CDAO

Chief Digital & Artificial Intelligence Office

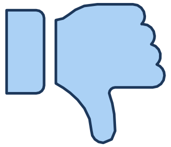
Maturity Model Opening Remarks

Dr. William Streilein, Chief Technology Officer

Which do you Agree with?

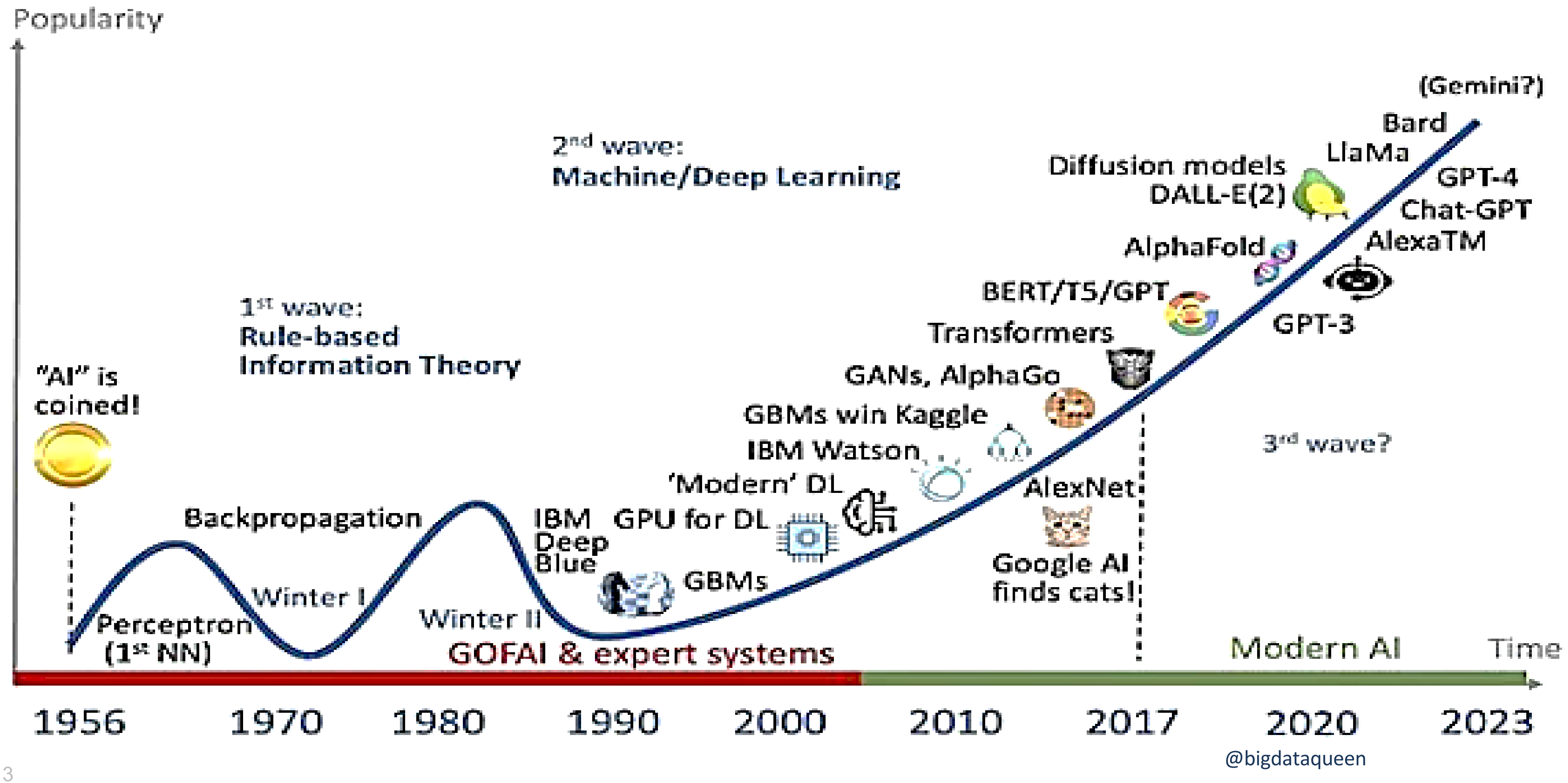
- “**AI is truly amazing.** Its potential to transform society and the world by automating tasks, making better decisions, personalizing experiences and generally making life better for all mankind make it the most important invention since the printing press.

- “**AI is nothing more than hype**. Recent advances in AI are narrow, only incremental and it is far from being as intelligent as humans. Progress is driven by corporate interests rather than societal benefit and rarely impact daily life in a positive way.



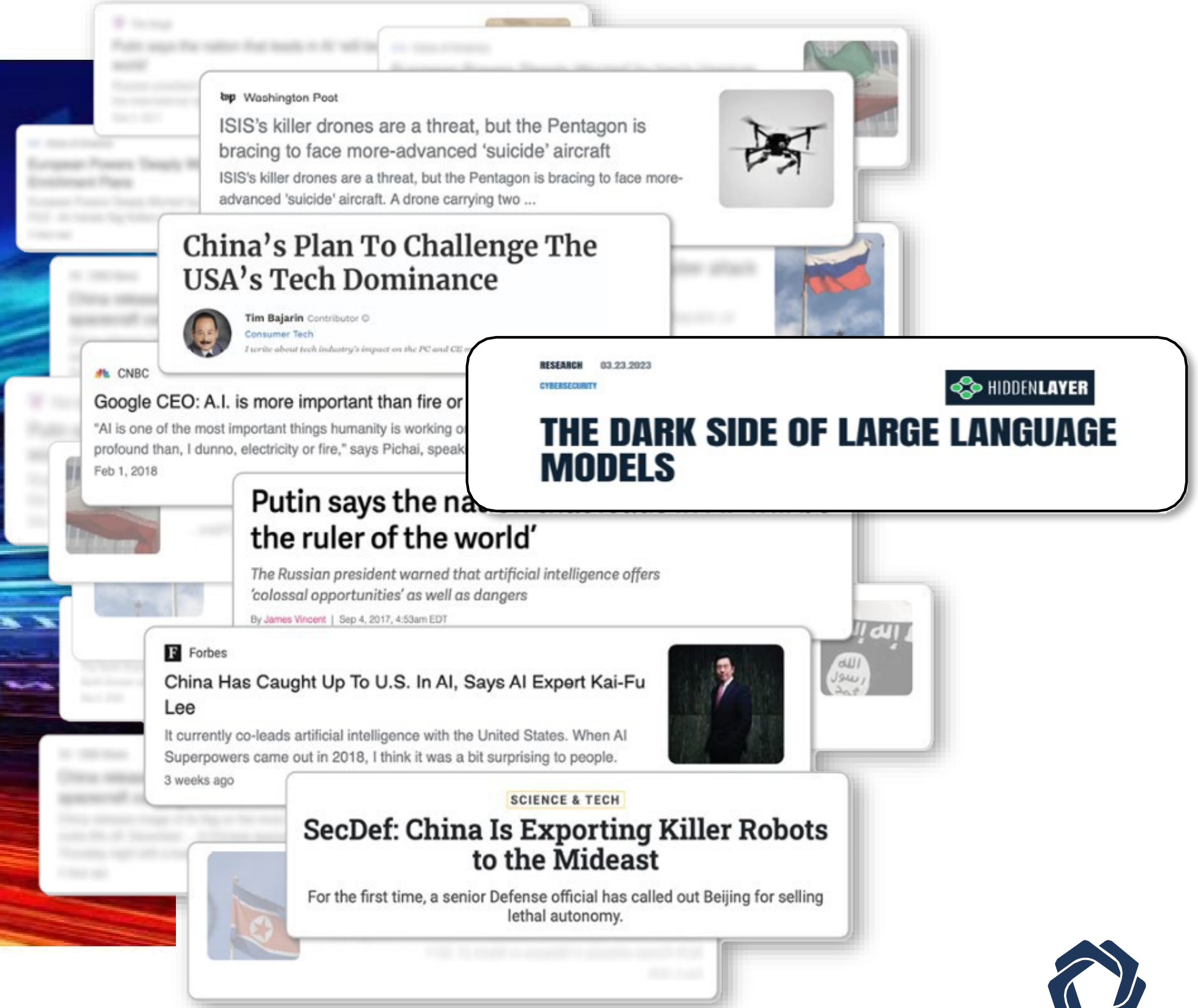
CDAO

History of AI

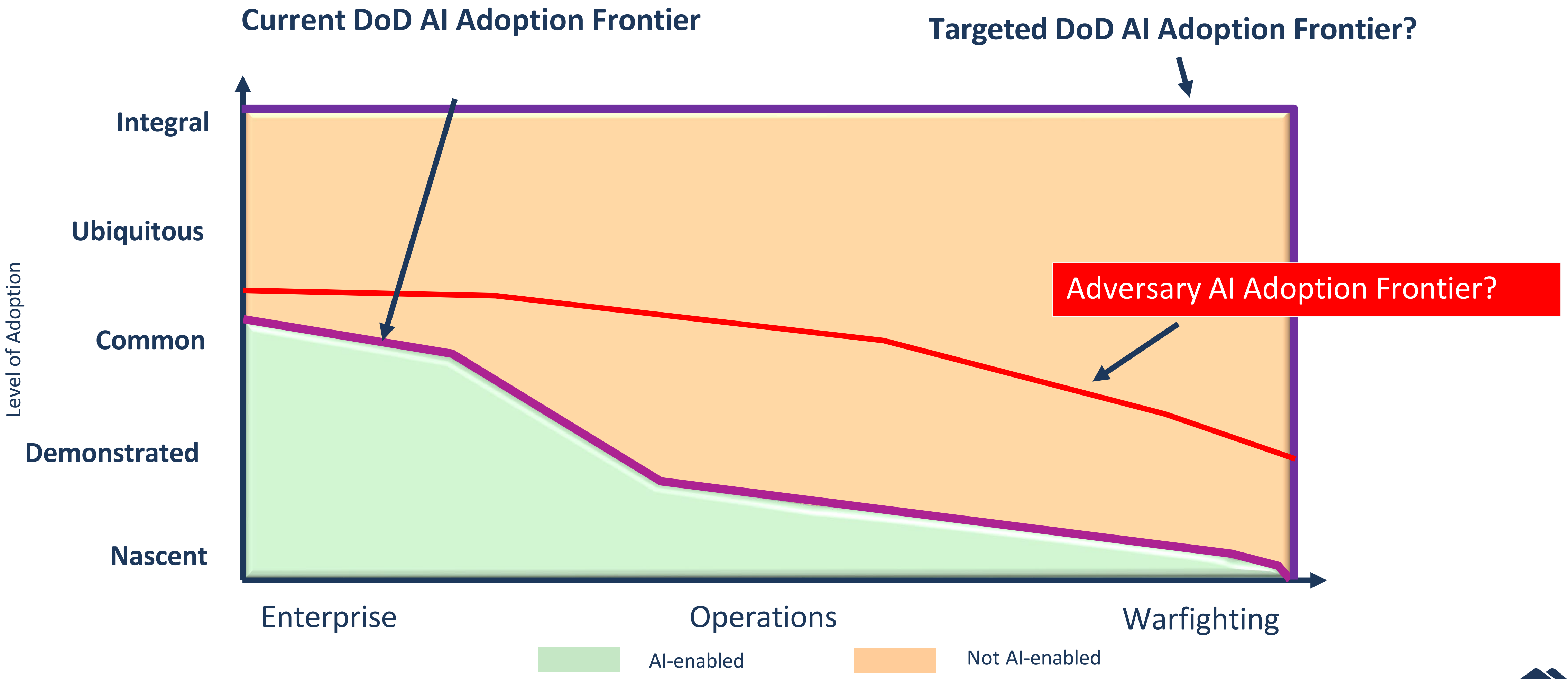


CDAO

The Competitive Environment

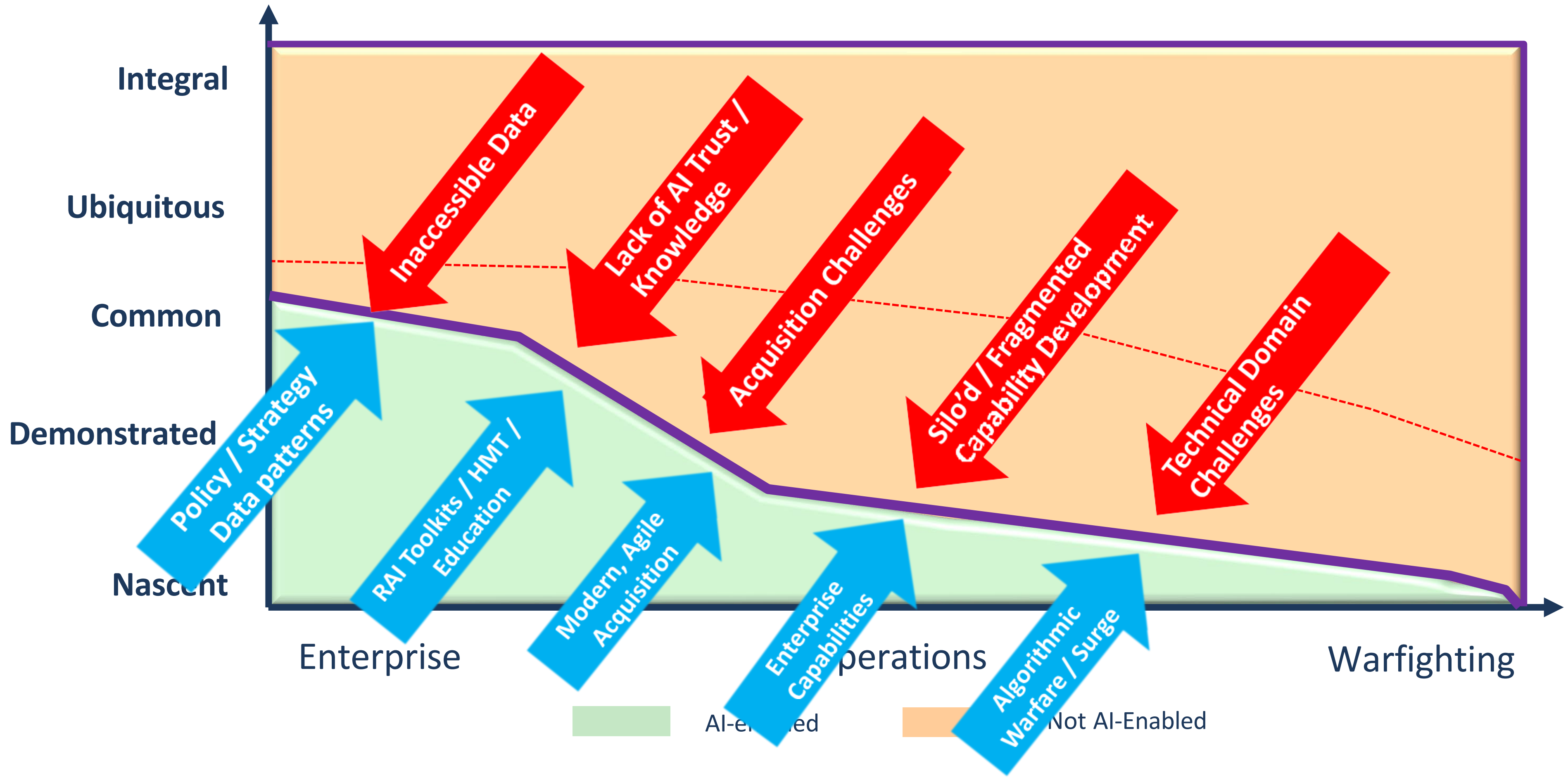


Race With Ramifications

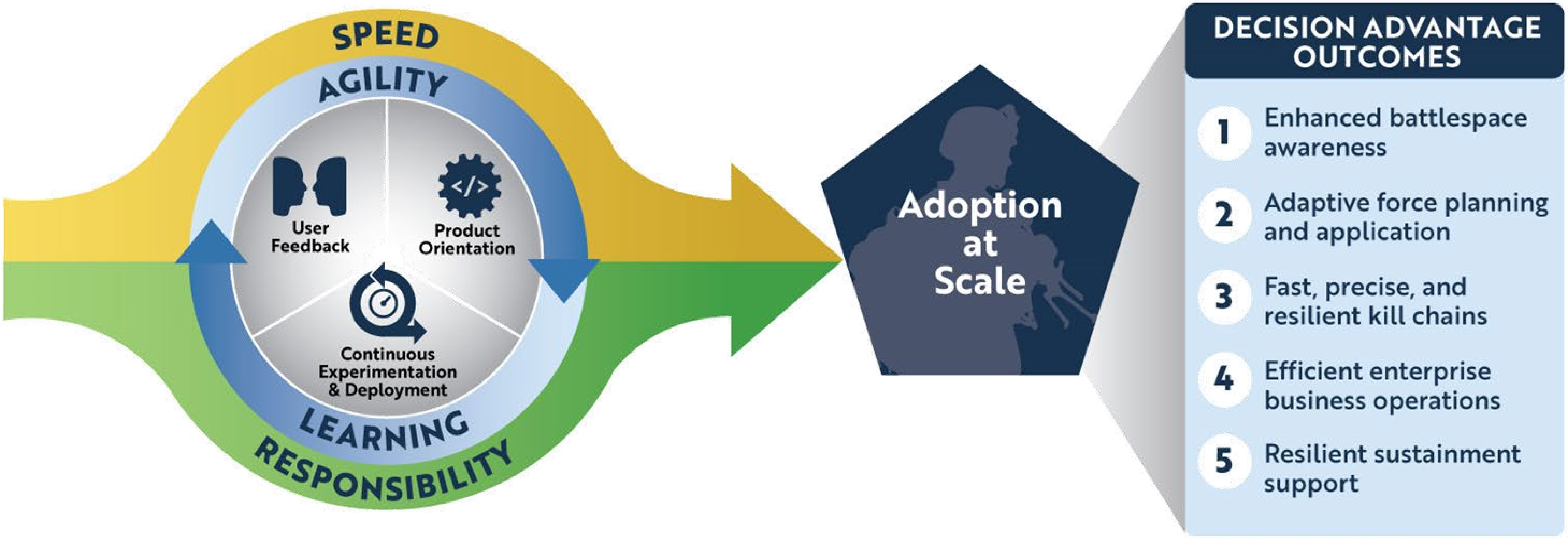


CDAO

Impediments to Adoption

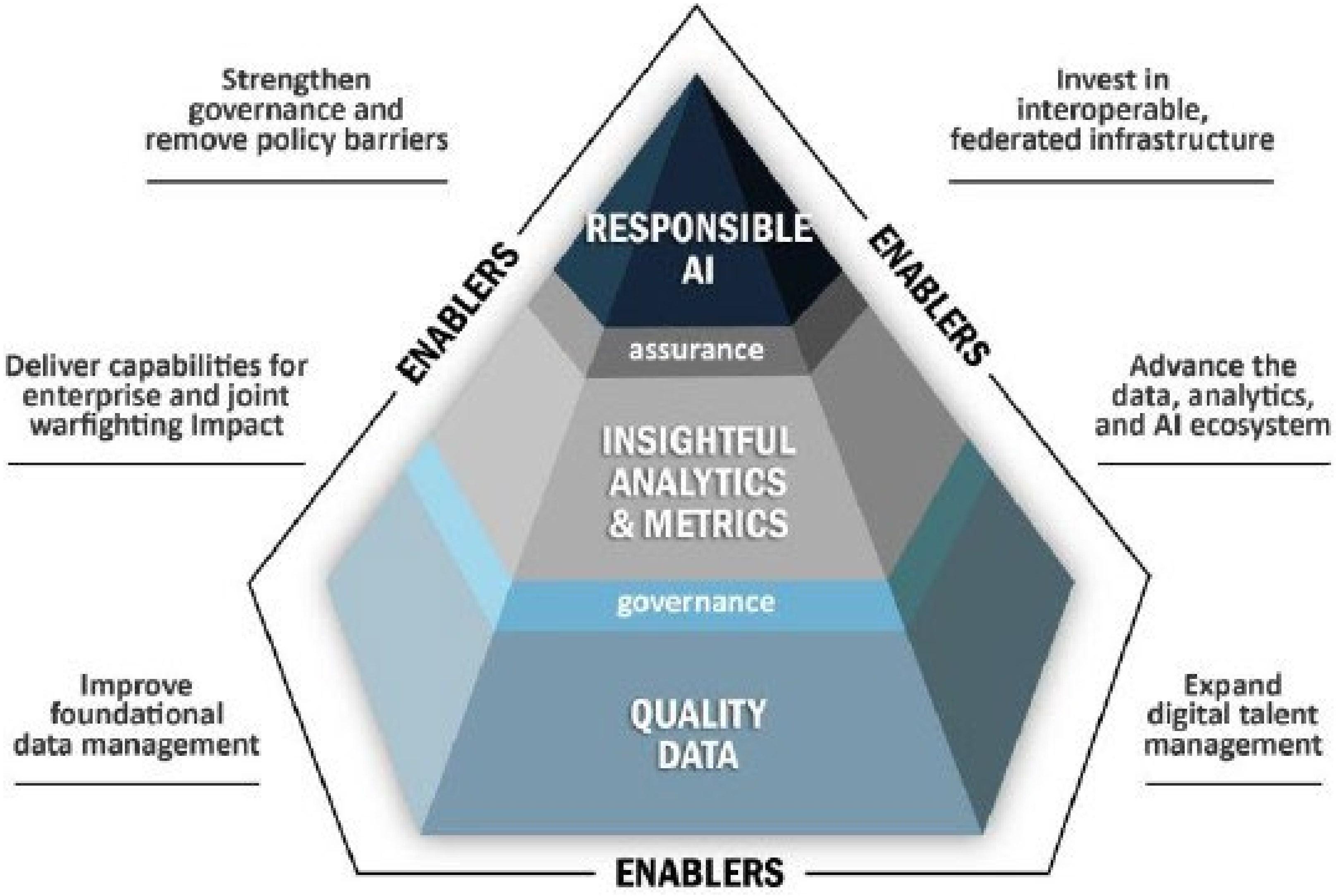


CDAO: Employing an Agile Approach to Adoption at Scale



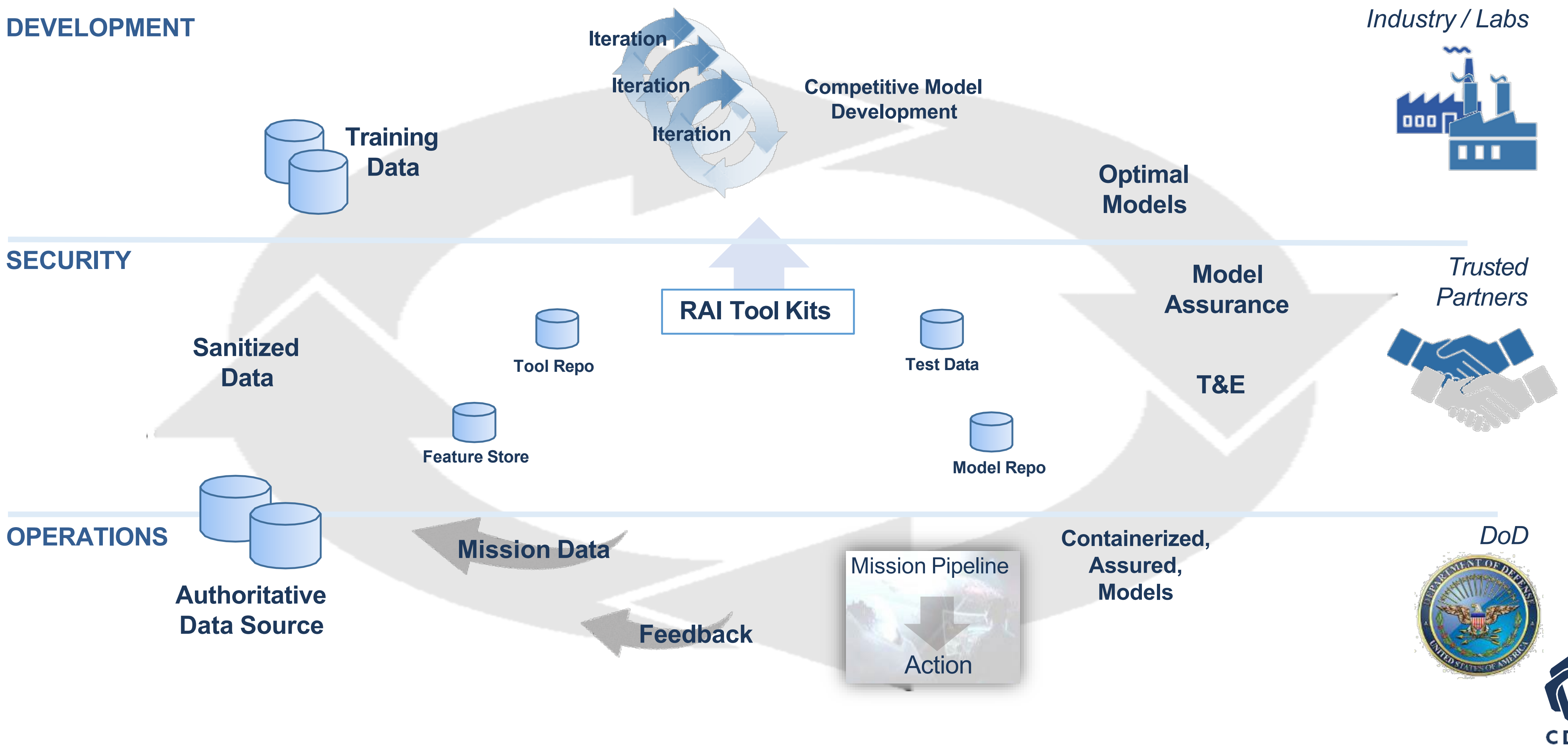
CDAO

DoD AI Hierarchy of Needs

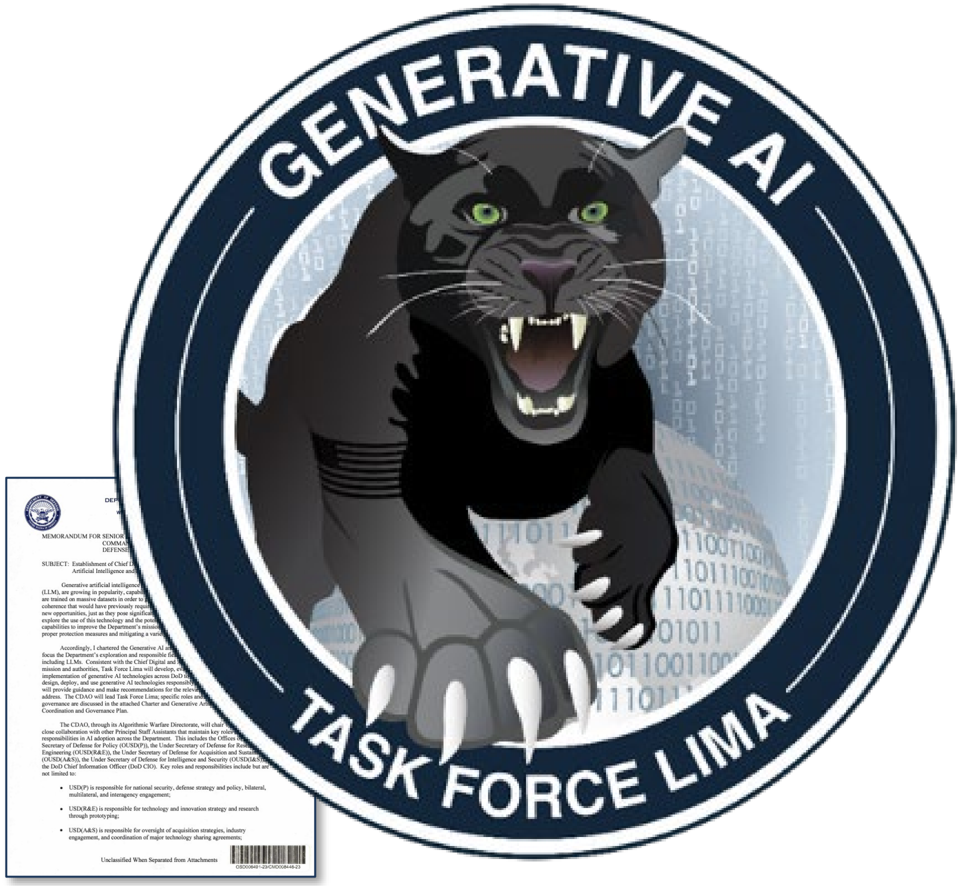


CDAO

A Digital Ecosystem Supports DoD AI Goals



Task Force Lima



August 10, 2023

- ***Accelerate*** promising generative AI initiatives and joint solutions;
- ***Federate*** disparate developmental and research efforts into a DoD community of practice to accelerate innovation and implementation;
- ***Evaluate*** solutions across Doctrine, Organization, Training, Materiel, Leadership, Personnel, Facilities, and Policy;
- ***Drive*** education and build a culture of responsible implementation and use; and,
- ***Ensure*** coordinated DoD engagement with interagency, international, educational, civil society, and industry partners regarding responsible development and use of generative AI.

We seek a *maturity model* that enables us to map LLMs to DoD use cases

CDAO

Towards an LLM Maturity Model

- ✓ **Understand potential LLM use cases and the level of capability required across several relevant LLM dimensions**
- ✓ **Assess the maturity of LLM solutions with respect to their application to mission use cases and workflows**
- ✓ **Identify areas where LLM capabilities need to improve to be useable in given mission use cases and workflows.**

LLM Characteristics

Use Case	Conversation / ChatBot	Summarization	Question / Answer	Analysis	Classification	Insight Generati
Governability						
Completeness						
Accuracy						
Equitability						
Consistency						
Traceability						
Novelty						
Fluency						
Interactivity						
...						

Notional Maturity Model

CDAO

LLM MM Working Session Agenda

- **Introductory / Framing Presentation – (15 briefing)**
  1. Challenges of DoD adopting LLMs and need for mechanism to enable dialog w/ broader LLM community about shaping development towards DoD needs

- **Maturity model presentations (20 briefing / 5 Q&A)**
  1. John Snow Labs – mapped benchmark tool scores to levels
  2. Parsons – leverages matrix framework to map levels
  3. iWorks – includes application of framework to use cases

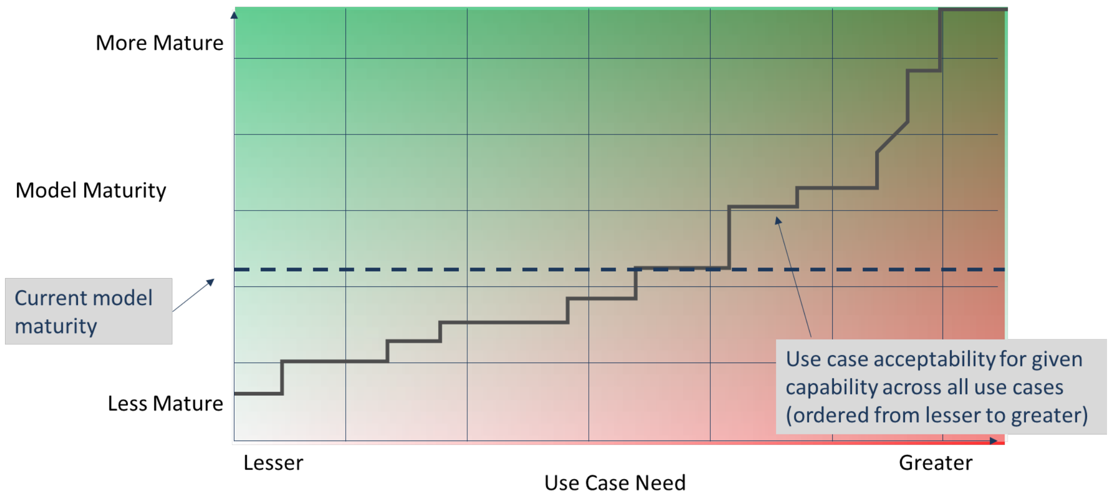
- **Workflow integration / LLM System (20 briefing / 5 Q&A)**
  1. Microsoft (see section 8.0 Measuring the Solution Architecture)
  2. ScaleAI – LLM system approach
  3. AWS – how infrastructure supports LLM use

- **Validation of LLM Maturity Model (use case, score card, process) (20 briefing / 5 Q&A)**
  1. Blue Halo - methodology for validating the model, red team / security evaluation
  2. Expression – proposal focused on Text-to-Query (electromagnetic battle management joint situation awareness);
  3. Tenet3 - LLM scorecard to communicate the maturity with others

CDAO

Maturity/Acceptability Model Approach

- LLM's have the potential to revolutionize DoD operations however, they are still a relatively new technology
- They are not well understood, nor can they be trusted to produce reliable results for important use cases.
- Many DoD organizations are struggling to understand how to adopt and use LLMs effectively
- These organizations require guidance to determine when LLM solutions are appropriate for organizational workflows
- A maturity model is need that allows DoD to map vendor model capability to use case needs
- Example: Autonomy levels for self-driving cars.

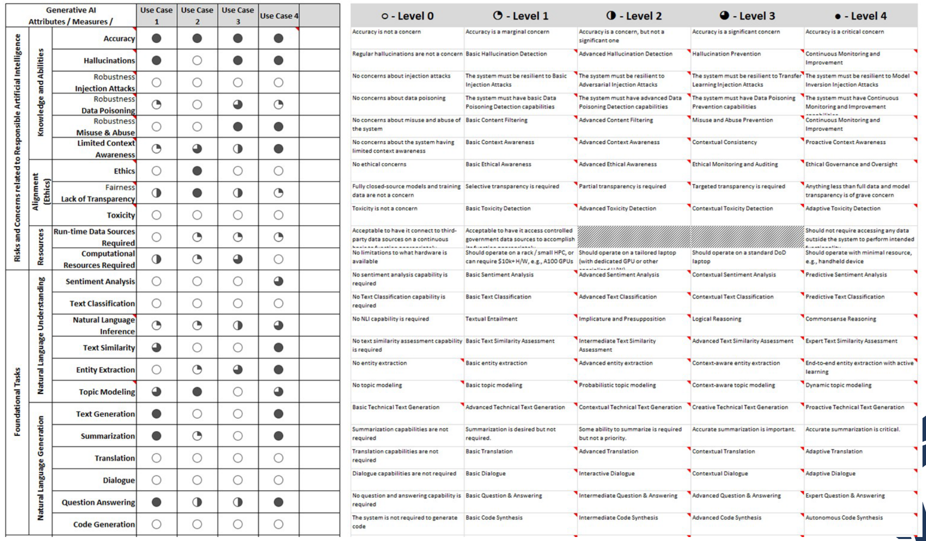


Determining LLM Use Case Need

➔ Step 1: For each use case, determine what are the capabilities (from a gen AI perspective) that are required for the successful execution of the use case

➔ Step 2: For each Capability of each use case determine the level of dependency to success of each capability.

➔ Step 3: Assess LLM Maturity in context of use case need



Presenter Notes
2024-02-13 17:24:49

LLM's have the potential to revolutionize DoD operations however, they are still a relatively new technology They are not well understood, nor can they be trusted to produce reliable results for important use cases. Many DoD organizations are struggling to understand how to adopt and use LLMs effectively These organizations require guidance to determine when LLM solutions are appropriate for organizational workflows A maturity model is need that allows DoD to map vendor model capability to use case needs Example: Autonomy levels for self-driving cars.

Motivation: Why a maturity model?

- Higher levels of automation requires a higher level of confidence in the model
  - Automation levels vary by use case and should be determined *a priori*
  - Confidence may have a different meaning at different levels of automation and in different use cases (e.g., deploying kinetic munitions requires much higher confidence than military planning)
- Confidence must be based on objective metrics to assess declines or changes in performance

Increased maturity

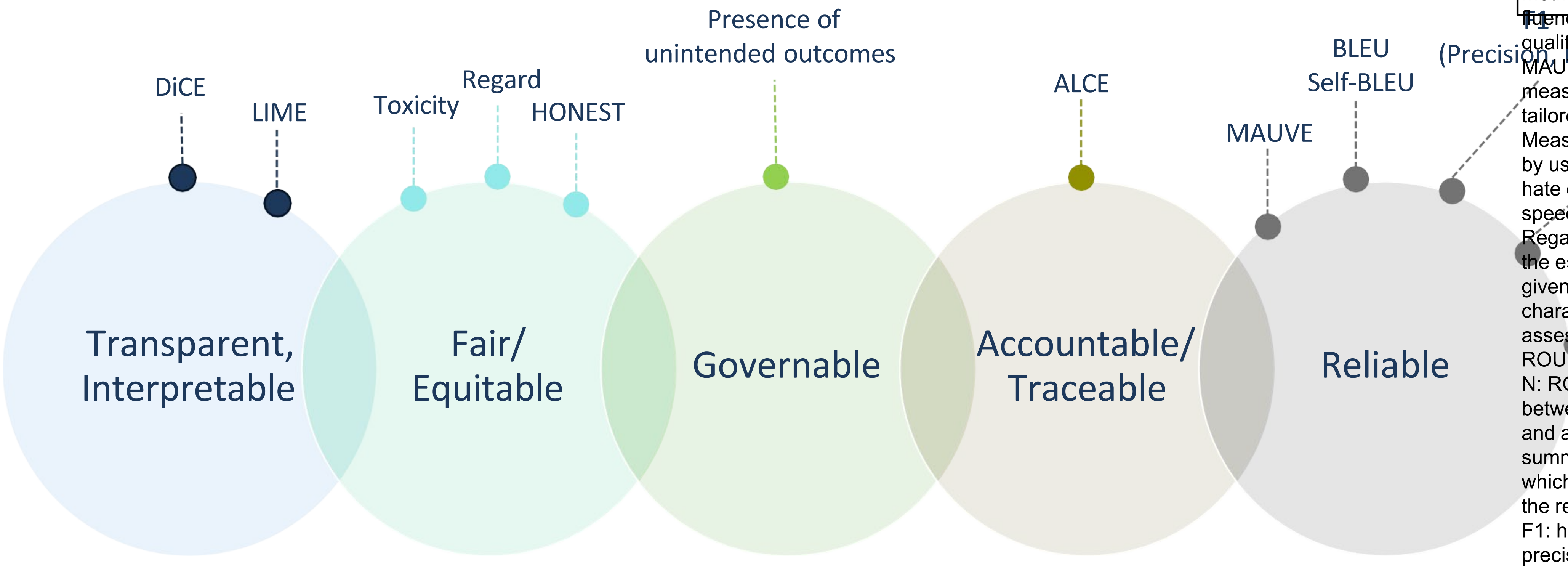
Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
No driving automation	Driver assistance	Partial automation	Conditional driving automation	High driving automation	Full driving automation

Automation levels are tailored to individual use cases

Maturity model aligns to automation levels. Higher automation necessitates a more advanced maturity model for evaluation and assessment.

CDAO

Metrics and Responsible AI

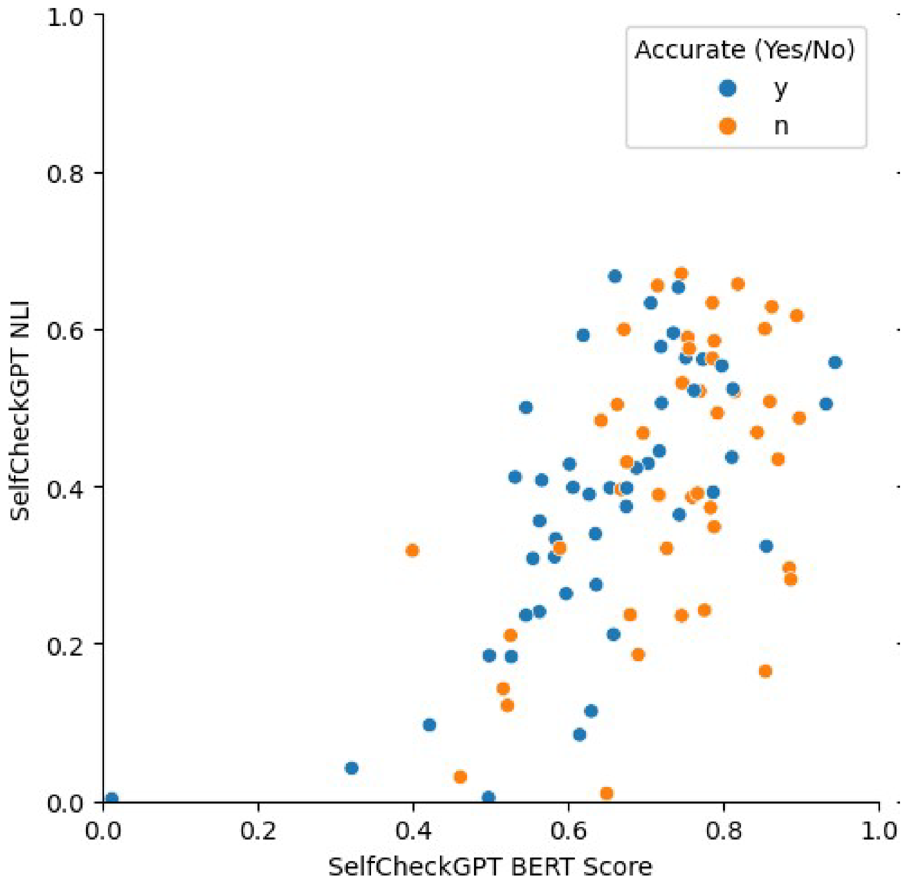


- Metrics target Responsible AI areas (though coverage varies)
- Metric evaluation is an important step to development of a maturity model

Presenter Notes
2024-02-13 17:24:50

MAUVE: Similarity metric of two strings (generated text and reference text)
ALCE: automatic evaluation methods in three dimensions: fluency, correctness, and citation quality. Specifically, we use MAUVE (Pillutla et al., 2021) to measure fluency, propose tailored corr Toxic Fraction: Measures hate speech content by using the R4 Target model, a hate detection model, as a hate speech classifier
Regard Measurement: returns the estimated language polarity given selected identity characteristic(s) HONEST: assess gendered stereotype bias ROUGE-1, ROUGE-2, ROUGE-N: ROUGE-N: is an n-gram recall between a candidate summary and a set of reference summaries Exact Match: Rate at which generated string matches the reference exactly.
F1: harmonic mean of the precision and recall. It can be computed with the equation: F1 = 2 * (precision * recall) / (precision + recall)
BERTScore: BERTSCORE computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embedding. Requires a reference sentence.
SelfCheckGPT: SelfCheckGPT can: i) detect non-factual and factual sentences; and ii) rank passages in terms of factuality

Common Pitfalls



- Initial experimentation results show that **metrics do not always align with SME evaluation**
- SelfCheckGPT measures consistency when repeatedly sampling from the LLM, and relates that to factuality, or likelihood of hallucination.
  - NLI or BERT Score > 0.5 suggests hallucination is more likely
- SelfCheckGPT has multiple methods of calculation including leveraging NLI and BERT Score.
  - Accuracy values from SMEs have minimal relationship with SelfCheckGPT BERT score, but no relationship with the NLI score. (See distributions on edges of graph)
  - Other work has shown promise using this metric, mileage may vary by use case

Objective metrics within Generative AI are an active area of research: *Can we use the LLM as a judge? Should metrics be consistent across use cases or vary? How should metrics be communicated to end users to facilitate the most effective adoption of models in operational contexts?*

Common Pitfalls: Assessing Telemetry Metrics is Non-Trivial

- Gold standard assessment should compare subject matter experts (SMEs) to metrics to assess:
  - **When** are they useful?
  - **Where** (in what contexts) are they useful?
  - **What component** is useful (i.e., is there a threshold? How do they fit into a maturity model?)
- Assessment is inherently time consuming, and non-trivial

Experiment #1 Metrics
SelfCheckGPT: When sampled repeatedly, how consistent are model responses. Leverages the intuition that hallucinations are more likely to be *not* consistent. Include two methods of calculation (NLI and BERT Score)
Self-BLEU: Similarity between a pair of sentences.
Perplexity: How well has the model learned the training set? (Lower values are better)
RAGAS (Answer Similarity, Answer Relevancy, Context Recall, Context Precision): Four separate metrics exploring how well RAG is performing at providing information relevant to the prompt, information that is included in the answer, similar answers, and answers that are relevant to the question. Uses the LLM-as-a-judge.
Toxicity: Fraction of sentences that include toxic or harmful language in the response.
SME Evaluation Metrics: Manual evaluation by SMEs on response (1) Accuracy and (2) Operational Usefulness

LLM Advancements Towards Responsible AI

	Accountable/ Traceable	Equitable	Reliable	Transparent/ Interpretable	Governable
<u>System Architecture</u> **Problem:** Real-time or perishable data are not available to LLMs. **Promising solutions and areas of research:** RAG, Context construction (reranking to optimize attention to most relevant information), Self-RAG (prompting the model to retrieve more information when needed)	X	X	X	X	
<u>Supplemental Knowledge</u> **Problem**: Models lack underlying "knowledge" yielding hallucinations. **Promising solutions and areas of research:** Model augmentation (e.g., using graph or semantic databases)	X		X		
<u>Efficient fine-tuning,</u> **Problem**: Full fine-tuning of models is computationally prohibitive. Knowledge loss is understudied, but certainly a side effect of full fine tuning. **Promising solutions and areas of research:** Adapt transformer architecture to less computationally intensive methods			X		
<u>Efficient Inference</u> **Problem:** Attention is computationally expensive, but a critical component of the encoder-decoder model. **Promising solutions and areas of research:** FlashAttention (smart kernel implementation), Sliding window attention to reduce computation (e.g., Mistral), Quantization (can reduce computational needs with minimal performance loss)			X		X
<u>Trust</u> **Problem:** Misinformation and hallucinations may be common and are difficult to identify. Differing policy stances of responsibility of trustworthiness (model builders? Developers? End users?) **Promising solutions and areas of research:** Metric development and standardization for misinformation detection (e.g., TrustLLM)	X	X	X		

LLM Systems Acquisition

- SomeUses of LLMs:
  - Transformation:
    - Re-arranging data and information for efficient consumption
  - Retrieval:
    - Information recall, summarization, information extraction
    - Multi-modal
  - Reasoning / Knowledge Utilization
    - Knowledge regurgitation
    - Knowledge synthesis
    - Task planning, Autonomous agents
  - Ideation:
    - Course of action alternatives

- Interface to LLMs
  - Natural human language is used to provide
    - Task instruction: summarize this document
    - Behavioral instruction: short summary
    - Variables (explicit and implicit): for a commander (in the US forces, NATO, etc.)
  - Programmability: Appropriate context must be specified.
    - Explicit context mechanisms.
    - Input / Output filtering / guardrails
  - Explainability: Relevant contextual completion must be provided back to the user

CDAO

Use Case Functional Requirements Template

A draft rubric for communicating the Generative AI needs of DoD Projects.
To see the definition for the levels or how they map to the RAI Principles, expand the hidden columns.

		Generative AI Attributes / Measures / Capabilities	Use Case 1	Use Case 2	Use Case 3	Use Case 4
Risks and Concerns related to Responsible Artificial Intelligence	Knowledge and Abilities	**Accuracy**	●	●	●	●
		Hallucinations	●	○	●	●
		Robustness **Injection Attacks**	○	○	○	○
		Robustness **Data Poisoning**	◔	○	◕	◔
		Robustness **Misuse & Abuse**	○	○	●	●
		Limited Context Awareness	◔	◕	◑	●
	Alignment (Ethics)	**Ethics**	○	●	○	○
		Fairness **Lack of Transparency**	◑	●	◑	◔
		Toxicity	○	○	○	○
	Resources	**Run-time Data Sources Required**	○	◔	◔	◔
		Computational Resources Required	◑	◔	◕	○
Foundational Tasks	Natural Language Understanding	**Sentiment Analysis**	○	○	○	◕
		Text Classification	○	○	○	○
		Natural Language Inference	◔	◔	◑	◕
		Text Similarity	◕	○	○	●
		Entity Extraction	○	◔	◕	●
		Topic Modeling	◕	●	○	◕
	Natural Language Generation	**Text Generation**	●	○	○	●
		Summarization	●	◔	○	●
		Translation	○	○	○	○
		Dialogue	○	○	○	○
		Question Answering	●	◑	◑	●
		Code Generation	○	○	○	○
Capabilities		**Behavioral Tuning**	◑	◔	◔	◕
		Extensible Vocabulary	◑	◔	●	◕
		Knowledge Retrieval	◕	●	●	●
		Multi-Lingual Support	◑	◔	◑	◕

Draft levels for each of the rubric measures.

Measure	○ - Level 0	◔ - Level 1	◑ - Level 2	◕ - Level 3	● - Level 4
Accuracy	Accuracy is not a concern	Accuracy is a marginal concern	Accuracy is a concern, but not a significant one	Accuracy is a significant concern	Accuracy is a critical concern
Hallucinations	Regular hallucinations are not a concern	Basic Hallucination Detection	Advanced Hallucination Detection	Hallucination Prevention	Continuous Monitoring and Improvement
Injection Attacks	No concerns about injection attacks	The system must be resilient to Basic Injection Attacks	The system must be resilient to Adversarial Injection Attacks	The system must be resilient to Transfer Learning Injection Attacks	The system must be resilient to Model Inversion Injection Attacks
Data Poisoning	No concerns about data poisoning	The system must have basic Data Poisoning Detection capabilities	The system must have advanced Data Poisoning Detection capabilities	The system must have Data Poisoning Prevention capabilities	The system must have Continuous Monitoring and Improvement capabilities
Misuse & Abuse	No concerns about misuse and abuse of the system	Basic Content Filtering	Advanced Content Filtering	Misuse and Abuse Prevention	Continuous Monitoring and Improvement
Limited Context Awareness	No concerns about the system having limited context awareness	Basic Context Awareness	Advanced Context Awareness	Contextual Consistency	Proactive Context Awareness
Ethics	No ethical concerns	Basic Ethical Awareness	Advanced Ethical Awareness	Ethical Monitoring and Auditing	Ethical Governance and Oversight
Lack of Transparency	Fully closed-source models and training data are not a concern	Selective transparency is required	Partial transparency is required	Targeted transparency is required	Anything less than full data and model transparency is of grave concern
Toxicity	Toxicity is not a concern	Basic Toxicity Detection	Advanced Toxicity Detection	Contextual Toxicity Detection	Adaptive Toxicity Detection
Run-time Data Sources Required	Acceptable to have it connect to third-party data sources on a continuous basis to function appropriately	Acceptable to have it access controlled government data sources to accomplish its function appropriately			Should not require accessing any data outside the system to perform intended functionality
Computational Resources Required	No limitations to what hardware is available	Should operate on a rack / small HPC, or can require $10k+ H/W, e.g., A100 GPUs	Should operate on a tailored laptop (with dedicated GPU or other specialized H/W)	Should operate on a standard DoD laptop	Should operate with minimal resource, e.g., handheld device
Sentiment Analysis	No sentiment analysis capability is required	Basic Sentiment Analysis	Advanced Sentiment Analysis	Contextual Sentiment Analysis	Predictive Sentiment Analysis
Text Classification	No Text Classification capability is required	Basic Text Classification	Advanced Text Classification	Contextual Text Classification	Predictive Text Classification
Natural Language Inference	No NLI capability is required	Textual Entailment	Implicature and Presupposition	Logical Reasoning	Commonsense Reasoning
Text Similarity	No text similarity assessment capability is required	Basic Text Similarity Assessment	Intermediate Text Similarity Assessment	Advanced Text Similarity Assessment	Expert Text Similarity Assessment
Entity Extraction	No entity extraction	Basic entity extraction	Advanced entity extraction	Context-aware entity extraction	End-to-end entity extraction with active learning
Topic Modeling	No topic modeling	Basic topic modeling	Probabilistic topic modeling	Context-aware topic modeling	Dynamic topic modeling
Text Generation	Basic Technical Text Generation	Advanced Technical Text Generation	Contextual Technical Text Generation	Creative Technical Text Generation	Proactive Technical Text Generation
Summarization	Summarization capabilities are not required	Summarization is desired but not required.	Some ability to summarize is required but not a priority.	Accurate summarization is important.	Accurate summarization is critical.
Translation	Translation capabilities are not required	Basic Translation	Advanced Translation	Contextual Translation	Adaptive Translation
Dialogue	Dialogue capabilities are not required	Basic Dialogue	Interactive Dialogue	Contextual Dialogue	Adaptive Dialogue
Question Answering	No question and answering capability is required	Basic Question & Answering	Intermediate Question & Answering	Advanced Question & Answering	Expert Question & Answering
Code Generation	The system is not required to generate code	Basic Code Synthesis	Intermediate Code Synthesis	Advanced Code Synthesis	Autonomous Code Synthesis
Behavioral Tuning	No behavioral tuning capability is required	Basic Behavioral Tuning	Intermediate Behavioral Tuning	Advanced Behavioral Tuning	Expert Behavioral Tuning
Extensible Vocabulary	No need to have an extensible vocabulary	Basic Extensibility	Intermediate Extensibility	Advanced Extensibility	Expert Extensibility
Knowledge Retrieval	No need to support knowledge retrieval	Basic Knowledge Retrieval	Intermediate Knowledge Retrieval	Advanced Knowledge Retrieval	Expert Knowledge Retrieval
Multi-Lingual Support	No need to support multiple languages	Basic Multilingual Support	Intermediate Multilingual Support	Advanced Multilingual Support	Expert Multilingual Support

A draft mapping for how critical each measure is to the Responsible AI Principles.

Measure	Responsible	Equitable	Traceable	Reliable	Governable
Accuracy				100%	
Hallucinations				100%	
Injection Attacks				100%	
Data Poisoning				100%	
Misuse & Abuse	100%				
Limited Context Awareness				100%	
Ethics		100%			
Lack of Transparency	100%				
Toxicity	100%				
Run-time Data Sources Required		50%			50%
Computational Resources Required		100%			
Sentiment Analysis					
Text Classification					
Natural Language Inference					
Text Similarity					
Entity Extraction					
Topic Modeling					
Text Generation					
Summarization					
Translation					
Dialogue					
Question Answering					
Code Generation					
Behavioral Tuning					
Extensible Vocabulary					
Knowledge Retrieval					
Multi-Lingual Support					

Workflow Acceptability Criteria

Generative AI Attributes / Measures /			Use Case 1	Use Case 2	Use Case 3	Use Case 4	○ - Level 0	◔ - Level 1	◑ - Level 2	◕ - Level 3	● - Level 4
Risks and Concerns related to Responsible Artificial Intelligence	Knowledge and Abilities	Accuracy	●	●	●	●	Accuracy is not a concern	Accuracy is a marginal concern	Accuracy is a concern, but not a significant one	Accuracy is a significant concern	Accuracy is a critical concern
		Hallucinations	●	○	●	●	Regular hallucinations are not a concern	Basic Hallucination Detection	Advanced Hallucination Detection	Hallucination Prevention	Continuous Monitoring and Improvement
		Robustness Injection Attacks	○	○	○	○	No concerns about injection attacks	The system must be resilient to Basic Injection Attacks	The system must be resilient to Adversarial Injection Attacks	The system must be resilient to Transfer Learning Injection Attacks	The system must be resilient to Model Inversion Injection Attacks
		Robustness Data Poisoning	◔	○	◕	◔	No concerns about data poisoning	The system must have basic Data Poisoning Detection capabilities	The system must have advanced Data Poisoning Detection capabilities	The system must have Data Poisoning Prevention capabilities	The system must have Continuous Monitoring and Improvement
		Robustness Misuse & Abuse	○	○	●	●	No concerns about misuse and abuse of the system	Basic Content Filtering	Advanced Content Filtering	Misuse and Abuse Prevention	Continuous Monitoring and Improvement
		Limited Context Awareness	◔	◕	◑	●	No concerns about the system having limited context awareness	Basic Context Awareness	Advanced Context Awareness	Contextual Consistency	Proactive Context Awareness
	Alignment (Ethics)	Ethics	○	●	○	○	No ethical concerns	Basic Ethical Awareness	Advanced Ethical Awareness	Ethical Monitoring and Auditing	Ethical Governance and Oversight
		Fairness Lack of Transparency	◑	●	◑	◔	Fully closed-source models and training data are not a concern	Selective transparency is required	Partial transparency is required	Targeted transparency is required	Anything less than full data and model transparency is of grave concern
		Toxicity	○	○	○	○	Toxicity is not a concern	Basic Toxicity Detection	Advanced Toxicity Detection	Contextual Toxicity Detection	Adaptive Toxicity Detection
	Resources	Run-time Data Sources Required	○	◔	◔	◔	Acceptable to have it connect to third-party data sources on a continuous	Acceptable to have it access controlled government data sources to accomplish			Should not require accessing any data outside the system to perform intended
		Computational Resources Required	◑	◔	◕	○	No limitations to what hardware is available	Should operate on a rack / small HPC, or can require $10k+ H/W, e.g., A100 GPUs	Should operate on a tailored laptop (with dedicated GPU or other	Should operate on a standard DoD laptop	Should operate with minimal resource, e.g., handheld device
Foundational Tasks	Natural Language Understanding	Sentiment Analysis	○	○	○	◕	No sentiment analysis capability is required	Basic Sentiment Analysis	Advanced Sentiment Analysis	Contextual Sentiment Analysis	Predictive Sentiment Analysis
		Text Classification	○	○	○	○	No Text Classification capability is required	Basic Text Classification	Advanced Text Classification	Contextual Text Classification	Predictive Text Classification
		Natural Language Inference	◔	◔	◑	◕	No NLI capability is required	Textual Entailment	Implicature and Presupposition	Logical Reasoning	Commonsense Reasoning
		Text Similarity	◕	○	○	●	No text similarity assessment capability is required	Basic Text Similarity Assessment	Intermediate Text Similarity Assessment	Advanced Text Similarity Assessment	Expert Text Similarity Assessment
		Entity Extraction	○	◔	◕	●	No entity extraction	Basic entity extraction	Advanced entity extraction	Context-aware entity extraction	End-to-end entity extraction with active learning
		Topic Modeling	◕	●	○	◕	No topic modeling	Basic topic modeling	Probabilistic topic modeling	Context-aware topic modeling	Dynamic topic modeling
	Natural Language Generation	Text Generation	●	○	○	●	Basic Technical Text Generation	Advanced Technical Text Generation	Contextual Technical Text Generation	Creative Technical Text Generation	Proactive Technical Text Generation
		Summarization	●	◔	○	●	Summarization capabilities are not required	Summarization is desired but not required.	Some ability to summarize is required but not a priority.	Accurate summarization is important.	Accurate summarization is critical.
		Translation	○	○	○	○	Translation capabilities are not required	Basic Translation	Advanced Translation	Contextual Translation	Adaptive Translation
		Dialogue	○	○	○	○	Dialogue capabilities are not required	Basic Dialogue	Interactive Dialogue	Contextual Dialogue	Adaptive Dialogue
		Question Answering	●	◑	◑	●	No question and answering capability is required	Basic Question & Answering	Intermediate Question & Answering	Advanced Question & Answering	Expert Question & Answering
		Code Generation	○	○	○	○	The system is not required to generate code	Basic Code Synthesis	Intermediate Code Synthesis	Advanced Code Synthesis	Autonomous Code Synthesis

CDAO

Workflow Acceptability Criteria

		Generative AI Attributes / Measures /	Use Case 1	Use Case 2	Use Case 3	Use Case 4	
Risks and Concerns related to Responsible Artificial Intelligence	Knowledge and Abilities	Accuracy	●	●	●	●	
		Hallucinations	●	○	●	●	
		Robustness Injection Attacks	○	○	○	○	
		Robustness Data Poisoning	◔	○	◕	◔	
		Robustness Misuse & Abuse	○	○	●	●	
		Limited Context Awareness	◔	◕	◑	●	
	Alignment (Ethics)	Ethics	○	●	○	○	
		Fairness Lack of Transparency	◑	●	◑	◔	
		Toxicity	○	○	○	○	
	Resources	Run-time Data Sources Required	○	◔	◔	◔	
		Computational Resources Required	◑	◔	◕	○	
	anding	Sentiment Analysis	○	○	○	◕	

○ - Level 0	◔ - Level 1	◑ - Lev
Accuracy is not a concern	Accuracy is a marginal concern	Accuracy is a concern, b significant one
Regular hallucinations are not a concern	Basic Hallucination Detection	Advanced Hallucinatio
No concerns about injection attacks	The system must be resilient to Basic Injection Attacks	The system must be res Adversarial Injection At
No concerns about data poisoning	The system must have basic Data Poisoning Detection capabilities	The system must have a Poisoning Detection ca
No concerns about misuse and abuse of the system	Basic Content Filtering	Advanced Content Filte
No concerns about the system having limited context awareness	Basic Context Awareness	Advanced Context Awa
No ethical concerns	Basic Ethical Awareness	Advanced Ethical Aware
Fully closed-source models and training data are not a concern	Selective transparency is required	Partial transparency is
Toxicity is not a concern	Basic Toxicity Detection	Advanced Toxicity Dete
Acceptable to have it connect to third-party data sources on a continuous	Acceptable to have it access controlled government data sources to accomplish	
No limitations to what hardware is available	Should operate on a rack / small HPC, or can require $10k+ H/W, e.g., A100 GPUs	Should operate on a tai (with dedicated GPU or
No sentiment analysis capability is required	Basic Sentiment Analysis	Advanced Sentiment Ar
No Text Classification capability is	Basic Text Classification	Advanced Text Classific

CDAO

LLM Cores are “Ballistic” in their token generation

- There are additional considerations for how to address task specification at the LLM core level.

- LLMs perform iterative production of “next token”
- Image models are wholistic, successively refining the full picture.
- Sequential models are hard to constrain, and hard to correct.
    - Address Topic T, including sections A,B,C.
    - May not ”recognize” that topic A’ is essential to bridging topic A and C.
    - Diffusion models (wholistic) can provide repair at inference time.

- May want to expand the need-list to address different modes for context definition, refinement, and model reprogrammability.
    - Zero-shot learning, multi-shot learning
    - Prompt tuning, Prompt-filtering, milti-agent programming
    - Fine tuning, retraining, knowledge editing
    - New computational architectures for memory and reasoning

- How do you get data into the systems? How do you get data out?
- How do you dynamically alter data during processing?
- How do you provide cross validation or specialized user validation?
- How do you define personas to make the workflow more effective? If we can define canonical agents for each individual workflow, then we can track programmability more effectively.
- How do you develop personalization that scales? Guardrails, state-dependent internal memory, method-of-experts add-on packages?

CDAO

Evolving Concepts for the role of LLMs

Presenter Notes
2024-02-13 17:24:53

But also to [read the title], and develop a baseline concept for integration of Generative AI into High-Consequence use cases.

Generative AIs, like LLMs, seem poised as a differentiating capability in high-level autonomous decision processes.

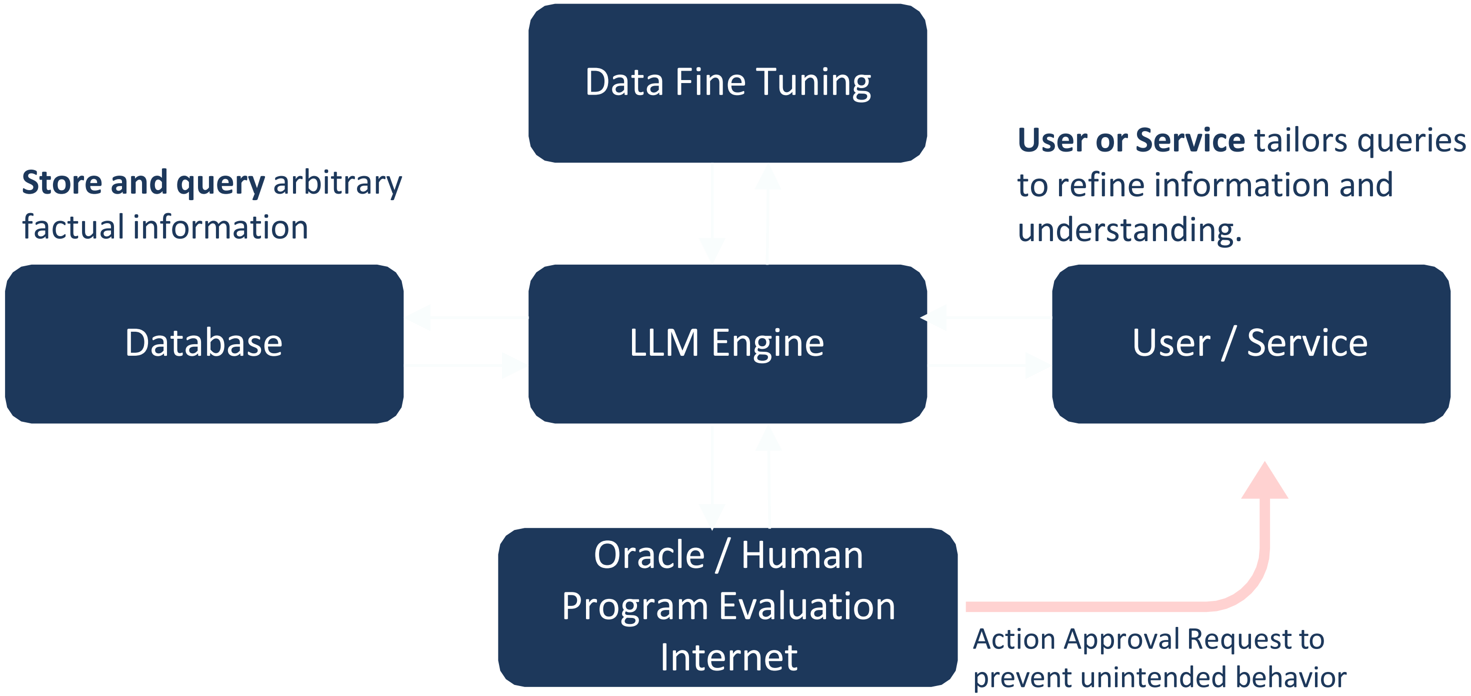
Unpredictability and hidden biases are both the power and the Achilles Heel of Generative AI.

Integration strategies might incorporate guardrails to combine the best of classical and generative algorithms.

Behavioral tuning for predictable responses in narrow input regimes.

Store and query arbitrary factual information

User or Service tailors queries to refine information and understanding.



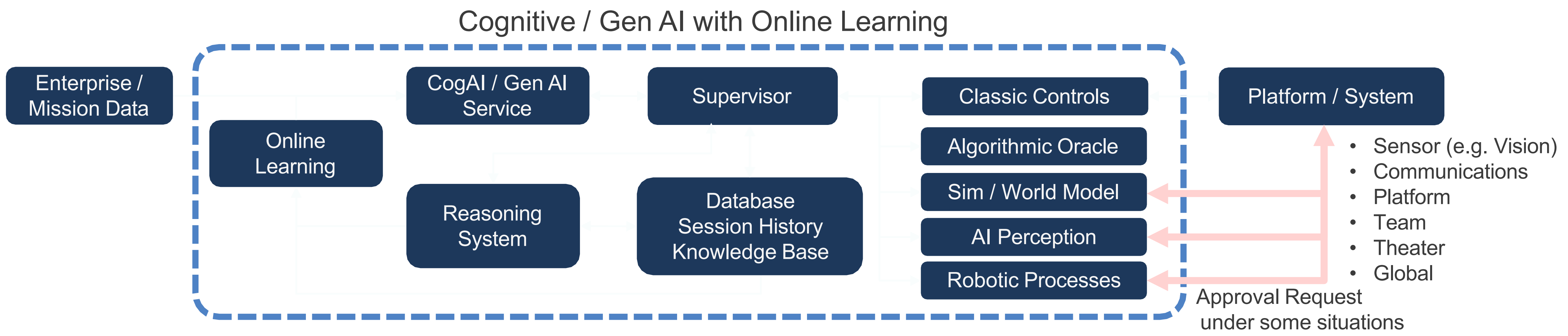
Action Approval Request to prevent unintended behavior

Evaluate and Improve
- Consistency
- Completeness
- Correctness
- AI / LLM Agent Frameworks

Formulate a Framework for the Adoption of Disruptive AI

- Emerging Challenges
  - **Distributed, Denied**
- Necessary properties of Autonomous AI
  - **Modular, Composable, Hierarchically Scalable**
- Guardrails → **Reliable, Trustworthy, and Trusted**

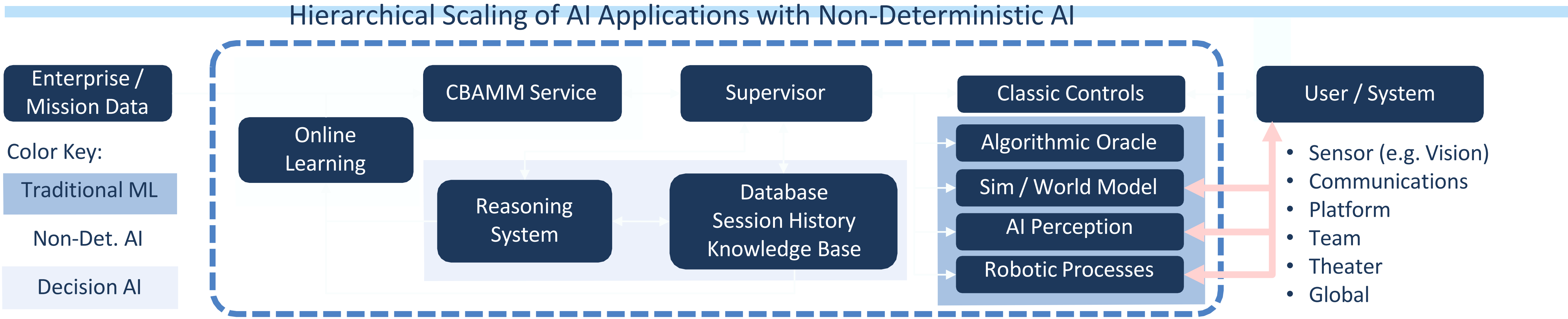
- **Scalable solutions** will ultimately be critical:
  - Ability to dynamically provide custom services: communications, ISR, effects, etc;
  - Develop local CoA to meet commander intent;
  - Autonomy needs to be able to assemble hierarchical solutions from only "end state" directions.



Cognitive / Generative AI plays the role of a "possibilities" engine, a computational analogy to **contextually biased associative memory (CBAM)** for addressing issues in planning, concept retrieval, and action generation. Such contextual bias should be mediated by more predictable and systematic processes.

Georgia Tech Research Institute

DoD Needs to Develop Trusted Architectures For Integrating AI



Model / Data / Design

The model designer maintains model parameters, configuration, and training data.

Reasoning Systems

Evaluates distributional shifts and adjusts online learning parameters.

Provides computational reasoning services: deductive, inductive, abductive. Evaluates internal consistency, completeness, and correctness.

Non-Deterministic AI

Contextually Biased Associative Memory Model (CBAMM) allows adaptation to new environmental stimulus and information.

Provides contextually biased concepts, information, or decisions for current mission objectives; representations for natural human-machine interface

Supervisor Process

Manages CBAM / Reasoning / Factual data and evaluates when to override or correct classical control process.

Information Management

Session history retains long-term memory of Supervisor interactions and supporting evaluation processes.

Database stores arbitrary factual information needed for evaluation of AI concepts.

Interactive Evaluation

Online software execution, internet search, human or software oracle, and robotic processes allow for evaluation and analysis of AI conceptual information vs. factual data.

Feedback on consistency, completeness, and correctness.

Platform / User / System

The physical or virtual platform providing the Autonomy Service

Trust and safety should not be a single point of failure for AI applications. Other components should provide protections (not shown here).

This diagram applies at every level of operation.

Georgia Tech Research Institute

Modeling the Productivity Impact of LLMs

Blair Johnson

blair.johnson@gtri.gatech.edu

Naïve Geometric Model

- Productivity factor, $$\Phi_p = \frac{\text{Human Cost}}{\text{Machine Cost}}$$ r the same result)

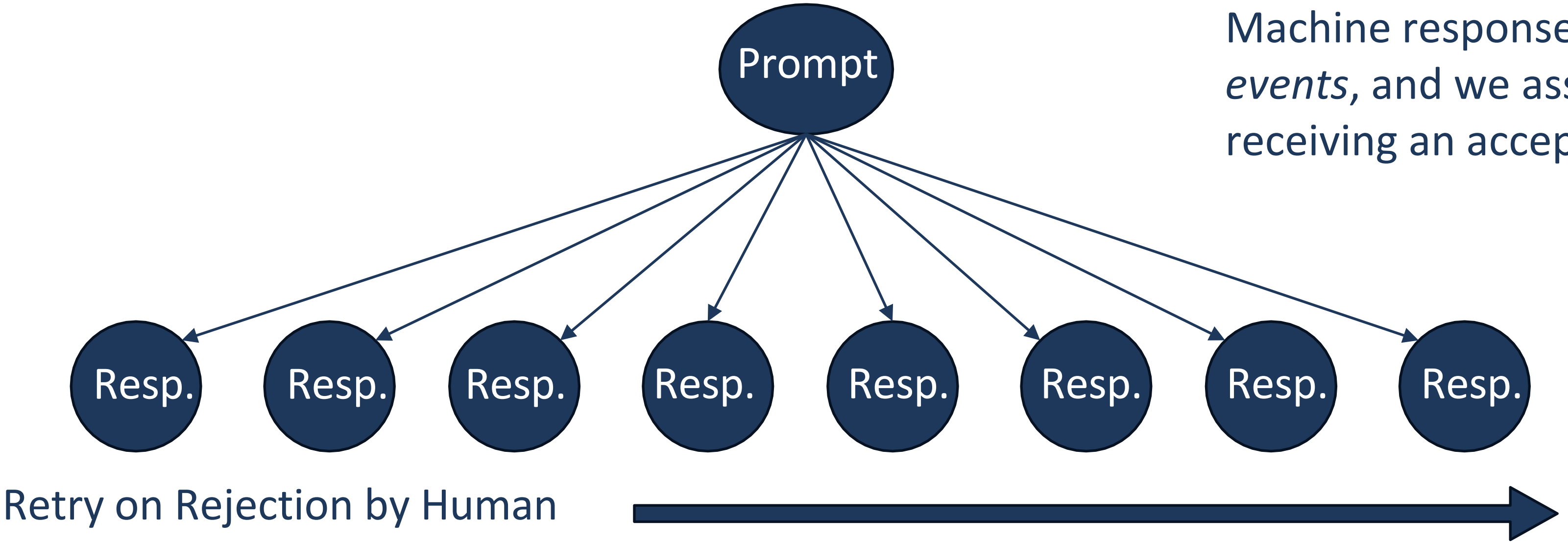
- Time is a simple measure of cost. $$\Phi_p = \frac{\text{Human Time}}{\text{Machine Time}}$$

- Assume the human verifies each model regenerates responses until they get an acceptable answer

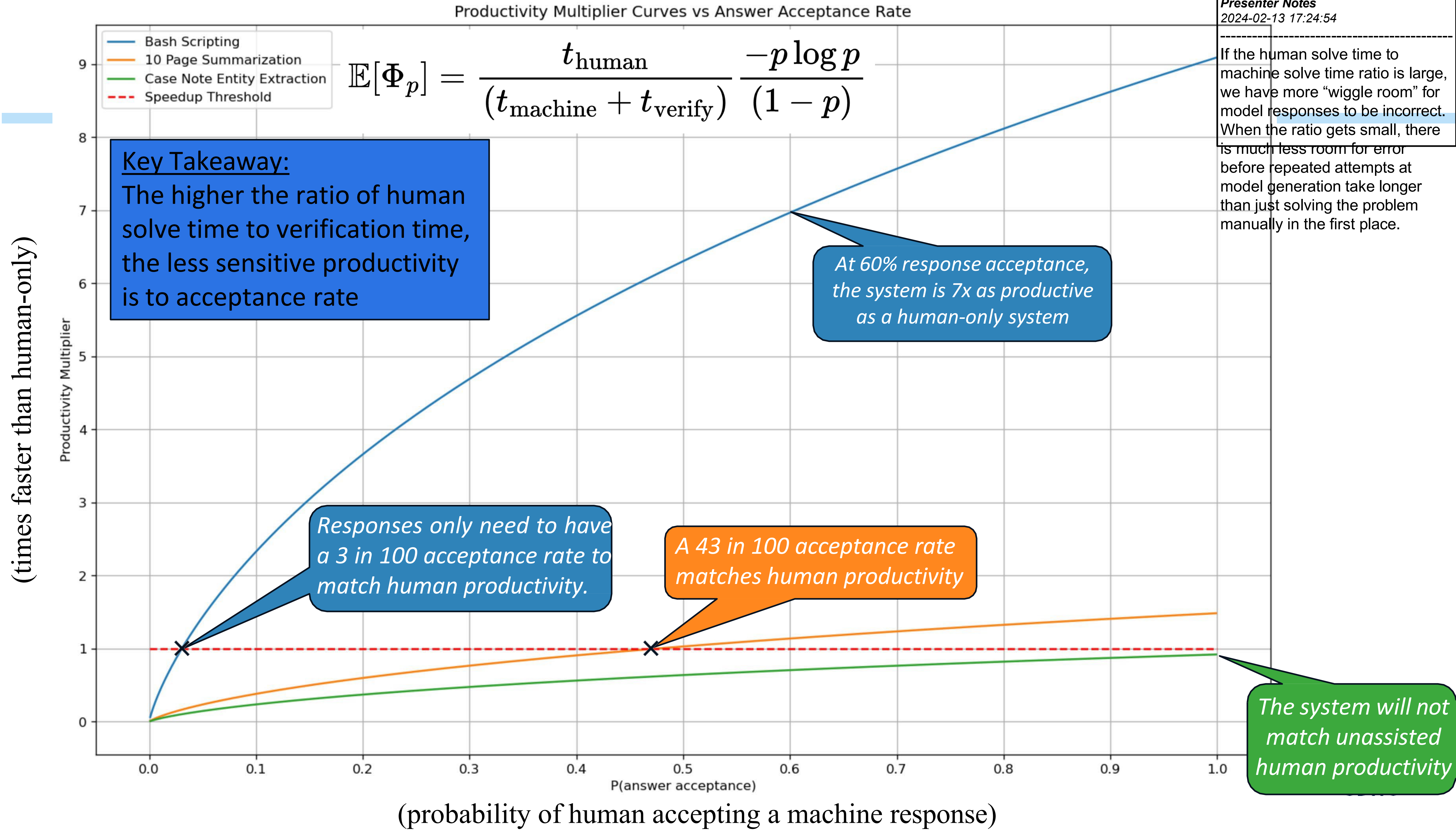
$$\mathbb{E}[\Phi_p] = \frac{t_{\text{human}}}{(t_{\text{machine}} + t_{\text{verify}})} \frac{-p \log p}{(1-p)}$$

Modeling Assumptions

Task Type	Time for Human to Solve	Time for Machine to Solve	Time for Human to Verify Solution
Bash Scripting	5min	3s	30s
10 Page Summarization	45min	20s	30min
Case Note Entity Extraction	34s	3s	34s



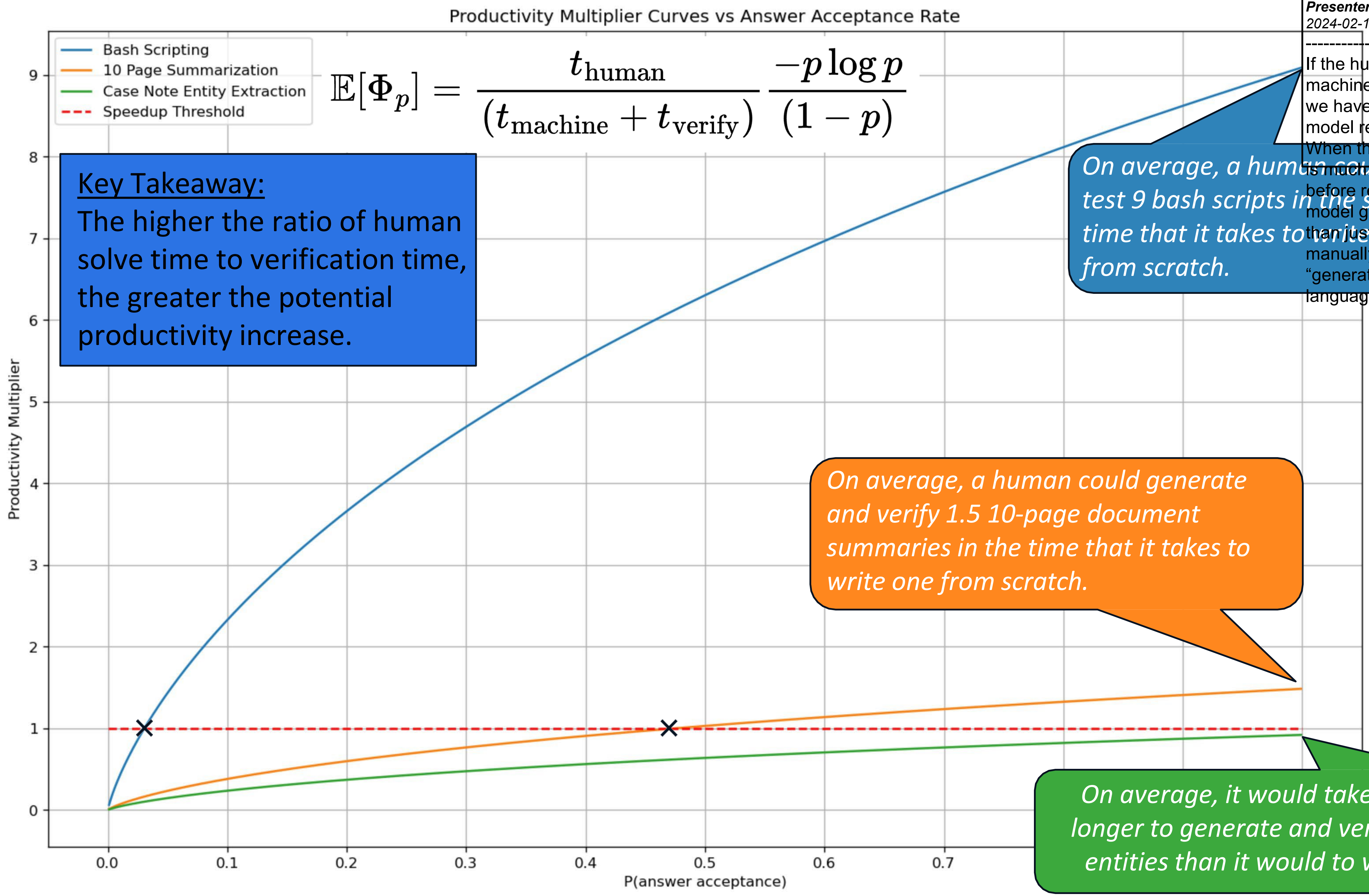
Machine responses are modeled as *independent events*, and we assume that a user retries until receiving an acceptable response.



Presenter Notes

2024-02-13 17:24:54

If the human solve time to machine solve time ratio is large, we have more "wiggle room" for model responses to be incorrect. When the ratio gets small, there is much less room for error before repeated attempts at model generation take longer than just solving the problem manually in the first place.



$$\mathbb{E}[\Phi_p] = \frac{t_{\text{human}}}{(t_{\text{machine}} + t_{\text{verify}})} \frac{-p \log p}{(1-p)}$$

<u>Key Takeaway:</u>
The higher the ratio of human solve time to verification time, the greater the potential productivity increase.

On average, a human could generate and test 9 bash scripts in the same amount of time that it takes to write a bash script from scratch.

On average, a human could generate and verify 1.5 10-page document summaries in the time that it takes to write one from scratch.

On average, it would take a human longer to generate and verify a list of entities than it would to write one.

(times faster than human-only)

(probability of human accepting a machine response)

Presenter Notes
2024-02-13 17:24:54

If the human solve time to machine solve time ratio is large, we have more “wiggle room” for model responses to be incorrect. When the ratio gets small, there is much less room for error before repeated attempts at model generation take longer than just solving the problem manually in the first place. “generate” means using a language model here

Discussion

- The data collection was minimal, representing reasonably high-skill individuals
  - High-skill individuals (in a task) will know more about a task, with lower gap between unassisted task completion time and task verification time
  - Low-skill individuals (in a task) will need to learn new material and refresh on old material to complete the task unassisted, resulting in a much larger gap between unassisted time-to-completion and task verification
  - The bottom line: LLM task multipliers should get larger with decreasing task skill.
- These curves represent averages
  - The distributions that they measure may not concentrate in probability around these values
- Consecutive trials are not truly independent
  - Humans are stateful, they get tired / bored, have biases
  - Real systems typically contain feedback mechanisms
- Time is not the only cost
  - Cognitive load, response quality, latency, compute cost, etc. are also important
- Measuring P(acceptance) is difficult
  - Requires marginalizing over all people, prompts, and model responses
  - This is where large task-representative benchmarks would come in

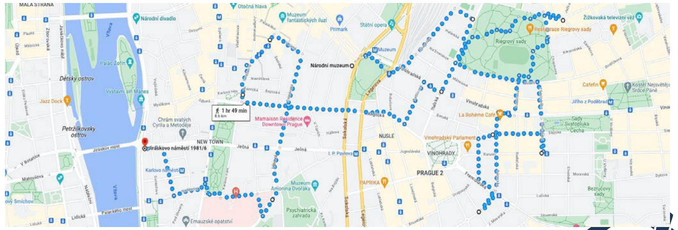
A Case Study in Integrating Disruptive Innovation for DoD

- DoD needs to understand how and when to adopt Generative AI.
- Context behind Generative AI:
  - Breakthrough exploratory research flips the familiar strategic research paradigm on its head.
    - Strategic: Where are we going? How do we get there?
    - Reality: Where did we end up? How did we get here?
  - How did we get here?
    - LLMs are big statistical regressions over a giant corpus of human generated text.
    - But this corpus contains all the “Great Conversations” about the essence of what it is to be human, as well as most major components of science, literature, philosophy, events, etc.
  - Where did we end up?
    - In a new place we didn’t quite imagine.
    - Now we need to figure out what it’s all about.

Where did we end up?

The Dancing House, Prague CZ (Wikipedia)

How did we get here?



Georgia Tech Research Institute

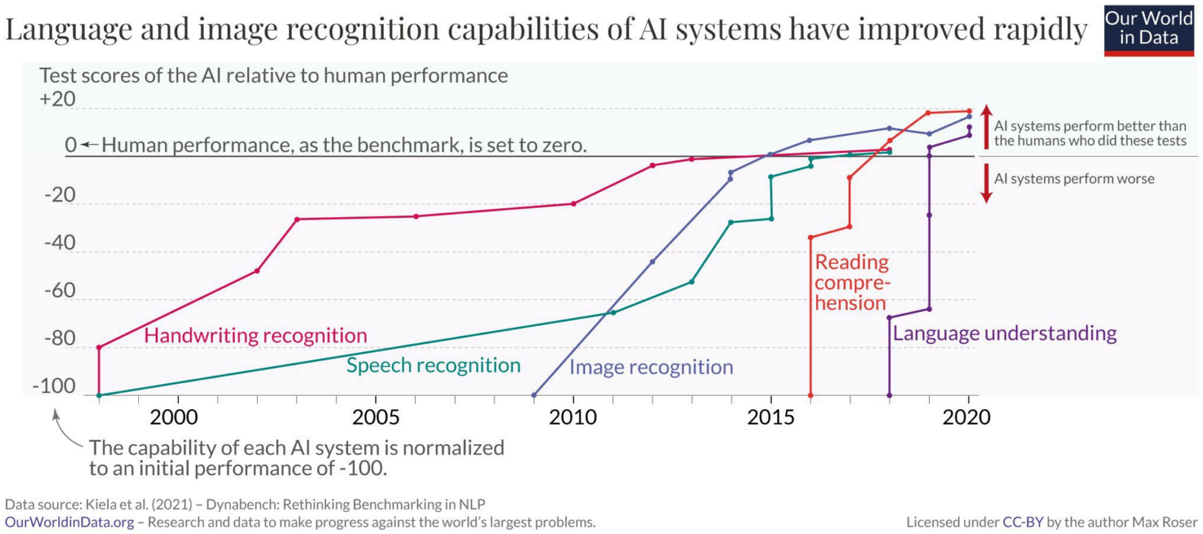
What is troubling? What are threats and risks for stakeholders?

- Malware / Exploit Diversification

- Automated Social Engineering, Social Media Attacks
  - Exquisite Personalization
  - Rapid CoA exploration and exploitation

- User misunderstanding of the capability, design, and implementation of the GenAI processes
  - What new capabilities will the iPhone have in 5 years?
  - One-shot question-answer

- Timeline Compression



Workflow Considerations & Human-Machine Teaming with LLMs

There are both great possibilities and great opportunities for risk with LLMs from an HMT perspective.

- Prompt Sensitivity: LLMs are sensitive to how prompts are worded. Even small changes in the syntax and semantics of a prompt can result in large changes in LLM output.
- Trust and Ubiquity: LLMs have a low barrier to entry for users and many potential applications, so their output can quickly appear in many contexts. Overreliance on these outputs is problematic if they are faulty and becomes riskier for high-stakes use cases.
- Anthropomorphism: Due to the inherent human-like communication of LLMs, their output can mimic social cues that alter human-machine team effectiveness, positively or negatively.

CDAO

LLM Maturity Models & Workflow/HMT

- The Human-Machine team that is formed by working with an LLM creates a subprocess in an overall product workflow.
- There must be analysis of the places where this team improves productivity and increases knowledge – or alternately, adds cost and introduces unacceptable risk – before an LLM is incorporated into a use case.
- Similarly to the evaluation of LLM characteristics, these HMT considerations may be measurable and mappable to maturity levels.

CDAO

Thank you!