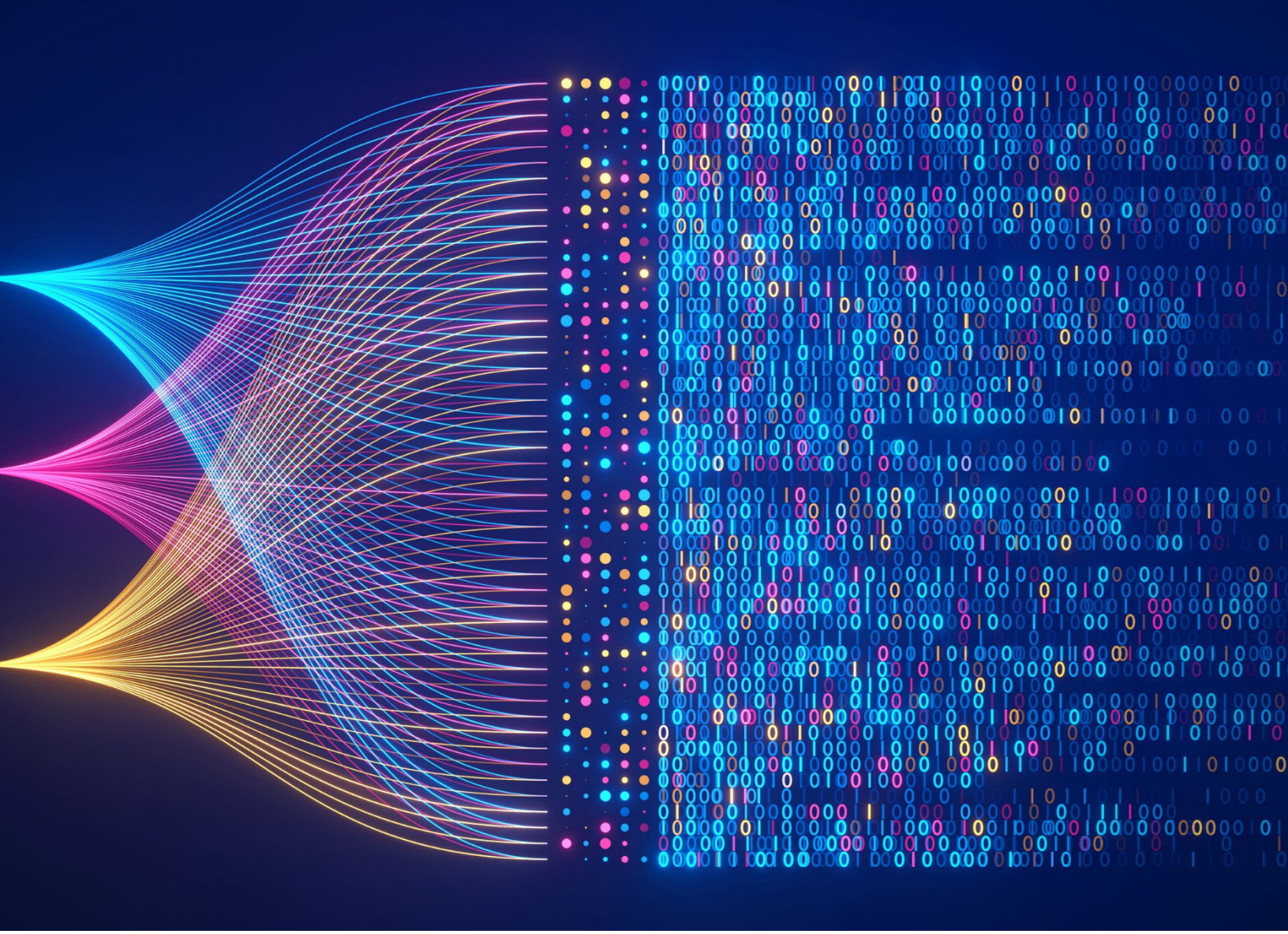


Foundation Models at the Department of Homeland Security:

Use Cases and Considerations

April 2023

Science and Technology

Executive Summary

This report is the first in a series of studies that examines the impacts and opportunities of emerging technologies on the missions of the U.S. Department of Homeland Security (DHS). The focus of this analysis is on the opportunities presented by foundation models (FMs)—the underlying basis for large language models (LLMs). FMs lie at the intersection of generative artificial intelligence (AI) and big data and characterize new approaches to how data can be utilized in many disparate use cases. LLMs have brought attention to how FMs play an important role in uses of data and AI with respect to language. Today, many of the use cases build on industry-created, language-based FMs. But the value extends more deeply to richer classes of data. Use cases in new domains—from building on genomic data sets, cyber-related information, or volumetric scans of baggage, cargo, and vehicles—requires more effort in building and training the FM. Furthermore, many of the approaches (hardware and software) designed and optimized for the flourish in language responses in LLMs need to be understood in the new contexts. At the same time, simpler entry points that build on top of industry provided LLMs can provide more immediate practical experience and value.

This report reflects discussions among DHS's overarching science and technology mission leads, select DHS components, private sector representatives on the current state of the art of FMs, national laboratories and academia on the underlying capability of the current LLMs. More broadly, it addresses potentially transformational approaches to the use of AI in government. It builds on a workshop hosted by the DHS Science and Technology Directorate (S&T) and IBM Research on April 20, 2023, on FMs for DHS. This and subsequent reports are also intended to help inform the DHS AI Task Force announced by Secretary Mayorkas[1] and the increasing importance of AI to DHS missions. The report introduces FM concepts in the context of DHS, reviews the technical underpinnings of FMs relevant for DHS consideration, offers observations to the successful operationalization of FMs, contemplates a host of homeland security mission use cases, and suggests opportunities in research and development needed to establish the FM ecosystem.

Amy Henninger, Ph.D.
Senior Advisor for Advanced Computing
Science and Technology Directorate

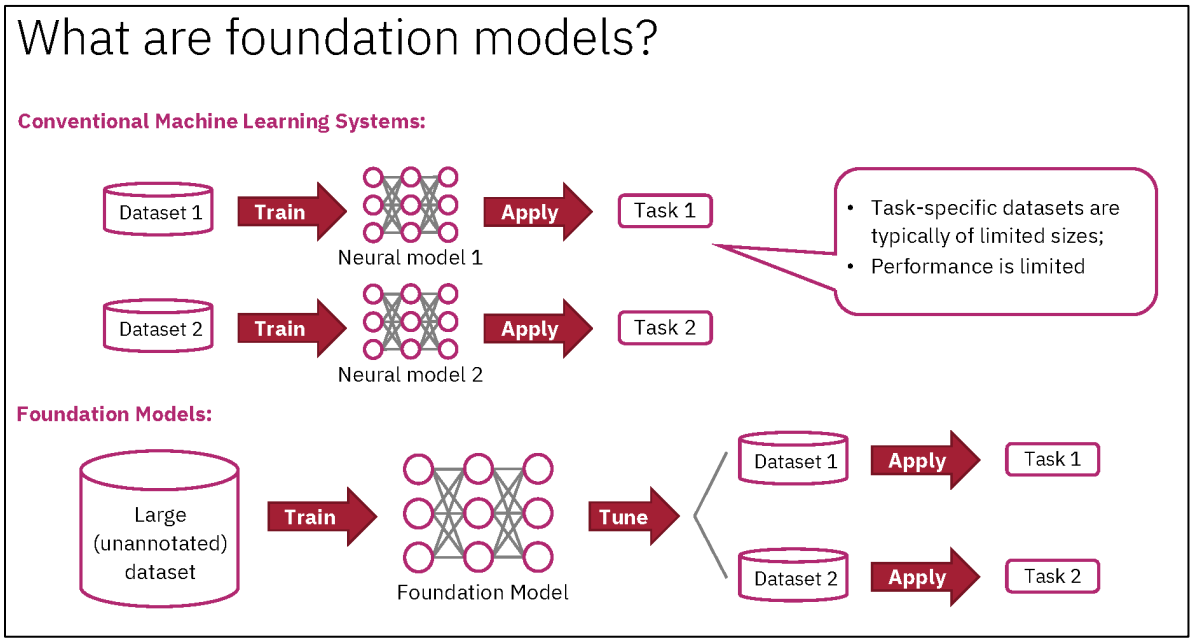
Dimitri Kusnezov, Ph.D.
Under Secretary for Science and Technology

[1] Artificial Intelligence-DHS Science and Technology

CONTENTS

1 Introduction

Considered the modern-day backbone of artificial intelligence (AI), a foundation model[2] (FM) is a type of machine learning model that is trained on a broad set of general domain data for the purpose of using that model as an architecture on which to build multiple specialized AI applications. The versatility of FMs sets them apart from previous iterations of AI models, which have traditionally been customized for a specific task or application. As illustrated in Figure 1.1, by collapsing data and technology across use cases, FMs benefit from increases in the scale and scope of datasets to become more capable and from economies of scale in workflow to become more efficient.[3]



Credit IBM. Reprinted with Permission

Figure 1.1. What are Foundation Models? [4]

Though large language models (LLMs) are currently the best-known examples of FMs, the concept is much broader and emerging applications of these models are being developed for vision and multi-modal AI. These emerging FM classes are trained on other types of data (e.g., imagery, video, protein structures, tabular data, audio, sensor readings, etc.) to accommodate new types of tasks and systems in the future. Table 1.1 provides examples of the kinds of tasks these different classes of FMs can support and the types of data on which the models in these FM classes are trained.

[2] See for example, R. Bommasani, et al, "On the Opportunities and Risks of Foundation Models," Stanford University (2021)

[3] For example, imagine two independent AI-based tools, a speech-to-text application that converts speech to text and a text analytics application that performs summarization. Developers could daisy-chain the systems to provide a speech summarization capability. This would be like concatenating Task 1 and Task 2 under the conventional machine learning systems section from Figure 1.1, and it would come with all the imperfections that accompany the unification of tasks that were not engineered from a common baseline. Alternatively, developers could adopt a foundations model approach, as in Figure 1.1, and use a common engineering baseline, which could support both tasks.

[4] DHS Science and Technology Directorate (S&T) AI workshop with IBM on Foundation Modeling. "What's Next—FMs DHS," page 3, April 20, 2023.

Table 1.1. Examples of Tasks and Data for Different Classes of FMs

FM Class & Maturity	Examples of Tasks	Examples of Data	Examples of Data at DHS
Most Mature ↑ **LLM**	• Sentiment analysis • Information extraction • Text summarization • Text classification • Translation • Question answering • Text generation • Short-form copy • Instruction following • Code generation • Table summarization	• Scrape internet for text • Text data licensed from third-party providers • Code repositories • Books, library holdings • Labeling data provided by humans (for reinforcement learning)	• Reports of Investigations, Suspicious Activity Reports, 911 calls • Cybersecurity and Infrastructure Security Agency's (CISA) 10 terabytes (TB) day of threat intelligence; applications and content; system, logs; system configuration and state, scripts, source code, binary files, IT tickets • Weather reports, social media, news reports • Genomic data • Encounters data, U.S. Customs and Border Protection (CBP) seized asset data, CBP Air and Marine Operations (AMO) Interdiction data, CBP drug seizure data, manifests • Millions of applications for persons seeking to visit or reside in the U.S.
Other Mode[(b)]-informed LLM	• Image description • Image labeling, captioning • Text-based image generation, manipulation • Image retrieval • Visual question answering • Detection of hate speech (including images) • Text-based object detection • Automated video understanding[(a)]	• Image/text pairs from internet • Protein structures • Tabular data	• Forensic sketches • Scanning imagery • Video
Vision	• Image generation • Image classification • Object detection • Video action recognition • Semantic segmentation • Depth estimation • Upsampling/super-resolution	• Images • Video	• Homeland Security Investigations (HSI) videos, forensic sketches • Geospatial imagery • Transportation Security Administration (TSA) five million (M) images/day • CBP imagery: cars, trucks, buses, cargo containers • X-ray diffraction, streaming video, tomography

Least Mature ↓ **Multi-modal**	• Potential for combinations across classes above as well as others…
	  ○ Content Generation[(a,c)]
	  ○ Multi-modal dialogue
	  ○ Cross-modal retrieval
		Anything above plus:
		• Audio to include voice
		• Some sensor readings, such as:
		  ○ Infrared radiation
		  ○ Inertial measurement units
		  ○ Depth estimation (3D)
		  ○ Radar/lidar
		  ○ Time-series
		• Haptic, olfactory, brain fMRI[(a)]
			Anything above plus:
			• Audio from captured phones
			• Voice recordings gathered only by CBP radio frequency (RF) sensors
			• Radar/Lidar AMO Center intelligence, surveillance and reconnaissance (ISR)
			• CBP ISR

[(a)] These are all stretch goals and require significant further development of many technologies to achieve.
[(b)] This other mode is usually vision-based.
[(c)] Songs (LLMs and audio); Movies (LLMs, vision, and audio); Simulation Scenarios (LLMs, vision, audio, and computer code).

The following sections expand on Table 1.1 by reviewing these four popular classes of FMs in the private sector (e.g., LLMs, Other Mode-informed LLMs, Vision-based Models, and Multi-modal Models), discussing their levels of maturity and examples of applications in the private sector. These sections then build on these private sector examples to postulate how the class of FMs might be applied to DHS missions.

1.1 Large Language Models as Foundation Models

LLMs are the most mature class of FMs to date and have set the standards for distinguishing FMs from very large models that only support single tasks. Private sector developed LLM FMs and the tasks they support include well-known models like the Generalized Pre-trained Transformer (GPT) family from OpenAI, providing the horsepower behind popular applications like ChatGPT. GPT-4, the most recently released GPT, is a transformer-style model pre-trained to predict the next token in a document. Other private sector models include BigScience's Bloom, the world's largest open multilingual language model and similar in architecture to GPT-3. Bloom provides the backbone to applications like Bloomchat, which is capable of multiple tasks in the LLM space, as is ChatGPT. As shown in Table 1.1, LLM FMs, such as GPT and Bloom, can support a variety of downstream tasks. Because the private sector has demonstrated successful applications of the technology, it is trivially easy to imagine counterparts supporting DHS missions, such as:

- *The manually intensive and time-consuming tasks of processing millions of applications and petitions for persons seeking to visit or reside in the U.S., which includes dependencies on immigration attorneys or translators to assist non-English speaking applicants, were semi-automated, freeing up valuable human resources to focus on more complex issues.*
- *HSI agents could quickly access and make sense of more than tens of millions of reports through ad hoc, unstructured queries over a voice interface.*
- *Promotion of cultural diversity and inclusion through preservation of indigenous languages by allowing for the creation of more accurate and comprehensive language resources, such as dictionaries, grammar, and language models.*

However, while feasible, there are important unknowns in this space for DHS yet to be investigated. [5]

One of the important considerations in forging a path forward in the development or adoption of FMs will be the business models behind the offerors of capabilities in this space. Options range from develop-your-own FM from scratch, to adopt an open-source pre-trained FM, to adopt a privately developed, pre-trained FM. For example, GPT-4's OpenAI is proprietary, providing no mechanism for practitioners to obtain consistent information on features like model size, hardware, training compute, dataset construction, training methods, etc. Given its proprietary nature, even a highly motivated, well-resourced team would not be able to replicate its training process. FMs like Bloom are open models, both in terms of ownership and visibility. Given the open-source nature and the files available on GitHub, training Bloom is a repeatable process for any highly motivated well-resourced team. These factors, in addition to others, such as the decision to adopt private instances of LLM FMs provided through hyper-scaler services, have significant implications for risks of vendor lock-in, requirements for explainability, privacy and security considerations, or in transparency needed to determine a model's fitness.

[5] Because the model development process is incremental and iterative in nature, as much "art as science," and new to DHS, uncertainties for DHS include things like: "What is the required scope and scale of the model such that it will be 'fit for purpose'?", "How much and what kinds of data will be needed to support the model development and testing such that it is 'fit for purpose'?", "What is the trade space between computing infrastructure and time to achieve required scope and scale?", How does one define and measure "fit for purpose" in the context of the homeland security use case?", "Is it better to build the base FM or to adopt an FM and tailor it to DHS use cases?", etc. These issues are discussed more thoroughly in sections 4 and 5.

1.2 Other Mode-informed LLMs

LLM FMs informed through other modes, usually imagery, are a popular class of FMs that combine LLMs with other models as a way of "grounding" (e.g., learning image features aligned with text) the LLMs to real-world concepts.[6] As a class, these FMs are less mature than LLM FMs, but more mainstream than vision FMs or multi-modal FMs, particularly in the context of meeting multiple use cases. As shown in Table 1.1, these Other Mode-informed hybrids can support a variety of downstream tasks in the text domain as related to images, ranging from image description and labeling or the creation of images using prompts.[7]

Popular examples of FMs in this class work within the realm of AI-driven image generation and manipulation. DALL-E, for example, a (quasi) multi-modal implementation of GPT-3 that swaps text for pixels, can generate images from textual descriptions. While DALL-E excels in generating realistic images from textual prompts, other generative image models transform visuals or enhance them by removing noise and restoring low-resolution or degraded images. The other mode-informed LLM does not always need to be image-based. For example, PaLM-E[8] integrates the 540-billion (B) PaLM LLM and the 22B Vision Transformer (ViT), incorporating real-world raw streams of robot sensor data into language models and establishing a link between words and percepts for use in task learning applications.

As enumerated in Table 1.1, Other Mode-informed hybrid FMs like these two examples can support a variety of downstream tasks. Because the technology has been successfully demonstrated in the private sector, it is trivially easy to imagine counterparts supporting DHS missions, such as:

- *An automated forensic sketching capability enabling witnesses to interact with a computer in developing sketches of persons of interest.*
- *A hate speech detection capability based on image-text pairings scraped from the internet.*

1.3 Vision-based Foundation Models

Vision-based FMs share all the advantages of LLM FMs but apply them to the creation or analysis of images and videos. As a class, vision-based FMs are much less mature than LLM FMs and in their infancy as a technology,[9] though their pattern of growth is expected to follow that of LLM FMs. Currently, the state of practice in the private sector is an array of small models making use of different data to support different simple tasks (e.g., autonomous vehicle navigation). Cutting-edge vision FMs, designed to be used as the backbone for several vision-related tasks (as described in Table 1.1), tend to work both at the image level and at the pixel level. DINOv2[10] is one such example of an open-source FM in this class being developed by Meta AI Research Lab. Serving as a backbone for several vision-related tasks (e.g., instance retrieval, semantic segmentation, and depth estimation), DINOv2 is more aligned with the multi-task support traditionally understood to be a feature of FMs. As another example, Google Research is developing ViT-22B to support research in understanding the scaling of vision transformers as enablers in supporting computer vision research areas in feature extraction that can be used in image recognition, dense prediction (semantic segmentation, depth estimation), video action recognition, etc. To date, the training methods developed demonstrate the potential for achieving "LLM-like" scaling in vision models. In a public-private sector initiative, the National Aeronautics and Space Administration (NASA) and IBM are creating AI FMs to analyze petabytes of text and remote-sensing data to make it easier to

[6] These are classified differently by different researchers, some considering these models LLM-centric, others considering these early pre-cursors to more capable vision FMs, and yet other researchers classifying them as (quasi) multi-modal FMs.

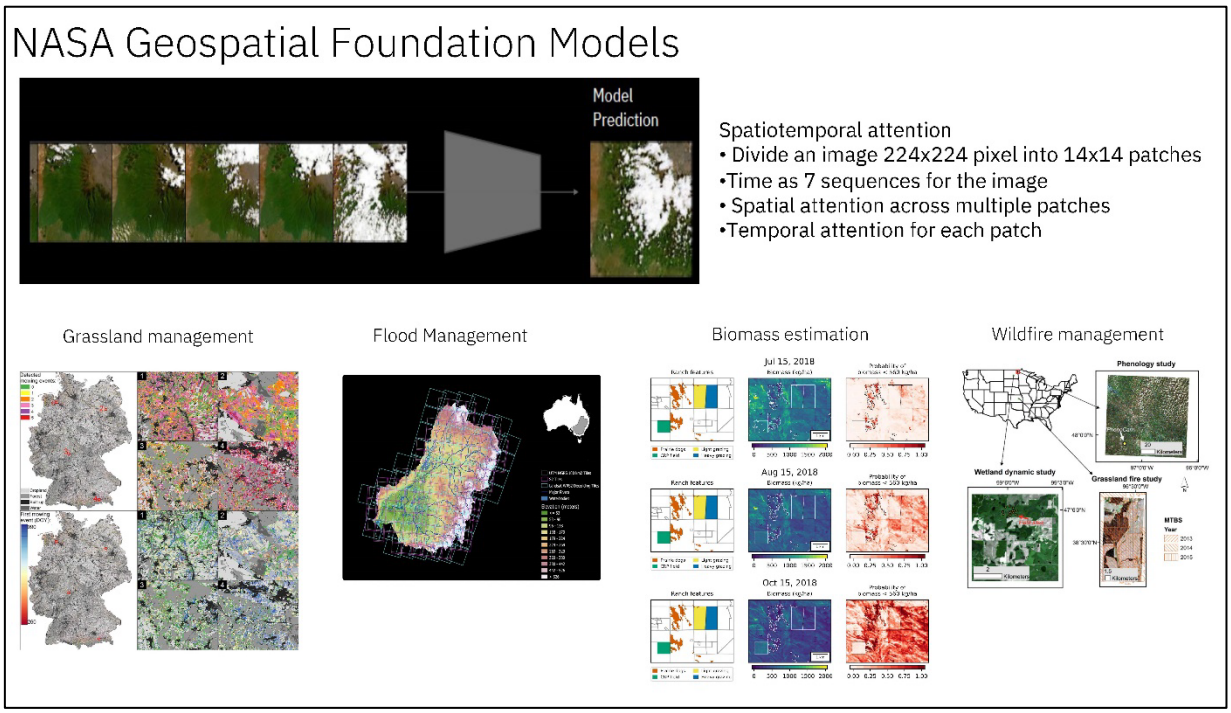
[7] C. Wu, et al., "Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models," (March 2023). Arguably, capabilities like these make use of LLM FMs and add imagery, but do not fully meet the requirements of being an "image-based FM."

[8] D. Driess, et al, "PaLM-E: An Embodied Multimodal Language Model," (March 2023).

[9] Compared to language and image FMs, current video FMs have limited support for video and video-language tasks. An active research community is working on this topic. However, more work is needed before we can develop video understanding systems that are robust and reliable enough to be used in real-world applications.

[10] M. Oquab, "DINOv2: Learning Robust Visual Features without Supervision," (April 2023).

build AI applications tailored to specific questions and tasks (see Figure 1.2 focused on the remote-sensing, data-focused model). One of the promising results from this effort is insight into the utility of transformer architectures, classically used in LLMs, to vision-based modeling.



Credit IBM. Reprinted with Permission

Figure 1.2. Example of Vision-based Foundation Model[11]

With progress in the development of models with these types of capabilities and especially given the mass of image-based data available at DHS for non-intrusive inspection, we can imagine counterpart models like these someday supporting DHS missions with significant computer vision components, doing instance retrieval, dense prediction, etc. In fact, it might even be possible to imagine DHS as leading the way in this space for the entire nation, given the wealth of data available. And, based on NASA's geospatial FM, to include some post-processing work, it is easy to imagine counterparts supporting DHS missions, such as:

- *People from underserved communities impacted by disasters seeking federal assistance could upload imagery to receive preliminary assessments of damage, guidance on relevant services, and have confidence that their case is being routed properly and that all seeking help are treated equitably and fairly.*
- *A novel approach to transportation security based on access to a near limitless amount of stream-of-commerce data in all areas of transportation security, all of which can be virtually guaranteed to be contraband free, and that would understand both X-ray and computed tomography, or CT, images much the way ChatGPT understands languages.*

1.4 Multi-modal Foundation Models

Multi-modality is one of the hottest trends in FMs. As indicated in Table 1.1, multi-modality encompasses all the tasks and data expressed thus far as well as other forms of data (e.g., audio, sensor readings, etc.), enabling models to create, reason about, and optimize content in new ways and to meet the

[11] DHS S&T AI workshop with IBM on Foundation Modeling, "Foundation models are…" page 18, April 20, 2023.

requirements of new tasks. Once more and larger multi-modal models come online, our world will look fundamentally different.[12]

One such example in the class, ImageBind,[13] being developed by Meta AI Research Labs, is an open-source visual FM more closely aligned with the multi-task support traditionally understood to be a feature of FMs. ImageBind enables a rich set of compositional tasks across different modalities (i.e., audio, depth, images and video, text, inertial measurement units (IMU), thermal) and incorporates modalities such as audio into existing models, enabling cross-model retrieval and audio-to-image generation. Especially in domains rich with multi-modal data, such as emergency management or law enforcement, with datasets spanning text-based reports, imagery, videos, speech, digital transactions, etc., multi-modal FMs are a natural way of fusing all the relevant information and discovering strong associations within and across data types.

1.5 DHS Opportunities

Given that real-world use cases and applications of FMs in the private and open-source sectors are accelerating exponentially and DHS is still nascent in its adoption of even conventional machine learning systems, it makes sense to shift our focus to the development of FMs trained on the enormous quantities of data available to the Department. These FMs could be key in providing us with leap-ahead AI capabilities supporting critical mission priorities. In parallel, advances in foundation modeling technology will also likely result in higher-quality, machine-generated content that will be easier to create and personalize for misuse purposes.[14] As such, it will be critically important for DHS to understand the evolution of the technology and its potential for harm.

As the driving force of innovation for the Department, the DHS Science and Technology Directorate (S&T) is evaluating FMs for their potential applications in homeland security use cases and to better understand the threat they will pose when exploited by America's adversaries. Building on initial technical discussions at a DHS S&T AI workshop with IBM Research on FMs,[15] the remainder of this document:

- Reviews the technical underpinnings of FMs,
- Offers observations to the successful operationalization of FMs,
- Contemplates a host of homeland security mission use cases, and
- Suggests a roadmap for the research and development needed to establish the FM ecosystem.

2 Technical Underpinnings of Foundation Models

Driven by growing datasets, increases in model size, and advances in model architectures, FMs offer previously unseen abilities. The key enabler to making FMs work is scale: scale of data, scale of modeling, and scale of computing.[16]

[12] Constrained now by compute, imagine a future Pandora that generates songs tailor-made to your preferences.

[13] R. Girdhar, "ImageBind: One Embedding Space to Bind Them All," (May 2023).

[14] The continued evolution and adoption of this technology in our daily lives must, therefore, also be understood as a serious threat, as it can be used for nefarious purposes.

[15] S&T and IBM Research held a joint workshop on April 20, 2023, to better understand the current and potential use cases for FMs based on S&T's role as the science and technical advisor to DHS and IBM's position as an industry leader in model architecture development.

[16] J. Kaplan, et al., "Scaling Laws for Neural Language Models," (2020). Performance ∝ Data Size x Parameter Size x Compute Size.

2.1 Scale of Data

Data is the lifeblood of FMs; the training data of these models largely determines what capabilities these models can acquire. While there is not yet a standard recommendation for how much data is necessary to train a model, there is a general belief that more data is better than less. The size of the dataset required to train a model depends on what tasks the model will perform. Additionally, the quality of the FM's output is directly related to the quality of the data it is trained on. While current training data selection practices can sometimes lack clear principles, transparency, traceability, and operate with rampant ad hocism, many researchers are advocating for a more principled approach, adopting a data hub-like construct and purposefully managing the selection, curation, documentation, quality assessment, and legal regulations surrounding the training data.[17]

Table 2.1 presents training data-related information on some of the FMs reviewed in sections 1.1 through 1.4, as an example of scale of data FMs might train on.

Table 2.1. Extending Table 1.1 with Information on Scale of Data

Class	Name	Data
LLM	GPT-3[(a)] GPT-4[(a)]	GPT-3 was trained on data, ranging from 17 gigabytes (GB) to 570 GB of data. GPT-4 was trained on between 45 GB of training data to 1 petabyte of training data.
	Bloom	46 natural languages and 13 programming languages. 1.6 terabytes (TB) pre-processed text was converted into 350B unique tokens as Bloom's training datasets.
Other Mode-informed LLM	DALL-E[(a)]	DALL-E, based on an implementation of GPT-3, was trained on a dataset of 12M text-image pairs from the internet (beyond training GPT-3 and with Contrastive Language-Image Pre-Training, or CLIP).
	PaLM-E	PaLM-E[18], integrates the 540B PaLM LLM and the 22B ViT, incorporating real-world raw streams of robot sensor data into language models.
Vision	DINOv2	DINOv2 is pretrained on a curated data set of 142M images.
	ViT-22B	Fine-tuned on ImageNet.
	NASA Geospatial	Harmonized Land Sat and Sentinel-2[19] trained on five years of data from the Southeastern United States. 30m granularity. 1 TB total data volume.
Multi-modal	ImageBind	Image-paired – (image, X) where X is one of text, audio, depth, IMU or thermal data.

[(a)] Since closed model, it is impossible to know, and available estimates vary.

[17] Such as FAIR (findable, accessible, interoperable and reusable) data principles to provide a vision for good data management.
[18] D. Driess, et al., "PaLM-E: An Embodied Multimodal Language Model," (March 2023).
[19] M. Claverie, Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., & Justice, C., "The Harmonized Landsat and Sentinel-2 surface reflectance data set," *Remote Sensing of Environment*, (2018): 219, 145-161.

2.2 Scale of Modeling

The rise of FMs is attributable to the often-repeated mantra of "bigger is better" in machine learning. FMs and transformer architectures owe their origin to Google Brain and the University of Toronto.[20,21] An early popular transformer model, which resulted from that paper, was Bidirectional Encoder Representations from Transformers (BERT).[22] Since the release of BERT, other popular models such as the GPT family (from OpenAI) and even larger models have since been released.[23] Table 2.2 presents model scaling related information on some of the FMs reviewed in sections 1.1 through 1.4.

Table 2.2. Extending Table 1.1 with Information on Model Scale

Class	Name	Model
LLM	GPT-3[(a)] GPT-4[(a)]	GPT-3 was made up of 175 B parameters. GPT 4, the replacement for GPT-3, with up to 100 times more capability, also has at least 175B parameters.
	Bloom	176B parameters.
Other Mode-informed LLM	DALL-E[(a)]	DALL-E, based on an implementation of GPT-3, used 12B parameters to "swaps text for pixels" that trained on text-image pairs from the internet.
	PaLM-E	PaLM-E up to 562B parameters, integrates 540B PaLM LLM and the 22B ViT into a large vision-language model.
Vision	DINOv2	• ViT-S (21M params): Patch size 14, embedding dimension 384, 6 heads, MLP FFN. • ViT-B (86M params): Patch size 14, embedding dimension 768, 12 heads, MLP FFN. • ViT-L (0.3B params): Patch size 14, embedding dimension 1024, 16 heads, MLP FFN. • ViT-g (1.1B params): Patch size 14, embedding dimension 1536, 24 heads, SwiGLU FFN.
	ViT-22B	Modified transformer architecture with 22B parameters.
	NASA Geospatial	ViT architecture.
Multi-modal	ImageBind	The architecture of ImageBind consists of three main components: a modality-specific encoder; cross-model attention module; a joint embedding space.

[(a)] Since closed model, it is impossible to know, and available estimates vary.

2.3 Scale of Computing

Increases in capabilities and efficiencies from the development of FMs come with increases in other computing needs.[24] Studies show, and Figure 2.1 corroborates, that the compute required for training the

[20] A. Vaswani, et al., "Attention is all you need," (2017).

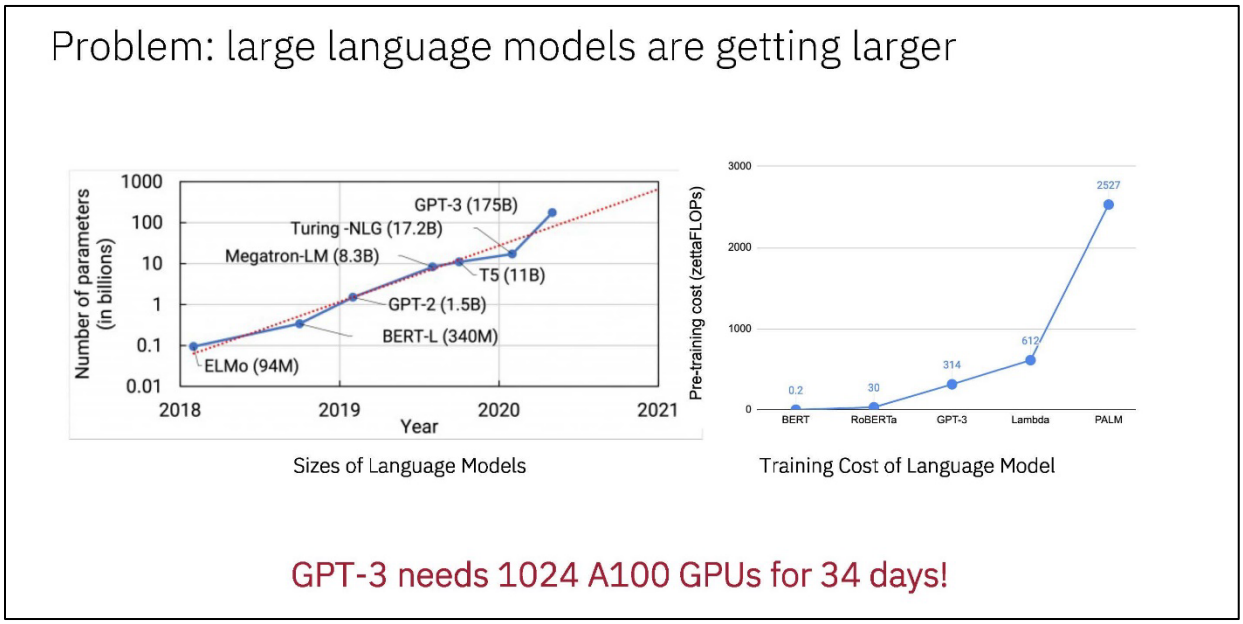
[21] There are numerous types of FMs beyond generative pre-trained transformers (GPTs), such as generative adversarial networks, variational auto-encoders, multi-modal models, among others.

[22] BERT was also a bidirectional model—understanding the context using the entire sequence before making predictions.

[23] Recently, startups such as Hugging Face have started releasing open-source models, which are further accelerating customer adoption.

[24] FMs need highly elastic compute, memory, storage networking for training, along with feasible energy requirements.

models has been increasing exponentially, that model performance scales with the amount of compute,[25] and organizations are limited by the performance of existing hardware solutions and training costs. Industry is rapidly stepping up to meet these requirements with special purpose chips, technology stacks, specialized domain models, and systems, both on-premises and cloud-based, optimized for the development and deployment of very large-scale FMs. Traditional hardware companies have continued to advance graphics processing unit (GPU) architectures.[26] This type of hardware is built around a logic-memory architecture where memory is located off-chip and high-bandwidth memory is used to shuttle the data between logic and memory, creating a significant bottleneck that slows down calculations (i.e., data latency). These architectures also have high power consumption and result in long training times that may take several weeks to converge. Moreover, training large models across many GPUs requires additional code and system configuration that is time-consuming to set up, and the cost and complexity is beyond the capacity of most enterprises. To close this gap, newer tech startups in this space are developing next generation novel hardware designs (e.g., SambaNova Reconfigurable Data Units (RDUs)[27] and DataScale SN30, Cerebras CS-2 and Cerebras-G42 Condor Galaxy 1 (CG-1),[28,29] etc.) with large amounts of distributed memory on-chip, providing significantly greater in memory processing,[30] enabling faster model convergence and lower training times. Not unlike how GPUs proved to be superior to CPUs for graphics tasks, these novel AI Accelerator hardware designs are proving to be superior to GPUs for large-scale machine learning.



Credit IBM. Reprinted with Permission

Figure 2.1. Compute Required for Training FMs is Growing Exponentially[31]

[25] OpenAI, for example, has reported that the amount of compute used in training the largest AI doubles every 3.4 months, significantly outpacing Moore's Law.

[26] GPUs were originally intended for graphics processing and not necessarily machine learning.

[27] The Reconfigurable Data Unit from SambaNova Systems is built with 1TB off chip, dynamic random-access memory, or DRAM, using double data rate (DDR) transmissions and 640 megabytes on chip static random-access memory, or SRAM. In contrast, NVIDIA H100 provides up to 80 GB high-bandwidth memory and Google's Tensor Processing Units (TPUs) 32 GB.

[28] Condor Galaxy is a network of nine interconnected supercomputers, each comprised of a CG-1 that links 64 Cerebras CS-2 systems together into a single, easy-to-use AI supercomputer, with an AI training capacity of 4 exaFLOPs and 54M AI-optimized cores.

[29] Cerebras Introduces Its 2-Exaflop AI Supercomputer

[30] On-chip memory near the compute, helps avoid the latency and overhead of large distributed systems like those using GPUs.

[31] DHS S&T AI workshop with IBM on Foundation Modeling, "What's Next—FMs DHS" page 14, April 20, 2023.

In the machine learning space, massive memory and on-chip memory solutions have a clear advantage over GPUs. Massive memory enables more parameters, higher resolutions, and larger embeddings, and on-chip memory near the compute helps to avoid latency and overhead associated with large distributed systems. For example, in the LLM space:

- ChatGPT, built on top of GPT-4, has a maximum sequence length of 2,000 words and token limits of 4,096, constrained by the reliance on (40-80G) high-bandwidth memory used by GPUs in its NVIDIA-based training compute infrastructure.
- Bloomchat, built on top of Bloom, has achieved a sequence length of 256K and a token count up to 2T, enabled by the RDUs used in the SambaNova DataScale SN30.
- CG-1, optimized for LLM and generative AI, comes with standard support for up to 600B parameter models and extendable configurations that support up to 100 trillion parameter models and offers native support for training with long sequence lengths of 50,000 tokens straight out of the box.

And, in the vision space:

- High-resolution convolution models are readily enabled with SambaNova solution, and seamlessly enable large image processing (within the limits available DDR memory 640 GB) eliminating the need for tiling at lower resolution on GPUs to avoid out-of-memory errors. To date, they have run 100K x 100K 2D images and 1K x 1K x 1K 3D images, while implementing such a high-resolution model architecture would take months or even years on a GPU.
- In contrast, consider running computer vision tasks on an NVIDIA A100 GPU, available with different amounts of video random access memory (VRAM), such as 40 GB and 80 GB. The maximum size of images that can be processed on that GPU is limited by the VRAM and extremely high-resolution images will exceed the VRAM capacity. As a result, tiling techniques are applied to divide the larger images into smaller tiles or patches that fit within the VRAM constraints. These tiling approaches, however, are notoriously lossy at the seams of the tile. Alternative approaches use down-sampling, but those approaches are also lossy with respect to fidelity. NVIDIA A100 GPU have introduced Tensor Cores, specialized hardware units designed to accelerate tensor operations commonly used in deep learning. Tensor Cores can improve performance in certain operations, but they also have specific memory requirements and data format constraints.
- On an image classification task training ResNet with ImageNet, the Time-to-Train (in minutes) on:
  - Low Resource Bow IPU[32] from Graphcore (Processor: AMD EPYC 7742 (2), Accelerator: Graphcore Bow IPU (16)) is 19.636.
  - A30 Tensor Core GPU (Processor: AMD EPYC 7742 (2), Accelerator: NVIDIA A30 (2)) from NVIDIA is 235.574.
  - Low Resource A100 Tensor Core GPU from NVIDIA (Processor: AMD EPYC 7742 (2), Accelerator: NVIDIA A100-SXM-80 GB (8)) is 28.685.

Lastly, as an indicator of power consumption:

- Cerebras has demonstrated one of its recent supercomputers, Andromeda, which puts 16 Wafer Scale Engine (WSE)-2 chips into one cluster with 13.5M AI-optimized cores, delivering up to one extra floating-point operations per second (FLOPS) of AI computing power, or at least one quintillion (1018) operations per second. The system uses 500 kilowatts (KW) of power.
- A single CS-2 system, containing the WSE-2 chips, uses 23kW for about 63 PetaFLOPS (2.7 TeraFLOPs/W), in contrast to a single TPU-4 processing unit from Google, which draws 192W while delivering 275 TeraFLOPS of performance (1.4 TeraFLOPs/W), or a single Ampere series

[32] Intelligence processing unit (IPU).

GPU from NVIDIA, which uses 300-400W delivering 624 TeraFLOPS of performance (1.8 TeraFLOPs/W).
- The Graphcore Bow IPU, AI processor to use Wafer-on-Wafer (WoW) 3D stacking technology, Bow Pod16 delivers over five times better performance than a comparable NVIDIA DGX A100 system, and up to 16% increase in performance per watt.

3 Observations in Operationalizing FMs

While many organizations are eager to adopt FMs, there a several challenges that must be considered and planned for in pursuit of this technology to use it properly and safely in operational or production use cases. FMs are computationally complex and expensive to run, so picking the right use cases and tasks the FM will support is key. It is equally important to understand the business model options, as discussed in section 1.1. Picking the right one is largely dependent on the nature of the task and the level of control or degrees of freedom you want to have in the development and use of your FM.

Other more technical challenges, and the foci of section 3, include improving our understanding of:
- How, what, and when models from industry can be leveraged for DHS missions as well as developing our own capability to develop these models for modalities and data streams where no capability exists. (AI Advancements: Building the Foundation for Foundation Models).
- How to test and evaluate, verify, and validate these models and developing DHS-specific benchmarks. (AI Assurance: Achieving, Ensuring, and Maintaining Model Fitness).
- How we build an appropriate level of trust between the users/analysts and these models. (Mission Assurance: Human Insight and Oversight will be Key).

3.1 AI Advancements: Building a Foundation for Foundation Models

In tune with the analysis in section 2, scale is key to making FMs work. Effective data processing and infrastructure are essential to handle the large-scale data sets. Robust preprocessing techniques, data augmentation, and efficient data storage are all crucial in maximizing the quality and quantity of data used for training. Model scaling research ensures that the FM can be adapted to various task complexities. Scaling up the model with more parameters allows it to capture intricate patterns, while scaling it down enhances its efficiency for low-resource environments. The compute infrastructure supports model scaling through leveraging cutting-edge hardware and harnessing the full potential of advanced compute infrastructure enables faster model training, significantly reducing development time and resource consumption. Multi-modality research empowers the FM to process and understand information from various sources, including text, images, and audio. Integrating multiple modalities enables a deeper understanding of complex data, making the model more contextually aware and capable of handling multifaceted tasks. Portability research ensures that the FM can seamlessly run across different hardware platforms and environments. A portable model is highly versatile, allowing it to be deployed in diverse settings, from edge devices to cloud servers, making AI solutions more accessible and widely applicable.

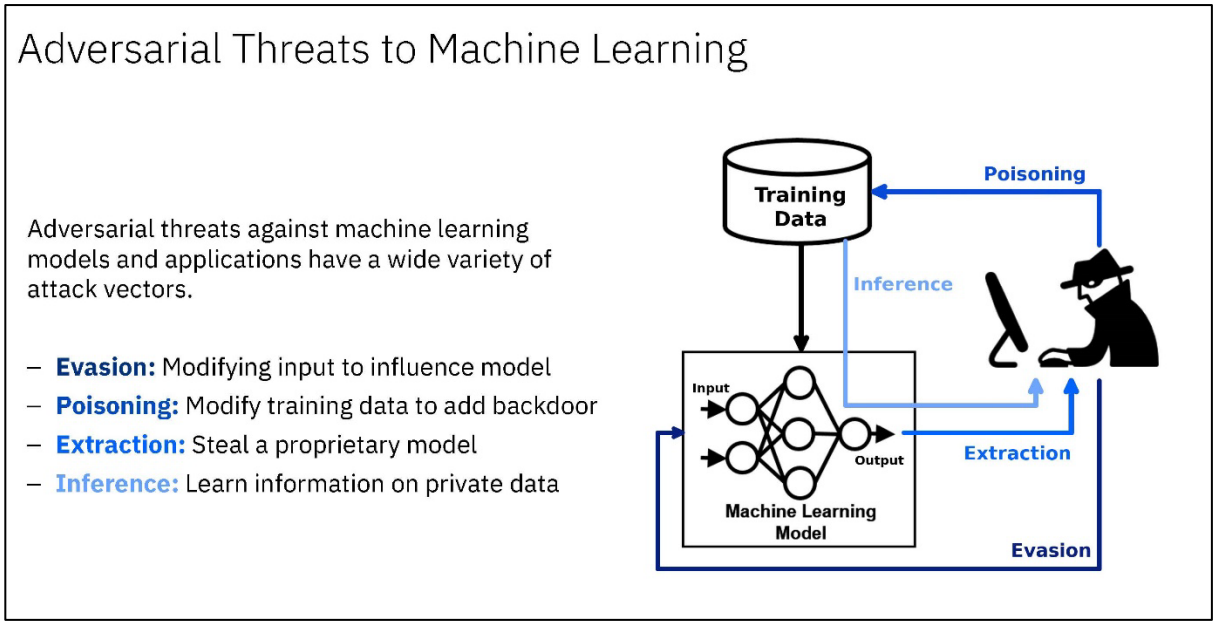
3.2 AI Assurance: Achieving, Ensuring, and Maintaining Model Fitness

Ensuring a model's performance, adaptability, and responsible use are key to knowing that the model is fit for purpose. This requires understanding how the model compares to existing state-of-the-art solutions as well as existing processes surrounding the use cases the FM is targeting, as well as developing the test infrastructure to assess. To maintain the model's relevance and accuracy and ensure it remains fit for purpose over time, it will be necessary to understand its drift, pace required for updating it with new data, and approaches to fine-tuning the model with that data. Model prompting techniques can guide the model's responses, making it more controllable and reliable, especially in critical applications. Finally, research in trustworthiness, to include mitigating biases, interpretability, and developing mechanisms for

adversarial attacks, is crucial to ensure the model's reliability and ethical use. A trustworthy FM will inspire confidence in its users and promote a responsible deployment.

3.3 Mission Assurance: Human Insight and Oversight will be Key

FMs are automation tools; they cannot completely replace human judgement and should not be thought of as having an opinion or personality. When utilized correctly, they can speed up or simplify complex analyses that free up human operators for other tasks. To strike that balance, users must have confidence in the model's performance, understand its behavior, and trust that it aligns with ethical principles. Thus, it will be important to find ways of making the model's decision process transparent to the user, as well as require usability assessments to ensure the model is user-friendly and easy to interact with, even for non-experts. Research in security and privacy will be needed to protect sensitive data and prevent unauthorized access. Moreover, it will be necessary to understand how to prevent, mitigate, respond, and recover to any number of adversarial attacks (see Figure 3.1) and unexpected variations in input data. Finally, to ensure effective integration and information exchange in a system-of-systems context, interoperability research will be required. This FM concept applies well to big decision spaces, so it will likely be better to have a common FM across multiple components than to have multiple competing FMs across multiple components.



Credit IBM. Reprinted with Permission

Figure 3.1. Selected Adversarial Threats to Machine Learning Aspects of Foundation Models.[33]

4 Use Case Deliberations

Throughout the workshop, participants identified potential opportunities to utilize FMs in support of DHS and component activities. This section offers a starting point for the Department to consider the application of FMs and to determine their potential impact on a variety of homeland security operations.

4.1 Law Enforcement to Include Digital Forensics and 911 Services

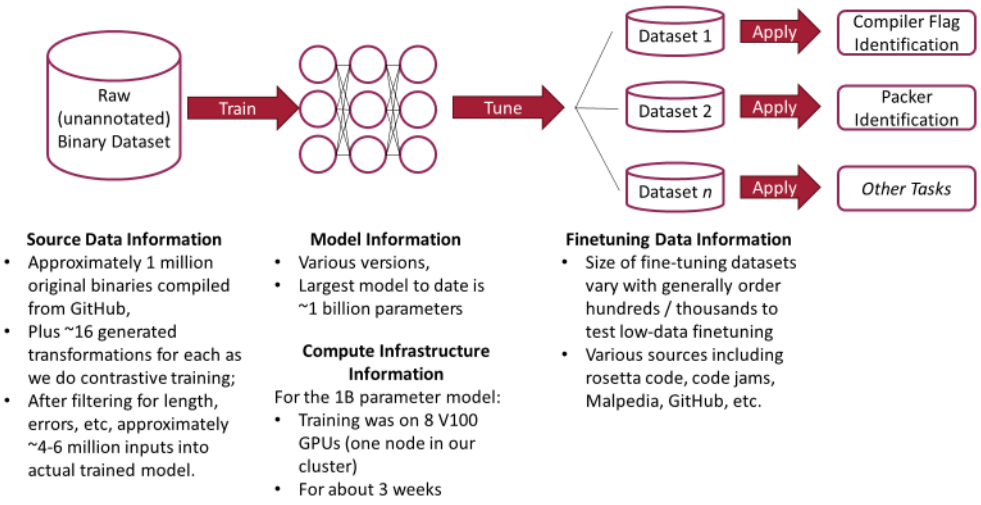
Digital forensic tools are incredibly useful when investigating a number of different types of cases (e.g., child exploitation, counter fentanyl, counterfeiting, etc.), and FMs could support efforts to solve these

[33] DHS S&T AI workshop with IBM on Foundation Modeling. "Adversarial AI" page 23. April 20, 2023.

cases. HSI case agents rely on a variety of reports including Reports of Investigations and Suspicious Activity Reports. These reports, over 40 million, comprise unstructured text narratives that describe fine-grained details about previous and ongoing investigations including the locations, dates, people, and artifacts related to those investigations. The task of retrieving and manually examining possibly several, complex reports, however, is tedious, time-consuming, and prone to error. This process could be significantly improved through use of state-of-the-art LLMs, such as GPT, fine-tuned on HSI-specific corpuses. FMs are also being used for violence detection,[34] and they can be used effectively by homeland security and law enforcement agencies. Next Generation 911 capabilities will one day make it possible to send text or photos to 911 centers, not just calls. As these capabilities are introduced, FMs could be trained on the data to help 911 operators and first responders coordinate their response.

4.2 Cybersecurity

DHS S&T has already begun work on FMs for the cybersecurity domain. A task under DHS S&T's Cyber Analytics and Platform Capabilities project called MIMISBRUNNR is creating an FM for binary analysis. MIMISBRUNNR creates an adaptable mathematical representation of binary semantics that aims to provide a more flexible and holistic description of a program, ultimately impacting the mathematical input to other machine learning tasks. To date, MIMISBRUNNR has created an initial research proof-of-concept FM for the field of binary analysis and showed initial tests on specific tasks (downstream benchmarks) within the binary analysis domain (e.g., compiler flag identification), and made sweeping architectural improvements in order to scale-up the size of the model, the type of data consumed, and the amount of data consumed, leading to drastic improvements in downstream performance.



Credit DHS S&T.

Figure 4.1. Overview of MIMISBRUNNR

This work under DHS S&T is the first of its kind; no one has built a generalizable FM based on binaries. Beyond novelty, this work has the potential to be applied to any binary analysis challenge with a solution that uses an underlying statistical model, without any domain knowledge, (e.g., our model is trained with no direct concept of compiler flags, malware, benignware, etc. Yet, with a limited downstream dataset (order of hundreds to thousands of binaries for each downstream task), it can be adapted to learn across myriad of tasks in binary analysis. Our FM can be used as an initial baseline for another team to improve

[34] X. Shaftgupta, et al, "A Vision Transformer Model for Violence Detection from Real-Time Videos," (December 2021).

upon for a specific downstream task, or as an improvement to an existing downstream task. Rather than starting from scratch with nothing for some very specific downstream tasks that require extreme domain knowledge, our FM can provide a baseline without any domain knowledge necessary. This has the potential to be extremely powerful across all tasks in binary analysis.

Other potential FM use cases spawned by IBM cybersecurity use cases include:

- Behavioral analytics: Categorize and structure user behavior patterns to identify abnormal activity across high volumes of data.
- Threat monitoring: Identify sources of vulnerability and security breach exposure, augmenting threat hunting and threat prevention.
- Automated incident response: Enable learning from each incident, providing feedback and remediation recommendations, including playbook automation.
- Augmented robotic process automation: Automate high volume, basic, repetitive tasks, enabling skilled analysts to focus on higher value work.

Because it is already in digital form and capable of being represented and integrated in purposeful ways, cybersecurity data (e.g., natural language and threat intelligence; applications and content; system, network and application logs; system configuration and state, scripts, source code, binary files, etc.) is a natural fit for FMs supporting a host of cybersecurity tasks. Given that CISA receives on the order of 10TB data per day, essentially for threat hunting, this seems like an area worthy of further exploration.

4.3 Emergency Management

When disasters occur, the Federal Emergency Management Agency (FEMA) needs accurate, real-time information to support response and recovery efforts. Data taken from a variety of sources, including weather reports, geospatial images, social media, and news reports, could be used by an FM to help FEMA quickly coordinate their actions with state and local officials. Translation capabilities would also be beneficial to help those on the ground when translators are not available. As people impacted by a disaster seek federal assistance, an FM could be helpful in directing them through the application process and understand what steps to take next. Additionally, historic flood data and current geospatial data could be combined to support the National Flood Insurance Program.

4.4 Genomics and Drug Discovery

FMs have proven useful at generating new molecules, reducing the time involved with the discovery process. Genomic data from national laboratories like the National Biodefense Analysis and Countermeasures Center could be used to train genomic-based FMs in identifying new zoonotic diseases and creating treatments. Chemistry and materials science literature holds massive amounts of multi-modal knowledge (e.g., text, tables, and images about experimental property measurements, synthesis procedures and methods, organic reaction pathways, etc.). These data could form an FM used to make predictions to discover novel synthesis mechanisms, reveal previously unknown molecular functionalities, and design new molecular structures through computer modeling.

4.5 Non-intrusive Inspection and Scanning

The TSA generates over five million images a day through its screening efforts at airports across the nation. With over seven thousand miles of border to cover, CBP generate a host of imagery (e.g., cars, trucks, buses, cargo containers). Land ports facilitated over $779B in trade between U.S. and Mexico in 2022, putting pressure on the number of inspections completed. Taken together, these sources represent a deep trove of data (e.g., X-ray diffraction, streaming video, tomography) that could be useful in training a visual FM to support the multitude of scanning use cases across components (e.g., instance retrieval, segmentation support, dense prediction, etc.). In fact, it might even be possible to imagine DHS as leading

the way in this space for the entire nation, given the wealth of data available. Beyond those common visual FM tasks, the model could be tailored to a multitude of specific tasks such as identifying suspicious items and flagging baggage for extra screening by TSA, gaining efficiencies over current processes. It is possible that FMs and generative AI could be used to consider potential threats "between bags" or "across bags," broadening a traditional "within bag" concept of operations to assist in identifying coordinated threats.

4.6 Smuggling, Trafficking, Exploitation, and Illegal Activities at the Border

With over seven thousand miles of border to cover, CBP must ensure resources are deployed where they will have the most impact. Fusing a host of data sources (e.g., multispectral data at the Southern border, text reports, imagery [such as cars, trucks, buses, cargo containers], sensor data, down-range media to include social/news media etc.), FM forecasting could help CBP anticipate the timing and location of potential surges in illegal border crossings, and historical data could be used to inform staffing and resource allocation decisions. Add automatic identification system (AIS) data, shipping manifests, data from captured cell phones, or combine with data described in section 4.5 to start building a multi-modal FM able to address multiple tasks.

4.7 Immigration Services

United States Citizenship and Immigration Services processes millions of applications and petitions for persons seeking to visit or reside in the U.S. and from permanent residents seeking to become U.S. citizens. These applicants routinely need help navigating the language barrier and depend on immigration attorneys or translators to assist them, adding to the complexity of the services provided. Also, there are many different forms to be completed, and while some are facilitated through digitized formats, many of these forms must be completed manually. Largely administrative in nature, the services offered range from collecting data, inputting data, processing data, and distributing data. Given these major tasks and a number of complementary tasks (e.g., language assistance), it seems that an LLM FM could be most useful in providing semi-automated support to staff, enabling the staff to focus on the hardest problems in immigration cases.

4.8 Biometrics

Biometrics modeling at DHS is currently focused on the assessment of vendor supplied models, as the AI and machine learning capabilities and expertise within the academic and private sector exceeds current capabilities available within DHS. Yet, DHS has access to large datasets, which remain under-utilized in this important problem space. In some cases, this is because many DHS datasets are sensitive and cannot be shared with researchers and technologists outside the Department. Advanced techniques in foundation modeling, such as transfer learning and retraining, make it possible for DHS to leverage large commercial and academic developed AI and machine learning models and attempt to quickly adapt them for specialized DHS applications. These techniques offer the promise of leveraging pre-trained models developed by academic and industry experts as well as make it possible to optimize and fine tune performance on sensitive non-public DHS data, greatly reducing the time and cost of developing models and improving performance for specialized DHS operations and data.

4.9 Business Applications

Countless business processes at DHS could be facilitated through the use of FMs:

- Software developers can use FMs to develop, debug and test software.[35]
- IT analysts can use FMs to manage help desk operations and IT tickets.

[35] See OpenAI's Codex, Amazon's CodeWhisperer, and IBM's CodeNet.

- Data scientists can use FMs to develop visualizations and code for analytics.[36]
- Data engineers can use FMs to assist in the burdensome but important task of data wrangling.[37]
- Lawyers can use FMs to assist in drafting legal opinions.[38]
- Doctors can use FMs to assist in diagnoses.[39]
- Communications specialists can use FMs to assist with drafts of press releases.[40]

In all of these cases, and many others, enterprise data can be leveraged in a model that would be foundational to automating and modernizing these, and myriad other business processes at DHS.

5 Recommendations and Conclusion

While DHS is still nascent in its adoption of conventional machine learning systems (i.e., previous iterations of AI models), shifting our focus to the adoption of FMs, could be key in providing us with leap-ahead AI capabilities supporting critical mission priorities. As automation tools, FMs can benefit DHS by making previously time-intensive analyses more capable and efficient. DHS, through the AI Task Force, should embark on a research campaign plan to accelerate human-AI (FM) response to emerging and evolving threats.

Table 5.1 summarizes DHS opportunities and falls into three concentrations: AI Advancements, AI Assurance, and Mission Assurance. Within each concentration, the table indicates considerations of research thrust—essentially the "spin" to be applied in expressing requirements and writing terms of reference. These recommended thrusts or considerations focus on a combination of consolidative activities, work on critical enablers, "sense-making" research, common-good systematic research, and fresh design work.

In an era defined by the convergence of advanced technology and security imperatives, DHS stands at the crossroads of innovation and protection. FMs support new approaches to versatile AI frameworks from which DHS can equip itself to navigate the intricate labyrinth of data-driven threats and opportunities. These models offer the Department a force-multiplier effect, enhancing anomaly detection, risk assessment, and threat mitigation strategies. With each new piece of data, DHS will be paving the way for a safer and more resilient nation.

[36] See IBM's CodeFlare tool and GitHub's DS-1000.
[37] A. Narayan, et al., "Can Foundation Models Wrangle Your Data?" (2022).
[38] Borrowing from the Law to Filter Training Data for Foundation Models.
[39] M. Moore, et al., "Foundation Models for Generalist Medical Artificial Intelligence," (2023).
[40] G. Bullard, "Smart Ways Journalists Can Exploit Artificial Intelligence," (2023).

Table 5.1. Summary Considerations[41]

Concentrations		*Considerations*
AI Advancements Foundation Models	Develop large-scale, multi-modal AI models for homeland security missions that leverage self-supervised learning on unlabeled data.	*Model Scaling.* Develop a capability to pre-train large-scale parameter models.
		Multi-modality. Create novel neural architectures and pre-training objectives to operate with heterogenous data streams.
		Portability. Explore model pruning and compression techniques to reduce model size, dimensionality and increase sustainability.
		Data and Compute. Establish and maintain shared data and compute resources using the best practices for data processing.
AI Assurance Test and Evaluation	Demonstrate how FMs can be rapidly adapted to many tasks without being explicitly trained to do so; show how task performance can be improved via fine tuning, continuous learning, and prompting.	*Use Case Benchmarks.* Create benchmarks, test harness software, and metrics to perform reproducible evaluation and trade-off analysis across research and development models.
		Fine Tuning and Continual Learning. Develop methods to detect when new knowledge is available and decide when and how to update or fine-tune models to ensure model generalizability.
		Model Prompting. Evaluate emerging behavior of multi-modal FMs using prompting techniques.
		Trustworthy and Responsible AI Ecosystem. Release unclassified models, FAIR and AI-ready data sets and benchmarks so they can be easily found, accessed, deployed, and reused.
Mission Assurance Human-AI Teaming	Develop theoretical foundations and practical methods to enhance the teamwork effectiveness and partnership between subject matter experts and AI.	*Transparency.* Build trust with subject matter experts through developing novel techniques to explain and interpret multi-modal FM behavior. Examine causal explanations, counterfactual explanations, and interpretability.
		Usability. Enhance user interactions and develop metrics to effectively communicate model performance.
		Security and Privacy. Develop methods to evaluate model resilience and susceptibility to adversarial AI.
		Robustness. Ensure models are robust to distribution (domain, task) shifts in multi-modal setting.
		Interoperability. Develop methods to understand impacts of foundational model enhancements on operations.

[41] Table 1.1 builds on discussions with the Department of Energy, National Nuclear Security Administration and is tailored to DHS mission use cases.