RESEARCH CHALLENGES FOR LARGE PRE-TRAINED MODELS
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)
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

Large Pre-Trained Models

DoD Model

Finetune on DoD Data

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
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

- 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)
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

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
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.

ATOMIC
Knowledge Model
Automatically trained from GPT-3

COMETdistill
~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

- Model Size (billion parameters, log scale)
- Realism Gap (PDJ lower is better)

Text Data: ____________________
Image Data: ____________________
Video Data: ____________________
Robotics Data: ____________________

generated image
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
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

(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”

Stable Diffusion + Prompt2Prompt

Edited Caption: “photograph of a girl riding a dragon”

Generated training examples:

“have her ride a dragon”

“Color the cars pink”

“Make it lit by fireworks”

“convert to brick”

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

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
Multimodal redundancy provides knowledge about the world

Exploration
Autonomous interactive exploration of environment leads to self-learning
**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.

---

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

**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 multi-turn 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 examples 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 PLM with a chain of thought improves ordinary state-of-the-art sample-ratio on the NAPALS arithmetic test from 1.5% to 13.2% and on a group of math word problems, surpassing even pretrained GPT-3 with a verifier.

**Let’s Verify Step by Step**

<table>
<thead>
<tr>
<th>Human Lighthouse</th>
<th>Prompted Humans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huang Liangle*</td>
<td>Roberta*</td>
</tr>
<tr>
<td>Vincent Kwekaje</td>
<td>Roberta*</td>
</tr>
<tr>
<td>Vera Rocules</td>
<td>Roberta*</td>
</tr>
<tr>
<td>Helen Swedde*</td>
<td>Roberta*</td>
</tr>
<tr>
<td>Karel Cikole*</td>
<td>Roberta*</td>
</tr>
</tbody>
</table>

**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 to other outcomes—such as feedback and prompt-similar reasoning. This public feedback 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 human outcomes. Recent work has already begun this comparison, but many questions still remain. We conduct our own investigation, finding that human feedback significantly outperforms human feedback for training models to solve problems from the challenging MATH dataset. Our prompt-feedback model solves 79% of problems from a representative subset of the MATH test set. Additionally, we show that active learning significantly improves the quality of prompt feedback. To support related research, we also release PROMERL, a complete dataset of 200,000 step-level human feedback labels used to train our best model.

---

**Toolformer: Language Models Can Teach Themselves to Use Tools**

**Abstract**

Language models (LMs) exhibit remarkable abilities to solve new tasks from just a few examples or natural 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 that takes a text to code policy and uses a language model to teach itself to use the tool.

*The New England Journal of Medicine is a registered trademark of [JAMA](https://www.jamanetwork.com) (The Journal of the American Medical Association)*.

Out of 1,400 participants, 400 (or 28.6%) passed the test.

The name derives from "InTurk" (the Spanish word for "portable") or "InTurk". 

---

**References**


---

*Images and figures are omitted for brevity.*
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

Tools

- dataset_exemplars
- text_to_image
- edit_image
- image_to_text
- log_experiment

What does unit X in this unknown network do?

• 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

```python
def run_experiment(system, tools):
    # Experiment 1: Start by identifying dataset exemplars to
    # characterize the neuron's behavior with real images
    activations, exemplars = tools.dataset_exemplars(system)
    # Document the results
    tools.log_experiment([{'dataset_exemplars'}*len(activations),
                          activations, exemplars])
```

The neuron is selective for complex visual scenes that include the conjunction "and" or the symbol ".&.", surrounded by colorful and decorative imagery. [LABEL]: Textual-semantic conjunction
• 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

1 Initializes search by computing prototypical behavior over large real datasets
2 Makes and tests individual hypotheses by synthesizing novel images
• 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

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

MAIA

The neuron is highly selective for the bright fluorescent green color in various contexts.

def run_experiment(system, tools):
    # Generating images with and without the bright fluorescent green color to test the hypothesis.
    prompt = "a red sports car on the road"
    # Editing prompts to introduce the bright fluorescent green color into the original images.
    editing_instruct = "change the color of the sports car to bright fluorescent green"
    # Generate original and edited images then test neuron activation for each.
    images = tools.edit_images(prompt, editing_instruct)
    activations, activation_maps = system.neuron(images)
    # Describe the edited images to ensure changes were correctly detected.

Bright fluorescent green color detection

UNCLASSIFIED

UNCLASSIFIED

 ResNet152 Layer 2 Unit 57
SYSTEMS-OF-SYSTEMS & GENERAL AI

- 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
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
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

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 - $450
- 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
DOD COMPUTE INFRASTRUCTURE

- Large Pre-Trained Models
- Adapt & finetune
- DoD Model
- Finetune on DoD Data

**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

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

Compute Infrastructure
DOD COMPUTE INFRASTRUCTURE

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

<table>
<thead>
<tr>
<th>Model size (# parameters)</th>
<th>Training data (# tokens)</th>
<th>Training compute (FLOPs)</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base (2018)</td>
<td>109M</td>
<td>250B</td>
<td>1.6e20</td>
</tr>
<tr>
<td>GPT-3 (2020)</td>
<td>175B</td>
<td>300B</td>
<td>3.1e23</td>
</tr>
<tr>
<td>PaLM (2022)</td>
<td>540B</td>
<td>780B</td>
<td>2.5e24</td>
</tr>
</tbody>
</table>

**Test Loss**

- Log-Log Scale
- Model size
- Training data
- Compute
DOD COMPUTE INFRASTRUCTURE

• 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)
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?