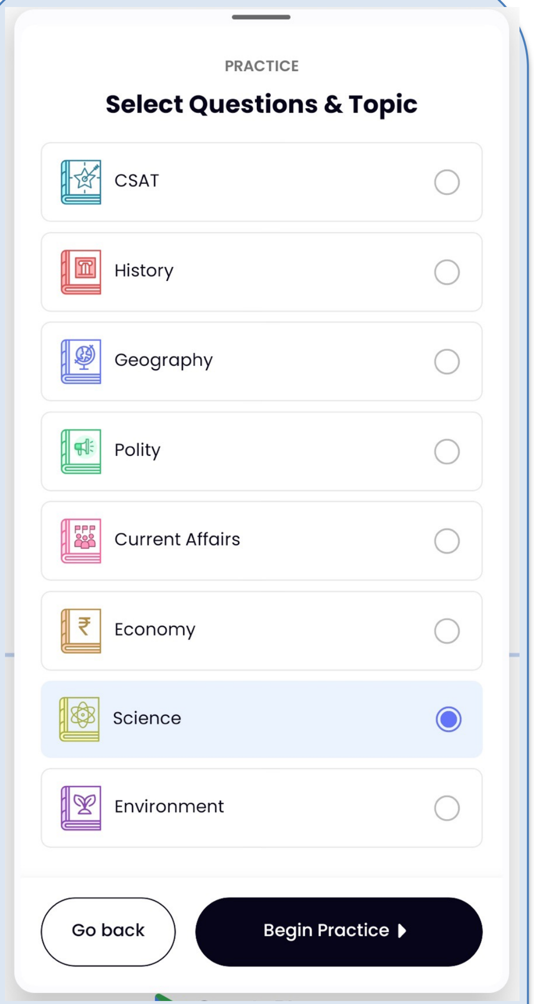
Personal Agents/Co-pilots

Machine Translation





Enterprise Agents/Co-pilots

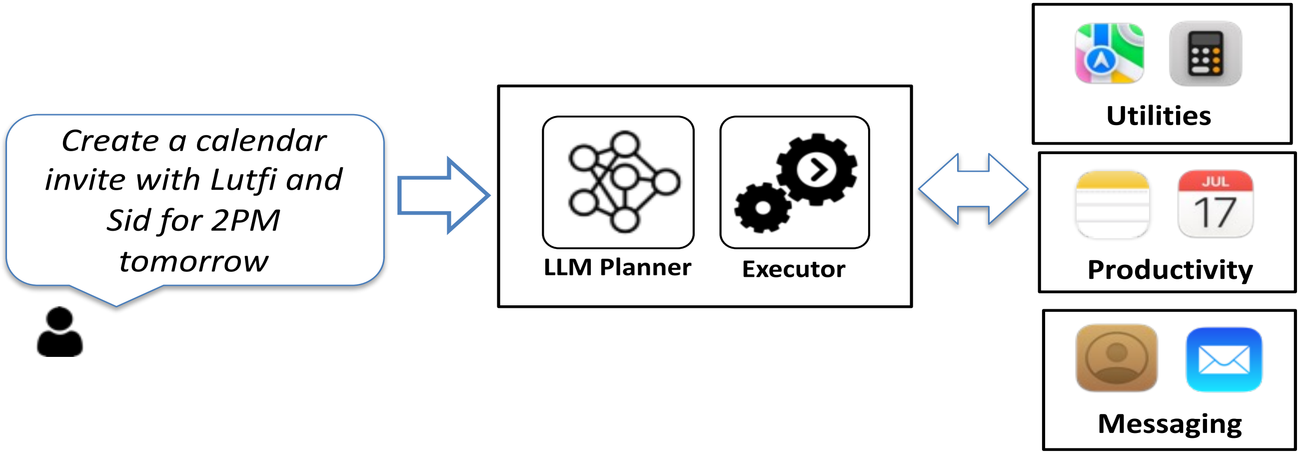


Personal Teacher

Challenges to All Large Language Models Compounded or Not

- Shortly I'll be contrasting these two approaches to applications
- But, as they are both built on LLMs, let's talk about common problems first.

VS



ML 3.0: Large Language Models

ML 4.0: Compound GenAI Systems

Common Characteristic #1: LLMs Have Relatively Similar Model Architectures

- Some architectural differences exist between LLMs:
  - *Different number of layers, number of attention heads, hidden dimension*
  - *Choice of positional embeddings – Eg. RoPE, absolute, ALiBi*
  - *Choice of activation function – Eg. GELU vs SwiGLU*
  - *Multi-head versus multi-query attention (MQA/GQA)*
  - *Mixture-of-experts for the FFN*
- **However, the core Transformer Decoder architecture is the same for nearly all LLMs, and has been relatively stable for the past few years**
  - *N stacked identical blocks, interleaving self-attention and feed-forward network subblocks*

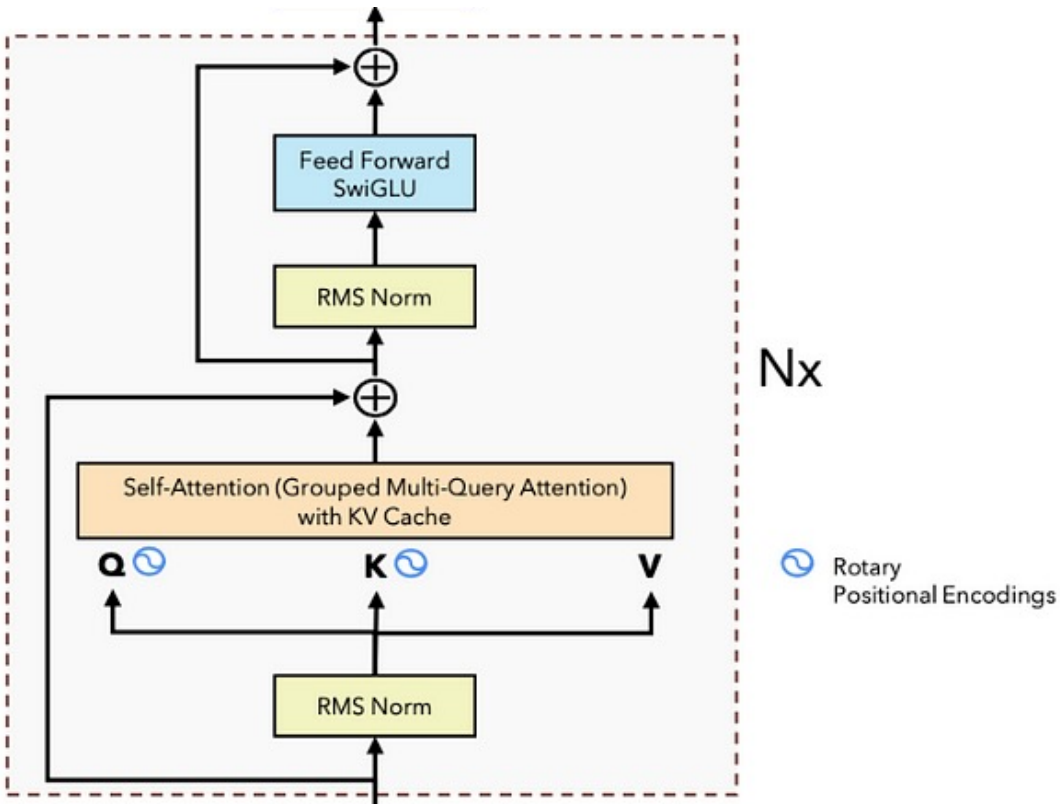
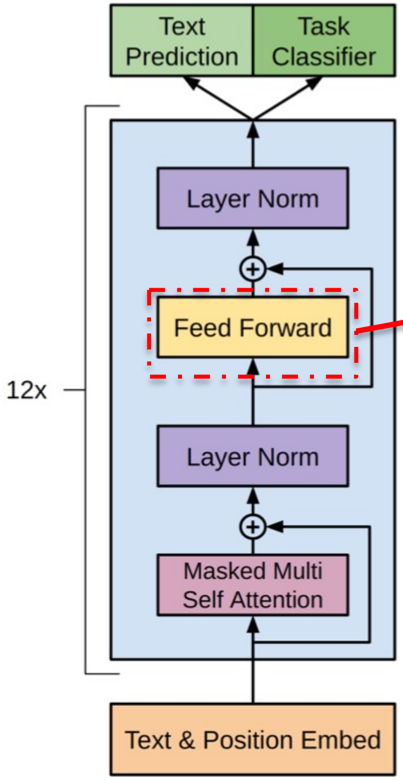
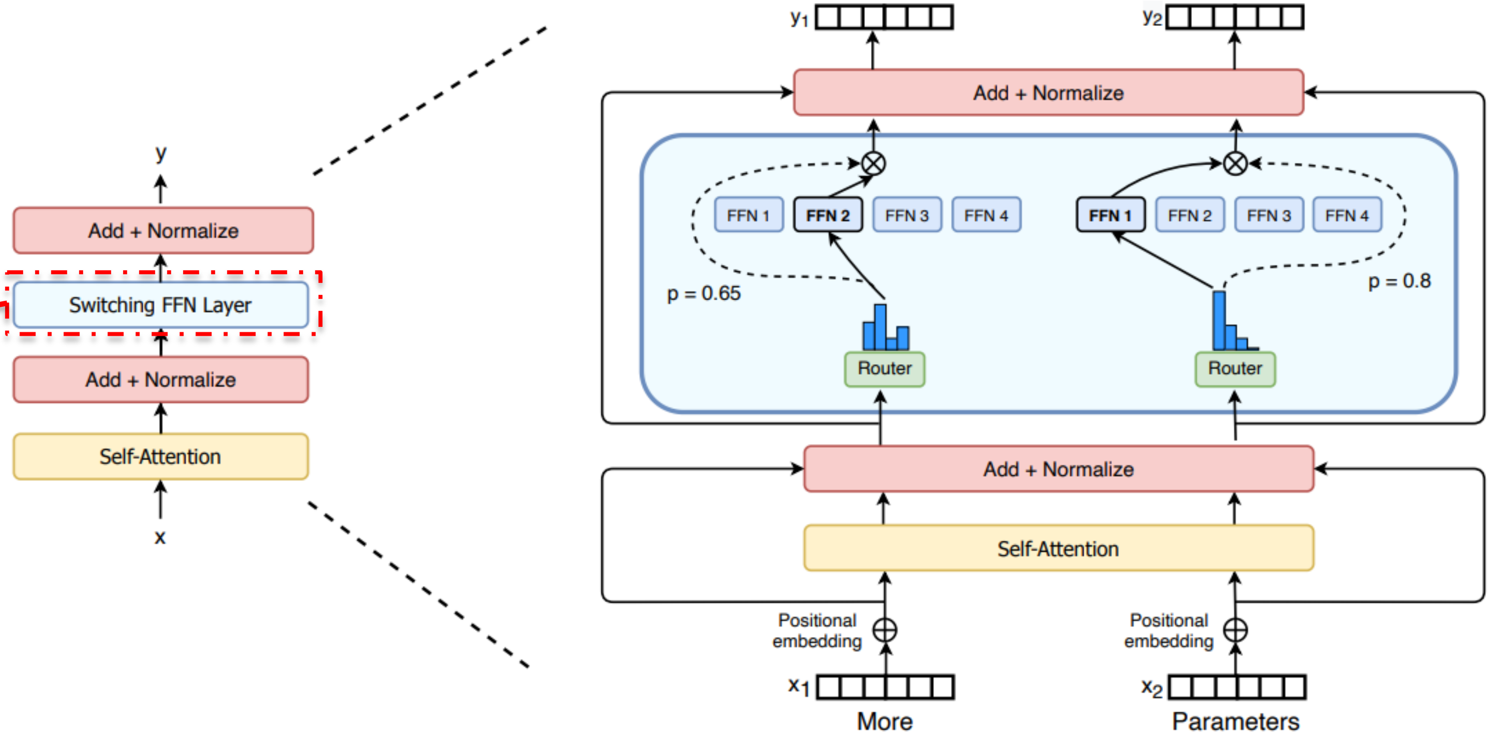


Diagram From: https://plainenglish.io/community/understanding-llama2-kv-cache-grouped-query-attention-rotary-embedding-and-more-9a79bd

Mixture of Experts



[GPT]



[MoE]

[GPT] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.

[MoE] Fedus, W., Zoph, B. and Shazeer, N., 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), pp.1-39. also *arXiv:2101.03961*.

Minor Architectural Variations (By Model Size)

Model	Year	Positional Encoding	Activation	Norm	Hidden Dim	# Heads	Head Dim	# Layers	MQA/GQA	MoE
Mistral (7B)	2023	RoPE	SwiGLU	RMSNorm	4096	32	128	32	Yes	No
Gemma (7B)	2024	RoPE	GeGLU	RMSNorm	3072	16	256	28	No	No
LLaMA (65B)	2023	RoPE	SwiGLU	RMSNorm	8192	64	128	80	No	No
LLaMA-3 (70B)	2024	RoPE	SwiGLU	RMSNorm	8192	64	128	80	Yes	No
Command R+ (104B)	2024	RoPE	SwiGLU	LayerNorm	12288	96	128	64	Yes	No
DBRX (132B)	2024	RoPE	SwiGLU	LayerNorm	6144	48	128	40	Yes	Yes
GPT-3 (175B)	2020	Absolute	GELU	LayerNorm	12288	96	128	96	No	No
Falcon (180B)	2023	RoPE	GELU	LayerNorm	14848	64	64	80	Yes	No
PaLM (540B)	2022	RoPE	SwiGLU	LayerNorm	18438	48	256	118	Yes	No

Minor Architectural Variations (By Year)

Model	Year	Positional Encoding	Activation	Norm	Hidden Dim	# Heads	Head Dim	# Layers	MQA/GQA	MoE
GPT-3 (175B)	2020	Absolute	GELU	LayerNorm	12288	96	128	96	No	No
PaLM (540B)	2022	RoPE	SwiGLU	LayerNorm	18438	48	256	118	Yes	No
Mistral (7B)	2023	RoPE	SwiGLU	RMSNorm	4096	32	128	32	Yes	No
LLaMA (65B)	2023	RoPE	SwiGLU	RMSNorm	8192	64	128	80	No	No
Falcon (180B)	2023	RoPE	GELU	LayerNorm	14848	64	64	80	Yes	No
Gemma (7B)	2024	RoPE	GeGLU	RMSNorm	3072	16	256	28	No	No
DBRX (132B)	2024	RoPE	SwiGLU	LayerNorm	6144	48	128	40	Yes	Yes
LLaMA-3 (70B)	2024	RoPE	SwiGLU	RMSNorm	8192	64	128	80	Yes	No
Command R (104B)	2024	RoPE	SwiGLU	LayerNorm	12288	96	128	64	Yes	No

Karpathy on Data

Andrej Karpathy
@karpathy

We see more significant improvements from training data distribution search (data splits + oversampling factor ratios) than neural architecture search. The latter is so overrated :)

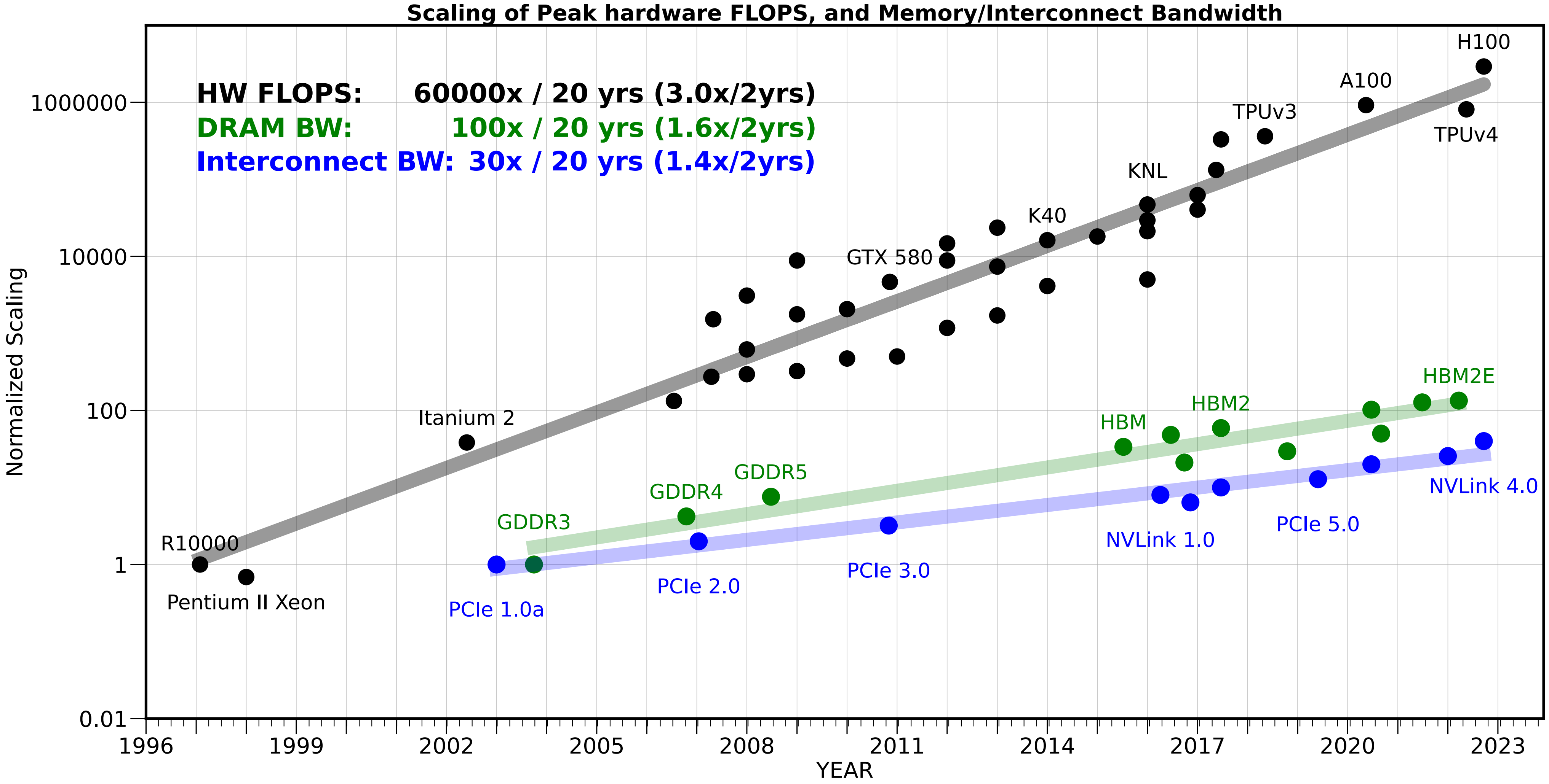
1:03 PM · Sep 20, 2019

The Bigger Difference in Models is the Training Data used, and How it is Used

Model	Year	Positional Encoding	Activation	Norm	Hidden Dim	# Heads	Head Dim	# Layers	MQA/ GQA	Training Data Used In Tokens
Mistral (7B)	2023	RoPE	SwiGLU	RMSNorm	4096	32	128	32	Yes	Unknown, but <8T Tokens speculated
Gemma (7B)	2024	RoPE	GeGLU	RMSNorm	3072	16	256	28	No	6 T tokens multilingual
LLaMA (65B)	2023	RoPE	SwiGLU	RMSNorm	8192	64	128	80	No	1.4 Trillion tokens CCNET (76%), C4 (15%), GitHub (4.5%)
LLaMA-3 (70B)	2024	RoPE	SwiGLU	RMSNorm	8192	64	128	80	Yes	15T tokens (5% multilingual)
Command R+ (104B)	2024	RoPE	SwiGLU	LayerNorm	12288	96	128	64	Yes	4T (speculative)
DBRX (132B)	2024	RoPE	SwiGLU	LayerNorm	6144	48	128	40	Yes	12T "carefully curated"
GPT-3 (175B)	2020	Absolute	GELU	LayerNorm	12288	96	128	96	No	300B Tokens
Falcon (180B)	2023	RoPE	GELU	LayerNorm	14848	64	64	80	Yes	3.5T Tokens
PaLM (540B)	2022	RoPE	SwiGLU	LayerNorm	18438	48	256	118	Yes	780B Tokens

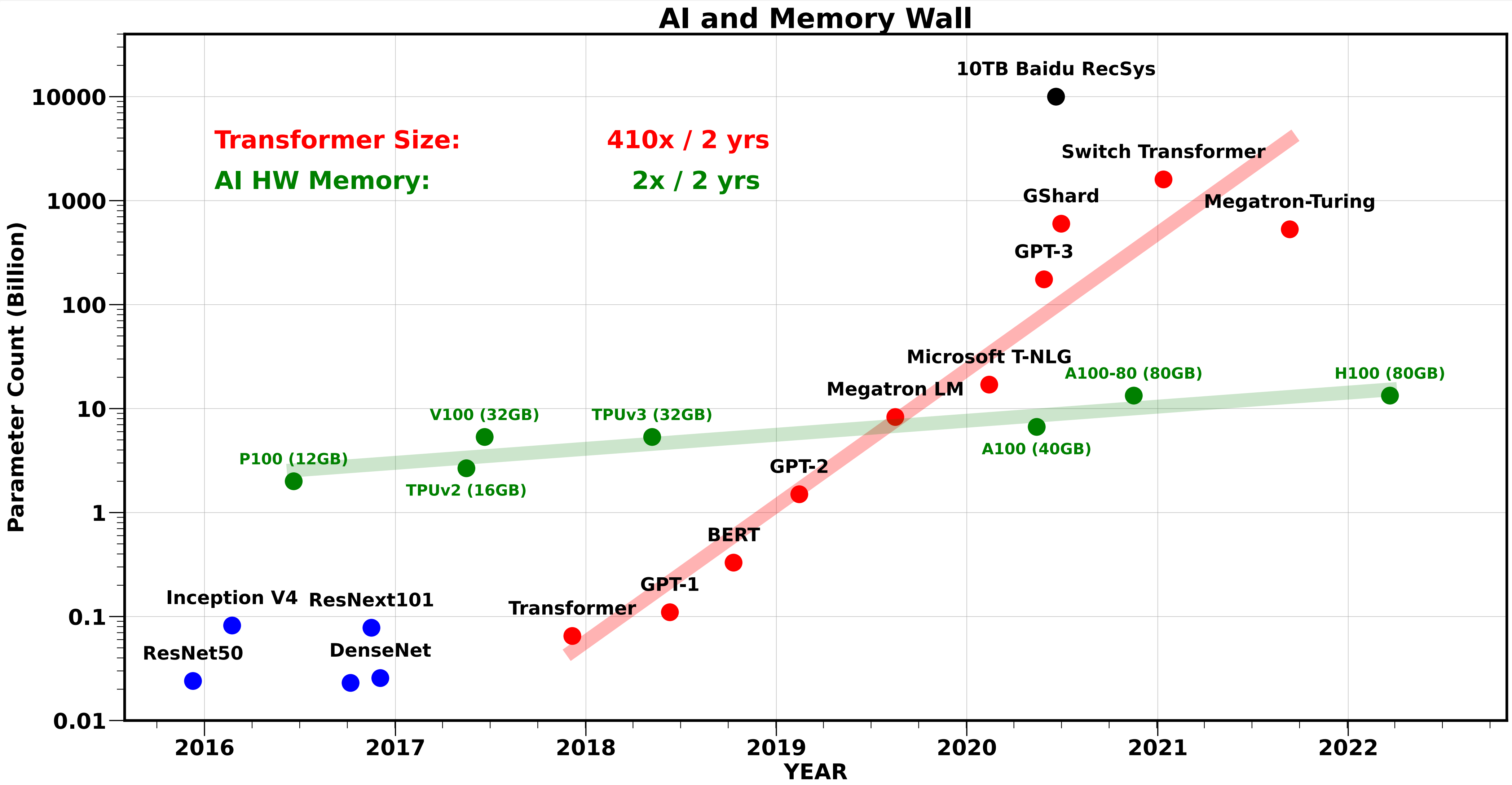
Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A. and Hennigan, T., 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Common Characteristic 2a, Facing the Memory Wall: Divergence Between Computation and Communication



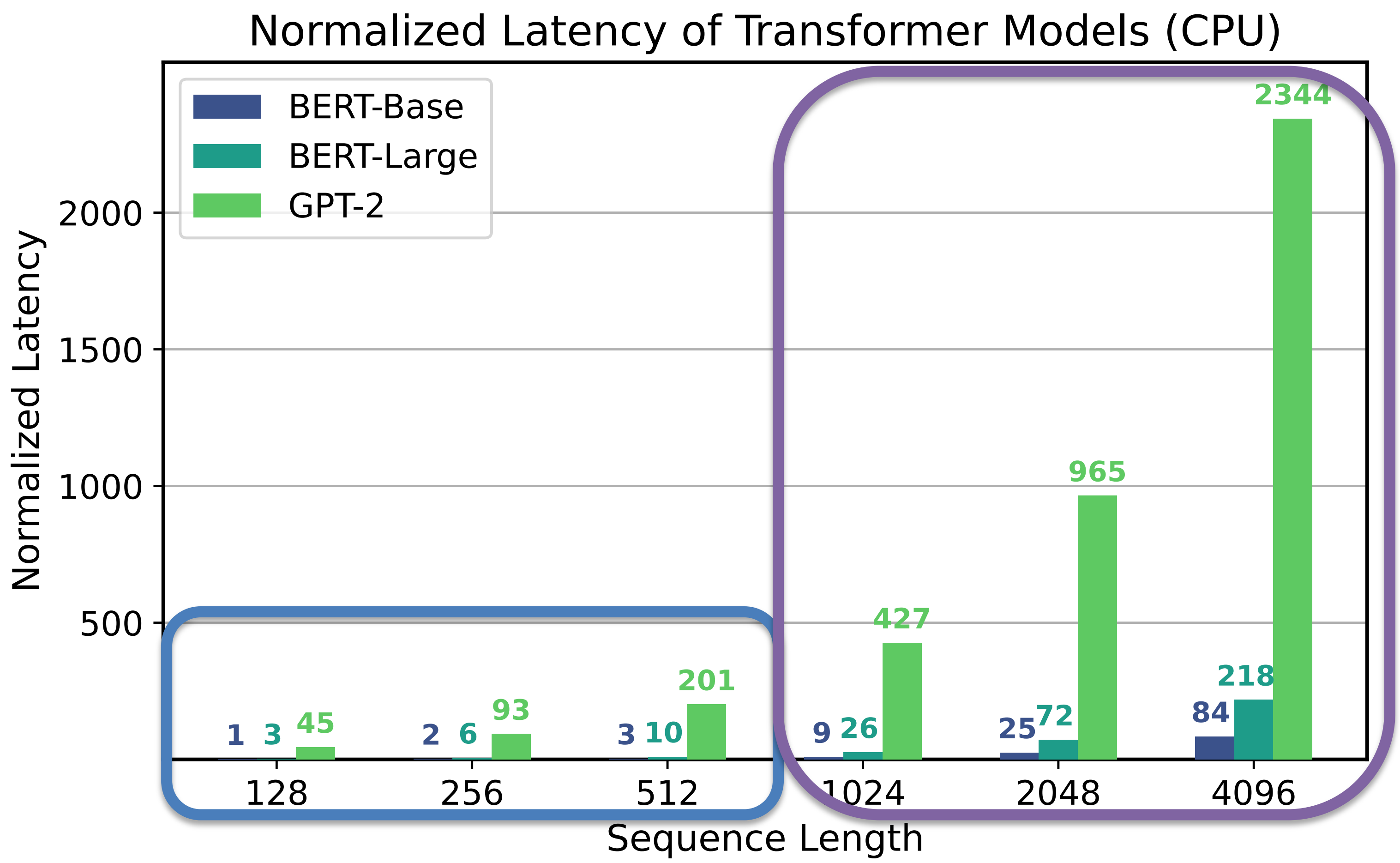
Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W. Mahoney, Kurt Keutzer, [AI and Memory Wall](), IEEE Micro, 2024.

Common Characteristic 2b:
Model Size of LLMs is Exacerbating the Memory Wall



Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W. Mahoney, Kurt Keutzer, AI and Memory Wall, IEEE Micro, 2024.

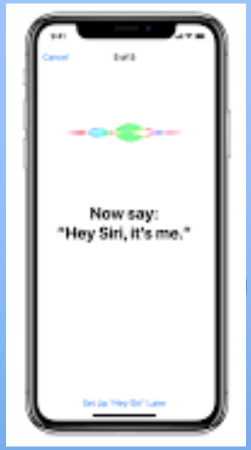
Memory Wall and Input Sequences



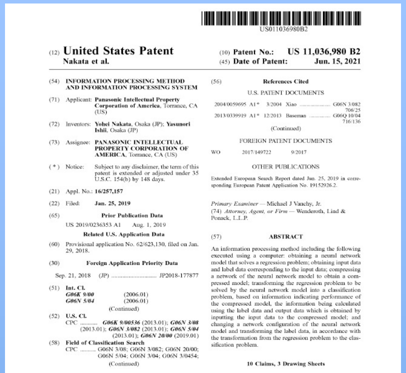
Kim, S., Hooper, C., Wattanawong, T., Kang, M., Yan, R., Genc, H., ... & Gholami, A. (2023). Full stack optimization of transformer inference: a survey. ASSYST Workshop, **ISCA 2023**.

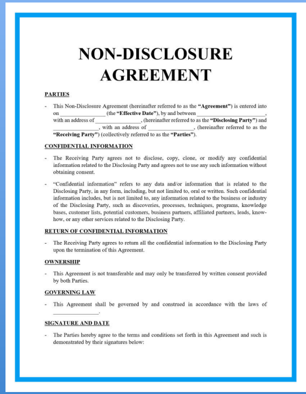
Differences in Sequence Length

- One key distinguishing feature is the typical size of the input sequence length
- Edge applications such as AR/VR glasses, in-car NLP, as well as consumer applications such as Tweets and FB posts may be very short
- B-2-B applications involving financial or legal documents, or results of RAG may be very long

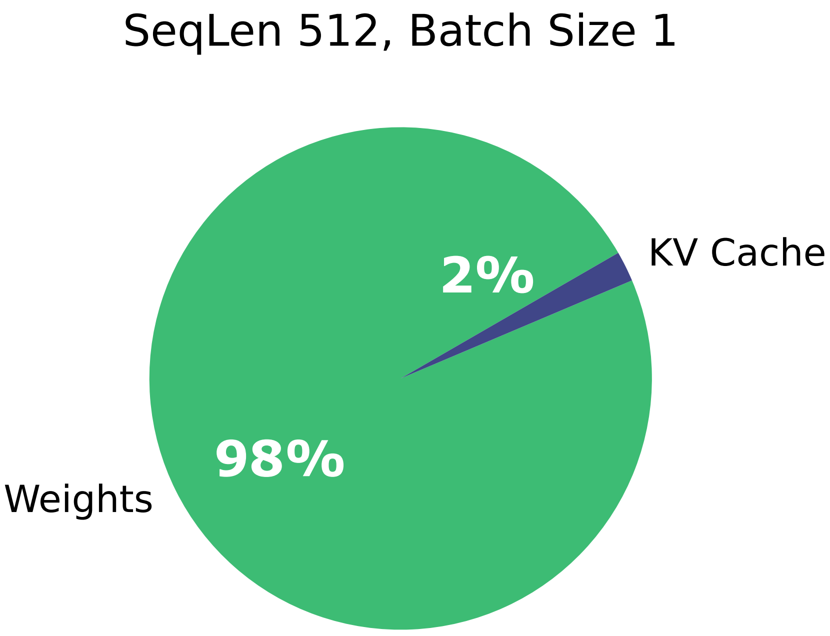


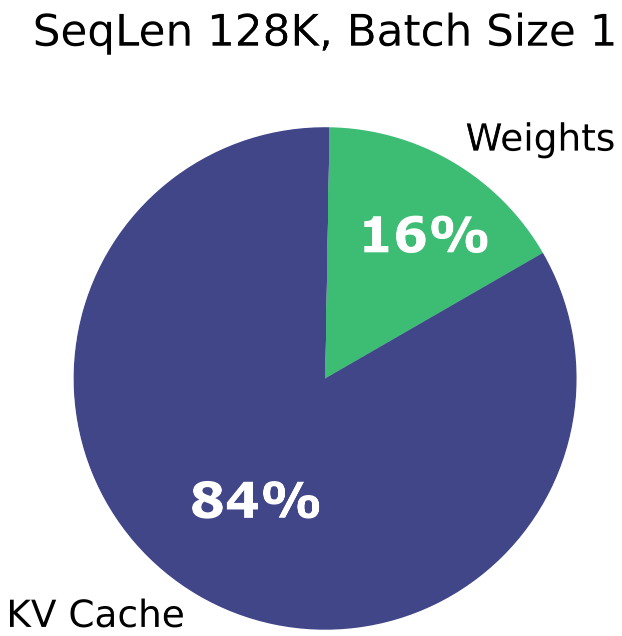

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
FORM 10-K
Apple Inc.


United States Patent
Nakata et al.
Patent No.: US 11,036,980 B2
Date of Patent: Jun. 15, 2021


NON-DISCLOSURE AGREEMENT

Memory Consumption for Long Sequence Lengths





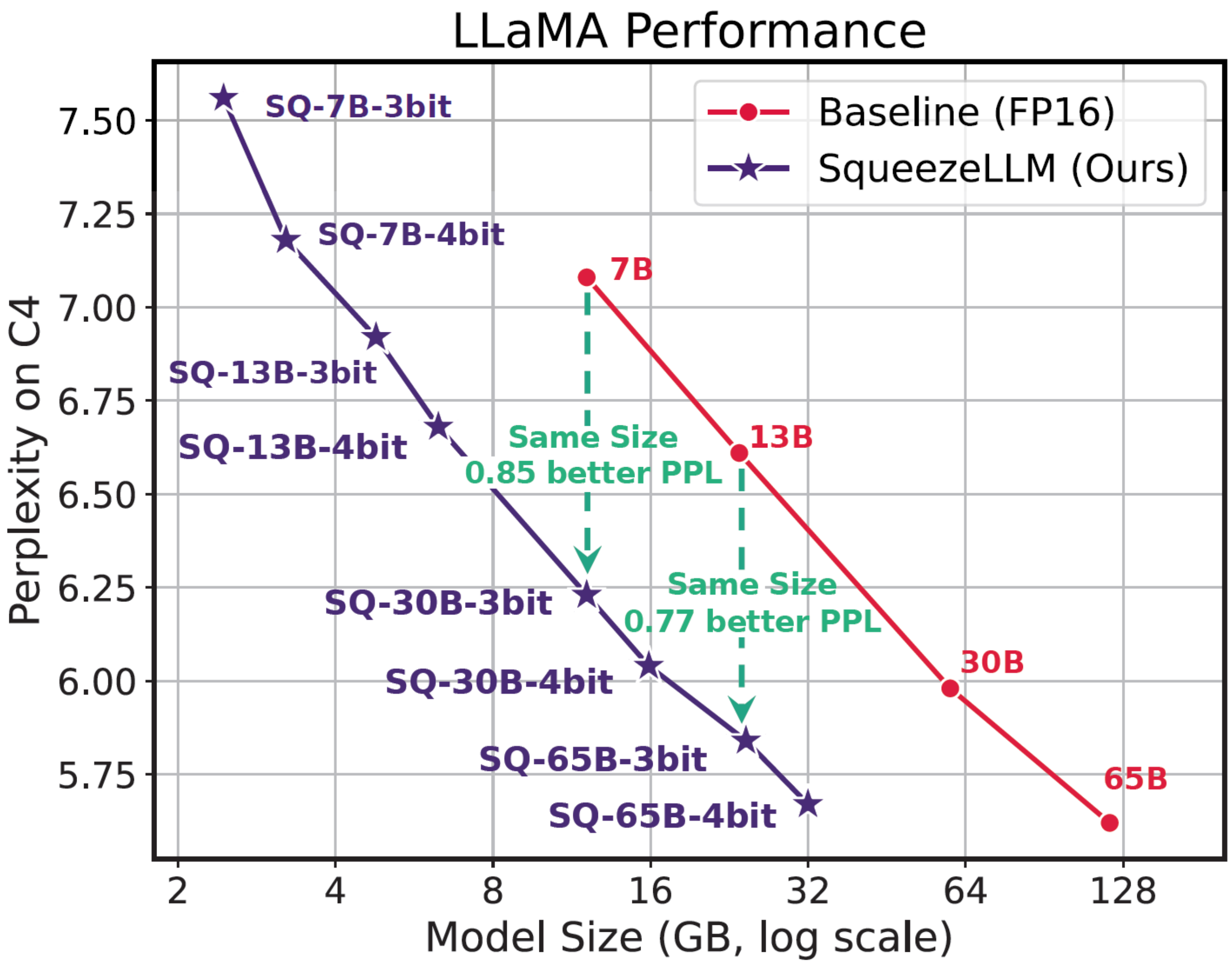
Short sequence length
Weights are the bottleneck

Long Sequence Lengths
KV Cache is the bottleneck

How Model Optimization Helps:
More Performant Model at the Same Comp Cost of a Smaller Model

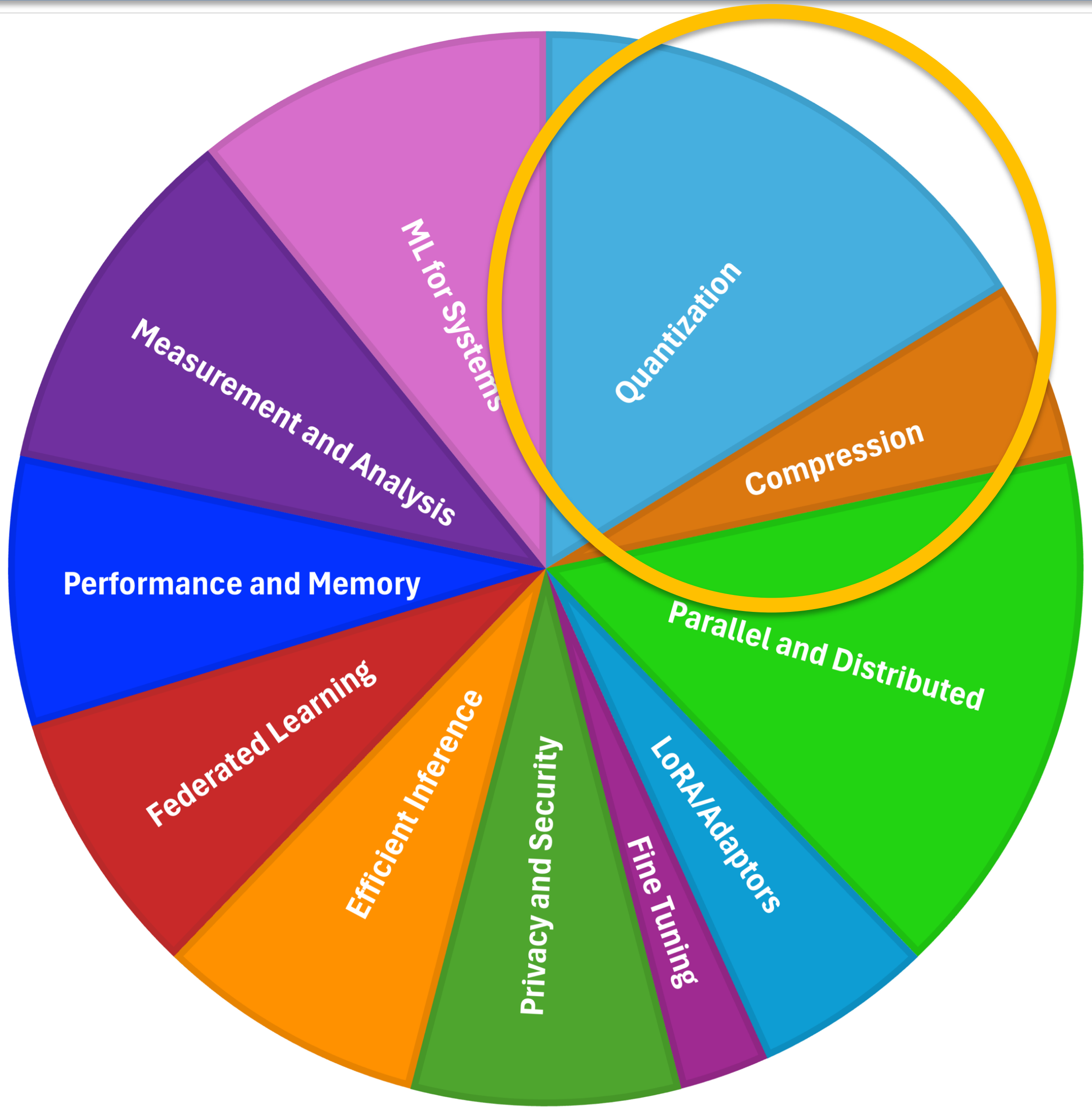
Perplexity

Better



Kim*, S., Hooper*, C., Gholami*, A., Dong, Z., Li, X., Shen, S., Mahoney, M.W. and Keutzer, K. SqueezeLLM: Dense-and-Sparse Quantization. *arXiv:2306.07629. ICML 2024 (to appear)*

Quantization and Compression Well Represented in MLSys 2024



Quantization and Compression 1

- AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration
- QMoE: Sub-1-Bit Compression of Trillion Parameter Models
- Atom: Low-Bit Quantization for Efficient and Accurate LLM Serving

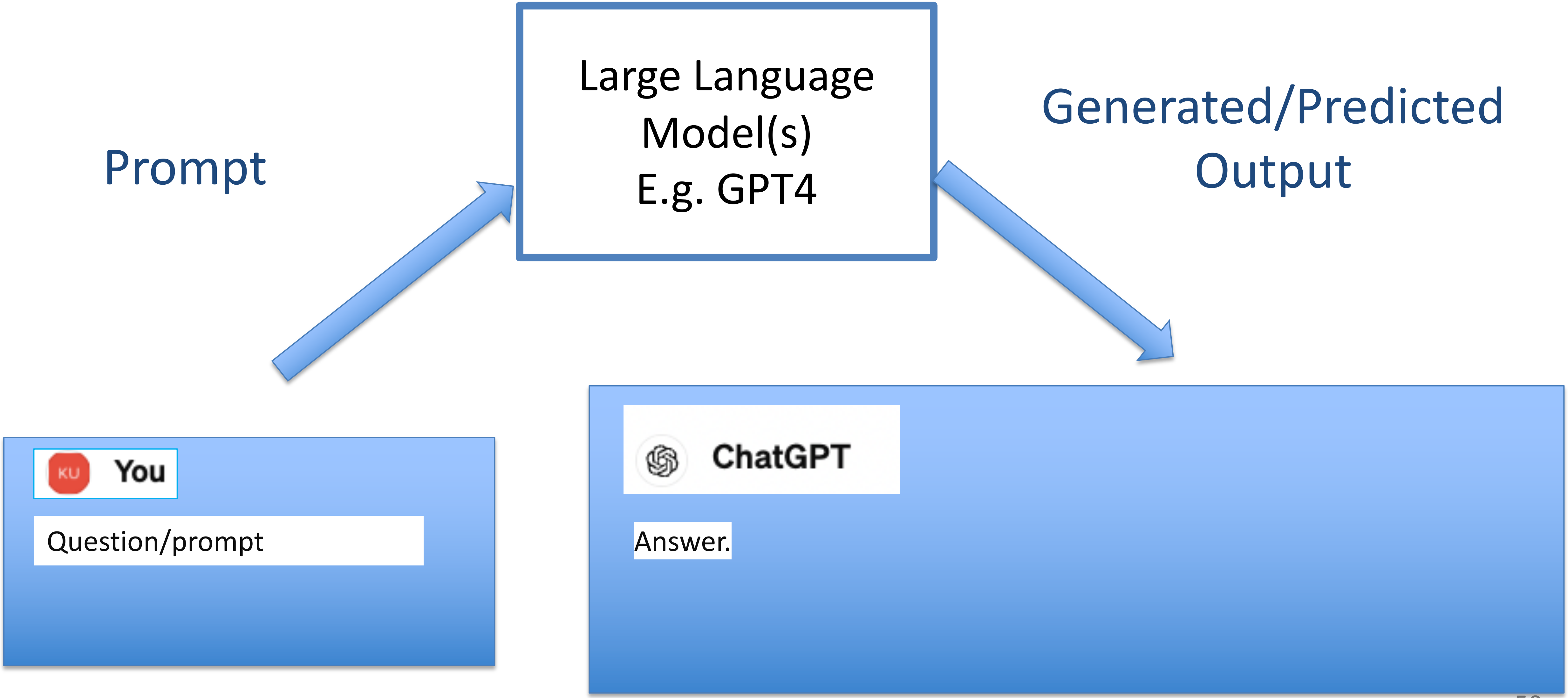
Quantization and Compression 2

▼

Posters 1:30-3:00

- JIT-Q: Just-in-time Quantization with Processing-In-Memory for Efficient ML Training
- Torch2Chip: An End-to-end Customizable Deep Neural Network Compression and Deployment Toolkit for Prototype Hardware Accelerator Design
- Schrodinger's FP Training Neural Networks with Dynamic Floating-Point Containers
- Efficient Post-training Quantization with FP8 Formats

Why Aren't Monolithic LLMs the Final Solution?



ChatGPT

Prompt

ChatGPT

Generated/Predicted Output

You

What is the most important machine learning systems conference?

ChatGPT

ChatGPT
Known to be Prone to Hallucination!

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

ChatGPT

Prompt

Generated/Predicted Output

You

What is the most important machine learning systems conference?

ChatGPT

The most important machine learning conference is generally considered to be the Conference on Neural Information Processing Systems (NeurIPS).

Perplexity.ai

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Prompt

perplexity

Generated/Predicted Output

what is the leading conference on machine learning systems?

Perplexity.ai
Ah, There's the Right Answer!

perplexity

Prompt

what is the leading conference on machine learning systems?

Generated/Predicted Output

Based on the search results provided, the leading conference on machine learning systems appears to be the Conference on Machine Learning and Systems (MLSys): The MLSys conference is described as targeting "research at the intersection of machine learning and systems" and aims to "elicit new connections amongst these fields, …

Perplexity.ai

perplexity

Prompt

Generated/Predicted Output

what is the most important machine learning systems conference?

Perplexity.ai
Oops, again.

perplexity

Prompt

Generated/Predicted Output

what is the most important machine learning systems conference?

The International Conference on Machine Learning (ICML) is widely regarded as the premier and most important conference in the field of machine learning systems.

Not Really so Bad, but …

- I'm confident I work with high school students who could do a superior conference information system using Compound GenAI Systems!
- Too much attention on GenAI has been placed on consumer-facing question-and-answer applications
- I'm here today because I feel the exciting way forward is in agents and co-piloted systems

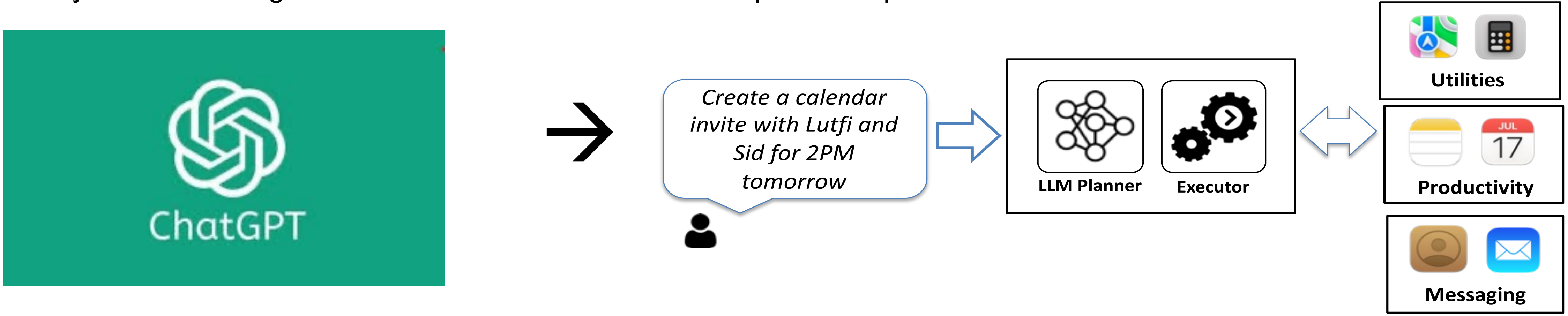
ML 4.0: Large Language Models

From ML 4.0 to ML 5.0
Compound GenAI Systems

"We define a Compound AI System as a system that tackles AI tasks using multiple interacting components, including multiple calls to models, retrievers, or external tools. "

– Zaharia, *et. al.* BAIR Blog 2/18/2024

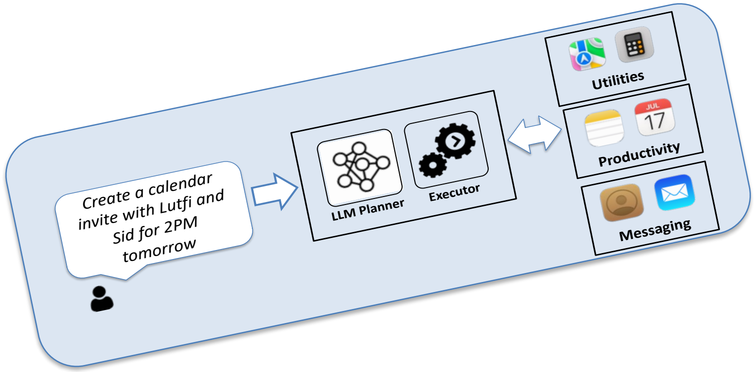
- In the future, almost every business-to-consumer and business-to-business activity will be co-piloted or intermediated by an agent build as a Compound GenAI System
- Compound GenAI Systems offer:
  - Superior performance for targeted applications
  - Privacy
  - Low-latency
  - Cost-effectiveness
- Compound GenAI systems not only exploit that natural language processing capability of LLMs at the front end, but they use the intelligence of LLMs at the back end to request and process additional information

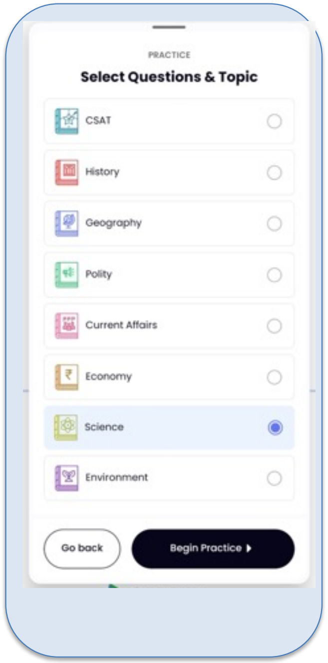


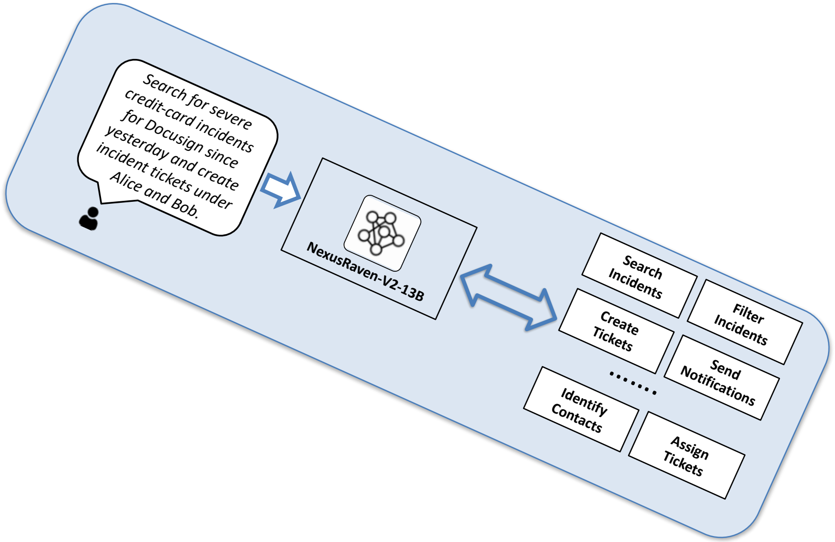
ML 3.0: Large Language Models

ML 4.0: Compound GenAI Systems

Characterizing Agents/Co-Pilots
The Poster Child of Compound GenAI Systems







A user-prompt driven GenAI System that:

- Uses one or more small to mid-size (<= 70B parameter) opensource LLMs (e.g. Llama X)
- Achieves superior results over monolithic proprietary LLMs (e.g. ChatGPT) through fine-tuning on proprietary data, Parameter Efficient Fine-tuning (PEFT), and/or prompting
- Accesses up to date, task-relevant information using:
  - Retrieval Augmented Generation (RAG)
  - Invocation of a variety of task-relevant tools
- Synthesizes results from all sources, and returns the result to the user

For the Young Professionals
If You're Really Interested in Efficiency

I can reduce latency and power 20% automatically with logic optimization

Kurt Keutzer

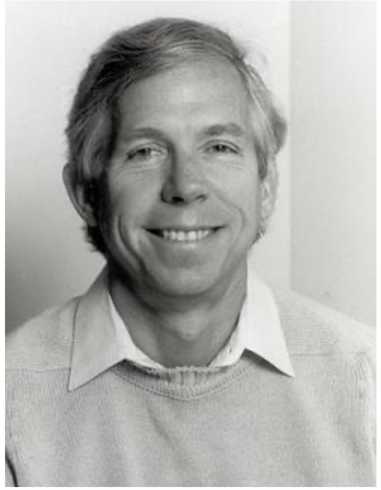
- International Workshop on Low Power Design in the early 90's

If You’re Really Interested in Efficiency
Understand the Applications (not just the benchmarks)

I can reduce latency and power 20% with logic optimization

I can reduce latency and power **2000X** at the application level

Kurt Keutzer



Prof. Bob Brodersen

- International Workshop on Low Power Design in the early 90’s

- It took me 10 years to finally learn that Bob was right, (then 18 years into my career).
- I wish I had been looking harder at applications from the beginning.

We Will Look at a Number of Compound GenAI Applications Today

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Image/video generation using **diffusion models**

Create a calendar invite with Lutfi and Sid for 2PM tomorrow

LLM Planner

Executor

Utilities

JUL 17

Productivity

Messaging

Personal Agents/Co-pilots

Machine Translation

Transliteration

Autodetect Devanagari Harvard-Kyoto Other

Output language

English Tibetan Sanskrit Other

Enter text to translate in Sanskrit, Tibetan, Buddhist scriptural Chinese or English...

Search for severe credit-card incidents for Docusign since yesterday and create incident tickets under Alice and Bob.

NexusRaven-V2-13B

Search Incidents

Filter Incidents

Create Tickets

Send Notifications

....

Identify Contacts

Assign Tickets

Enterprise Agents/Co-pilots

PRACTICE

Select Questions & Topic

- CSAT
- History
- Geography
- Polity
- Current Affairs
- Economy
- Science (selected)
- Environment

Go back

Begin Practice

Personal Teacher

To Beat ChatGpt
All GenAI Apps Must Show Sophisticated Use of LLMs:

Neuro-symbolic Programming

Full Finetuning

Partial tuning/Adapters/LORA

Prompting

Out of the box LL Model

Intelligent Model Selection

As Well as Sophisticated Use of Retrieval Augmented Generation (RAG) and Tools

Neuro-symbolic Programming

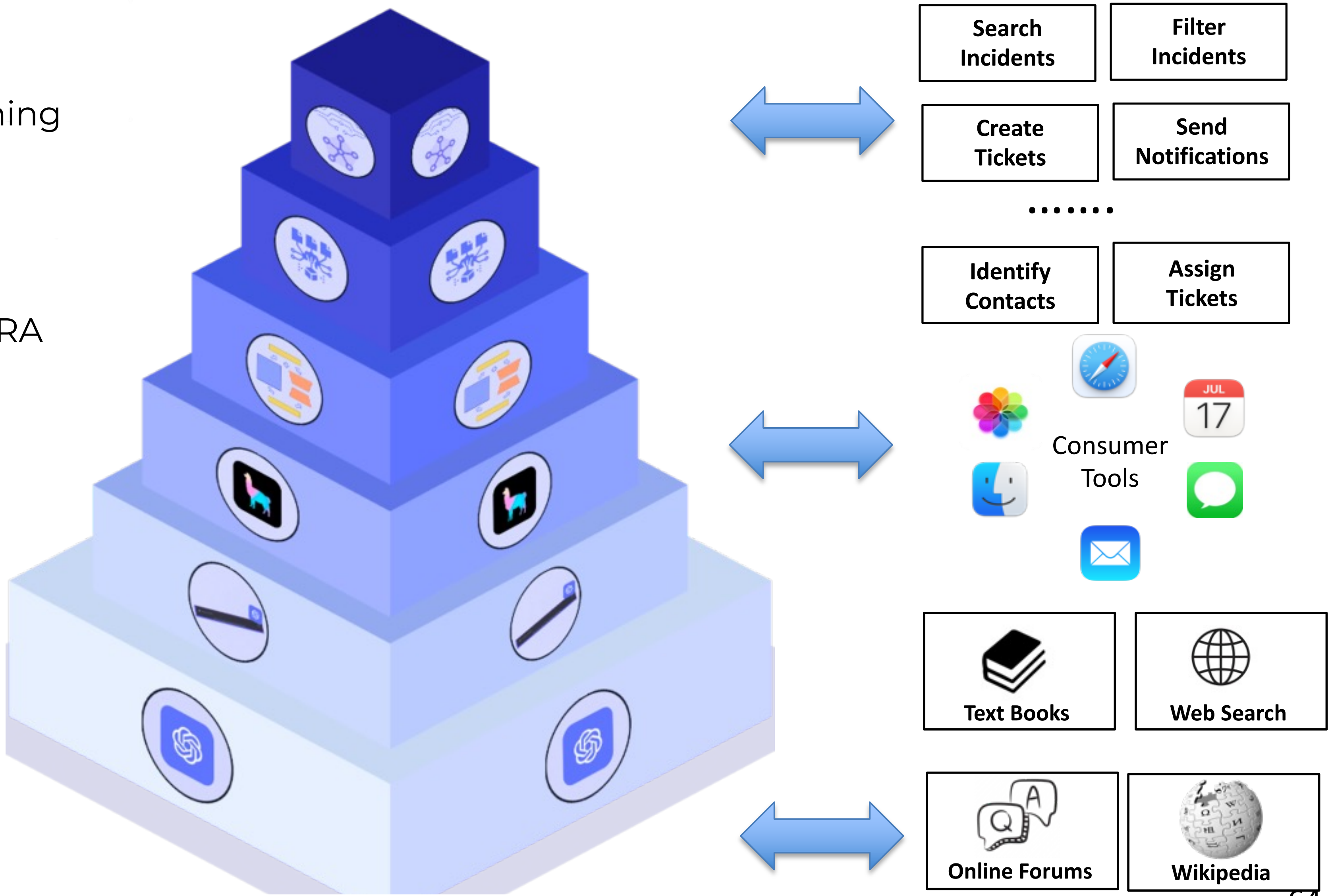
Full Finetuning

Partial tuning/Adapters/LORA

Prompting

Out of the box LL Model

Intelligent Model Selection



And … If You Can Show Do This
A Small Team Can Build Power Software/Agents

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Neuro-symbolic Programming

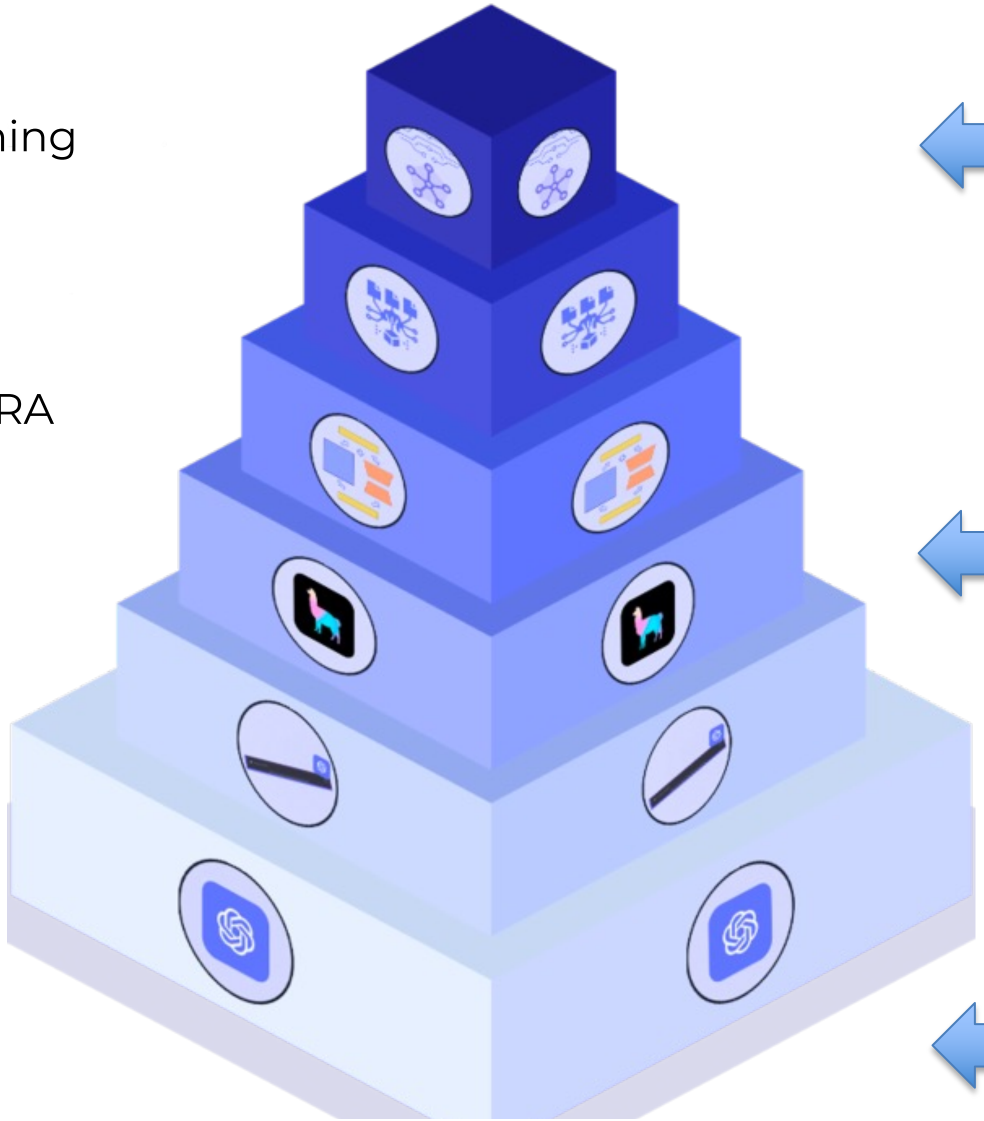
Full Finetuning

Partial tuning/Adapters/LORA

Prompting

Out of the box LL Model

Intelligent Model Selection



Search Incidents

Filter Incidents

Create Tickets

Send Notifications

·······

Identify Contacts

Assign Tickets

Consumer Tools

Text Books

Web Search

Online Forums

Wikipedia

LLM-centric GenAI Systems
Let's Look at them Individually

Image/video generation using **diffusion models**

Create a calendar invite with Lutfi and Sid for 2PM tomorrow

LLM Planner | Executor

Utilities

Productivity

Messaging

Personal Agents/Co-pilots

Machine Translation

Transliteration: Autodetect | Devanagari | Harvard-Kyoto | Other

Output language: English | Tibetan | Sanskrit | Other

Enter text to translate in Sanskrit, Tibetan, Buddhist scriptural Chinese or English...

Search for severe credit-card incidents for Docusign since yesterday and create incident tickets under Alice and Bob.

NexusRaven-V2-13B

Search Incidents | Filter Incidents

Create Tickets | Send Notifications

....

Identify Contacts | Assign Tickets

Enterprise Agents/Co-pilots

PRACTICE

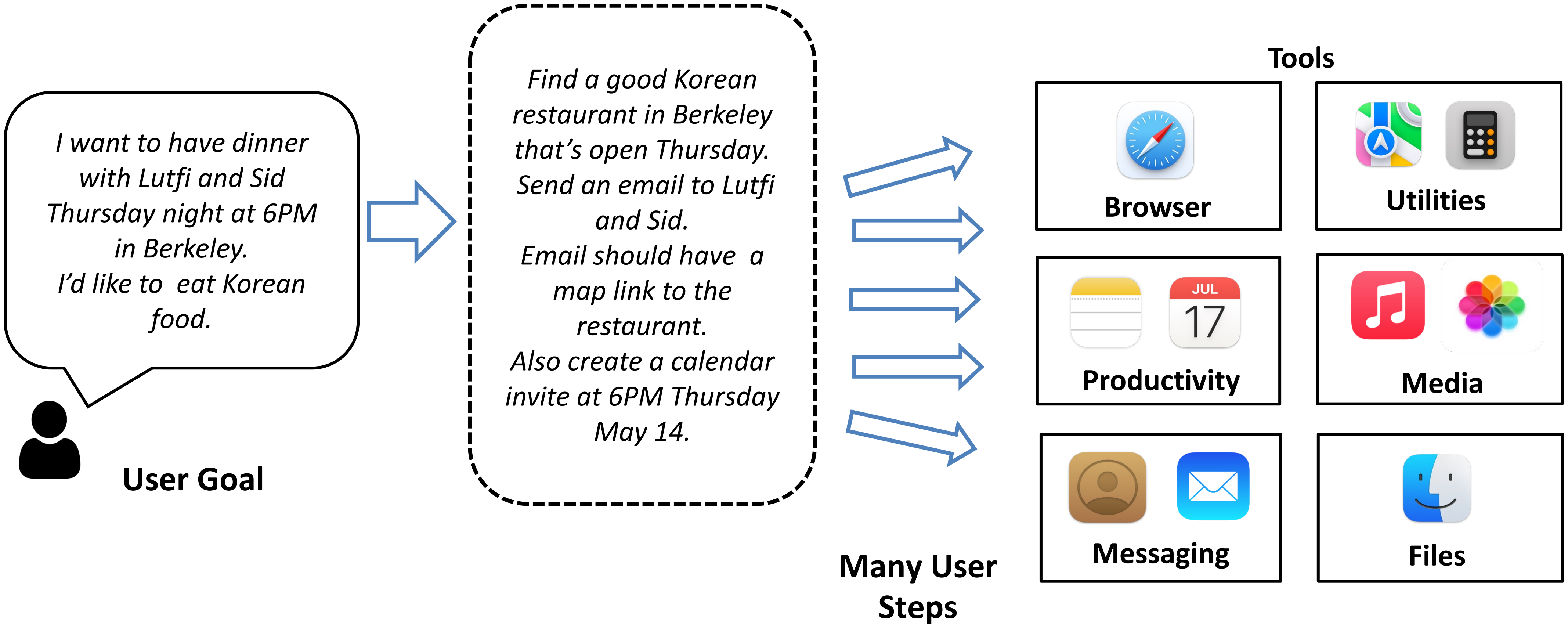
Select Questions & Topic

- CSAT
- History
- Geography
- Polity
- Current Affairs
- Economy
- Science
- Environment

Go back | Begin Practice

Personal Teacher

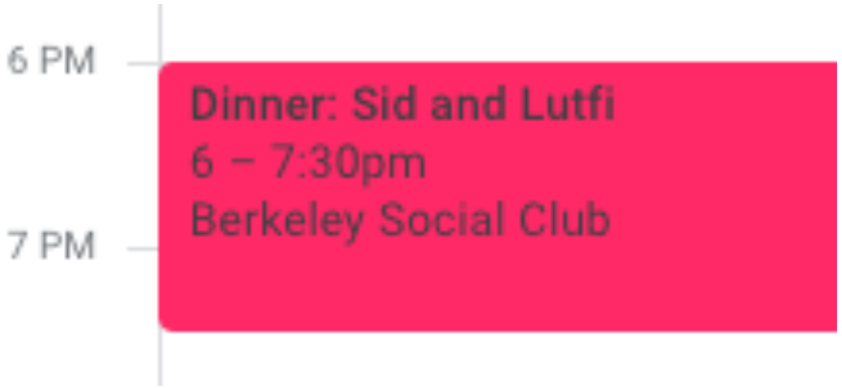
Why Does the Execution of Simple Tasks on a SmartPhone Take So Many Steps with Today's Apps?

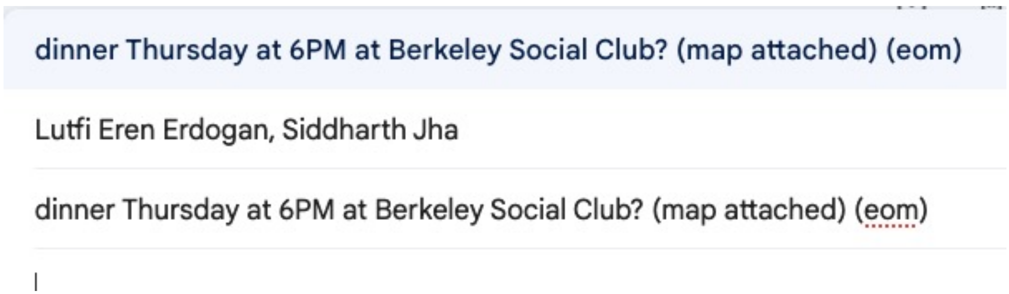


What We Want Seems so Simple

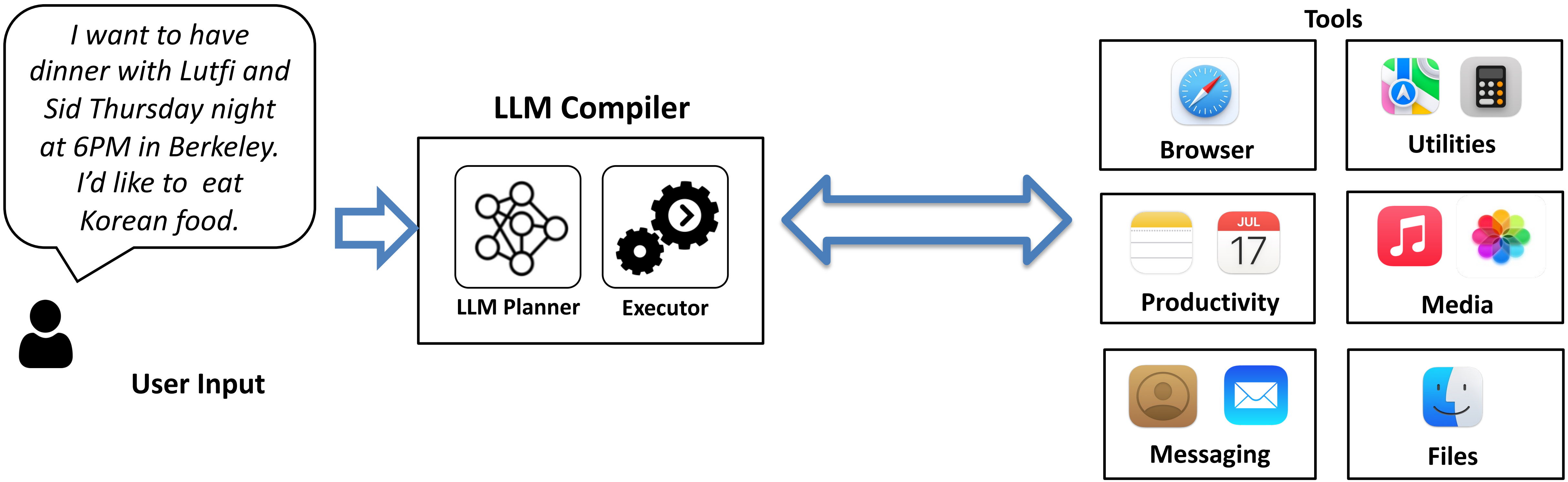
I want to have dinner with Lutfi and Sid Thursday night at 6PM in Berkeley.
I'd like to eat Korean food.

User Goal





TinyAgent + LLM Compiler Aim to Solve This: Managing Everyday Tasks Using Natural Language

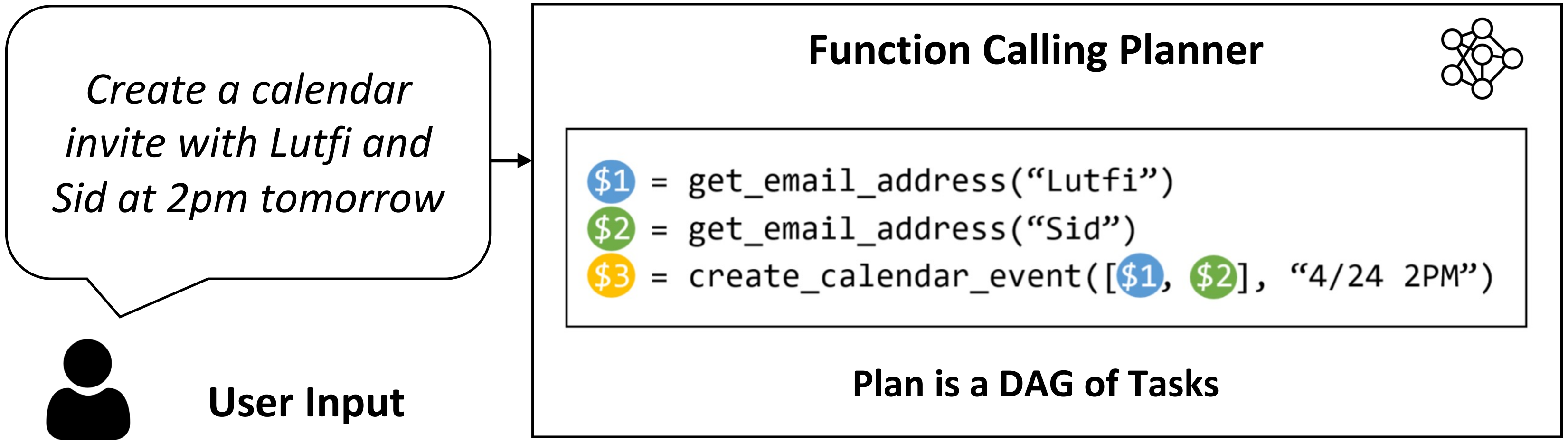


Kim, S., Moon, S., Tabrizi, R., Lee, N., Mahoney, M.W., Keutzer, K. and Gholami, A., 2023. An LLM compiler for parallel function calling. *arXiv preprint arXiv:2312.04511*. *To appear* ICML 2024.

Overview
LLM Compiler and Function Calling

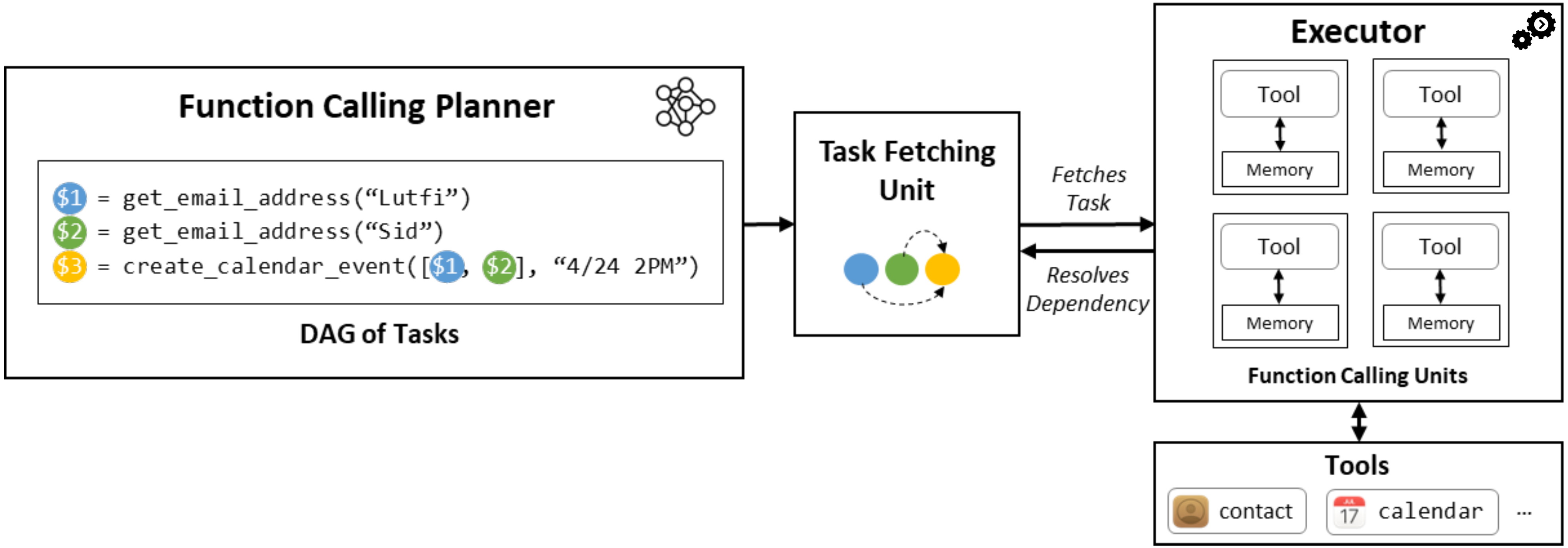
LLMCompiler enables function calling by decomposing the user queries into a series of function calls with the right set of arguments and execution order

- We use **LLM's reasoning capability** to build a Directed Acyclic Graph (DAG) from the user input



Overview
LLMCompiler: Parsing and Execution

Plans are then parsed and executed by the **Executor**



Function calling planner
- Creates DAG

Task Fetching Unit
- Schedules work
- Dispatches work

Executor
- Chooses tools
- Executes API calls

Let's Look at that Step by Step
Pre-Planning: Step 0

Step 0:
Trimming Tools List

Create a calendar invite with Lutfi and Sid at 2pm tomorrow

User Input

Which Tools are Needed?

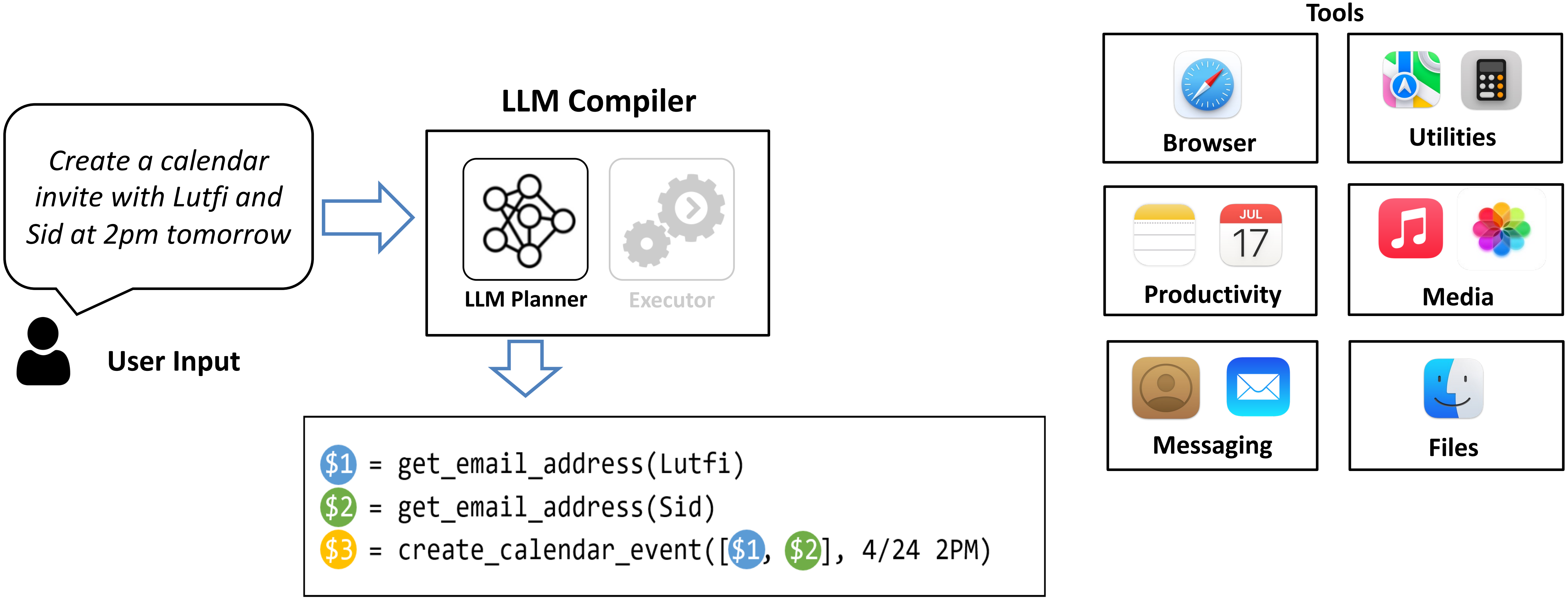
- ✅ **create_calendar_event**
- ❌ compose_new_email
- ❌ summarize_pdf
- ❌ reply_to_email
- ❌ get_zoom_meeting_link
- ❌ maps_show_direction
- ❌ create_note
- ❌ forward_email
- ❌ get_phone_number
- ❌ send_sms
- ❌ open_note
- ❌ web_search
- ❌ create_reminder
- ❌ append_note_content
- ❌ open_and_get_file_path
- ❌ maps_open_location
- ✅ **get_email_address**

Tools

- The prompt size can grow with the number of tools.
- To reduce the prompt size, we pre-process to select only relevant tools and provide them to the Planner.

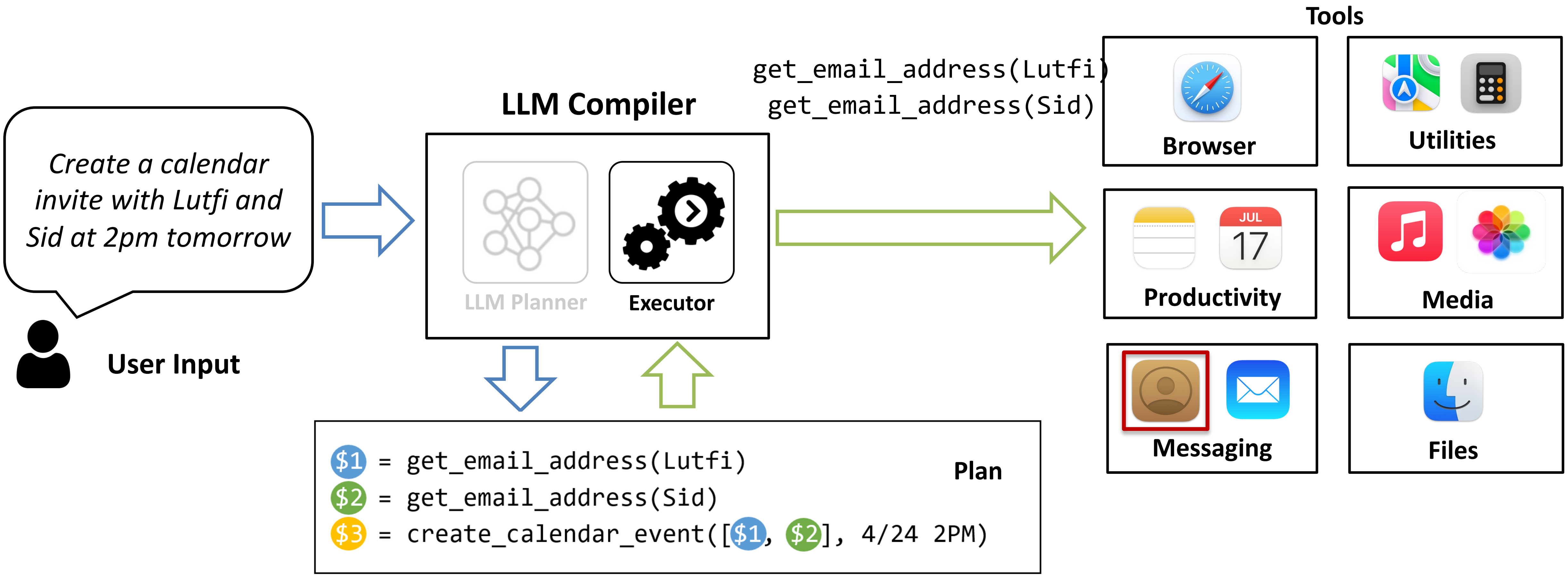
Planning: Step 1

Step 1:
LLM Planner provides the plan



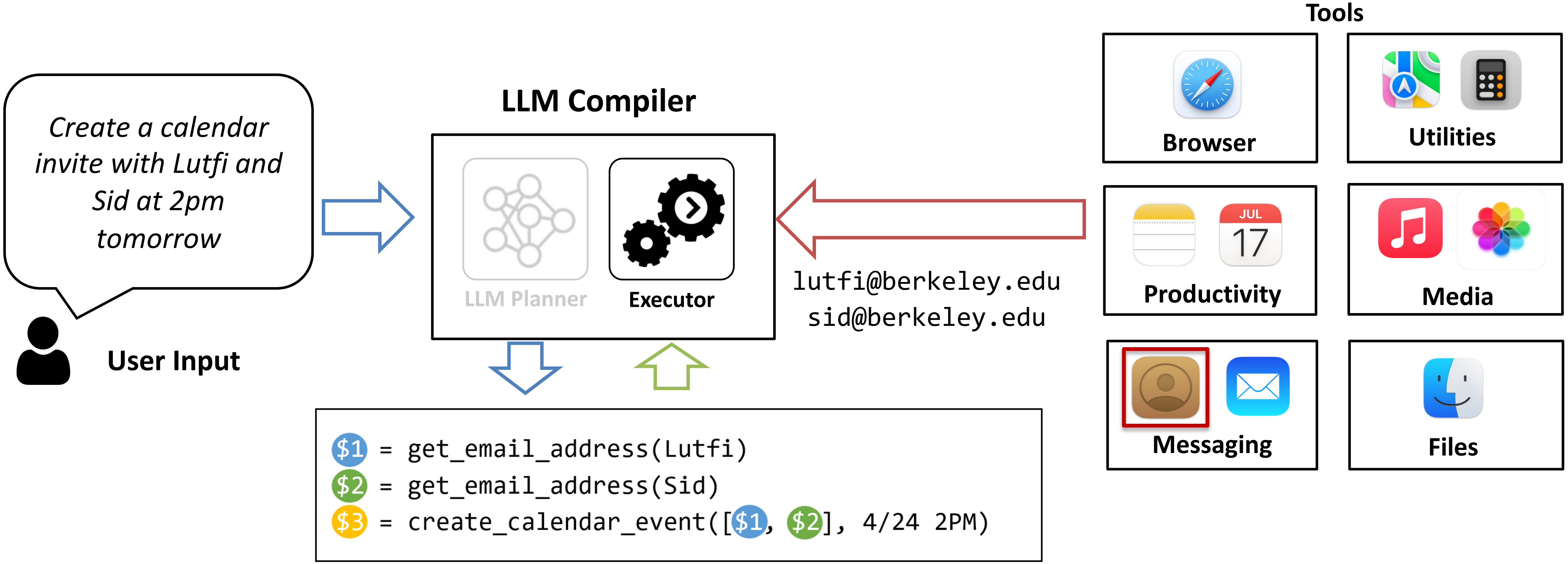
Execution: Step 2-1

Step 2-1:
Executor executes the plan



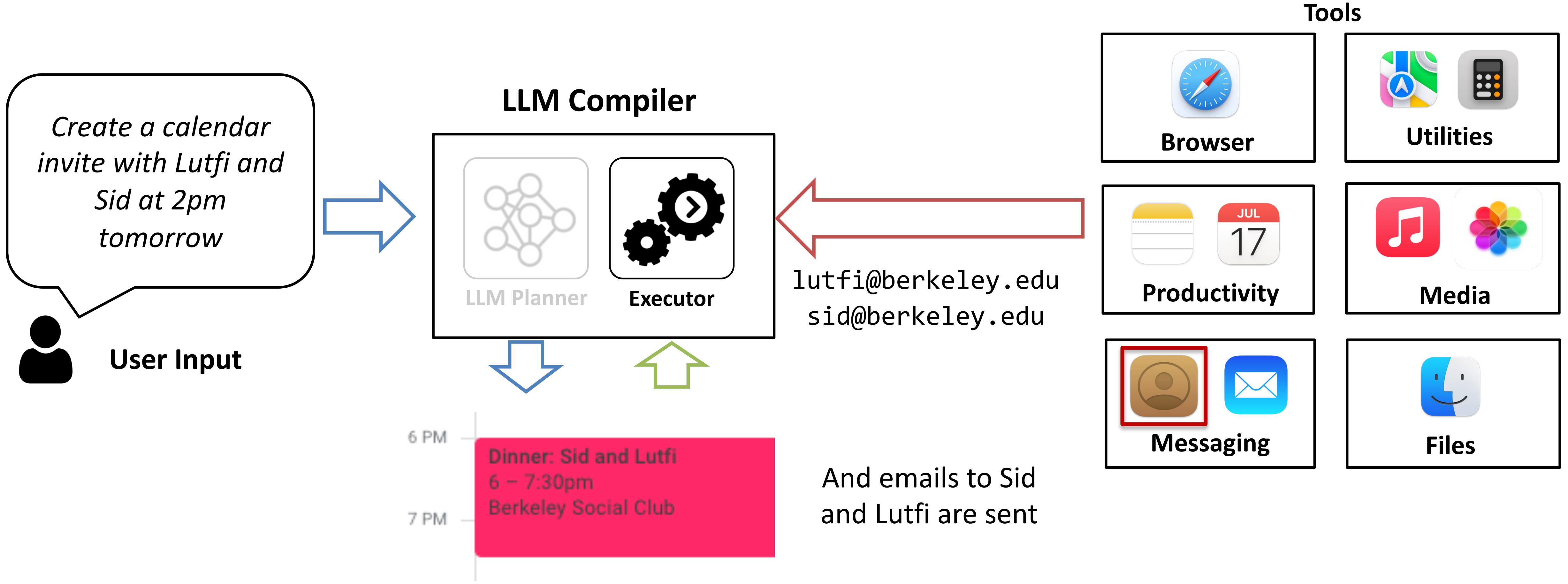
Execution Pt-2

Step 2-2:
Executor executes the plan



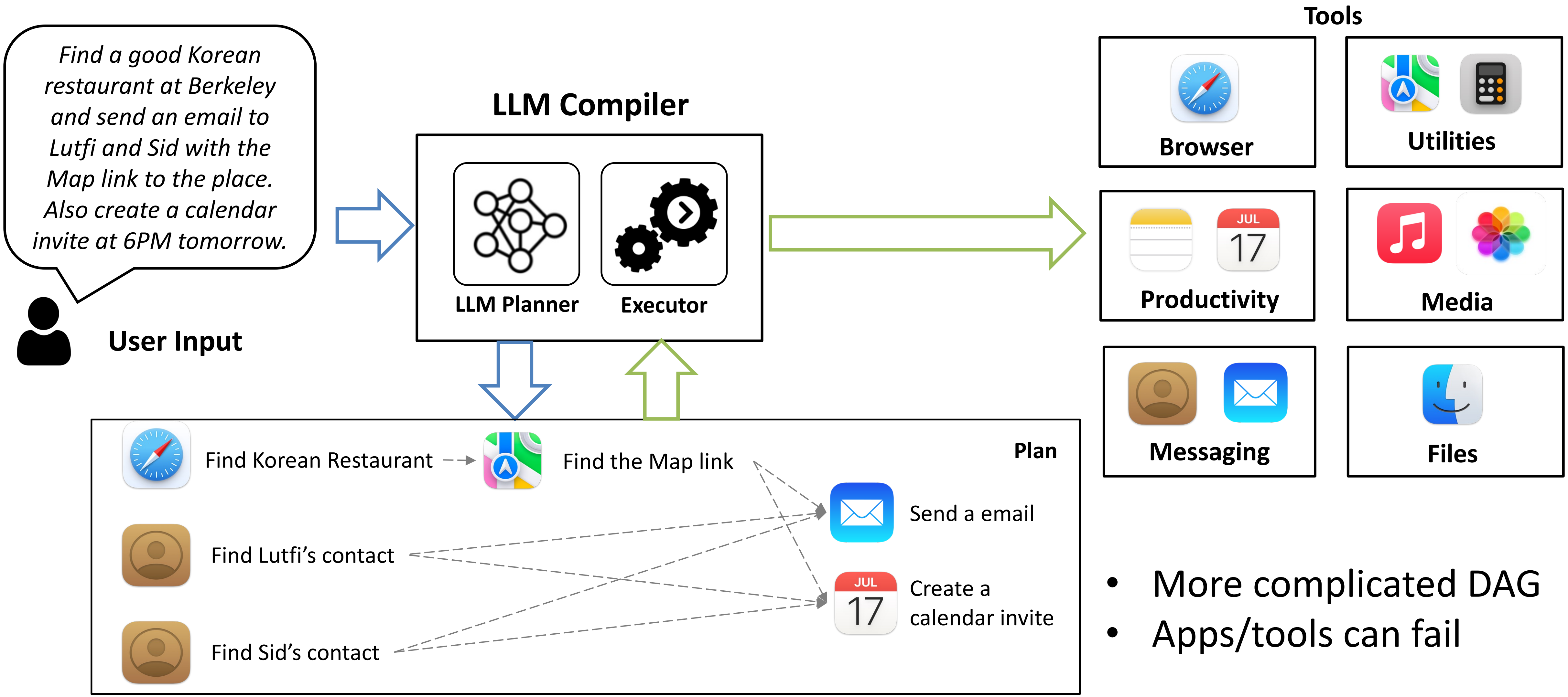
Execution Pt-3

Step 2-3:
Executor executes the plan



Things Can be a Bit More Complicated

More complicated Planning Example



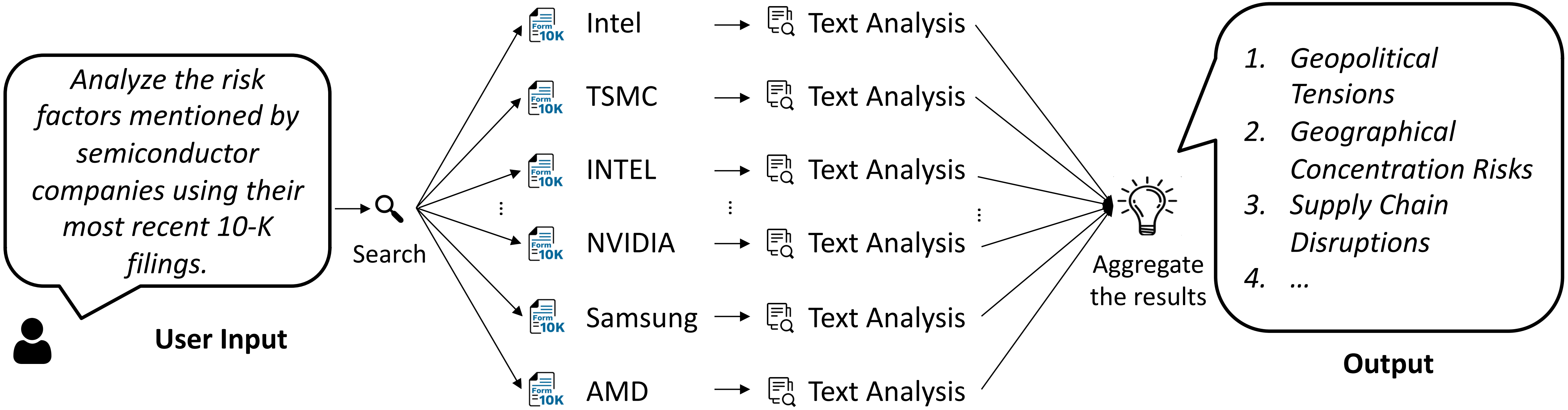
- More complicated DAG
- Apps/tools can fail

Even Simple Queries Can Get Complicated, Quickly
In General, We Have a Distributed Computing Problem

Prompt: Analyze the risk factors mentioned by major semiconductor companies using their most recent 10-K filings.

Parallel Operations:

- Identification of semiconductor companies
- Text extraction from multiple 10-K filings simultaneously.
- Natural language processing to identify and categorize risks in each filing.
- Aggregation and comparison of risk data across different companies.



LLM Compiler for Efficient Tool Planning and Execution

Neuro-symbolic Programming

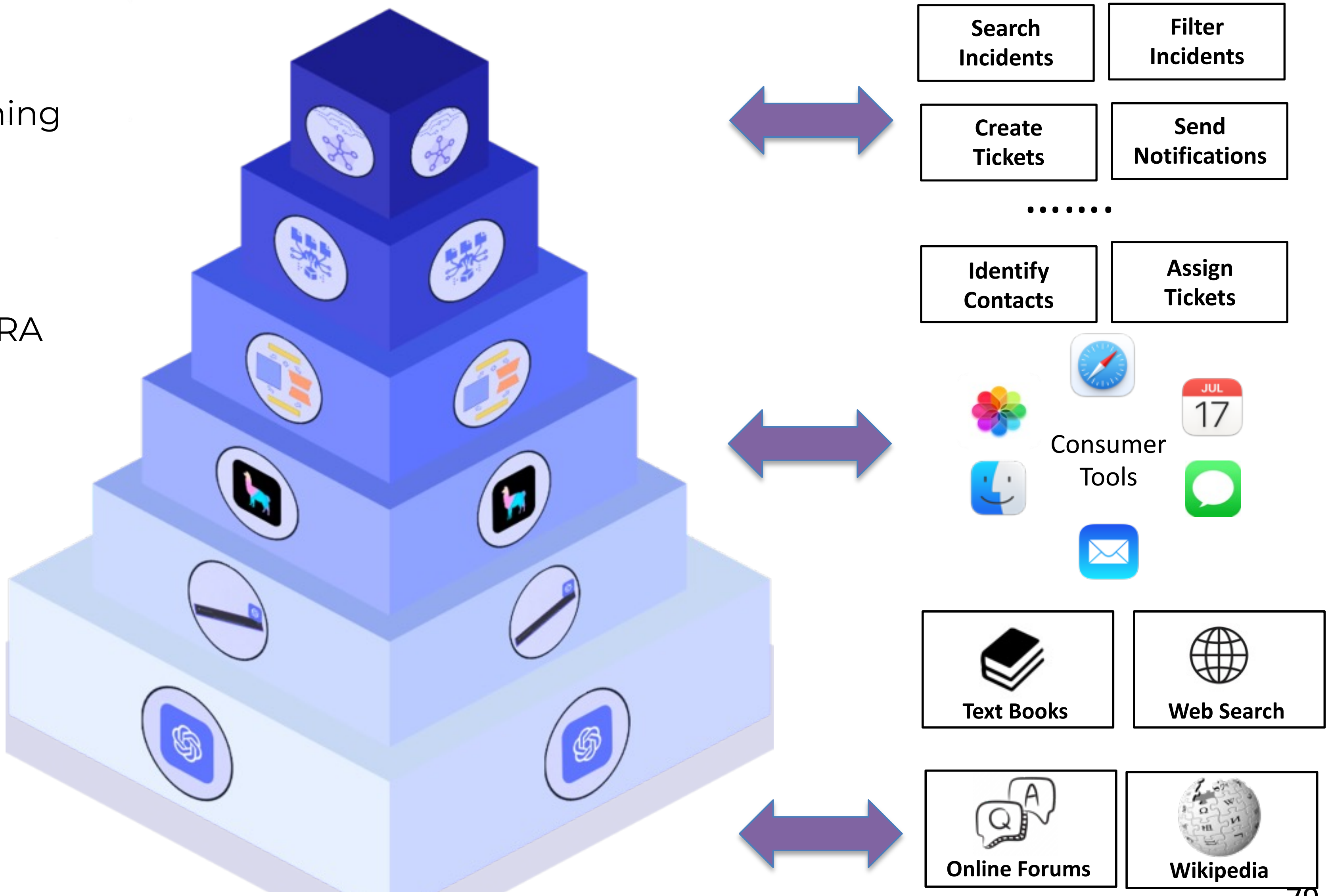
Full Finetuning

Partial tuning/Adapters/LORA

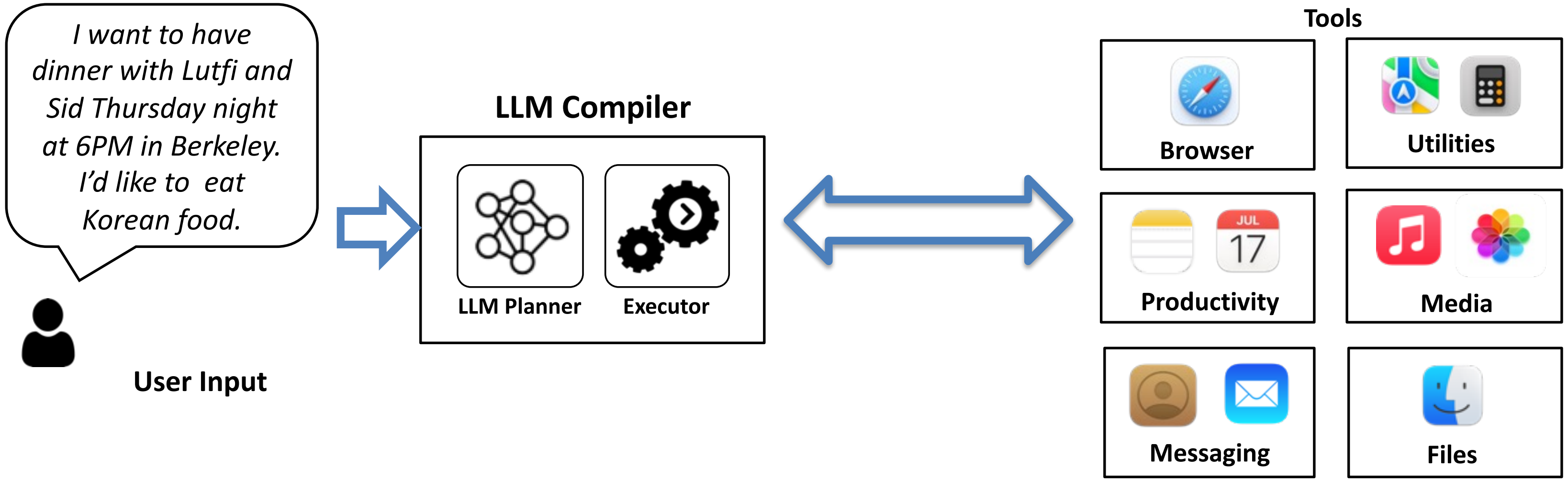
Prompting

Out of the box LL Model

Intelligent Model Selection



Summary
TinyAgent and LLM Compiler



- TinyAgent is our project aimed at investigating how agents can simplify management of everyday tasks
- LLMCompiler is a research outcome that anticipates that as function calling becomes more complicated, planning and execution of function calls becomes a distributed computing problem
  - Speed of execution dramatically improved through distributed computing
  - Robustness in the face of failed/delayed tool calls

Kim, S., Moon, S., Tabrizi, R., Lee, N., Mahoney, M.W., Keutzer, K. and Gholami, A., 2023. An LLM compiler for parallel function calling. *arXiv preprint arXiv:2312.04511*. *To appear* ICML 2024.

ML Sys in 2026? 2030?

MLSYS
or
NeurIPS etc

MLSYS

In the program today

- Quantization and Compression

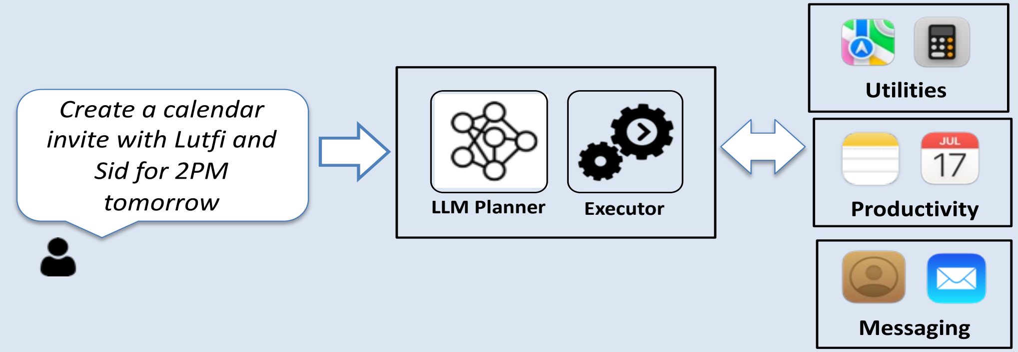
In the future?

- Efficient function calling
- Efficient RAG

Meet the real Lutfi and Sid
(and Monish) at their poster today
Retrieval Augmented Generation:
Challenges and Opportunities

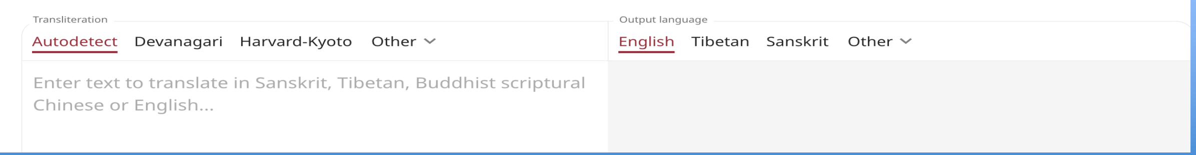
LLM-centric GenAI Systems
Let's Look at them Individually

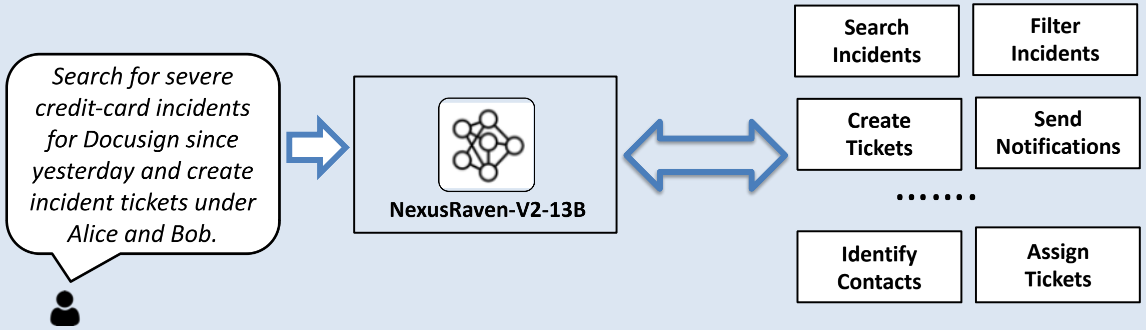
Image/video generation using **diffusion models**



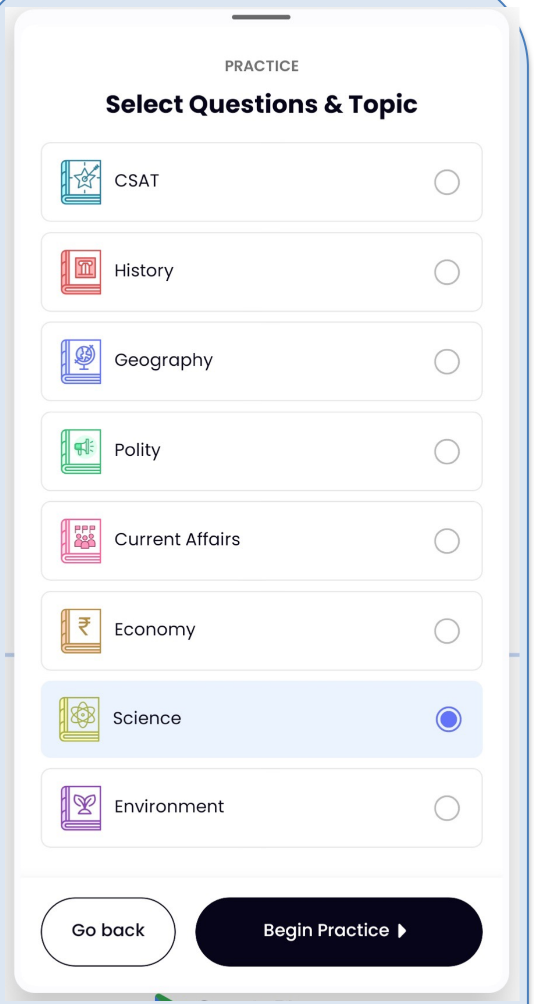
Personal Agents/Co-pilots

Machine Translation





Enterprise Agents/Co-pilots

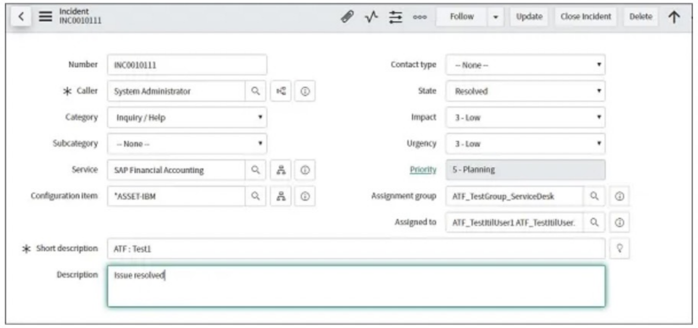


Personal Teacher

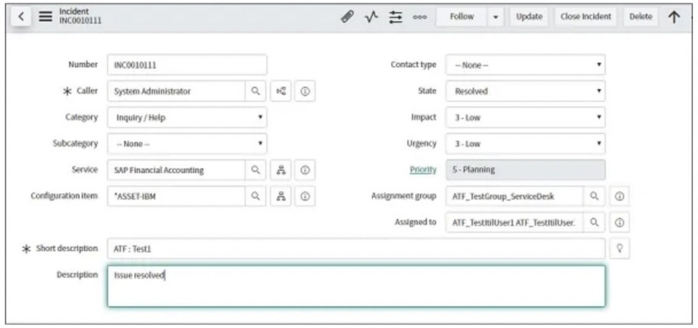
Why Can't a Security Analyst's Life Be This Easy?

Search for severe credit-card incidents for Docusign since yesterday and create incident tickets under security analysts Alice and Bob.

Security Analyst at a
Managed Security Services Provider

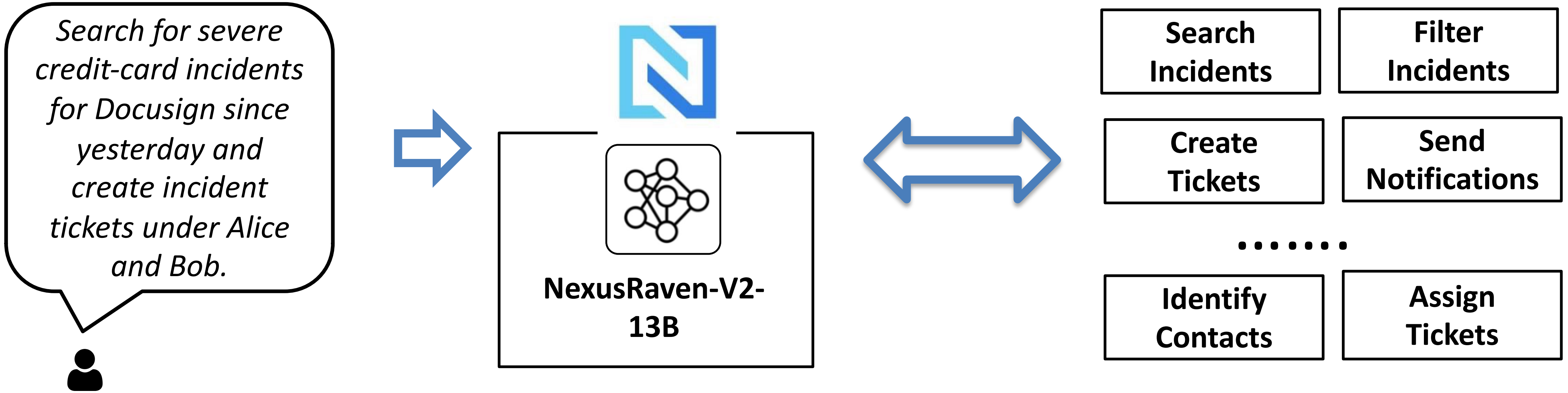


New Incident
Ticket for Alice



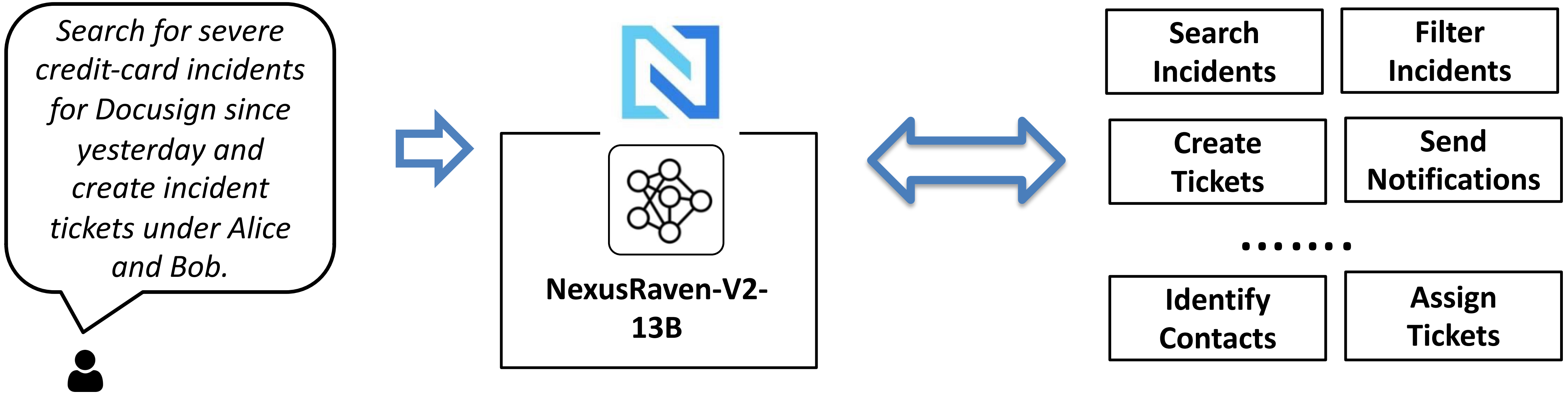
New Incident
Ticket for Bob

Nexusflow.ai
Enterprise-grade GenAI Software Agent



- Nexusflow.ai is creating a GenAI software agent that simplifies burdensome tasks in enterprises by providing a natural language interface
- In a Security Operation Center/Managed Security Services Provider, a single incident may require security analysts to deal with dozens of tools each with complicated APIs
- The first step is automating the translation of a query in natural language to a series of function calls to tools
- The information returned from the tools will then be used to satisfy the query.

Nexusflow.ai
Enterprise-grade GenAI Software Agent



- We've already seen TinyAgent do function calling, but …
- In an enterprise context, accurately, translating a single natural language command to function calls to tools has many challenges:
  - A single natural language query may result in ten or more tool invocations
  - Each security tool API may have 30 or more arguments
  - Each argument may have 100 or more alternative values
  - All this must be executed correctly to get the correct result
  - And … the user wants immediate (~1 second response) time
- Even a very strong model such as GPT-4 struggles to get >50% zero-shot accuracy
- How can a small model compete?

Requirements for Enterprise-Strength Function Calling

- Requirement of function-calling in an enterprise-strength GenAI model:
  - Few/no hallucinations during tool use
  - Generalizes to new APIs → 90% Zero-shot performance
  - Can orchestrate multiple complex APIs to execute a plan

Smaller Models Compete by Fine-Tuning
The Key Element of Fine-tuning is Data

Neuro-symbolic Programming

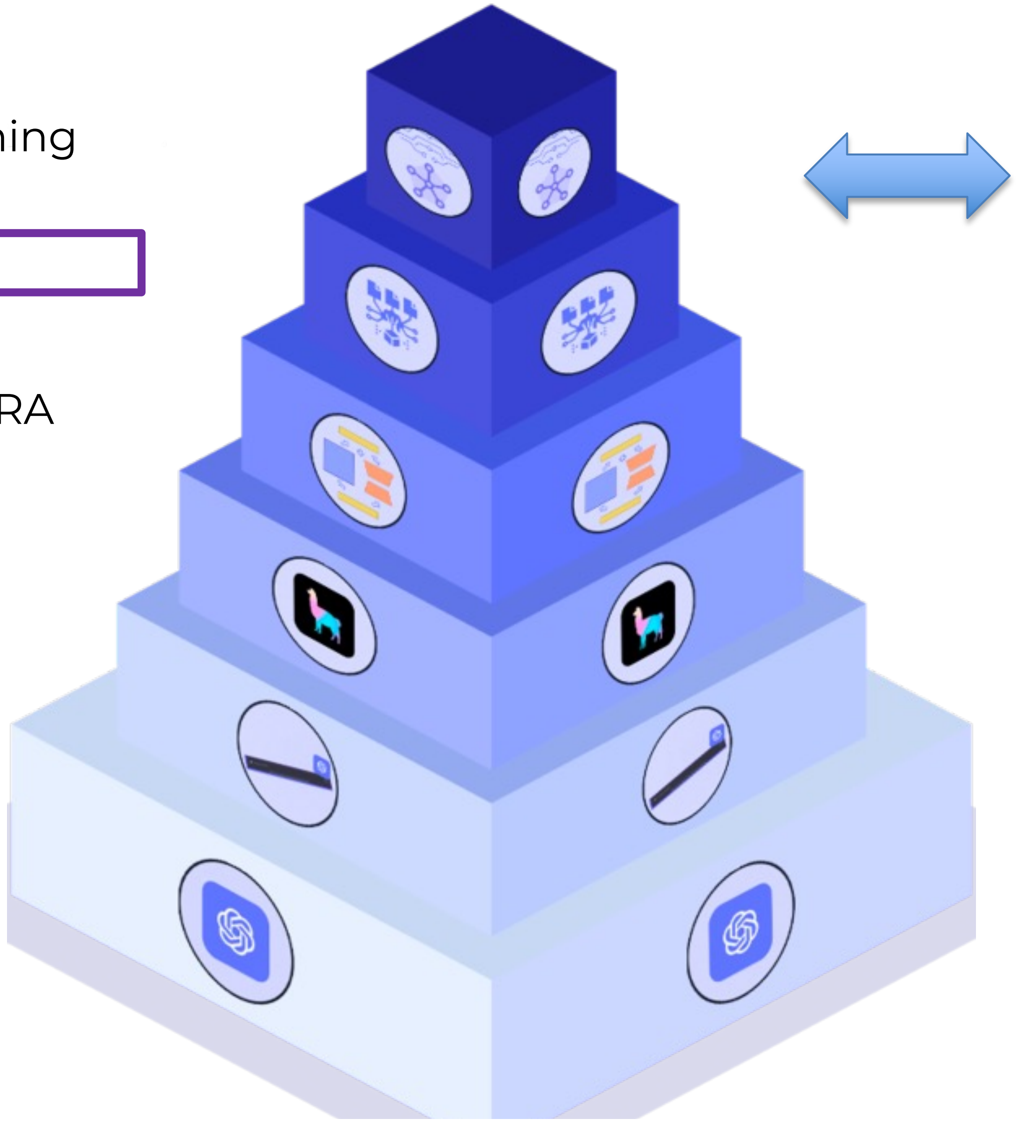
Full Finetuning

Partial tuning/Adapters/LORA

Prompting

Out of the box LL Model

Intelligent Model Selection

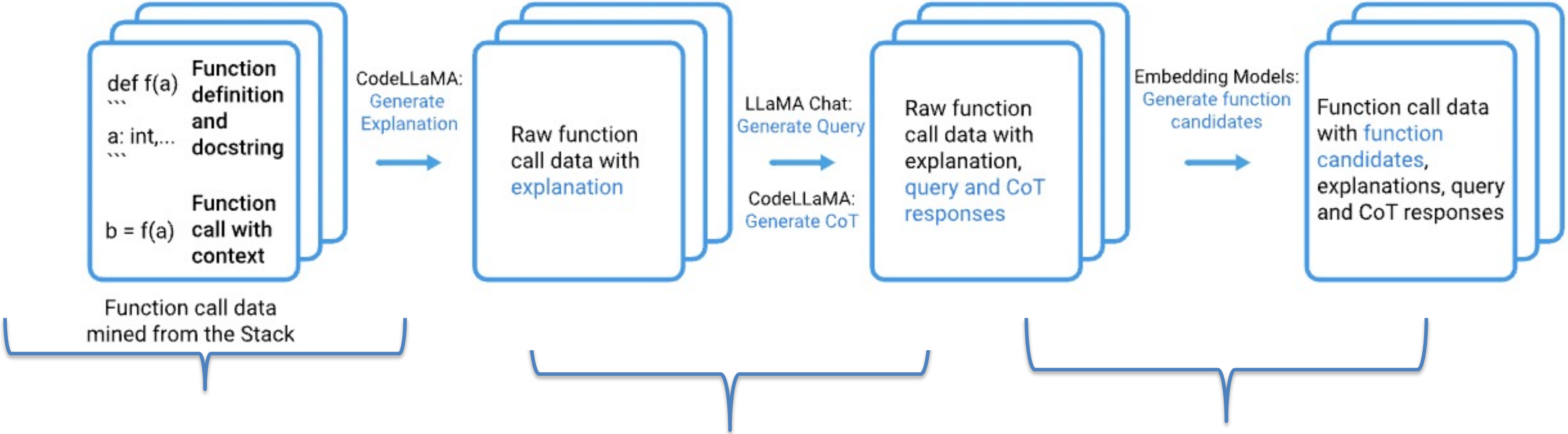


Search Incidents	Filter Incidents
Create Tickets	Send Notifications
…….	
Identify Contacts	Assign Tickets

Requirements on Data

- In developing a data set for in fine-tuning, the data should be:
  - High Quality:
    - Every detail in the completion of the function call must be mapped to information provided in the prompt explicitly.
  - Diverse:
    - Data needs to be from varied real-world (not synthetic) sources, capturing **diversity** of use cases.
  - Difficult:
    - End goal should require orchestration of multiple calls, where the outputs feed into each other: → Model learns to plan

Generating the Data Set for Fine-tuning



Mining data from The Stack[1] ensures **diversity**.

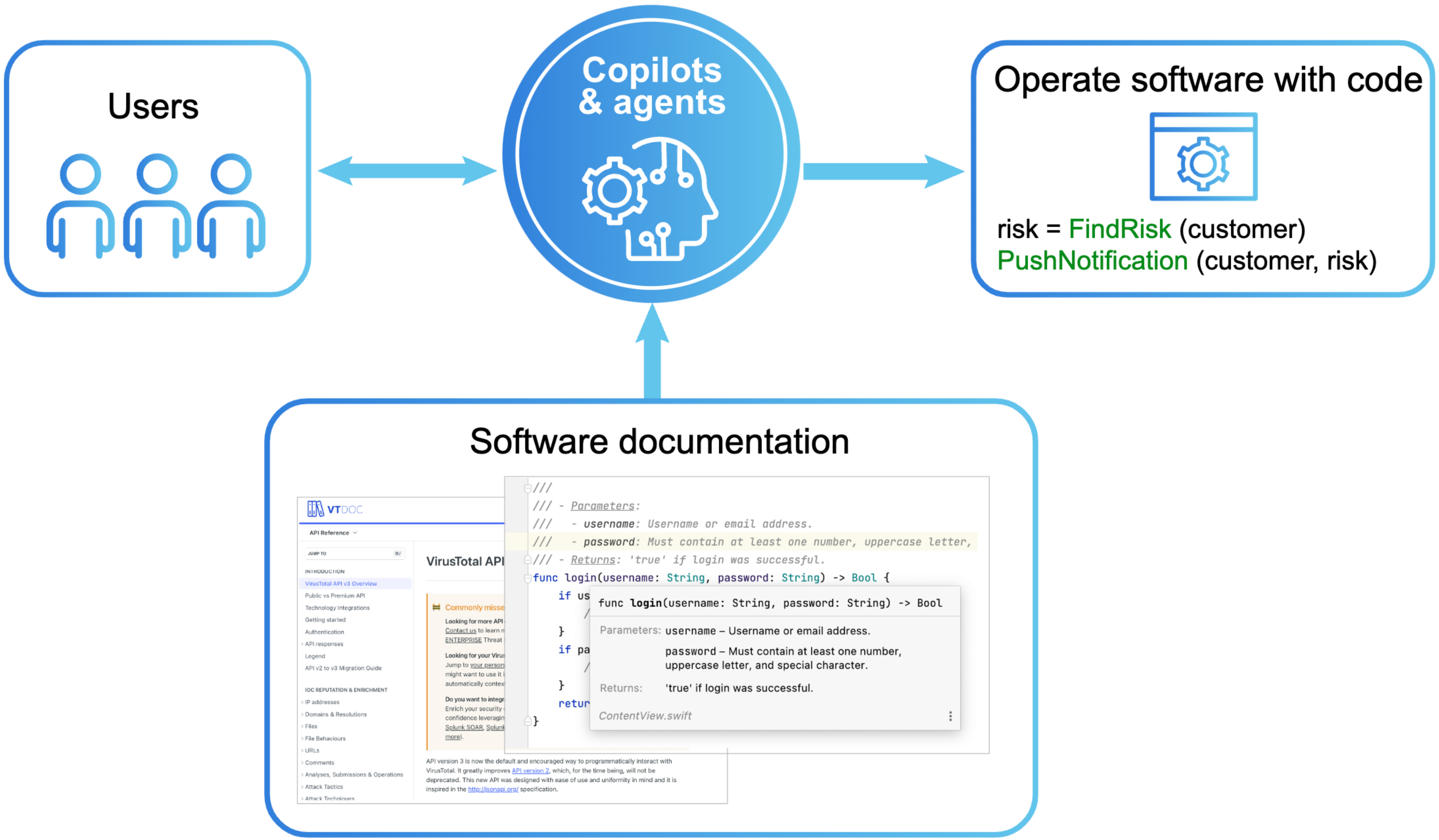
Mining deeply nested call chains from The Stack [1] ensures **difficulty**.

[1]huggingface.co/datasets/bigcode/the-stack-dedup

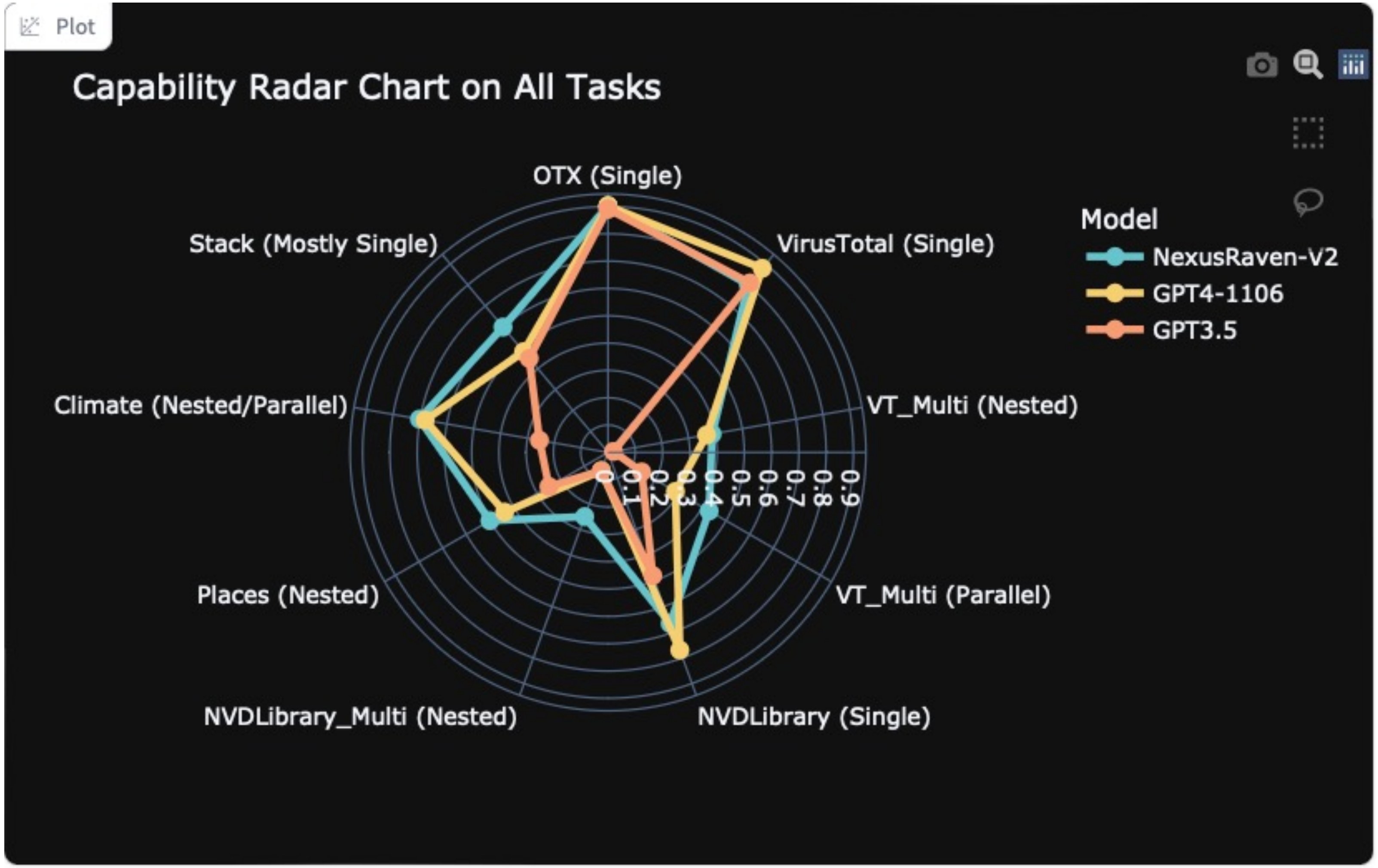
- Explanations teach the model where details in the completion are motivated from.
- Allows the model it to learn strong mappings between inputs and outputs, mitigating hallucinations.

Postprocessing the data to include irrelevant functions makes the function calling task more difficult, allowing for a stronger model. (**Difficulty**)

Building an Agent



State of the Art in Zero Shot Function Calling
The Problem is Far from Being Solved



Radar Chart of Accuracy Across Different APIs

NexusRaven-v2 surpasses GPT4 in accuracy

Model and Benchmark are Opensource

huggingface.co/Nexusflow

- Acceptable results on only two benchmarks and by only two tools
- We have a long way to go

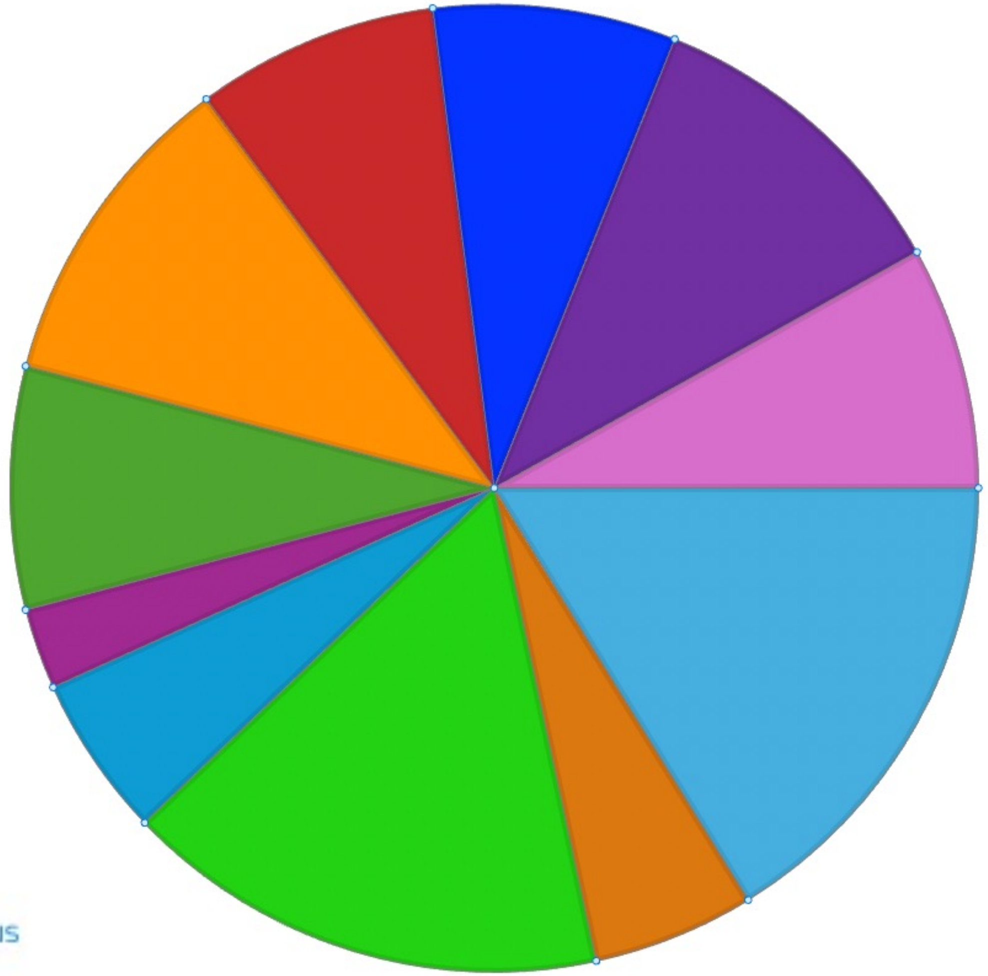
Research Directions for Function Calling

- Bring Zero Shot performance to 90% across a broad range of applications
  - Why Zero-Shot?
    - Want a highly portable model that doesn't require a lot of fine-tuning or prompting for every new customer and their APIs
- One key element: Improve the formalization of Function Call definition
  - Current *de facto* standard is OAS (OpenAI Specification)
  - Seems to be a genuine need for a more formal abstraction of functional calling and the languages used
- As always: reduce latency within user requirements

ML Sys in 2026? 2030?

MLSYS
or
NeurIPS etc

- Accurate Function Calling
- Data curation for fine-tuning



MLSYS

In the program today
- Quantization and Compression
- Fine-tuning

In the future?
- Efficient function calling
- Efficient RAG

- Fine-Tuning Language Models Using Formal Methods Feedback: A Use Case in Autonomous Systems

Personalized Teacher

Image/video generation using **diffusion models**

Create a calendar invite with Lutfi and Sid for 2PM tomorrow

LLM Planner

Executor

Utilities

Productivity

Messaging

Personal Agents/Co-pilots

Machine Translation

Transliteration

Autodetect Devanagari Harvard-Kyoto Other

Enter text to translate in Sanskrit, Tibetan, Buddhist scriptural Chinese or English...

Output language

English Tibetan Sanskrit Other

Search for severe credit-card incidents for Docusign since yesterday and create incident tickets under Alice and Bob.

NexusRaven-V2-13B

Search Incidents

Filter Incidents

Create Tickets

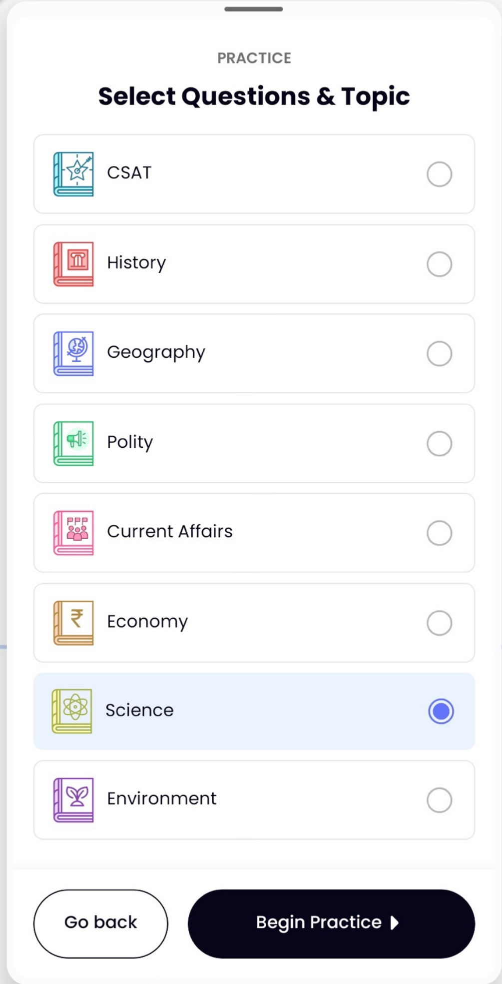
Send Notifications

....

Identify Contacts

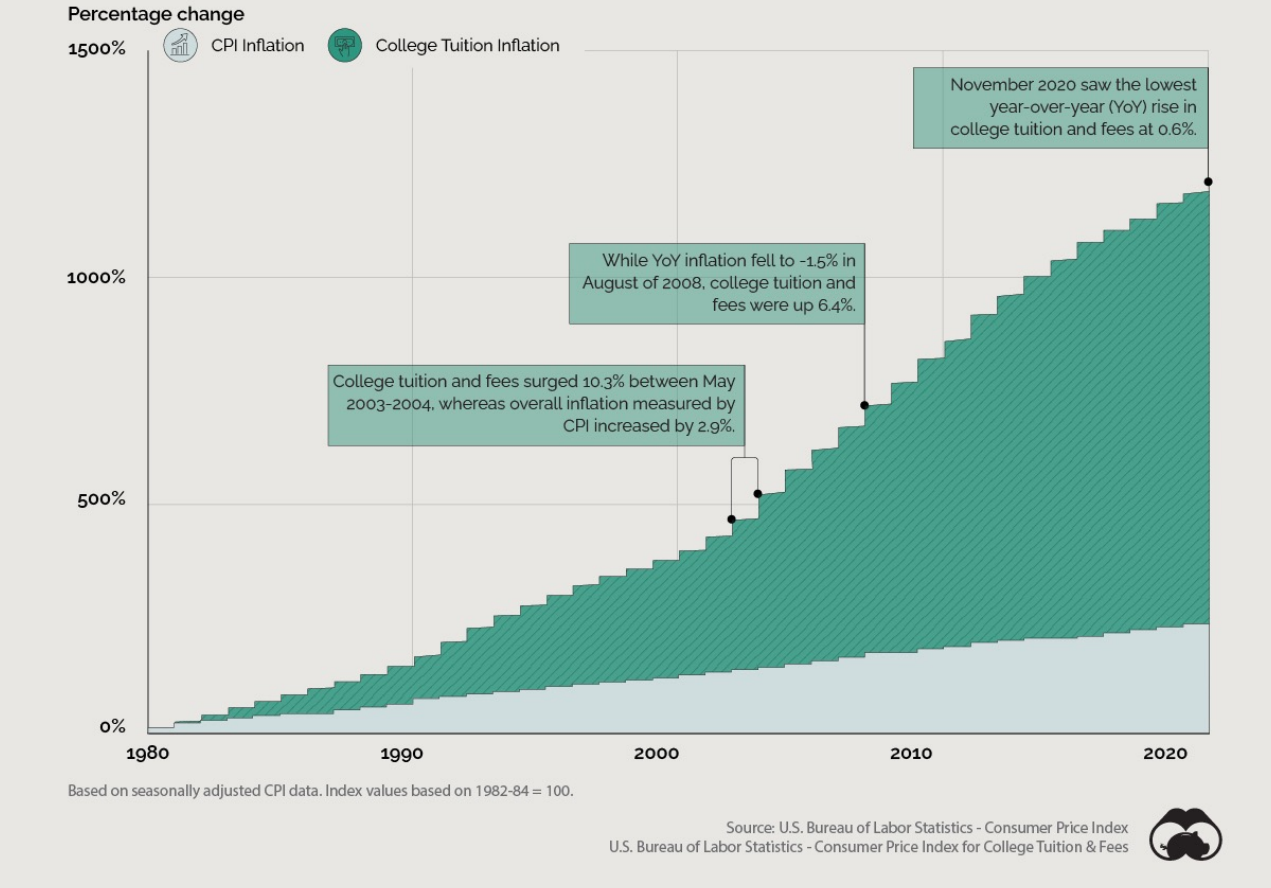
Assign Tickets

Enterprise Agents/Co-pilots



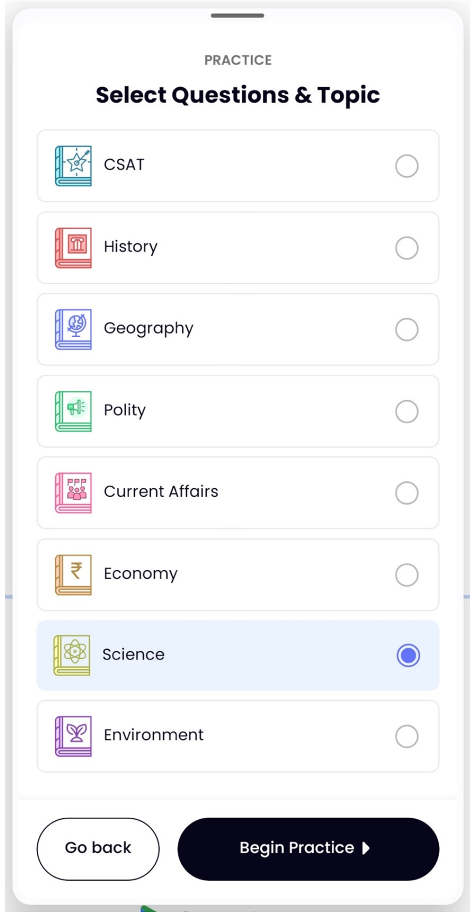
Personal Teacher

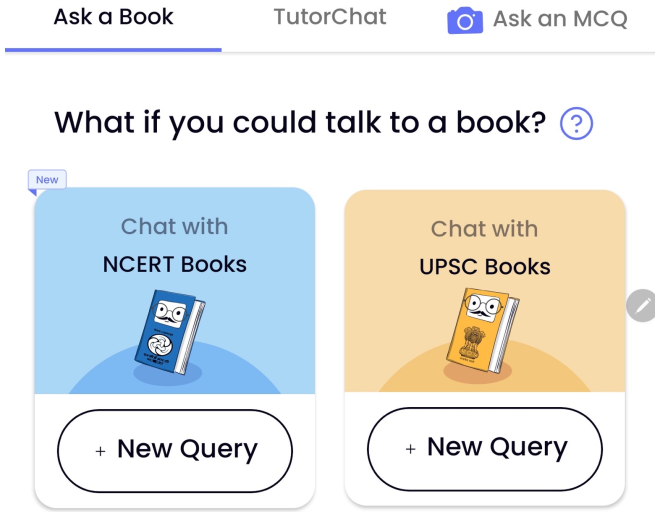
Education is Such a Great Human Need
How Can GenAI Help?

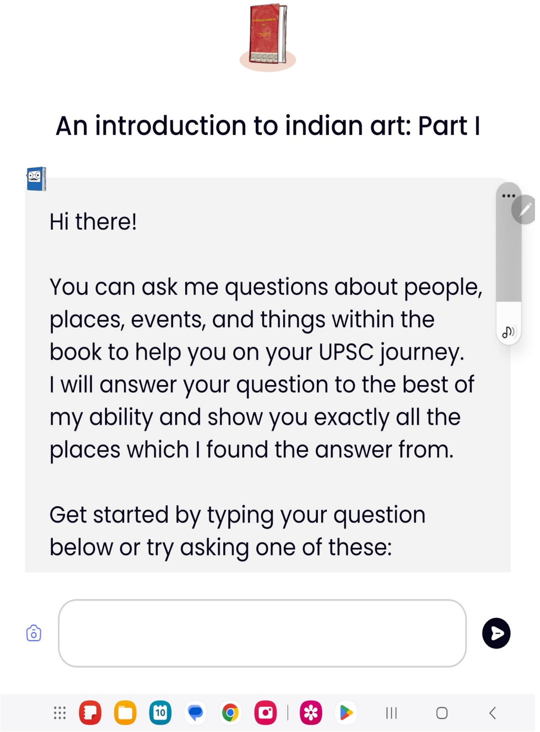


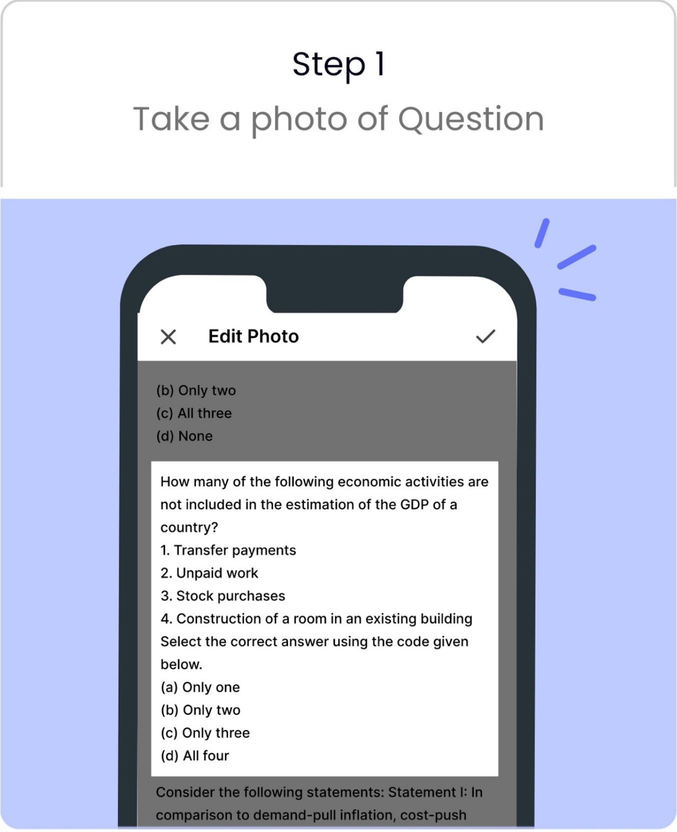
- Competition for a university education has never been higher
- Costs of education are rising exponentially
- Exams are still the gateway for many opportunities
- What if you could have your own personal tutor/teacher?

SigIQ.ai offers GenAI Agent: Personal Tutor/Teacher



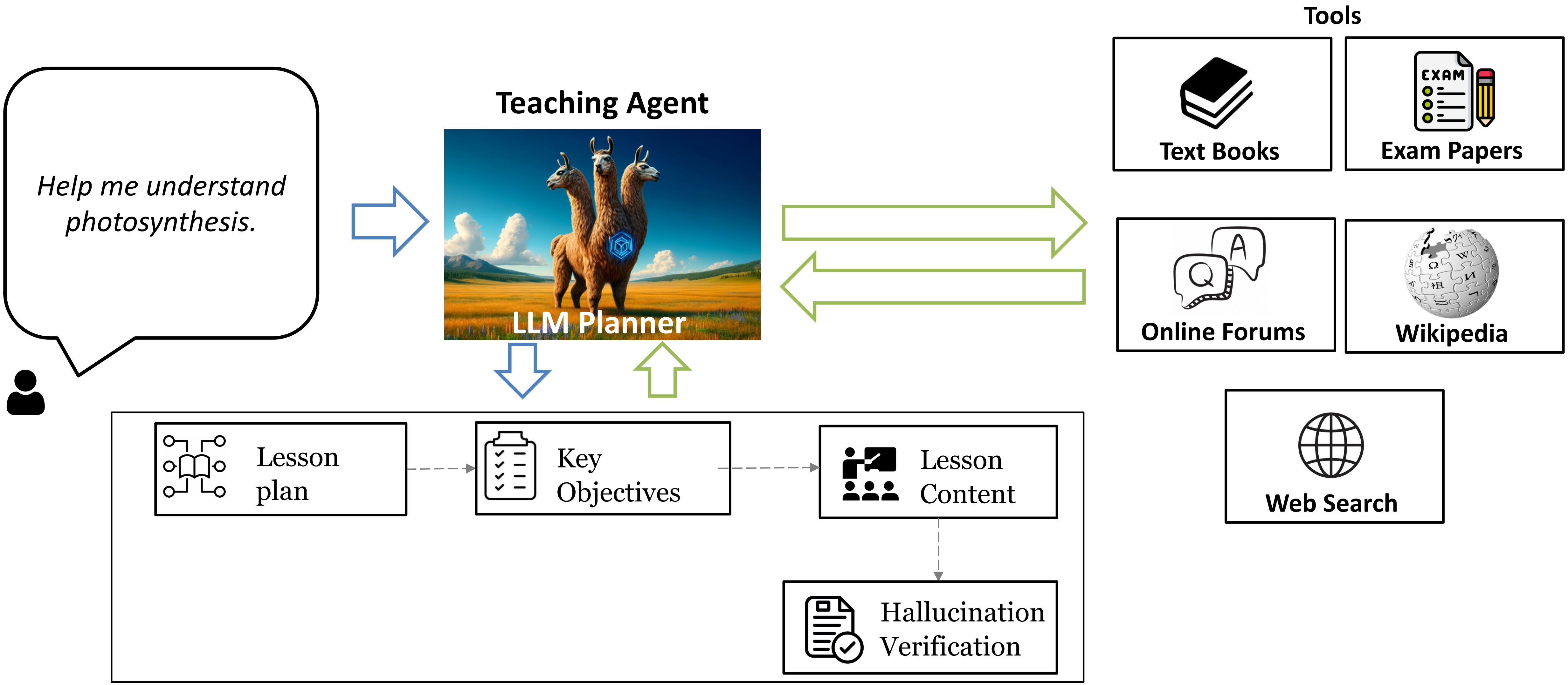




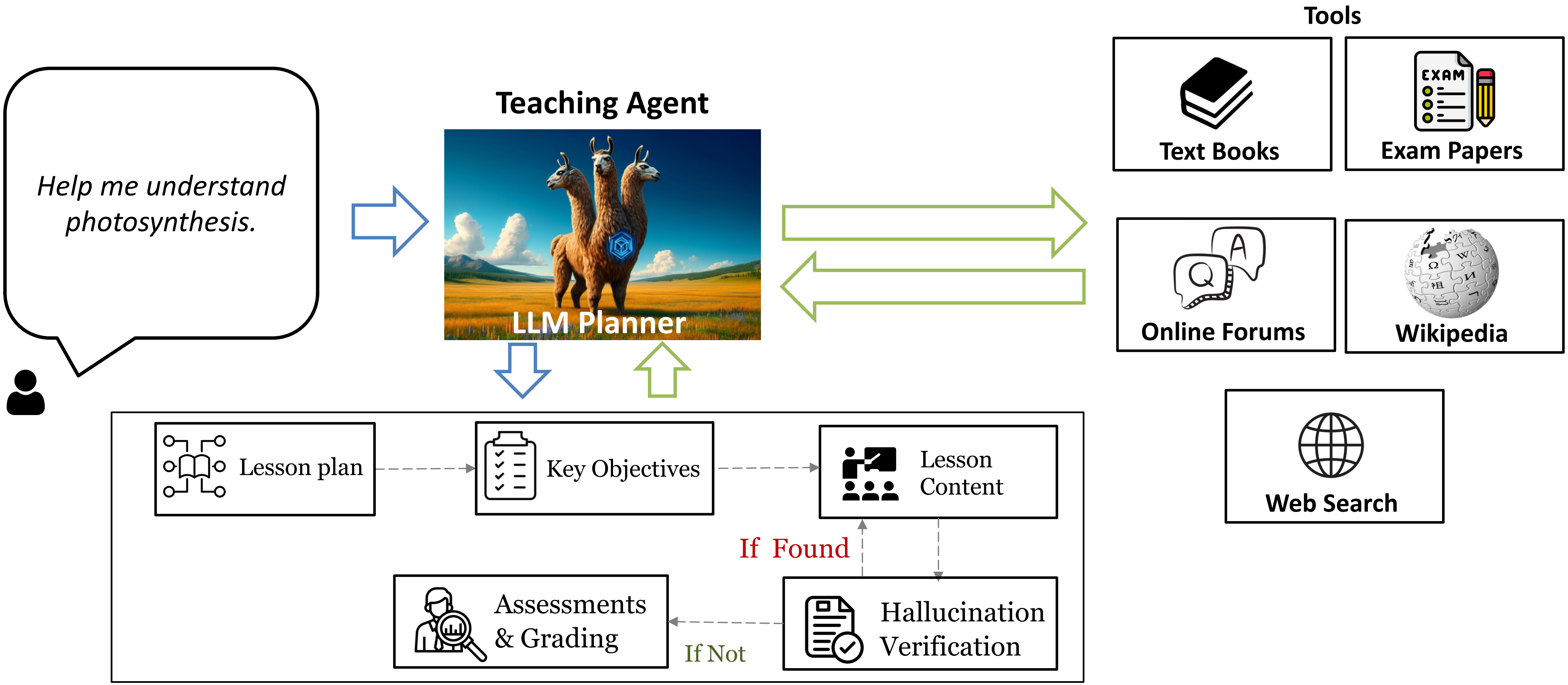


- SigIQ.ai has created a platform that creates GenAI agents that serve as personal tutors
- These agents are able to tailor tutoring to each individual user based on their behavior

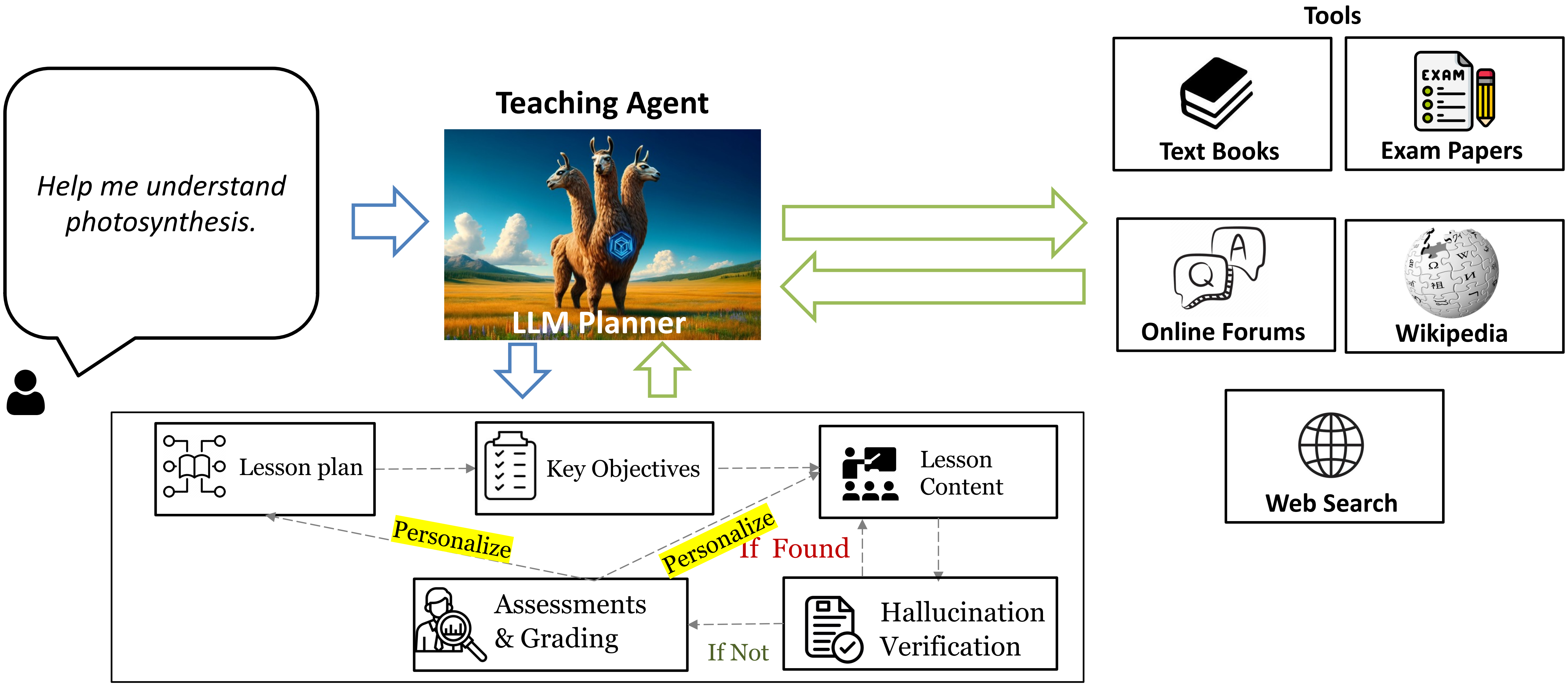
SigIQ.ai Personalized AI Teacher Responding to a Question



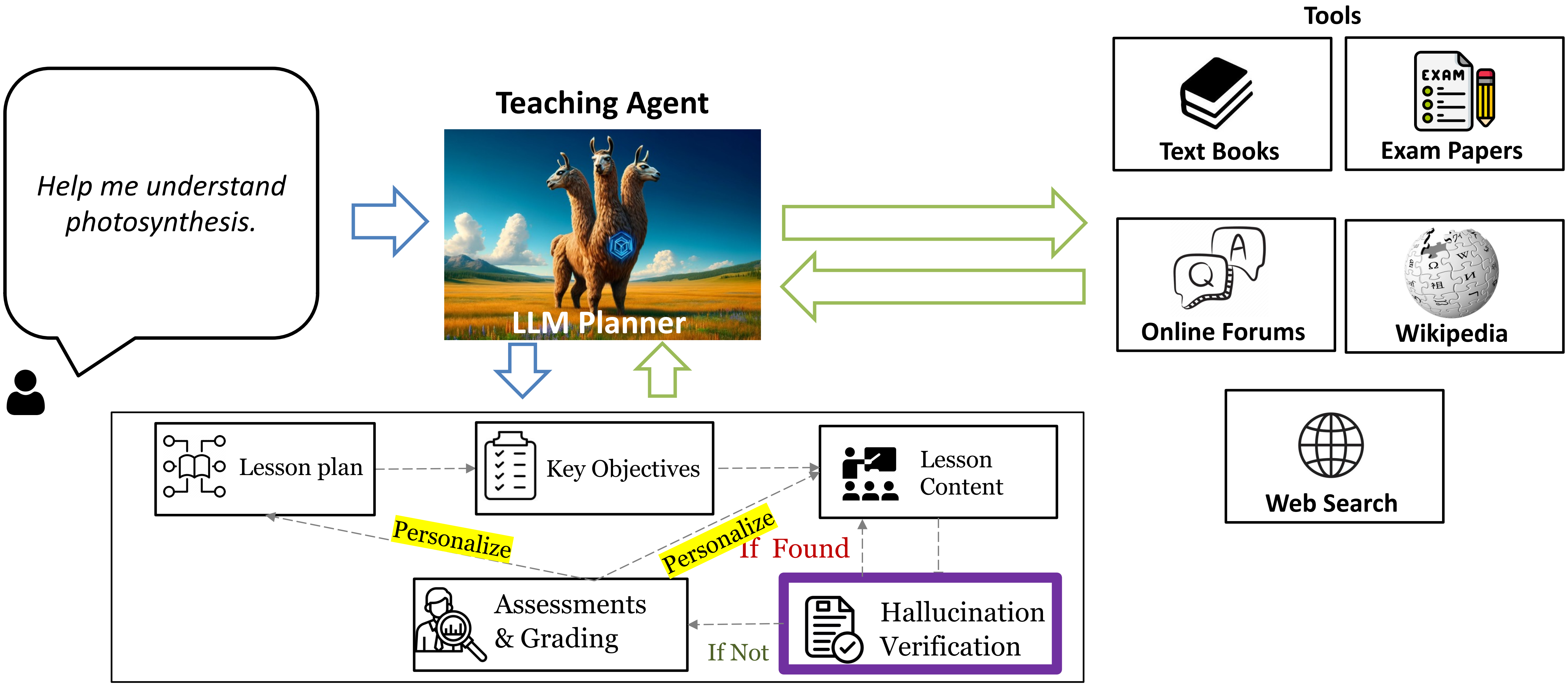
Checking for Hallucinations
Users Will Not Use an App with Inaccuracies



SigIQ.ai Personalized AI Teacher
Personalizing Plan and Feedback



Let’s Look at Hallucination Verification/Elimination



Neuro-symbolic Programming
Plus Function Calling and Retrieval Augmented Generation

Neuro-symbolic Programming

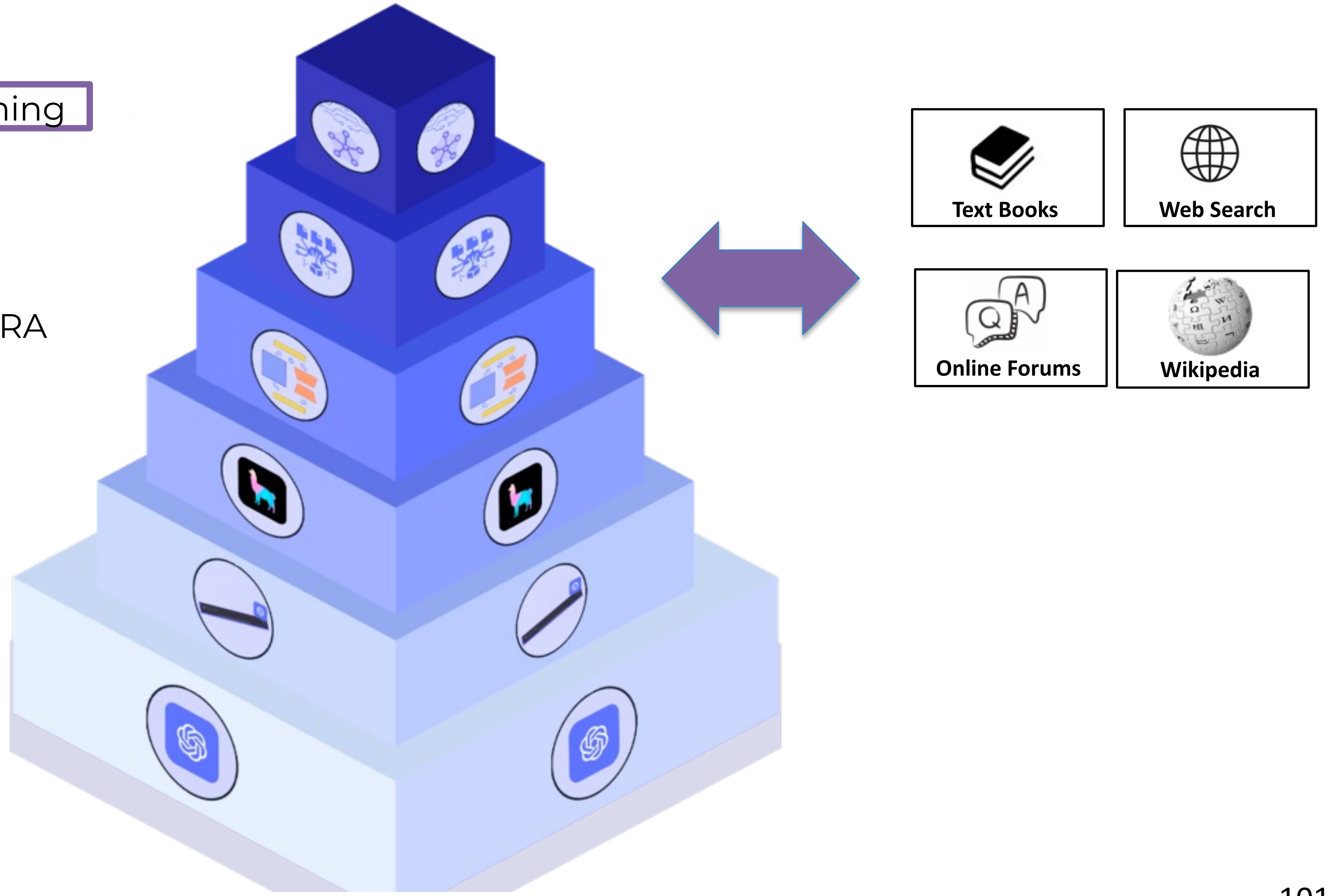
Full Finetuning

Partial tuning/Adapters/LORA

Prompting

Out of the box LL Model

Intelligent Model Selection

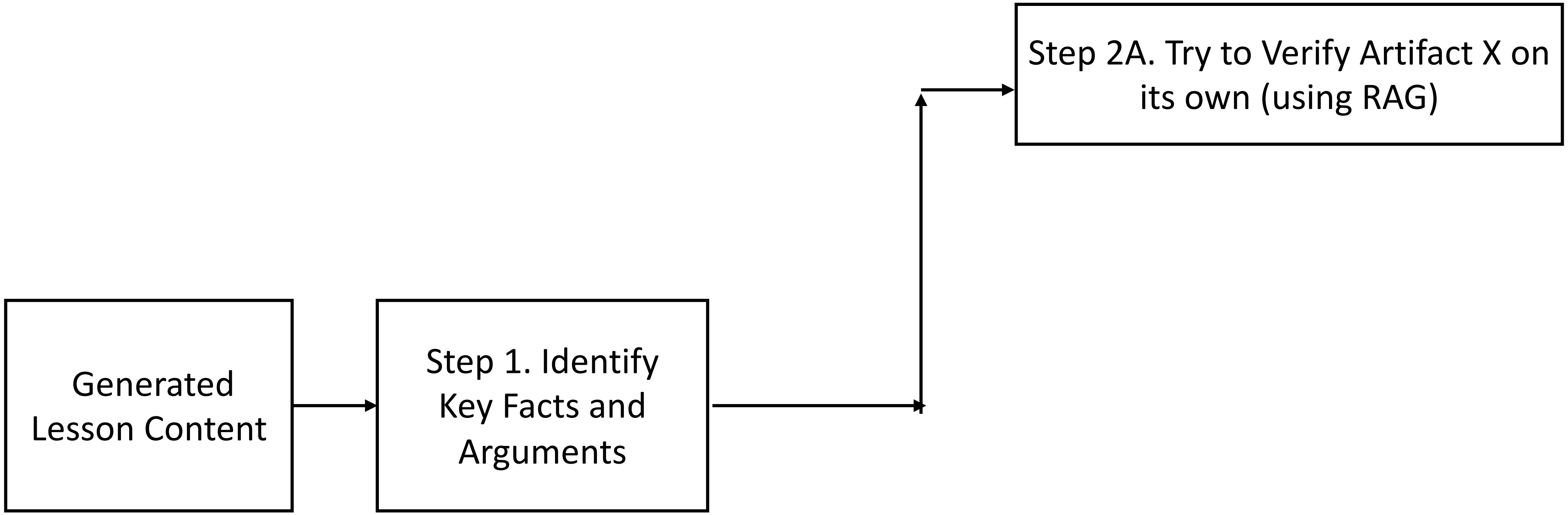


An Instantiation of Neuro-Symbolic Programming

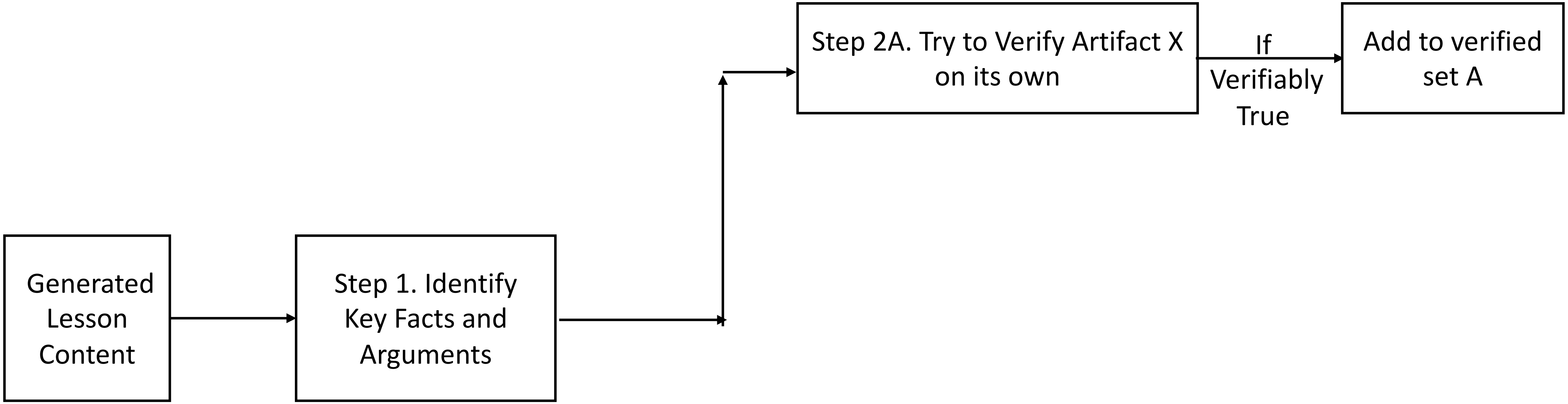
Generated Lesson Content

Step 1. Identify Key Facts and Arguments

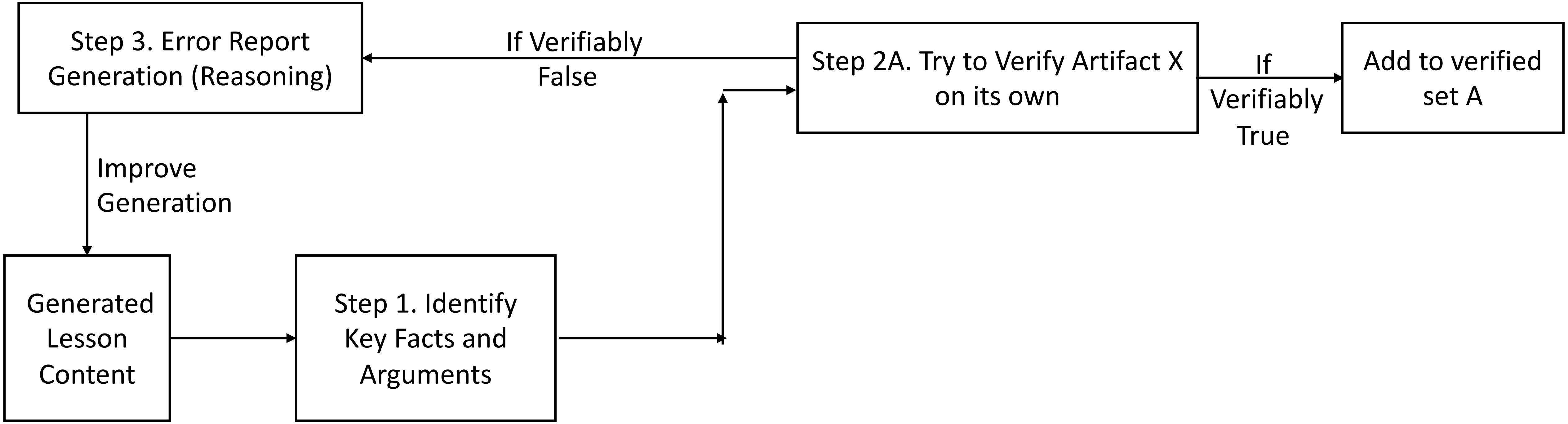
An Instantiation of Neuro-Symbolic Programming



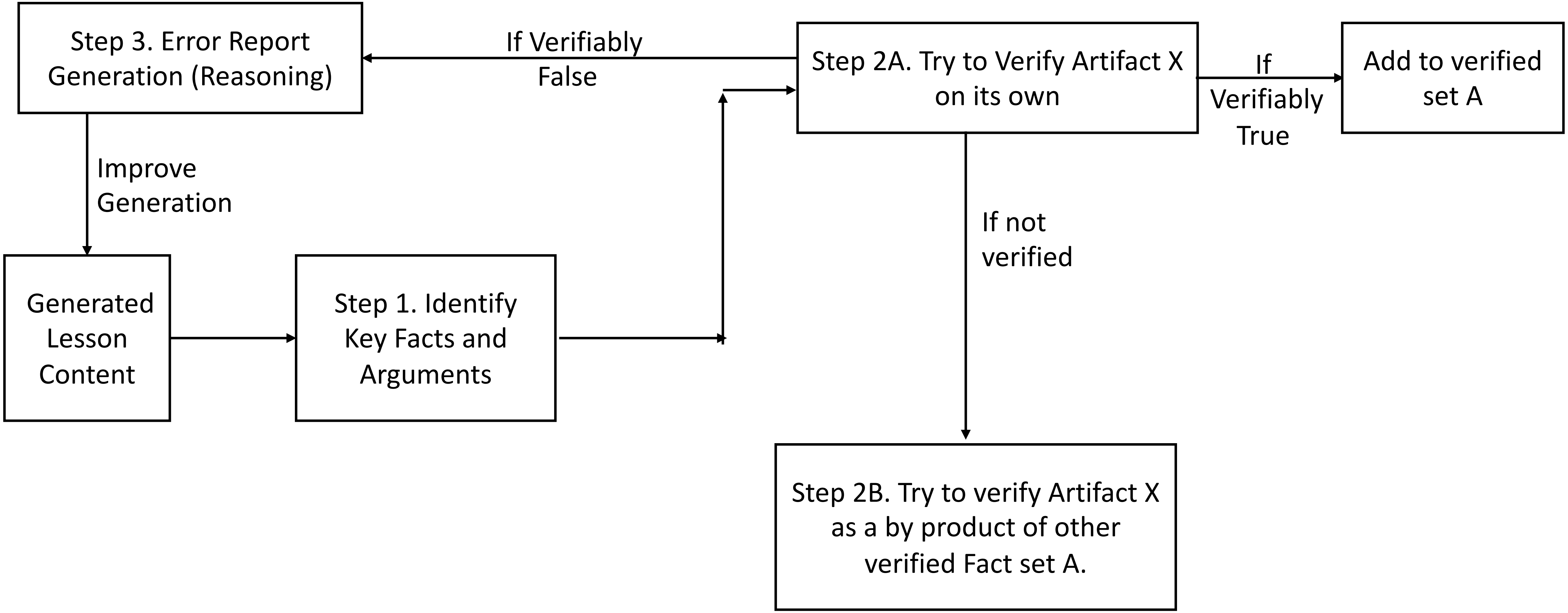
An Instantiation of Neuro-Symbolic Programming



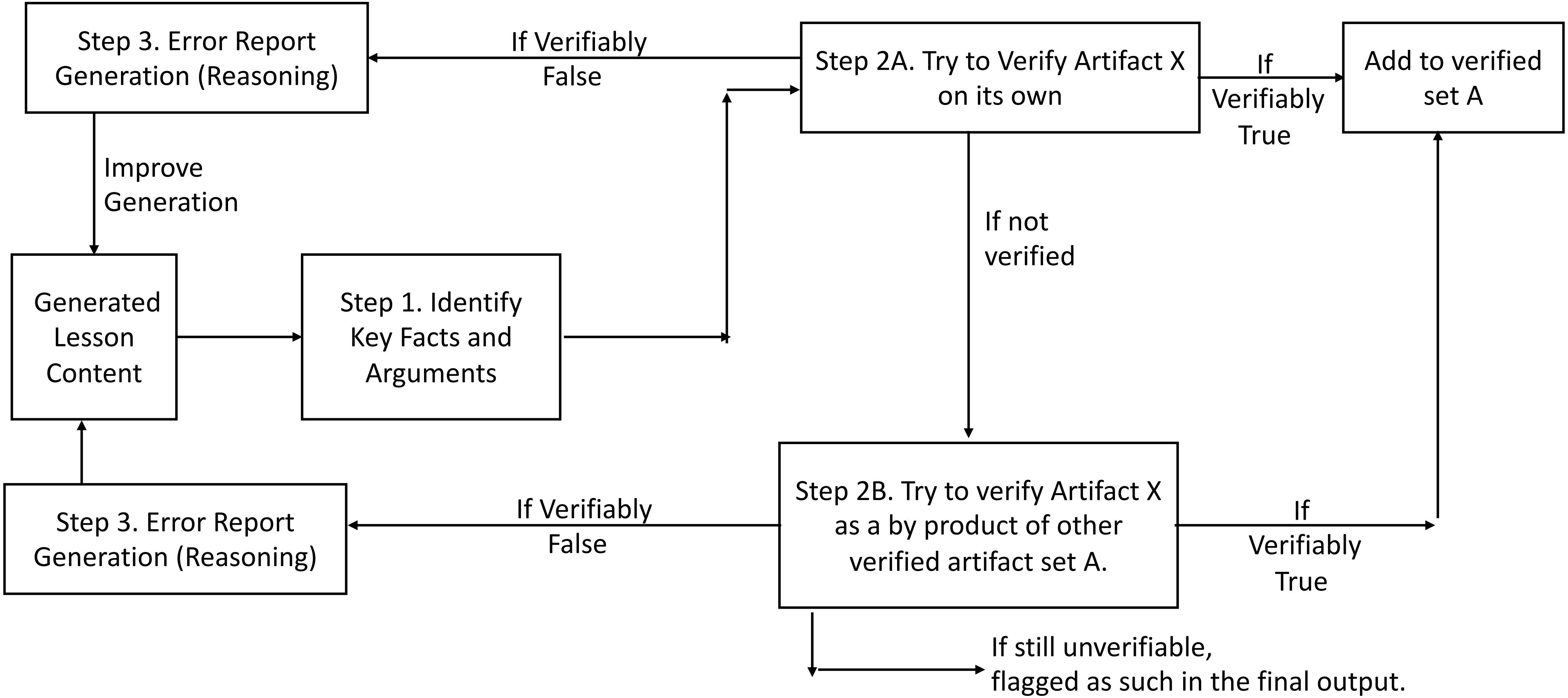
An Instantiation of Neuro-Symbolic Programming



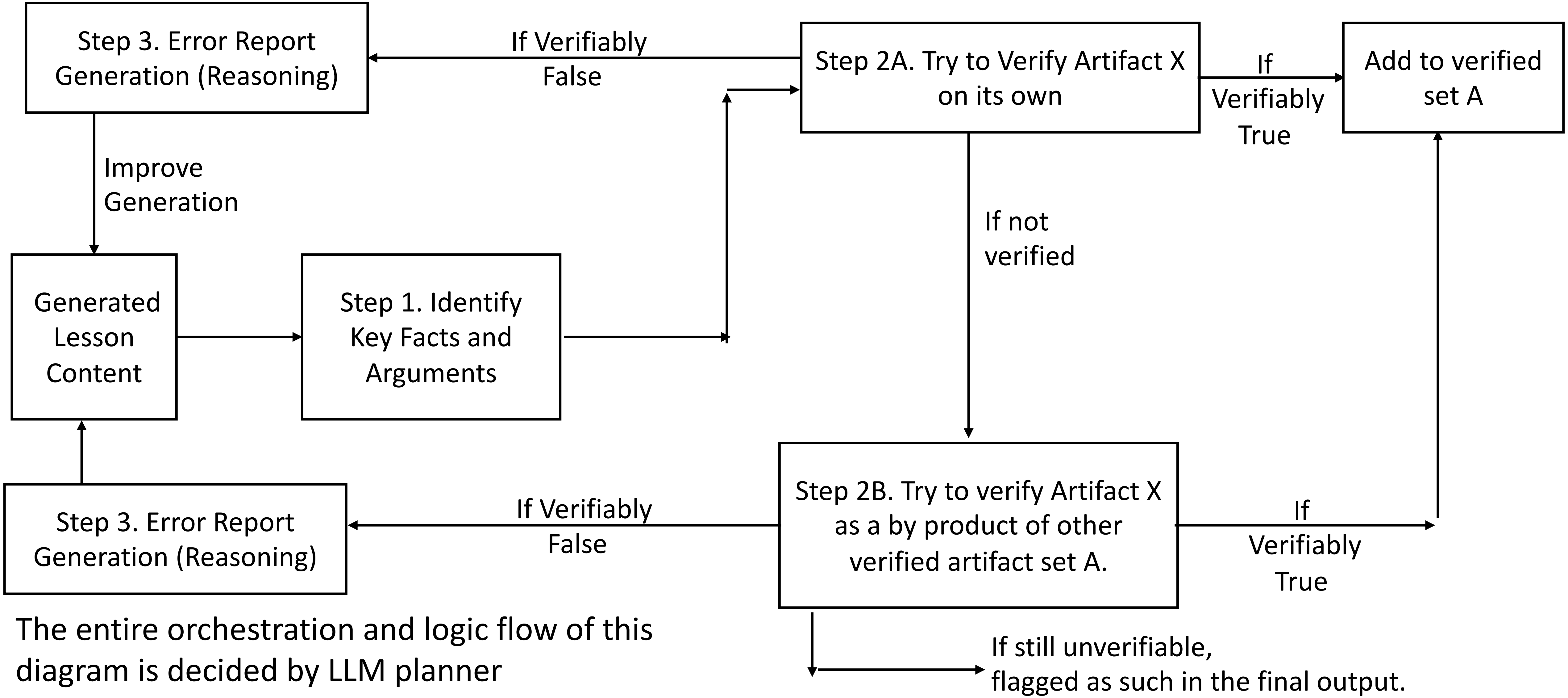
An Instantiation of Neuro-Symbolic Programming



An Example Instantiation of Neuro-Symbolic Programming



An Instantiation of Neuro-Symbolic Programming



The entire orchestration and logic flow of this diagram is decided by LLM planner

ML Sys in 2026? 2030?

MLSYS
or
NeurIPS etc

- Accurate Function Calling
- Data curation for fine-tuning
- Hallucination Elimination

MLSYS

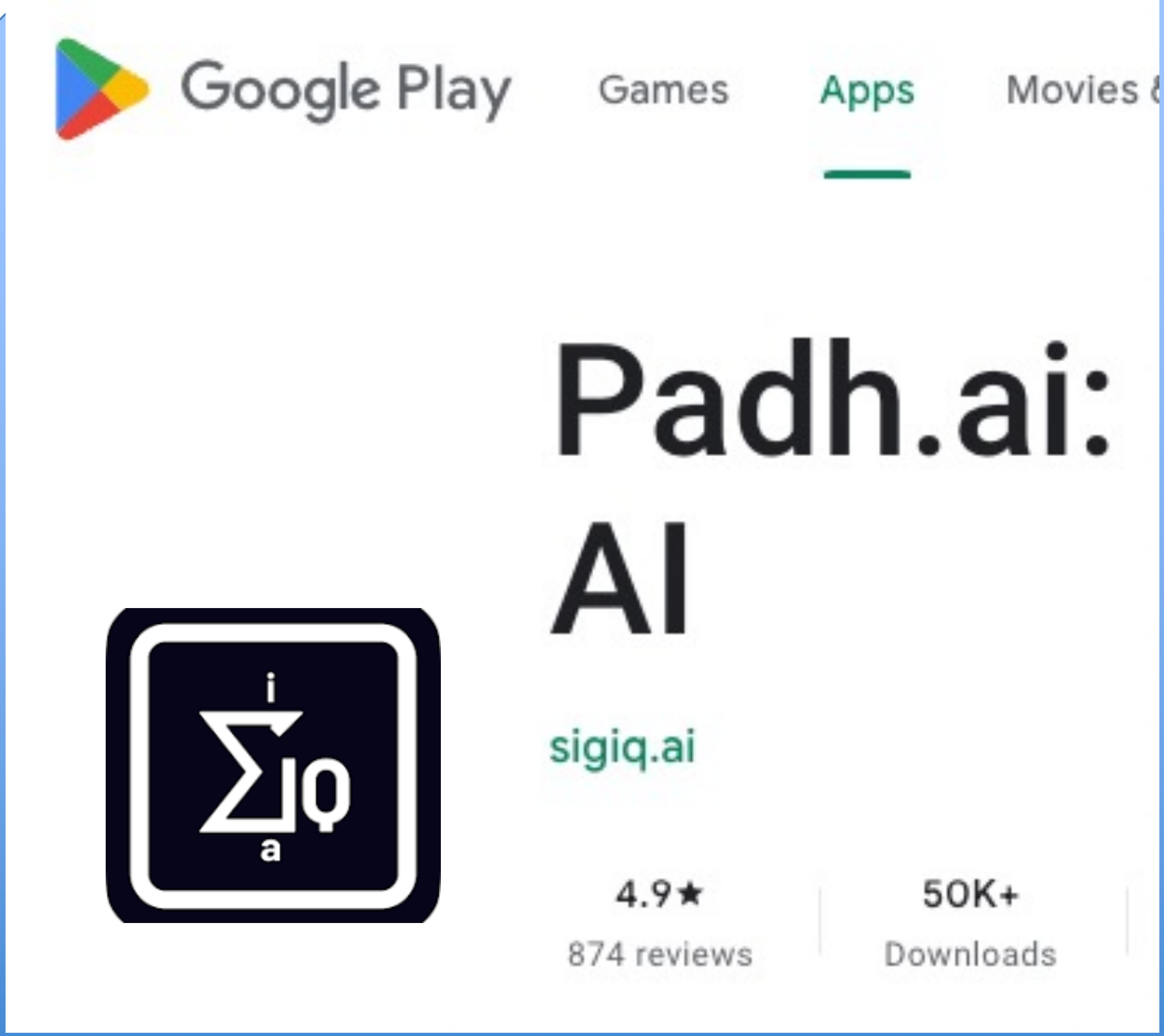
In the program today

- Quantization and Compression
- Fine-tuning

In the future?

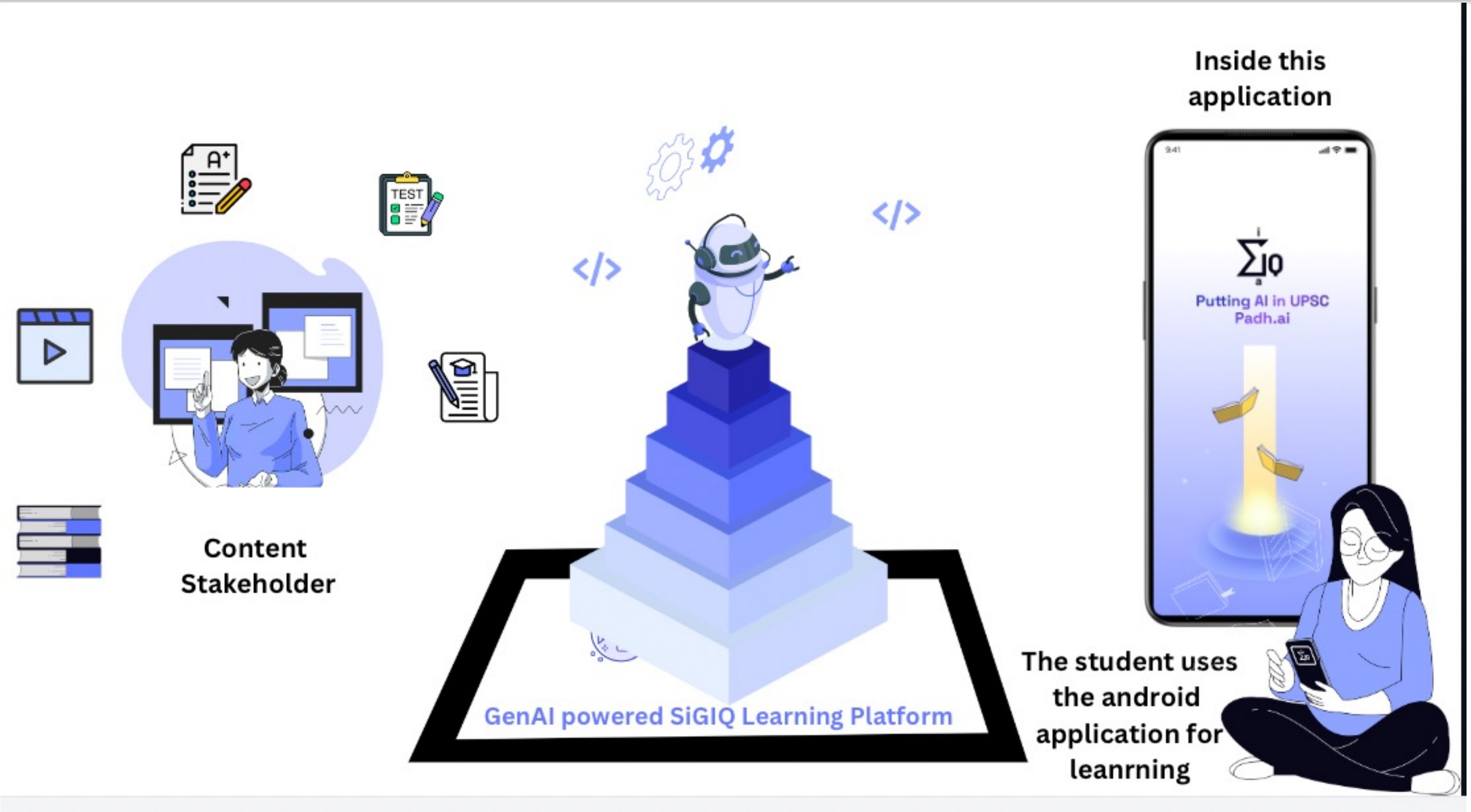
- Model Search for Compound Gen-AI Systems
- Efficient function calling/RAG
- Efficient Neuro-symbolic programming

Available at Google Play



•

SigiQ.ai is not an Application
It is an Application Generator!



"The internet reduced the cost of distribution to zero.
GenAI reduces the cost of *generation* to zero." Martin Casado, General Partner, a16z

For the Young Academic Professionals

- Perhaps you're already working on your own applications, on campus, in your labs, or at your startups

- But if not, where do you find one?:

- As for doing your own application you already have the:
  - GenAI expertise
  - Modest compute: 1-8 GPUs

- But what about the data?

Data is Everywhere Once You Know Where to Look

Many non-STEM (and some STEM) groups like

- Social scientists
- Humanities

Are sitting on lots of data, would love to have GenAI applications, but they don’t know how to create them

My Hobby Application:
Machine Translation for Low Resource Languages

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Image/video generation using **diffusion models**

Create a calendar invite with Lutfi and Sid for 2PM tomorrow

LLM Planner

Executor

Utilities

Productivity

JUL 17

Messaging

Personal Agents/Co-pilots

Machine Translation

Transliteration

Autodetect | Devanagari | Harvard-Kyoto | Other

Enter text to translate in Sanskrit, Tibetan, Buddhist scriptural Chinese or English...

Output language

English | Tibetan | Sanskrit | Other

Search for severe credit-card incidents for Docusign since yesterday and create incident tickets under Alice and Bob.

NexusRaven-V2-13B

Search Incidents

Filter Incidents

Create Tickets

Send Notifications

....

Identify Contacts

Assign Tickets

Enterprise Agents/Co-pilots

PRACTICE

Select Questions & Topic

- CSAT
- History
- Geography
- Polity
- Current Affairs
- Economy
- Science
- Environment

Go back

Begin Practice

Personal Teacher

Our Own Machine Translation System for Sanskrit, Tibetan and Scriptural Chinese

https://dharmamitra.org/

- Driven by Research Specialist Sebastian Nehrdich, this has been my pet project
- Involved in every aspect from low-level data cleaning to grand strategy
- Tibetan is the most mature and unique aspect of the project, but Sanskrit is the most sophisticated

A/The Key to Machine Translation is Data
Specifically in the form of sentence pairs

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Sanskrit Originals
- Unicode, pdfs etc

National Digital Library of India

BOOKS IN SANSKRIT ON AYURVEDA
HOME / BOOKS / SANSKRIT / AYURVEDA

eGangotri

MITRA VecAlign Text Aligner

~850,000 Sanskrit/English Sentence pairs

INTERNET SACRED TEXT ARCHIVE

WISDOM LIBRARY
peace · love · dharma

गीर्वाणि
giirvaaNi

English translations
- Unicode
- .epubs, pdfs
- Data cleaning and curation

Archive.org

Sentence pair:
<source, target>
<उद्यमो भैरवः,
Bhairava is the upsurge.>

- Identifying Sanskrit originals,
  - Downloading unicode
  - Downloading pdfs, OCR'ing, and editing
- Identifying English translations and similarly downloading, OCR'ing, and editing
- Using Sebastian Nehrdich's automated alignment system between Sanskrit and English

Our Professional Collaborators in India

Prof. Mitesh Khapra
IIT Madras, India
AI4Bharat

Prof. Pawan Goyal
Associate Professor,
IITKGP India

Dr. Jivnesh Sandhan
Visiting Assistant Professor
IIT Dharwad, India

Sujeet Kumar Jaiswal
PhD. IIT Kharagpur, India

MITRA Sanskrit Data Collection Team
Student (HS & UG) Volunteers: Berkeley

Present:

Kayshav Bhardwaj 

Kush Bhardwaj 

Aarnav Srivastava 

Rohan Sarakinti

Vinaya Sivakumar 

Varun Rao 

Devika Gopakumar 

Miranda Zhu, UG 

Frances Balleza 

Past:

Om Chandran

Siya Mehta 

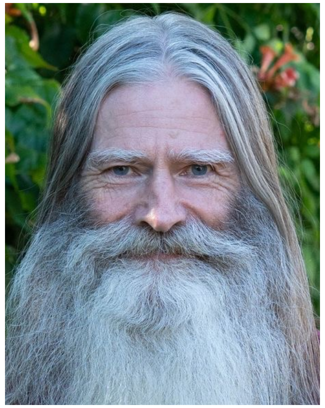
Rhea Rajendra 

Om Janamanchi 

Sriram Madanapalli 

Raj Virgink 

Sanjana Vipperla 



Team Lead:
Dr. David Allport
PhD, Sanskrit,
Oxford 1982

Fine-tuning Open-Source models
Now in its Third Generation Based on Llama 3

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

OliverHellwig / sanskrit

Sanskrit Texts in Unicode
Sanskrit texts with POS tagging

MITRA VecAlign
Text Aligner

INTERNET SACRED TEXT ARCHIVE

WISDOM LIBRARY
peace · love · dharma

गीर्वाणि
giirvaaNi

English translations

~850,000
Sanskrit/English
Sentence
pairs

Meta

LLAMA 3
Pre-trained on
30+ languages

Google

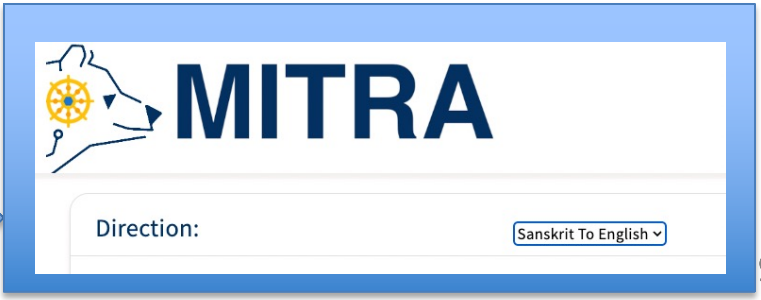
MADLad-400 , LLM
Pre-trained on
419 languages

Meta

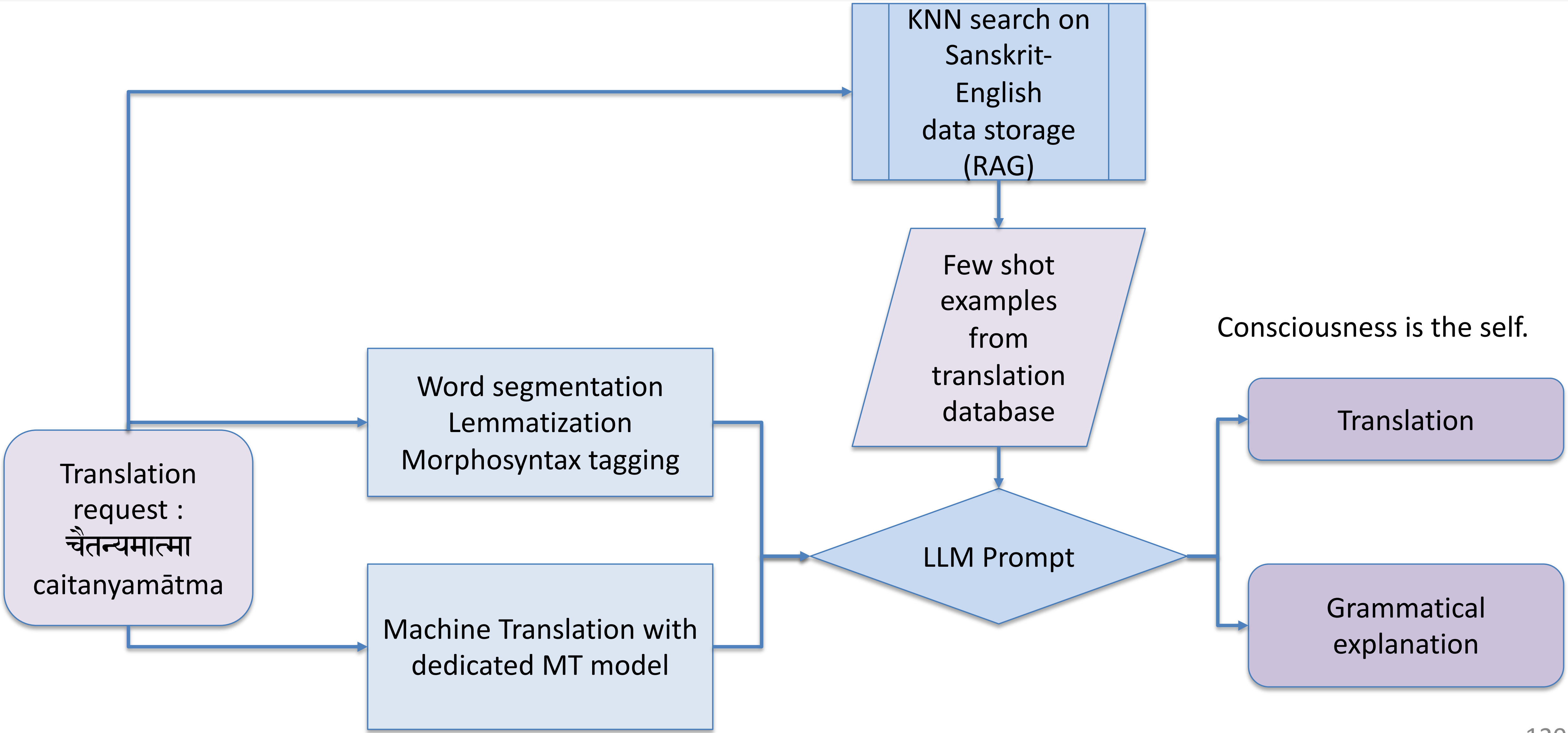
NLLB , LLM
Pre-trained on
200 languages

MITRA LLM
Fine-tuning

Sentence pair:
<उद्यमो भैरवः, Bhairava is the upsurge.>



The MITRA
Sanskrit Translation Workflow



Development of the Input Prompt

Prompt

Please translate this Sanskrit sentence into English:

रामो वनं गच्छति rāmo vanaṃ gacchati

Here is a list of reference sentences and their translation:

Sanskrit: ā enam śraddhā gacchati ainam yajñaḥ gachati ainam lokaḥ gacchati ainam annam gachati ainam annādyam gacchati yaḥ evam veda
Translation: Faith Sacrifice, the world, food and nourishment approach him who possesses this knowledge.
[…]

Here is a grammatical annotation of the sentence:
ram common noun case=nominative , number=singular, gender=masc
vana common noun case=accusative , number=singular, gender=neuter
gam finite verb tense=pres, mood=ind, person=plural, number=singular

Input comes from user:
रामो वनं गच्छति
rāmo vanaṃ gacchati

RAG: Using Semantic Similarity and kNN search we find 10-20 similar sentences in our translation database

Function calling: Using a morphological analyzer developed by Hellwig/Nehrdich, we give a morphological analysis

Can Supply Full Grammatical Analysis

Pali/Sanskrit

चैतन्यमात्मा

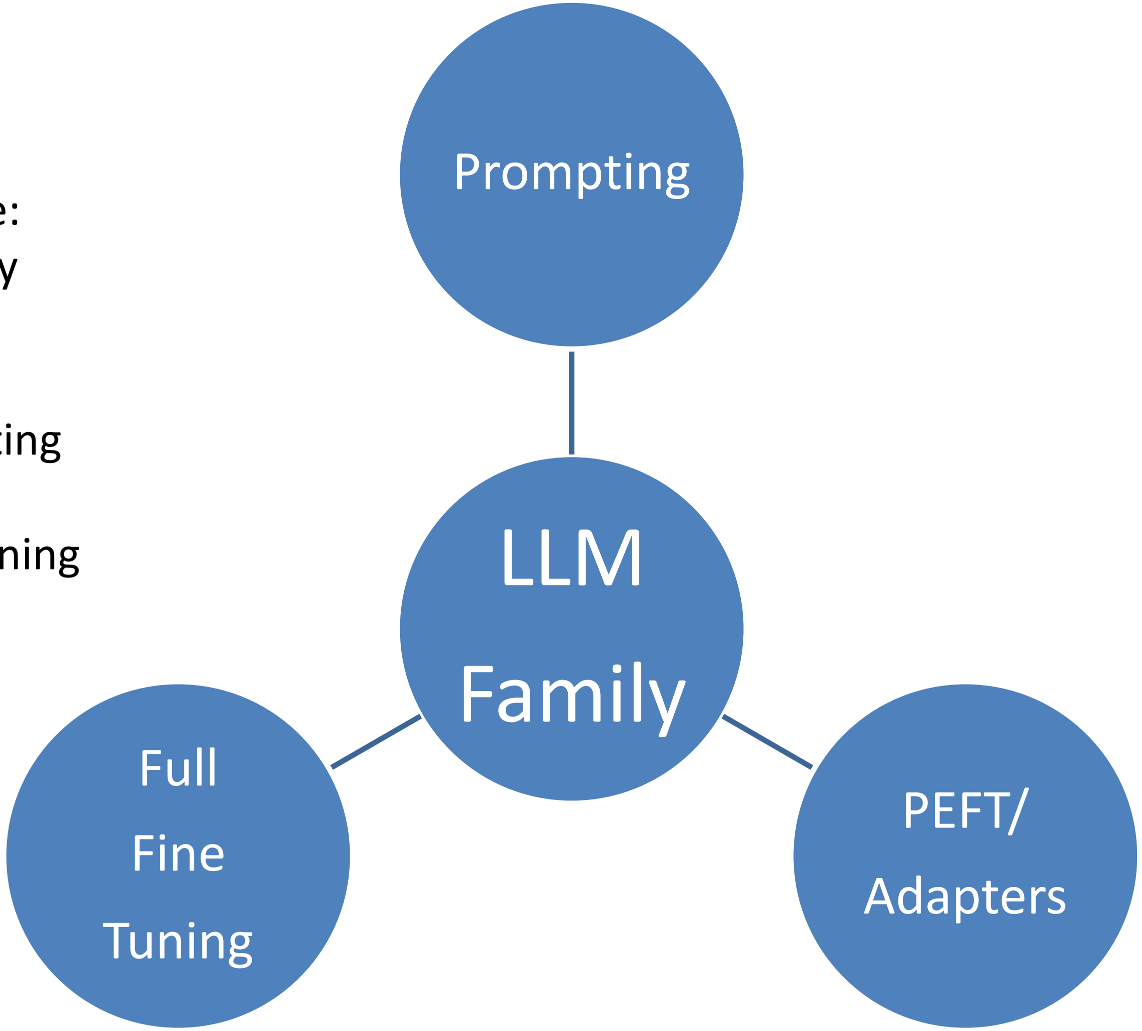
→ **Translation**

English Translation: Consciousness is the self.
Explanation of Structure and Meaning:- The Sanskrit compound "caitanyamātmā" is made up of two words: "caitanyam" and "ātmā."- "Caitanyam" comes from the stem "caitanya," which means "consciousness" or "sentience." It is in the nominative case, neuter gender, and singular number, indicating that it is the subject of the sentence.- "Ātmā" comes from the stem "ātman," which means "self" or "soul." It is in the nominative case, masculine gender, and singular number, also indicating that it is the subject of the sentence.- The compound "caitanyamātmā" is a Karmadhāraya compound, where the first member "caitanyam" (consciousness) qualifies the second member "ātmā" (self). In this type of compound, both words are in the nominative case because they refer to the same entity.- The translation "Consciousness is the self" reflects the grammatical relationship between "caitanyam" and "ātmā," where "consciousness" is equated with or identified as the "self." The sentence asserts that the essential nature of the self is consciousness.

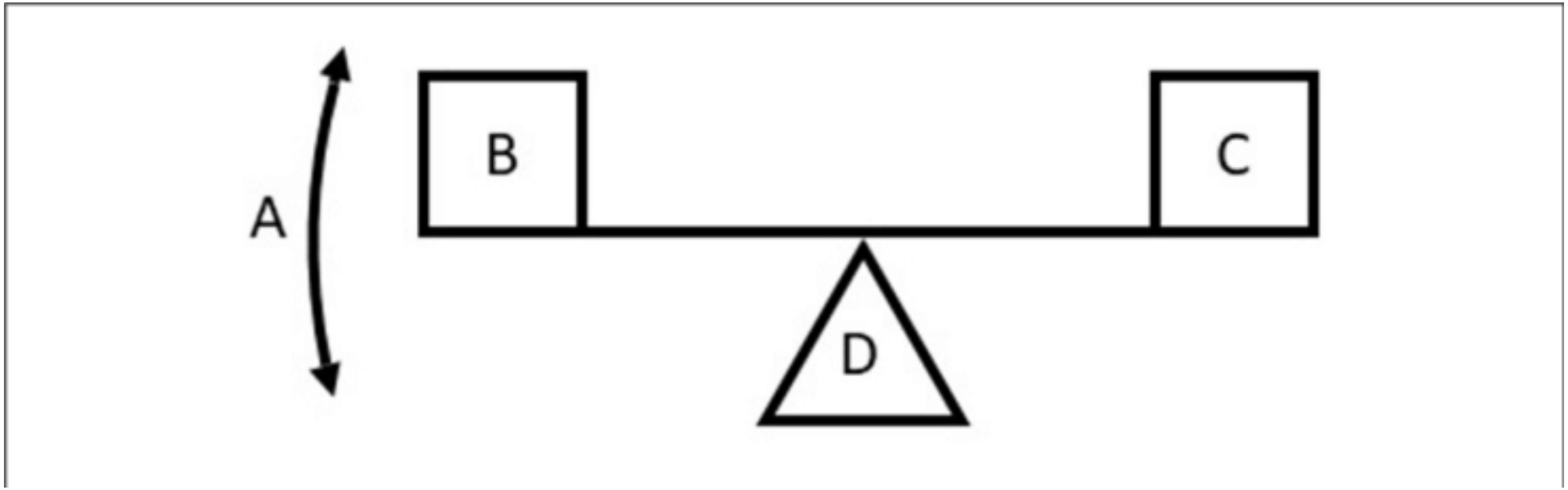
What Machine Translation is Teaching us About Design Space Exploration

Need to choose:

- Model family
- Model size
- Strategy
    - Prompting
    - PEFT
    - Fine-tuning



Trading off Application Performance and Computational Efficiency



- Applications inevitably bring
  - Requirements on application performance (aka accuracy)
  - Requirements on computational efficiency (e.g. latency)
  - Unfortunately, improving performance typically requires more computation

Choosing Model Size: Latency

Model	Year	Positional Encoding	Activation	Norm	Hidden Dim	# Heads	Head Dim	# Layers	MQA/ GQA	MoE	Time per token generated (s)*
Mistral (7B)	2023	RoPE	SwiGLU	RMSNorm	4096	32	128	32	Yes	No	0.01
Gemma (7B)	2024	RoPE	GeGLU	RMSNorm	3072	16	256	28	No	No	0.01
LLaMA (65B)	2023	RoPE	SwiGLU	RMSNorm	8192	64	128	80	No	No	0.06
LLaMA-3 (70B)	2024	RoPE	SwiGLU	RMSNorm	8192	64	128	80	Yes	No	0.06
Command R (104B)	2024	RoPE	SwiGLU	LayerNorm	12288	96	128	64	Yes	No	0.10
DBRX (132B)	2024	RoPE	SwiGLU	LayerNorm	6144	48	128	40	Yes	Yes	0.03
GPT-3 (175B)	2020	Absolute	GELU	LayerNorm	12288	96	128	96	No	No	0.16
Falcon (180B)	2023	RoPE	GELU	LayerNorm	14848	64	64	80	Yes	No	0.16
PaLM (540B)	2022	RoPE	SwiGLU	LayerNorm	18438	48	256	118	Yes	No	0.49

Latency of models above was estimated by:
Model_size_in_bytes/Nvidia_A100_bandwidth
or
2 *model_size /(2039) * (1024^3))

Choosing Strategy

Economy of Finetuning

- Prompting (more economical than)
- PEFT >
- Finetuning

Economy of Inference

- Fine-tuning> (more economical than)
- Prompting

Flexibility

- Prompting > (is more flexible than)
- PEFT >
- fine-tuning

Model Performance

- fine-tuning > (gives better performance than)
- PEFT >
- prompting (alone)

ML Sys in 2026? 2030?

MLSYS
or
NeurIPS etc

- Accurate Function Calling
- Data curation for fine-tuning
- Hallucination Elimination

MLSYS

In the program today
- Quantization and Compression
- Fine-tuning
- (PEFT) Parameter-efficient fine-tuning

In the future?
- Model Search for Compound Gen-AI Systems
- Efficient function calling/RAG
- Efficient Neuro-symbolic programming

There's So Much More to Talk About:Diffusion
A Big Portion of Future System Workloads

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Image/video generation using **diffusion models**

Create a calendar invite with Lutfi and Sid for 2PM tomorrow

LLM Planner

Executor

Utilities

Productivity

Messaging

Personal Agents/Co-pilots

Machine Translation

Transliteration
Autodetect Devanagari Harvard-Kyoto Other

Enter text to translate in Sanskrit, Tibetan, Buddhist scriptural Chinese or English...

Output language
English Tibetan Sanskrit Other

Search for severe credit-card incidents for Docusign since yesterday and create incident tickets under Alice and Bob.

NexusRaven-V2-13B

Search Incidents

Filter Incidents

Create Tickets

Send Notifications

....

Identify Contacts

Assign Tickets

Enterprise Agents/Co-pilots

PRACTICE

Select Questions & Topic

- CSAT
- History
- Geography
- Polity
- Current Affairs
- Economy
- Science
- Environment

Go back

Begin Practice

Personal Teacher

So Much I Didn't Talk About

- Diffusion: Images, Videos, soon 3D
- Partial Fine-tuning Methods:
  - Low-Rank Adapters
    - Large Language Models 1: Tuesday 1:30PM
      - SLoRA: Scalable Serving of Thousands of LoRA Adapters
- Retrieval Augmented Generation (RAG) on a Large Scale
  - Legal, financial, research
  - Large Context Windows
  - Prompting and Prompting Optimization
- Reinforcement Learning with Human Feedback
- GenAI at the Edge
- …

Summary:
For Young Professionals

- Always strive to find a new perspective
  – It’s worth 80 IQ points!
- How? Get close to real world applications
  – Don’t stop at standard benchmarks
- Develop a toolkit/playbook of techniques and use your perspective identify the right places to apply them
- Startup tip:
  – “The internet reduced the cost of distribution to zero.
  – GenAI reduces the cost of generation to zero.”
    - Martin Casado, General Partner, a16z

Summary:
GenAI is About More than Models -1

- Much of the focus on Large Language Models has been on their ability to process natural language input and generate an interesting output based on their parametric knowledge
- The real power of LLMs is shown when their ability to retrieve data (RAG) and manipulate tools (through function calling) to create Compound GenAI Systems
- Agents/co-pilots are the exemplar of this trend
  - Soon most of our daily and business lives will by co-piloted by agents

Summary:
GenAI is About More than Models - 2

- Explicitly I showed you a number of Compound GenAI Systems
- Perhaps more importantly, implicitly I showed you that the bar to creating valuable consumer and enterprise applications has never been lower
  - Even 5 years ago the systems I showed you would have required 5-10x more programing effort!

ML Sys in 2026? 2030?

MLSYS
or
NeurIPS etc

- Accurate Function Calling
- Data curation for fine-tuning
- Hallucination Elimination

MLSYS

In the program today

- Quantization and Compression
- Fine-tuning
- (PEFT) Parameter-efficient fine-tuning

In the future?

- Efficient diffusion
- Model Search for Compound Gen-AI Systems
- Efficient function calling/RAG
- Efficient Neuro-symbolic programming
- Larger context windows
- Prompt compression
- GenAI at the edge
- …

Thanks to My
Research Team in BAIR

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

As well as colleagues at:

- Nexusflow.ai: Venkat Srinivasan, Jian Zhang
- SigIQ.ai: Karttikeya Mangalam

A Final Question

Machine Learning/Deep Learning have rapidly evolved through a number of eras:

- ML Era 1: Orchestration of statistics gave us **Machine Learning**
- ML Era 2: Orchestration of Machine Learning algorithms gave us **Neural Nets**
- ML Era 3: Orchestration of Neural Net model functions/components gave us the **Transformer**
- ML Era 4: Orchestration of Transformers gave us **Large Language Models**
- ML Era 5: Orchestration of Large Language Models gives us **Compound GenAI Systems**
- **ML Era 6: Orchestration of Compound GenAI systems give us …. What?**