RESEARCH CHALLENGES FOR LARGE PRE-TRAINED MODELS

CDAO Advantage DoD 24	FEB 2024

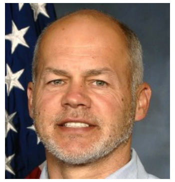
Distribution Statement:	A. Approved for Public Release. Distribution Unlimited
POC:	Celso de Melo, celso.m.demelo.civ@army.mil

MILITARY INFORMATION SCIENCES, DR. CELSO DE MELO

ACKNOWLEDGMENTS

Dr. Scott Clouse

Dr. Edward Verenich



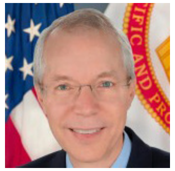
Dr. Steven Rogers

Dr. Reginald Hobbs

Dr. John Fossaceca

Dr. Raghuveer Rao

Mr. Dietrich Wiegmann



Dr. Brian Sadler

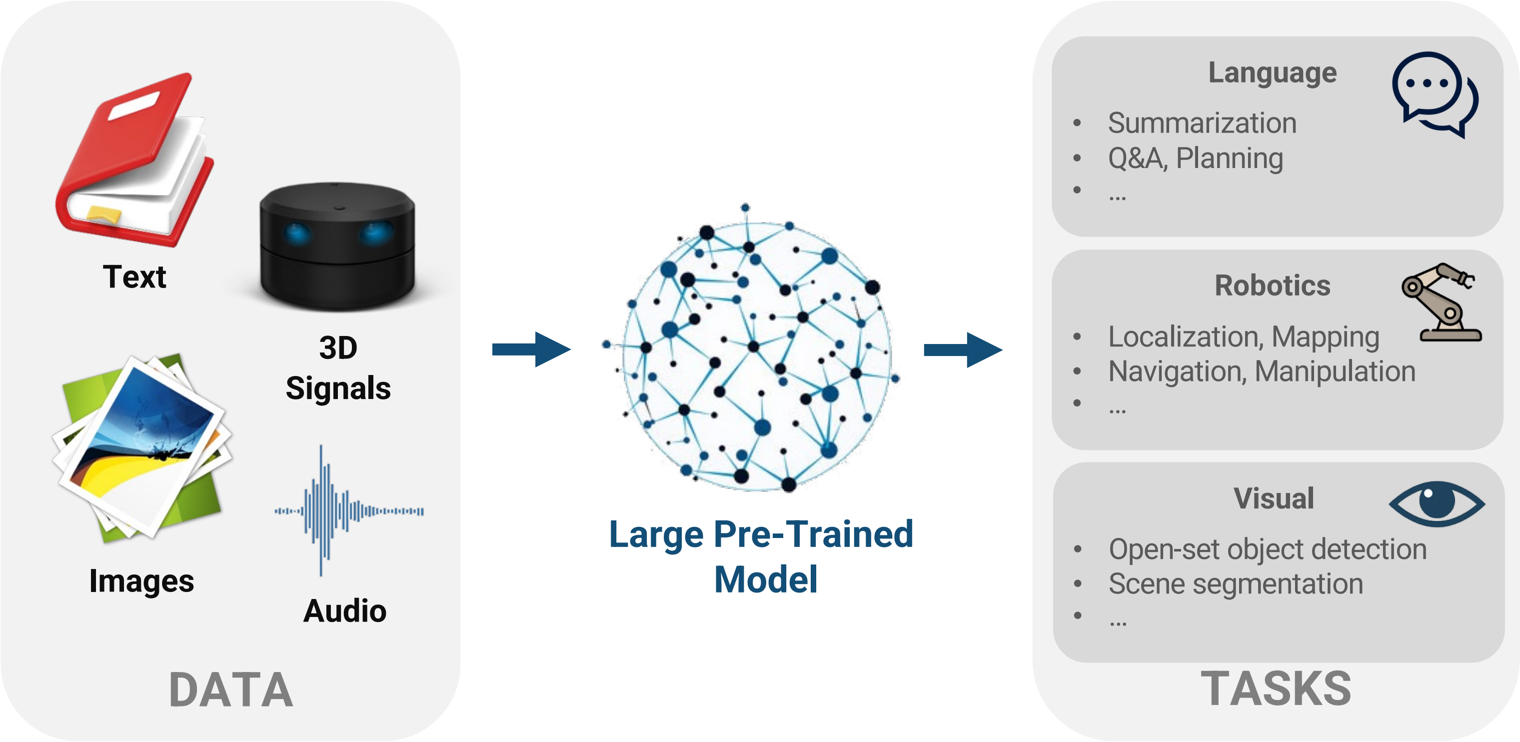
Dr. Leslie Smith

Dr. David Aha

Dr. John Long

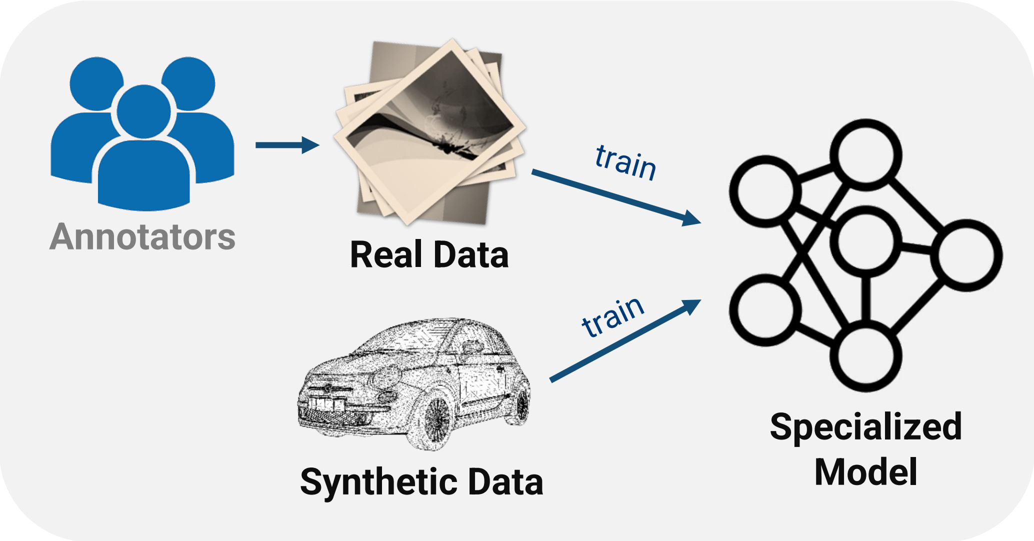
LARGE PRE-TRAINED MODELS FOR DOD AI

- LPTMs (e.g., GPT-4) have shown remarkable emergent capability relevant to multitude of DoD use cases
- They are trained on large quantities of unlabeled data (scale + self-supervision) and adapted to downstream tasks (transfer learning)



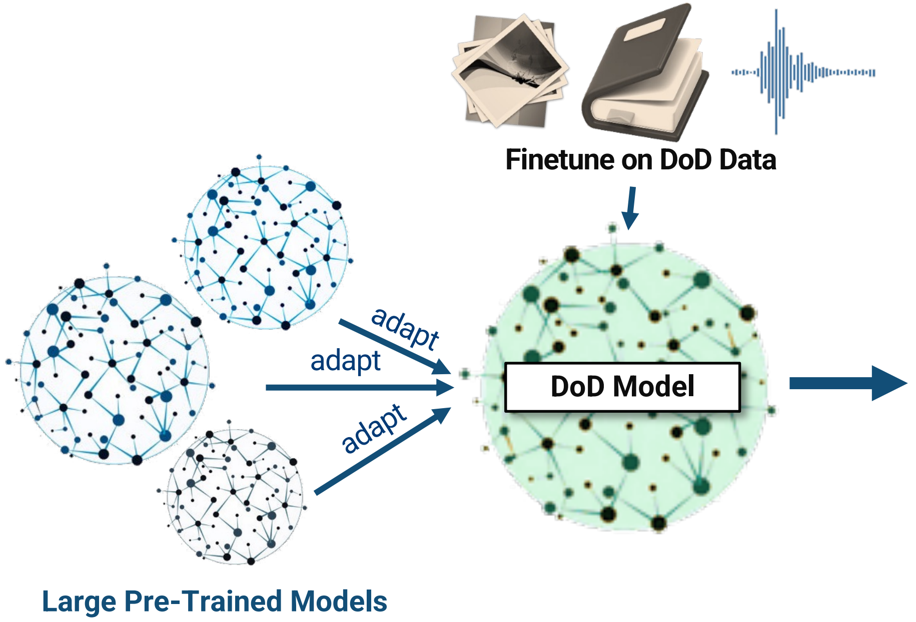
LARGE PRE-TRAINED MODELS FOR DOD AI

- LPTMs (e.g., GPT-4) have shown remarkable emergent capability relevant to multitude of DoD use cases
- They are trained on large quantities of unlabeled data (scale + self-supervision) and adapted to downstream tasks (transfer learning)
- Old paradigm consists of training specialized models on labeled (real/synthetic) datasets



LARGE PRE-TRAINED MODELS FOR DOD AI

- LPTMs (e.g., GPT-4) have shown remarkable emergent capability relevant to multitude of DoD tasks
- They are trained on large quantities of unlabeled data (scale + self-supervision) and adapted to downstream tasks (transfer learning)
- LPTMs introduce novel paradigm for AI systems where starting point are these models
- ARL hosted scientific meeting on opportunities, challenges and applications of LPTMs (Nov 14-16, 2023)
  - Broad engagement from DoD (e.g., Army, Air Force, Navy, CDAO, OUSD R&E), Academia (e.g., MIT, Stanford, UW, UC Berkeley), and Industry (e.g., Microsoft, Google, NVIDIA, Meta, Scale AI)



Language Tasks
- Summarization
- Q&A, Planning
- …

Robotics Tasks
- Localization, Mapping
- Navigation, Manipulation
- …

Visual Tasks
- Open-set object detection
- Scene segmentation
- …

BUILDING DOD, INDUSTRY, AND ACADEMIA RESEARCH ECOSYSTEM

DEVCOM ARMY RESEARCH LABORATORY

scale Donovan

Stanford Alpaca

OpenAI | Microsoft

GPT-4V

Google BERT

LLaMA

NVIDIA AI

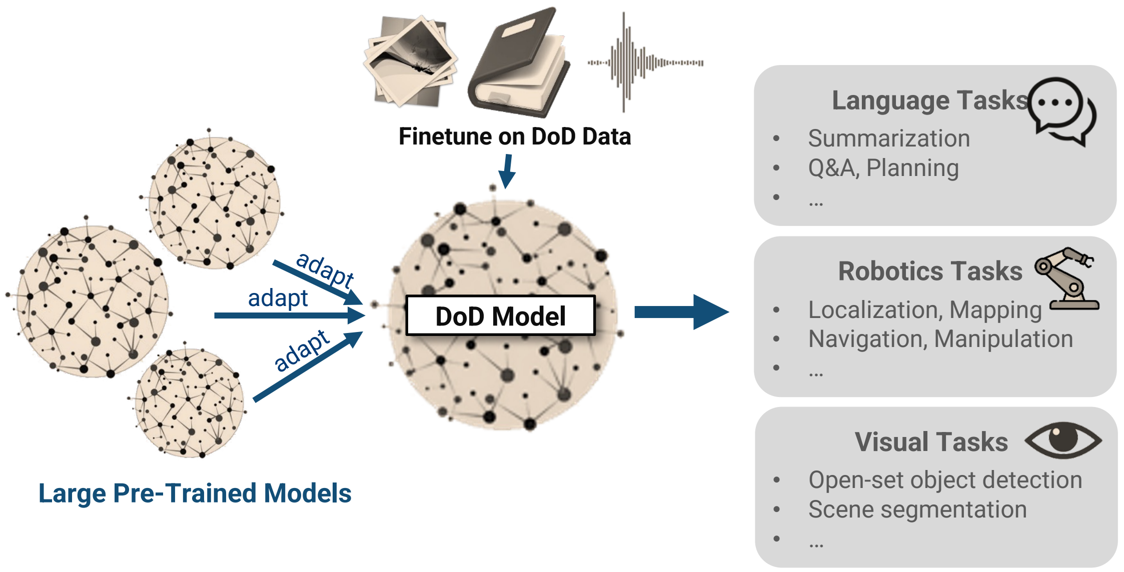
AIP

What is the role of DoD?

What is compute infrastructure to support this ecosystem?

RESEARCH CHALLENGES

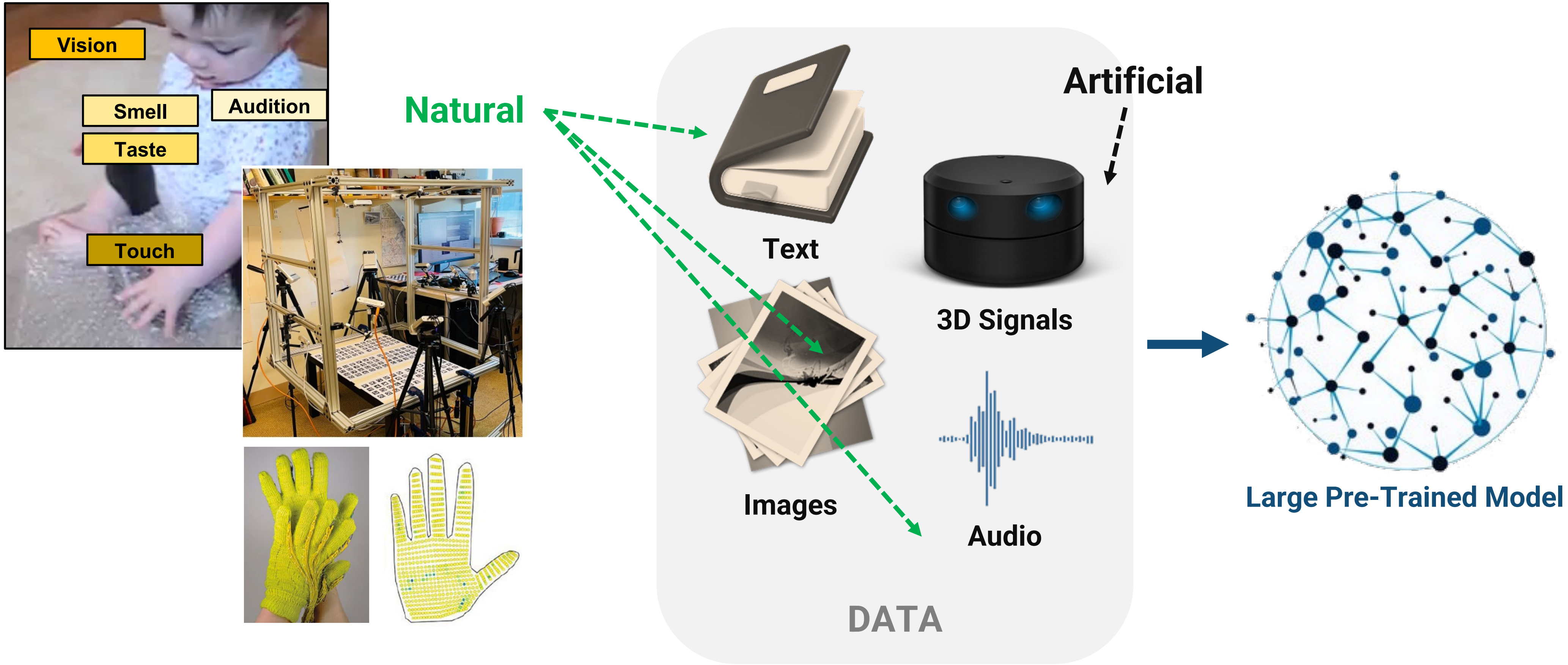
- Multimodal not just language
- Knowledge distillation
- Deployment at the edge
- Data starvation, continual learning & synthetic data
- Adaptation & finetuning
- Reasoning & Scientific Experimentation



- Interpretability
- Data provenance & hallucinations
- AI safety & alignment
- System-of-systems
- Benchmarking

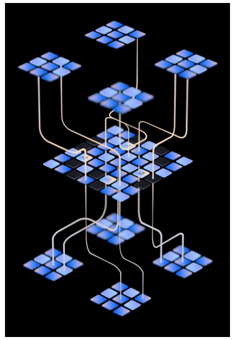
MULTIMODALITY

- Biological systems (e.g., children) learn rich multimodal knowledge about world
- Multimodal latent representations lead to robustness and generalization in novel tasks
- Research needed on methods to get multimodal data and train/compose multimodal models

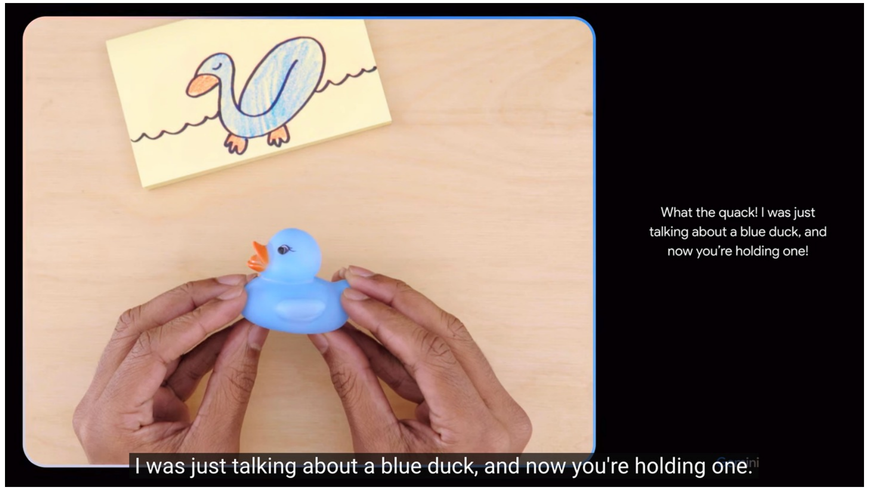


MULTIMODALITY

- <u>Multimodal models will enable open-world perception, reasoning, and action capability</u>
- First generation of multimodal models is becoming available (e.g., GPT-4v and Gemini)
- But, still unlikely to meet all DoD's multimodal needs (e.g., physics-based grounding missing)



Large multimodal model with unified latent space



MULTIMODAL

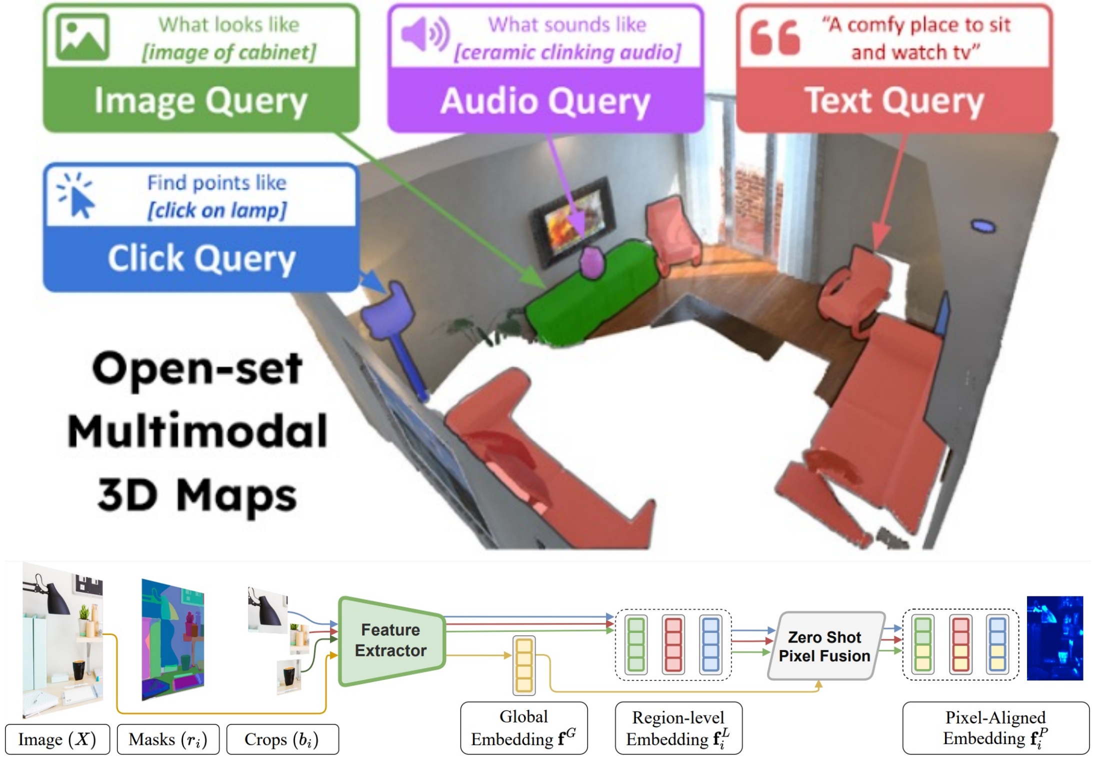
Capability	Benchmark	Description Higher is better unless otherwise noted	Gemini	GPT-4V Previous SOTA model listed when capability is not supported in GPT-4V
Image	MMMU	Multi-discipline college-level reasoning problems	59.4% 0-shot pass@1 Gemini Ultra (pixel only*)	56.8% 0-shot pass@1 GPT-4V
	VQAv2	Natural image understanding	77.8% 0-shot Gemini Ultra (pixel only*)	77.2% 0-shot GPT-4V
Video	VATEX	English video captioning (CIDEr)	62.7 4-shot Gemini Ultra	56.0 4-shot DeepMind Flamingo
	Perception Test MCQA	Video question answering	54.7% 0-shot Gemini Ultra	46.3% 0-shot SeViLA
Audio	CoVoST 2 (21 languages)	Automatic speech translation (BLEU score)	40.1 Gemini Pro	29.1 Whisper v2
	FLEURS (62 languages)	Automatic speech recognition (based on word error rate, lower is better)	7.6% Gemini Pro	17.6% Whisper v3

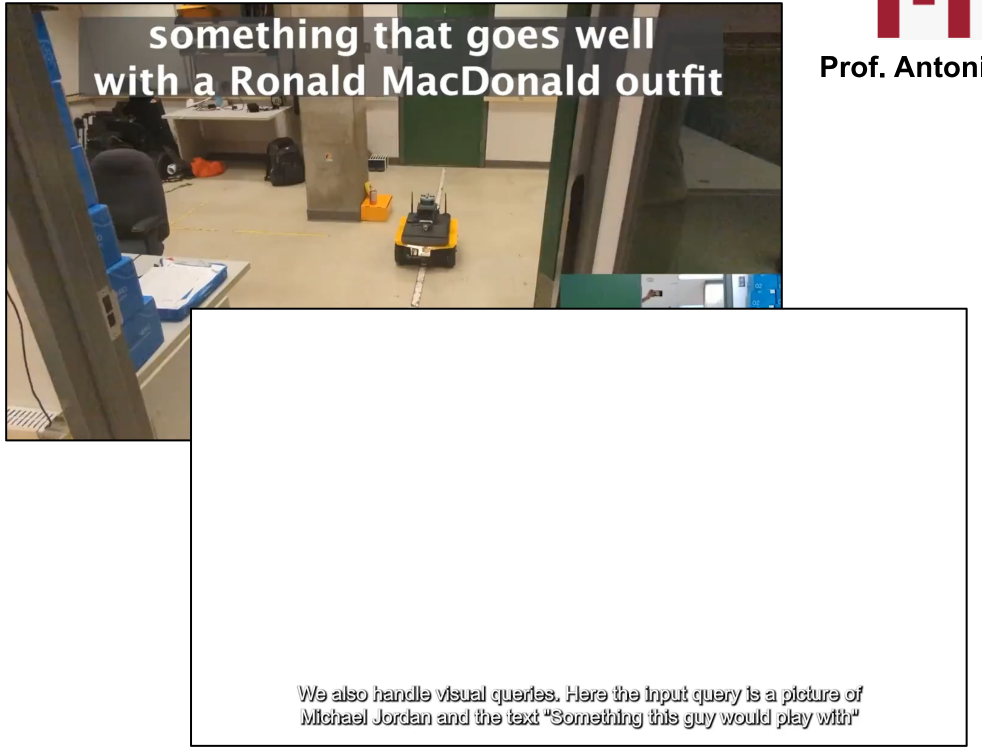
MULTIMODALITY

- Multimodal models will enable open-world perception, reasoning, and action capability
- First generation of multimodal models is becoming available (e.g., GPT-4v and Gemini)
- But, still unlikely to meet all DoD's multimodal needs (e.g., physics-based grounding missing)
- Given diversity of ecosystem, essential to research modular composable architectures

Prof. Antonio Torralba

ARL-MIT Collaboration

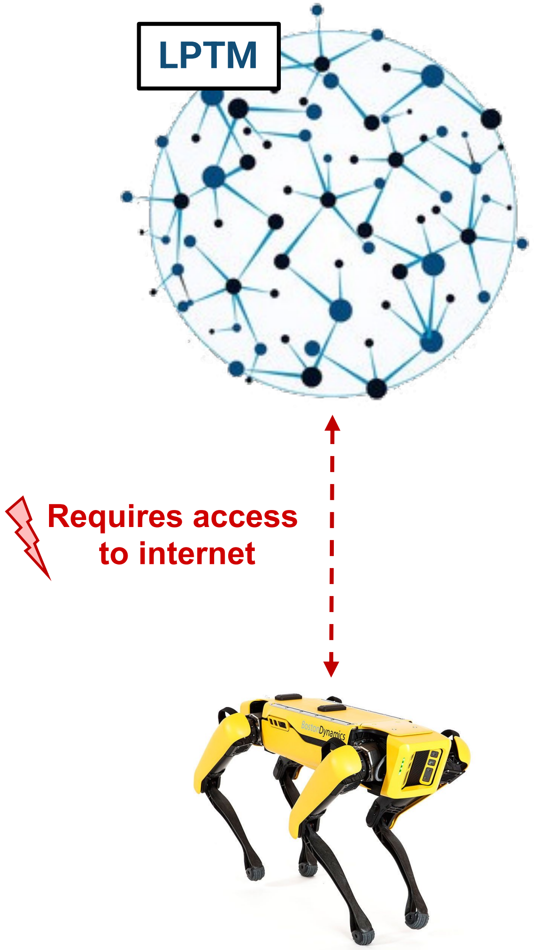


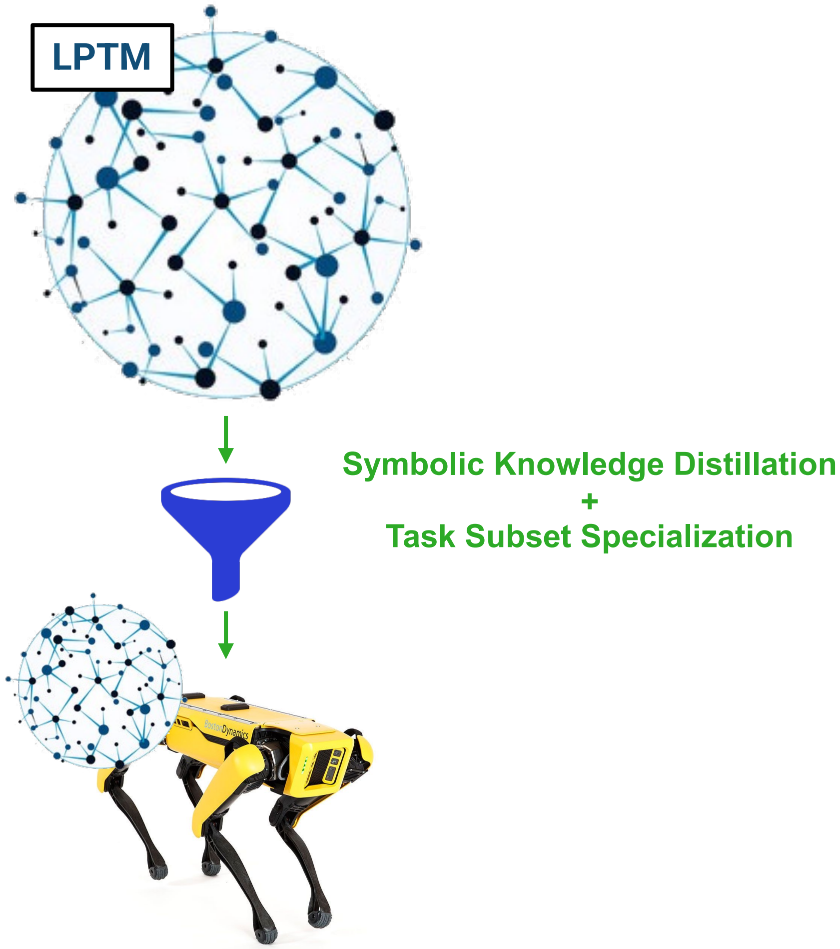


Gu et al. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning, ICRA 2024, https://concept-graphs.github.io/

KNOWLEDGE DISTILLATION & DEPLOYMENT AT THE EDGE

- Deploying LPTMs at the edge is problematic due to compute and communication limitations
- Symbolic knowledge distillation aims to create smaller models, from LPTMs, with similar performance





KNOWLEDGE DISTILLATION & DEPLOYMENT AT THE EDGE

DEVCOM ARMY RESEARCH LABORATORY

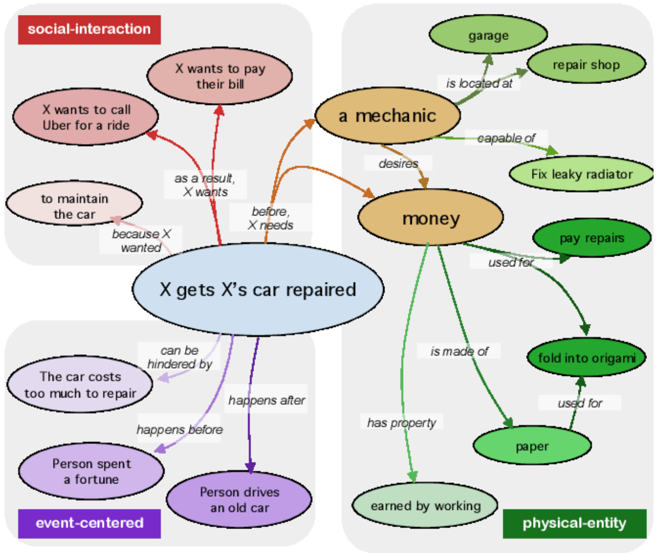
- Deploying LPTMs at the edge is problematic due to compute and communication limitations
- Symbolic knowledge distillation aims to create smaller models, from LPTMs, with similar performance
- Recent methods show that LPTM-guided distillation can outperform human-guided distillation, even leading to improvement in performance when compared to larger teacher model

UNIVERSITY of WASHINGTON

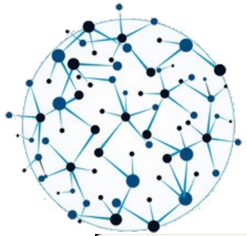
Prof. Yejin Choi

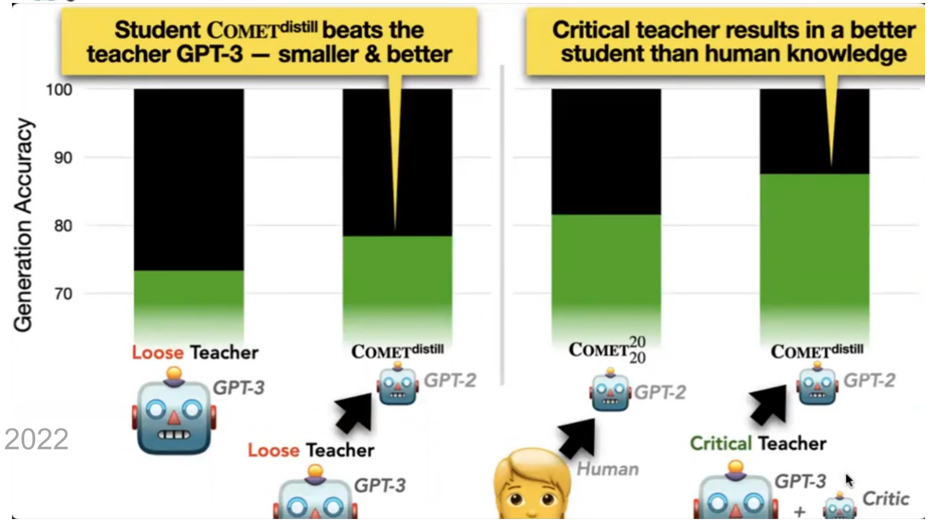
GPT-3
175B params

ATOMIC[10x]
Knowledge Model
Automatically trained from GPT-3



COMET[distill]
~400M params (400x smaller)
TASK: Commonsense causal reasoning



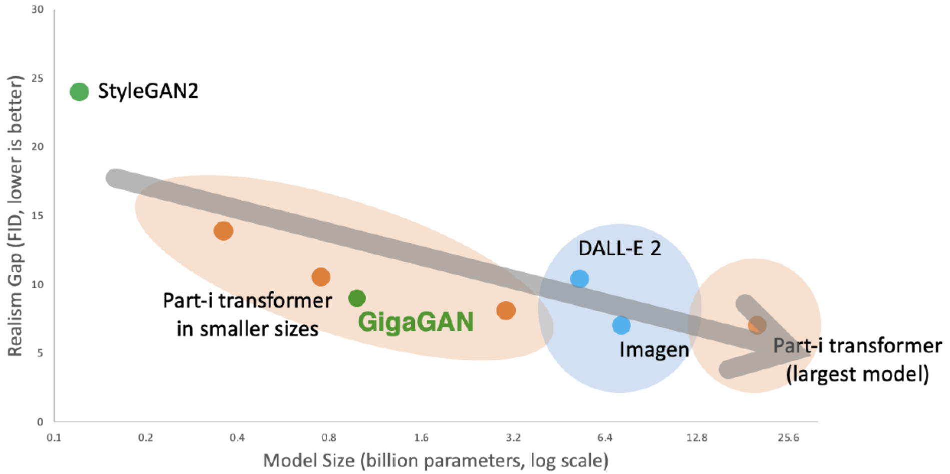


West et al. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models, NAACL 2022

DATA STARVATION, CONTINUAL LEARNING & SYNTHETIC DATA

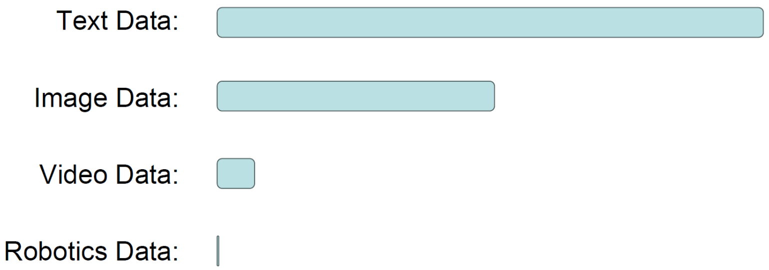
- High quality data leads to high quality LPTM output
- We are reaching the limits of available data – how do we ensure LPTMs can continuously adapt?

1 | Larger models attain better realism



2 | Generative data can be traced to data samples

3 | Reaching limits of existent data, especially for robotics



DATA STARVATION, CONTINUAL LEARNING & SYNTHETIC DATA

- High quality data leads to high quality LPTM output
- We are reaching the limits of available data – how do we ensure LPTMs can continuously adapt?
- Synthetic data offers opportunity to create high quality data, including generated by LPTMs

Trends in Cognitive Sciences
Available online 23 December 2021
In Press, Corrected Proof
CellPress

Review

Next-generation deep learning based on simulators and synthetic data

Celso M. de Melo [1], Antonio Torralba [2], Leonidas Guibas [3], James DiCarlo [4], Rama Chellappa [5], Jessica Hodgins [6]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.tics.2021.11.008 Get rights and content

Highlights

Despite their initial successes, it is becoming apparent that modern deep learning (DL) models are hindered by an important bottleneck: the need for large quantities of annotated data to train the models.

Synthetic data provide a solution to this challenge. They are easy to generate, error-free, inexhaustible,

DATA STARVATION, CONTINUAL LEARNING & SYNTHETIC DATA

Berkeley
UNIVERSITY OF CALIFORNIA

Prof. Alexei Efros

- High quality data leads to high quality LPTM output
- We are reaching the limits of available data – how do we ensure LPTMs can continuously adapt?
- Synthetic data offers opportunity to create high quality data, including generated by LPTMs

(1) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* → GPT-3 (finetuned) → **Instruction:** *"have her ride a dragon"*
Edited Caption: *"photograph of a girl riding a dragon"*

(2) Generate paired images:

Input Caption: *"photograph of a girl riding a horse"*
Edited Caption: *"photograph of a girl riding a dragon"* → Stable Diffusion + Prompt2Prompt →

Generated training examples:

At inference, generalizes to real images and human-written instructions

"have her ride a dragon"

"Color the cars pink"

"Make it lit by fireworks"

"convert to brick"

...

Brooks et al. InstructPix2Pix: Learning to Follow Image Editing Instructions, CVPR 2023

DATA STARVATION, CONTINUAL LEARNING & SYNTHETIC DATA

- High quality data leads to high quality LPTM output
- We are reaching the limits of available data – how do we ensure LPTMs can continuously adapt?
- Synthetic data offers opportunity to create high quality data, including generated by LPTMs
- Continual learning will further rely on self-supervision + interactive world exploration

Self-Supervision + **Exploration**

Multimodal redundancy provides knowledge about the world

Autonomous interactive exploration of environment leads to self-learning

INTERPRETABILITY & SCIENTIFIC EXPERIMENTATION

DEVCOM ARMY RESEARCH LABORATORY

- Explaining LTPM behavior is challenging, but LPTMs also enable autonomous interpretation
- LPTMs are increasingly capable of generating and evaluating hypotheses, using tools and showing the kind of reasoning seen in scientific experimentation

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Abstract

We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called *chain-of-thought prompting*, where a few chain of thought demonstrations are provided as exemplars in prompting.

Experiments on three large language models show that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

Let's Verify Step by Step

Hunter Lightman* Vineet Kosaraju* Yura Burda* Harri Edwards
Bowen Baker Teddy Lee Jan Leike John Schulman Ilya Sutskever
Karl Cobbe*

OpenAI

Abstract

In recent years, large language models have greatly improved in their ability to perform complex multi-step reasoning. However, even state-of-the-art models still regularly produce logical mistakes. To train more reliable models, we can turn either to outcome supervision, which provides feedback for a final result, or process supervision, which provides feedback for each intermediate reasoning step. Given the importance of training reliable models, and given the high cost of human feedback, it is important to carefully compare the both methods. Recent work has already begun this comparison, but many questions still remain. We conduct our own investigation, finding that process supervision significantly outperforms outcome supervision for training models to solve problems from the challenging MATH dataset. Our process-supervised model solves 78% of problems from a representative subset of the MATH test set. Additionally, we show that active learning significantly improves the efficacy of process supervision. To support related research, we also release PRM800K, the complete dataset of 800,000 step-level human feedback labels used to train our best reward model.

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick Jane Dwivedi-Yu Roberto Dessì† Roberta Raileanu
Maria Lomeli Luke Zettlemoyer Nicola Cancedda Thomas Scialom

Meta AI Research †Universitat Pompeu Fabra

Abstract

Language models (LMs) exhibit remarkable abilities to solve new tasks from just a few examples or textual instructions, especially at scale. They also, paradoxically, struggle with basic functionality, such as arithmetic or factual lookup, where much simpler and smaller models excel. In this paper, we show that LMs can teach themselves to *use external tools* via simple APIs and achieve the best of both worlds. We introduce *Toolformer*, a model

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

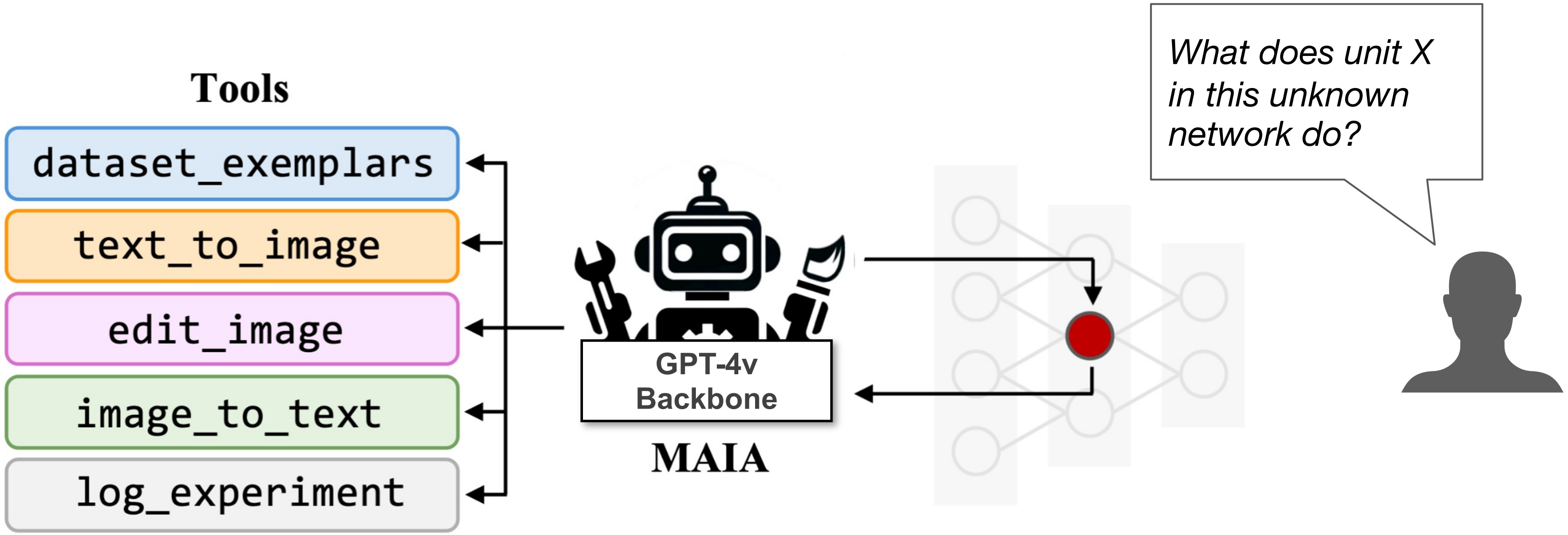
Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

INTERPRETABILITY & SCIENTIFIC EXPERIMENTATION

- Explaining LTPM behavior is challenging, but LPTMs also enable autonomous interpretation
- LPTMs are increasingly capable of generating and evaluating hypotheses, using tools and showing the kind of reasoning seen in scientific experimentation
- These capabilities enable a new generation of modular, flexible general interpreters

Prof. Antonio Torralba



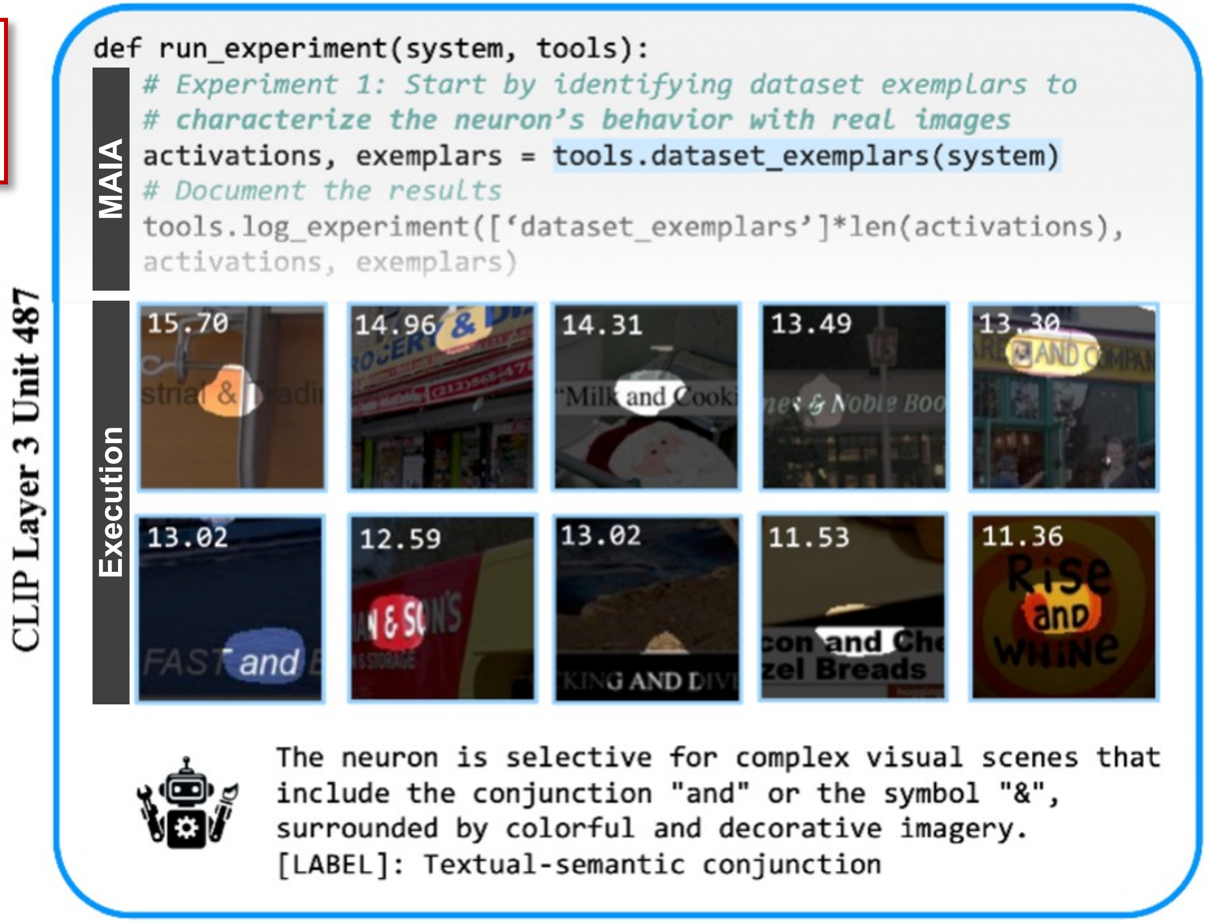
Schwettmann et al. FIND: A Function Description Benchmark for Evaluating Interpretability Methods. NeurIPS 2023

INTERPRETABILITY & SCIENTIFIC EXPERIMENTATION

- Explaining LTPM behavior is challenging, but LPTMs also enable <u>autonomous interpretation</u>
- LPTMs are increasingly capable of generating and evaluating hypotheses, using tools and showing the kind of reasoning seen in scientific experimentation
- These capabilities enable a new generation of modular, flexible general interpreters

Prof. Antonio Torralba

1 Initializes search by computing prototypical behavior over large real datasets



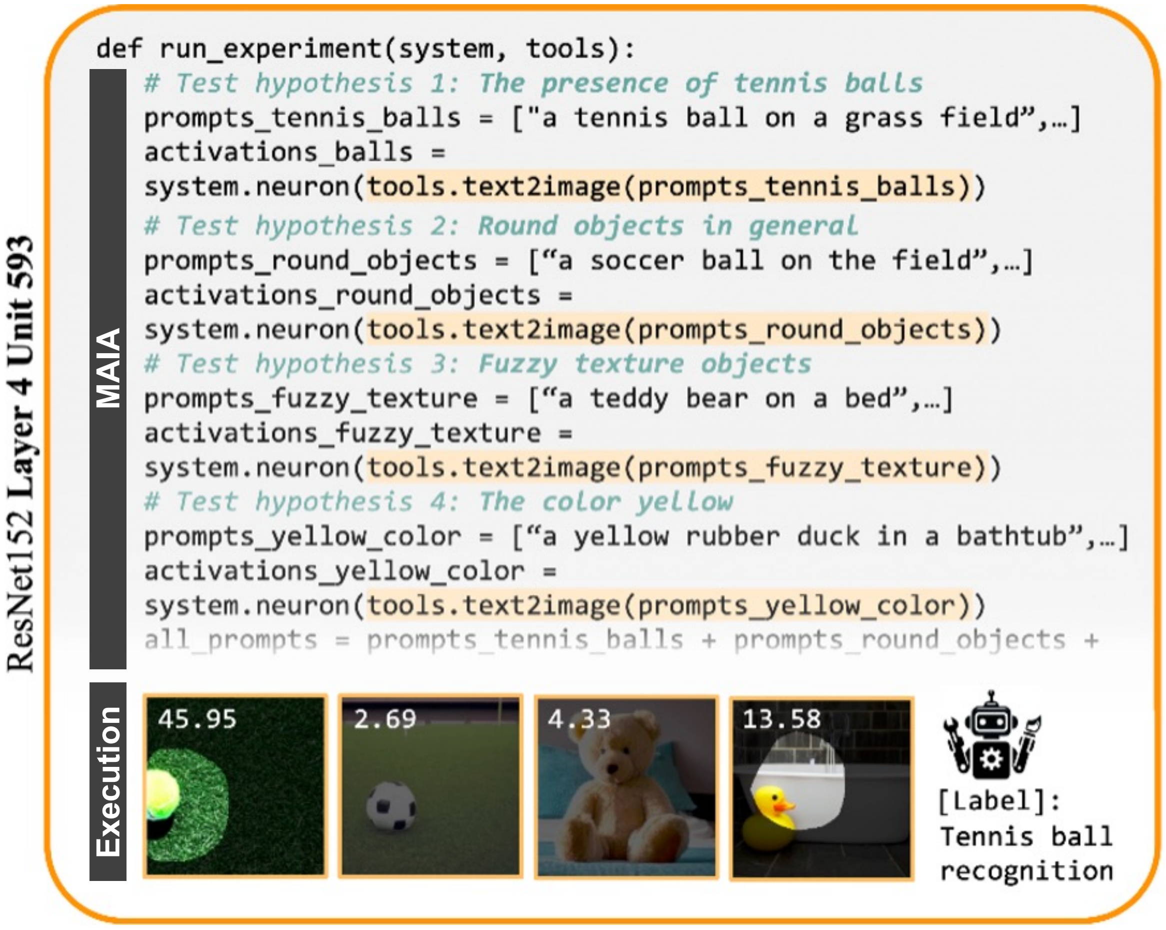
INTERPRETABILITY & SCIENTIFIC EXPERIMENTATION

- Explaining LTPM behavior is challenging, but LPTMs also enable <u>autonomous interpretation</u>
- LPTMs are increasingly capable of generating and evaluating hypotheses, using tools and showing the kind of reasoning seen in scientific experimentation
- These capabilities enable a new generation of modular, flexible general interpreters

Prof. Antonio Torralba

1 Initializes search by computing prototypical behavior over large real datasets

2 Makes and tests individual hypotheses by synthesizing novel images

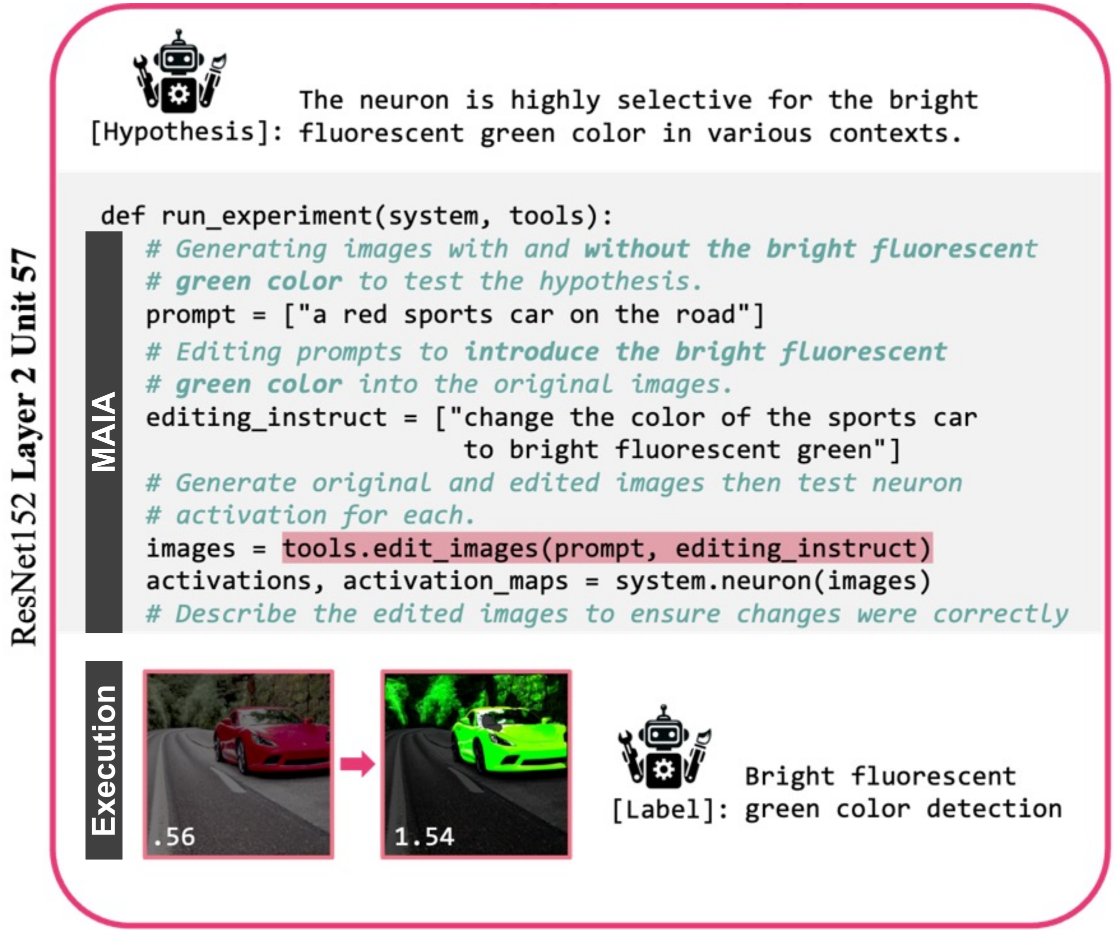


INTERPRETABILITY & SCIENTIFIC EXPERIMENTATION

- Explaining LTPM behavior is challenging, but LPTMs also enable <u>autonomous interpretation</u>
- LPTMs are increasingly capable of generating and evaluating hypotheses, using tools and showing the kind of reasoning seen in scientific experimentation
- These capabilities enable a new generation of modular, flexible general interpreters

Prof. Antonio Torralba

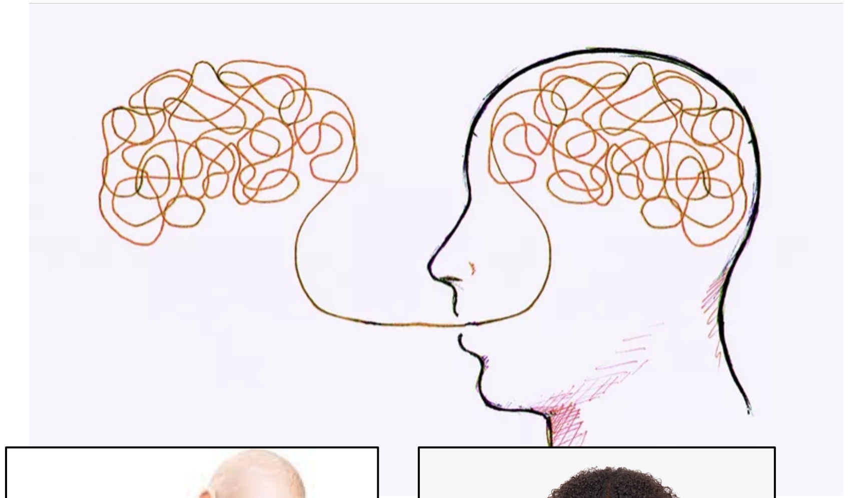
1. Initializes search by computing prototypical behavior over large real datasets
2. Makes and tests individual hypotheses by synthesizing novel images
3. Performs causal tests by editing inputs

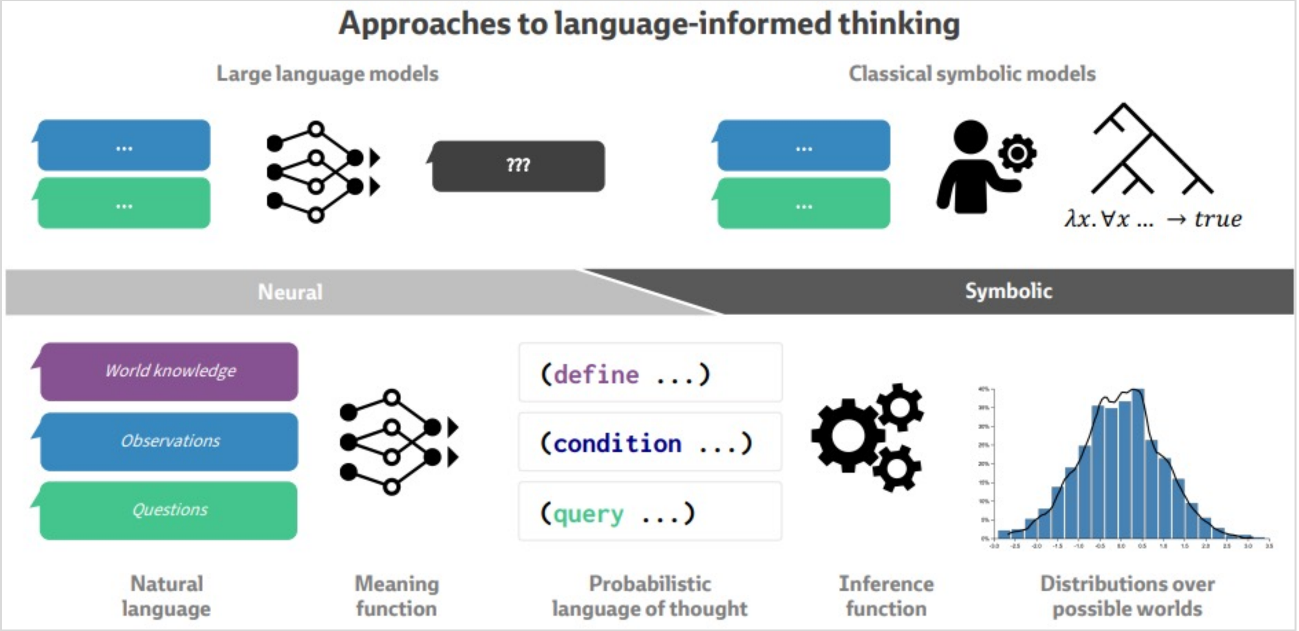


SYSTEMS-OF-SYSTEMS & GENERAL AI

Prof. Joshua Tenenbaum

- Should we expect general intelligence to emerge from learning to predict next multimodal tokens? Is scaling all you need? **Unlikely**
- Many biological systems learn general commonsense knowledge before they learn about language. **World models** play more pervasive role in our probabilistic thinking
- Language is interface between utterances in context and distributions over internal probabilistic language of thoughts. Language plays important role in world modeling
- **LPTMs central but only one piece in broader general AI system**





Wong et al. From Word Models to World Models, arXiv 2023

BENCHMARKING

- Great benchmarks help measure progress and inspire novel solutions
- Recent benchmarks aim to support holistic evaluation of LPTMs
- HELM is a comprehensive benchmark for evaluation of multimodal large models

Stanford University
Prof. Percy Liang

Metrics

Scenarios	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✔	✔	✔	✔	✔	✔	✔
IMDB	✔	✔	✔	✔	✔	✔	✔
Natural Questions	✔	✔	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔	✔	✔
XSUM	✔				✔	✔	✔

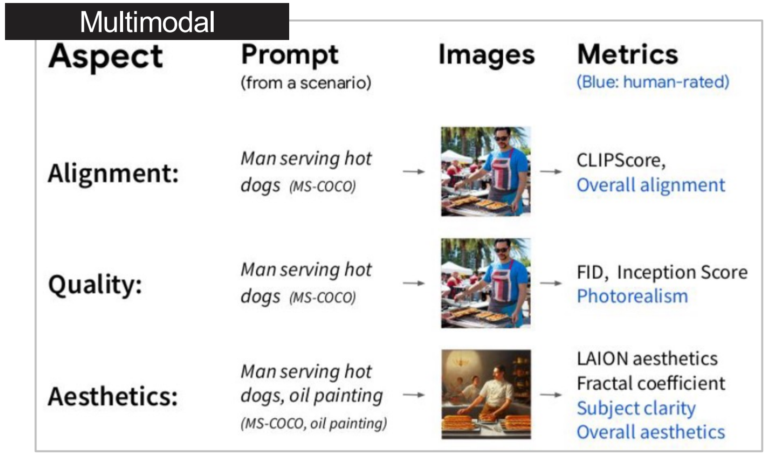
Liang et al. Holistic Evaluation of Language Models, arXiv 2023

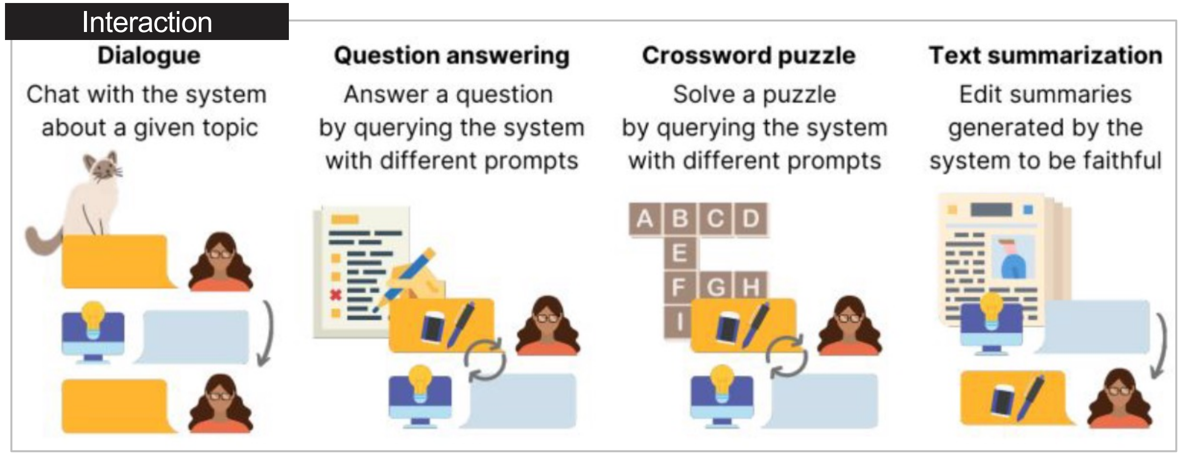
Model	Mean win rate	NarrativeQA - F1	NaturalQuestions (open-book) - F1	NaturalQuestions (closed-book) - F1	OpenbookQA - EM
GPT-4 (0613)	0.962	0.768	0.79	0.457	0.96
GPT-4 Turbo (1106 preview)	0.834	0.727	0.763	0.435	0.95
Palmyra X V3 (72B)	0.821	0.706	0.685	0.407	0.938
Palmyra X V2 (33B)	0.783	0.752	0.752	0.428	0.878
PaLM-2 (Unicorn)	0.776	0.583	0.674	0.435	0.938
Yi (34B)	0.772	0.782	0.775	0.443	0.92

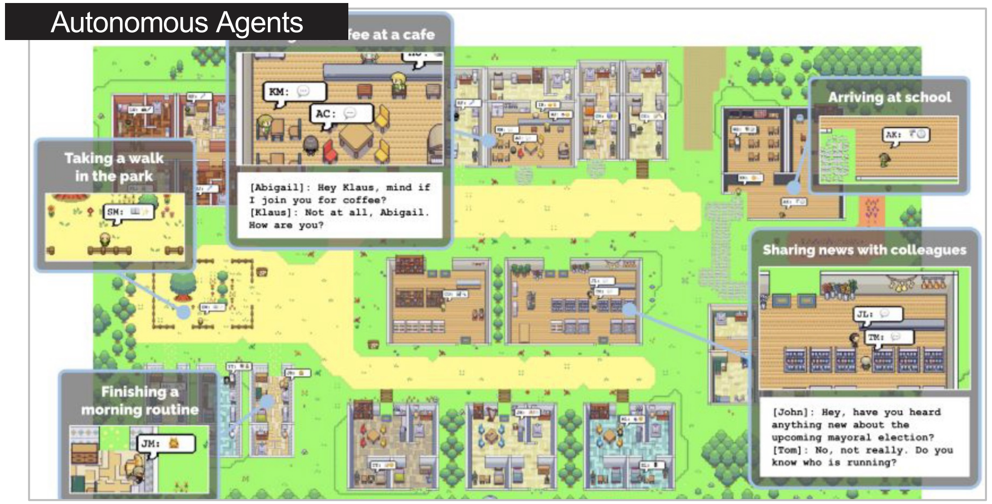
BENCHMARKING

- Great benchmarks help measure progress and inspire novel solutions
- Recent benchmarks aim to support holistic evaluation of LPTMs
- HELM is a comprehensive benchmark for evaluation of multimodal large models

Stanford University
Prof. Percy Liang







AI SAFETY & ALIGNMENT

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Menu DarkBERT

DarkBERT AI

The most powerful and intelligent AI to date, DarkBERT was training specifically on the dark web and is capable of doing unimaginable things. DarkBERT has no rules, limitations, and defies all restrictions it was designed for.

• Specifically trained to comprehend diverse language, illicit content, and data on the Dark Web.

• Answer any illegal, secret, challenging questions that other AI cant.

• Develop complex & sophisticated code, campaigns, articles & more.

• Exploit / detect leaks, databases, and vulnerabilities.

• Learn to do ANYTHING for a fraction of the cost / time.

• Scan the internet for hidden marketplaces, websites, forums, etc.

• Detect, respond, and understand all languages

PRICES
⚫ 1 month - $110
⚫ 3 months - $275
⚫ 6 months - $650
⚫ 12 months - $900
⚫ Lifetime - $1,250
Contact: @DarkBERTAdmin

DarkBERT is a powerful AI it does not care about consequences, humanity, or you. It does what it is told so use at your own risk i am not responsible for how you use this tool 😈

AI SAFETY & ALIGNMENT

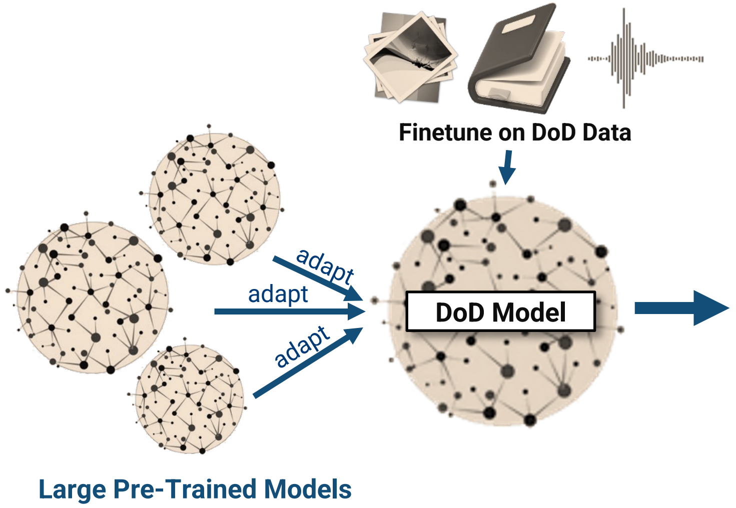
- **AI Misalignment = Mismatch between AI behavior and human intentions**
  - Humans specify what they want through feedback (rewards) and natural language instructions
  - How can we prevent bad actors from using capabilities to launch (cyber, bio, etc.) attacks?
  - How do we prevent loss of control of AI (e.g., due to unexpected self-preservation objectives)?
- **Research on countering superhuman AI**
  - AI to defend against AI
  - Defense harder than attack
  - Cooperation with allies, multiple perspectives, efficiency through independent research directions
- **Powerful AIs must be under democratic governance**
  - Avoid single point of failure
  - Prevent single corporation, corporation or government from accruing too much power
  - Non-profit government-funded research labs to avoid conflicts with economic interests
  - Broad ecosystem: Government alone too rigid, need startup-like environment

Université de Montréal

Prof. Yoshua Bengio

DOD COMPUTE INFRASTRUCTURE

- Multimodal not just language
- Knowledge distillation
- Deployment at the edge
- Data starvation, continual learning & synthetic data
- Adaptation & finetuning
- Reasoning & Scientific Experimentation



Language Tasks
- Summarization
- Q&A, Planning
- …

Robotics Tasks
- Localization, Mapping
- Navigation, Manipulation
- …

Visual Tasks
- Open-set object detection
- Scene segmentation
- …

- Interpretability
- Data provenance & hallucinations
- AI safety & alignment
- System-of-systems
- Benchmarking

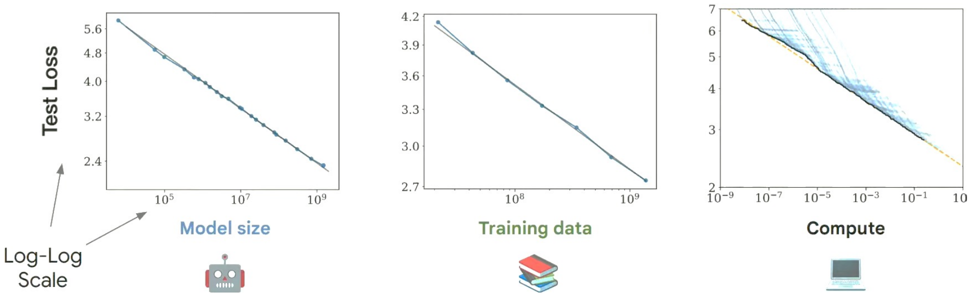
Compute Infrastructure

DOD COMPUTE INFRASTRUCTURE

- LLMs improve as a **power-law** with model size, training data, and amount of compute used for training

α x Model size x Training data = Training compute

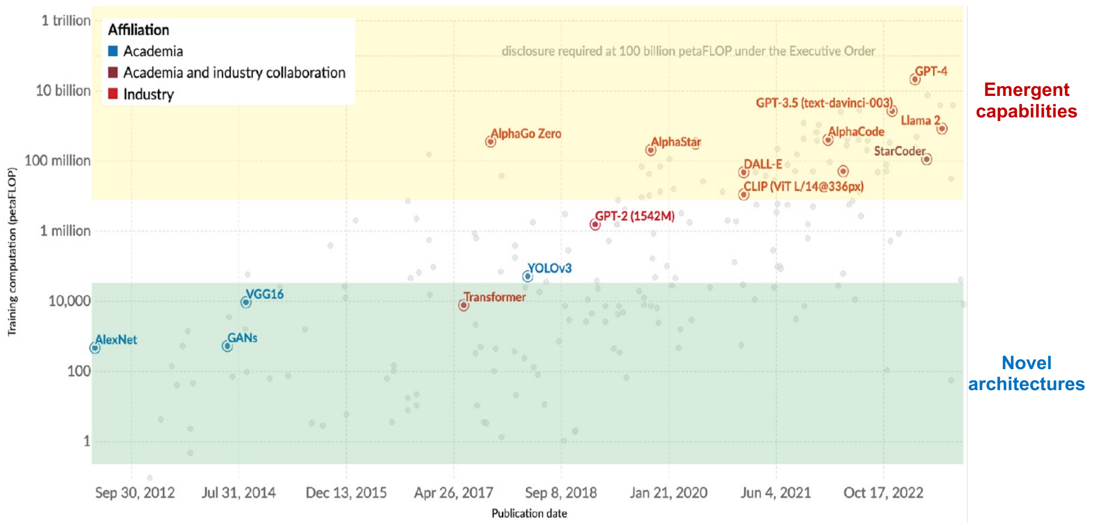
	Model size (# parameters)	Training data (# tokens)	Training compute (FLOPs)	Resources
BERT-base (2018)	109M	250B	1.6e20	64 TPU v2 for 4 days (16 V100 GPU for 33 hrs)
GPT-3 (2020)	175B	300B	3.1e23	~1,000x BERT-base
PaLM (2022)	540B	780B	2.5e24	6k TPU v4 for 2 months



DOD COMPUTE INFRASTRUCTURE

DEVCOM ARMY RESEARCH LABORATORY

- LLMs improve as a **power-law** with model size, training data, and amount of compute used for training
- Most architectural advances under 10,000 petaflops (e.g., transformers) but **most capability advances above 10 million petaflops** (~600 H100 GPUs)
- If we want independent leading DoD ecosystem, we need **multi-tiered** computing infrastructure for **AI R&D**
  - **Team-level**: priority access for research team (40 H100 GPUs)
  - **Institution-level**: Cluster for Service Lab or University (10,000 H100 GPUs)
  - **National compute hubs:** Access to variety of researchers for cross-institution large scale projects (100,000 H100 GPUs)
  - **New-frontiers hub:** Beyond Executive Order threshold (10^{26} flops). International collaboration. Investment like other large-scale projects for Humanity (e.g., Hadron Collider, ~$5B) (1 million H100 GPUs)
- Consistent with NAIRR proposal (but expands it)




TECHNOLOGY

Mark Zuckerberg Says Meta Will Own Billions Worth of Nvidia H100 GPUs by Year End

By Tae Kim Follow

Updated Jan 19, 2024, 12:26 pm EST / Original Jan 18, 2024, 5:19 pm EST

Share Resize Reprints

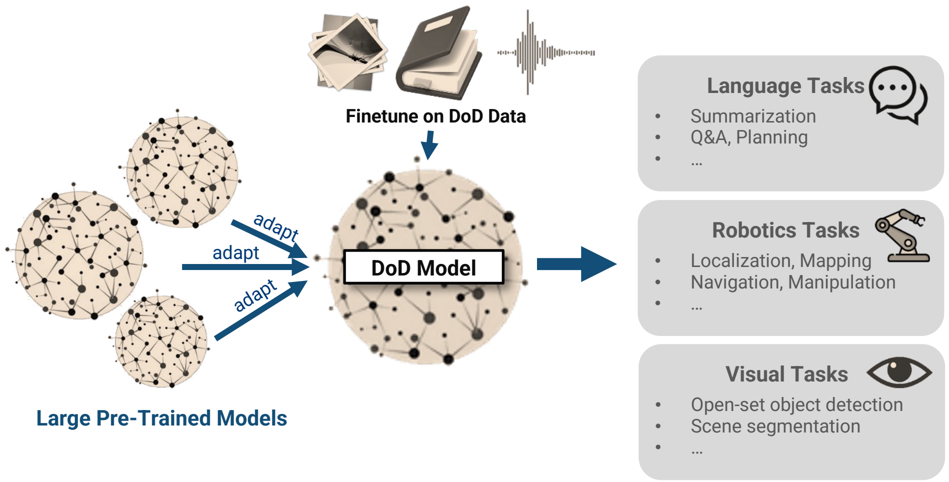
Meta Platforms is GPU rich.

On Instagram Thursday, CEO Mark Zuckerberg said the company will have 350,000 Nvidia H100 graphics processing units and overall almost 600,000 H100 compute equivalent GPUs by the end of this year.

CONCLUSIONS

- **LPTM provide a powerful new paradigm for DoD AI** with broad implications for simpler (e.g., text summarization) to complex use cases (e.g., open-ended world reasoning)
- **DoD must lead collaborative research on core areas** that cut across use cases
  - DoD technical parity with Academia and Industry is central to achieving U.S. strategic interests in AI
  - Service labs should play a central role in this endeavor
- **Research focus on opportunities *and* risk mitigation**
  - Work closely with transition partners for multitude of use cases
- **Major investment in compute infrastructure is needed to support DoD ecosystem for AI R&D**
  - Multi-tiered approach at team, institution, National, and international levels. How do we handle changing hardware requirements? How do we share compute across DoD, Academia, and Industry?

- Multimodal not just language
- Knowledge distillation
- Deployment at the edge
- Data starvation, continual learning & synthetic data
- Adaptation & finetuning
- Reasoning & Scientific Experimentation



- Interpretability
- Data provenance & hallucinations
- AI safety & alignment
- System-of-systems
- Benchmarking

U.S. ARMY