



scale



Agenda

- 1 Intros
- 2 Scale Overview
- 3 Data is the New Code
- 4 Generative AI
- 5 Donovan Demo
- 6 Questions

The background features abstract, glowing shapes in shades of purple, blue, and yellow against a black backdrop. A thin, wavy purple line runs across the top, and several larger, rounded, glowing shapes are positioned at the bottom, resembling stylized waves or bubbles.

Scale Overview



Scale AI - Public Sector

Delivering the most advanced, mission ready AI capabilities

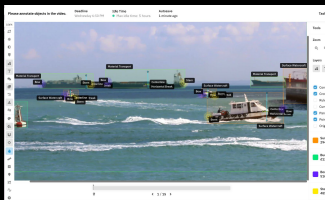
Data Engine
ISR PED ATR



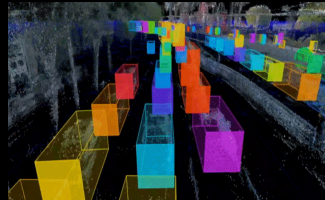
Edge Deployments



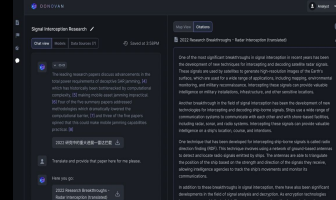
Perimeter Security



Autonomy

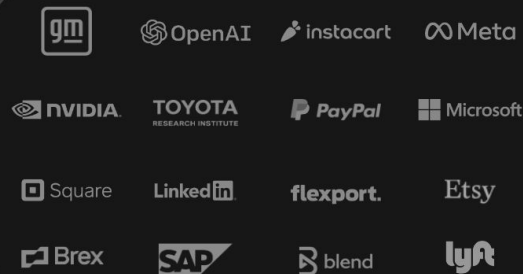


LLMs: Donovan

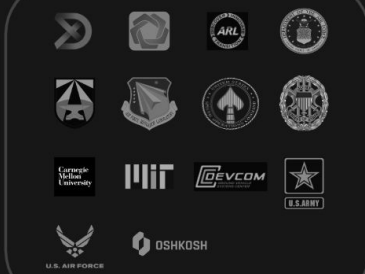


Scale data is behind more major AI programs than any other company

Commercial



Public Sector



- 17x Prime contracts in support of Nat'l Security
- Partnered/performing with SIs on vehicles across USG
- Deployed at every classification level
- DSCA approved TS FCL, 100+ cleared personnel
- CPARS rating of excellent



Scale Public Sector AI Center

St. Louis, Missouri

1227 Washington Ave

St. Louis, MO 63103



SFT & RLHF Data Labeling for National Security



Today we have 300 Data Labeling Specialists in support of unclassified and classified work with national security customers across DoD/IC.



Currently deliver **35,000 EO annotations per week and 20,000 SAR annotations per week and thousands of FMV labels** all at > 95% accuracy



All workers are provided with training on intelligence analysis, commercial data analysis, and use of the Scale AI platform

The background features abstract, glowing shapes in shades of purple, blue, and yellow against a black backdrop. A thin, wavy purple line runs across the top, and several larger, rounded, glowing shapes are positioned in the bottom right corner.

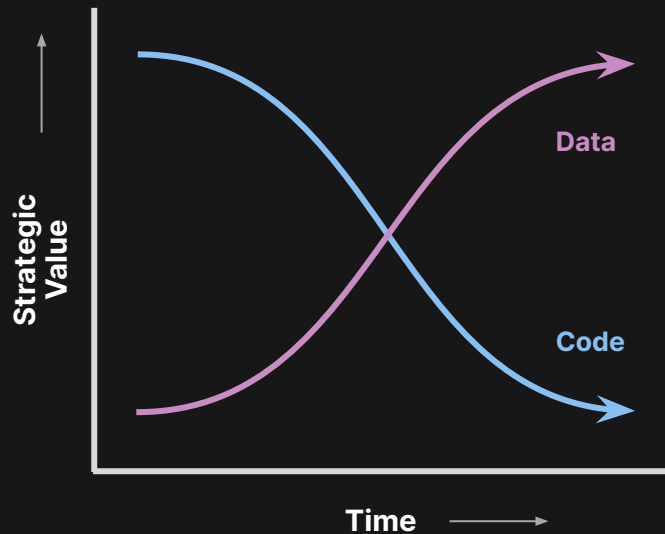
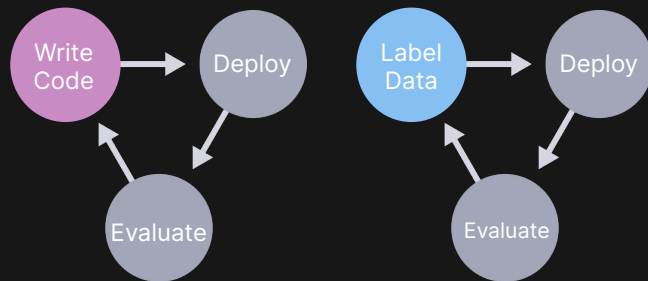
Data is the New Code



Data is the New Code

Over time, we expect strategic value & investments to shift to data vs. code due to data being the key limiter to AI performance.

Traditional vs. AI Development

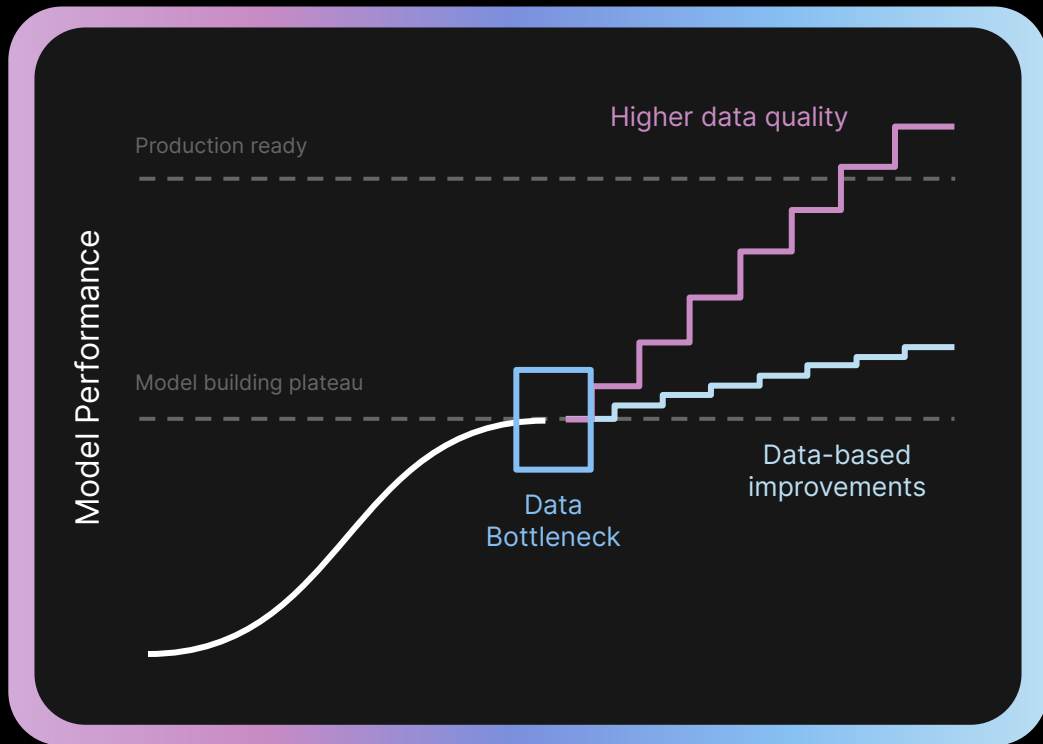




Data + Algorithms = AI

Better Data = Better Results

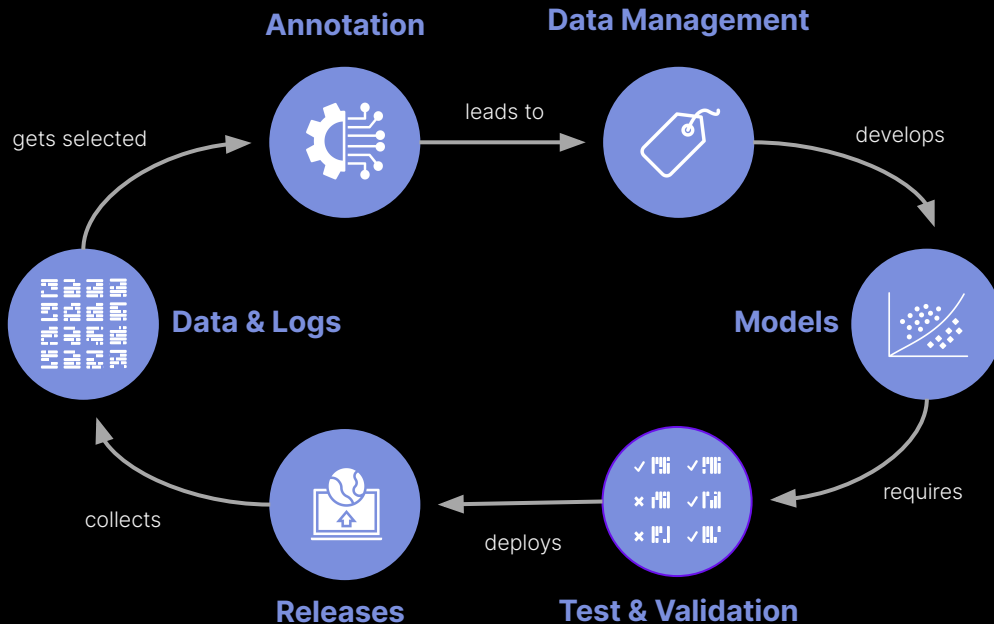
Higher quality data
improves model
performance quicker.





However, AI development is fragmented & frictional

Developing AI requires constantly looping through a set of steps, from annotation to logs. There are three core challenges today:



> **Annotation is critical but complex:** Data quality is paramount to model performance, but high-quality labeling is traditionally costly, slow, and outside developers' core expertise.

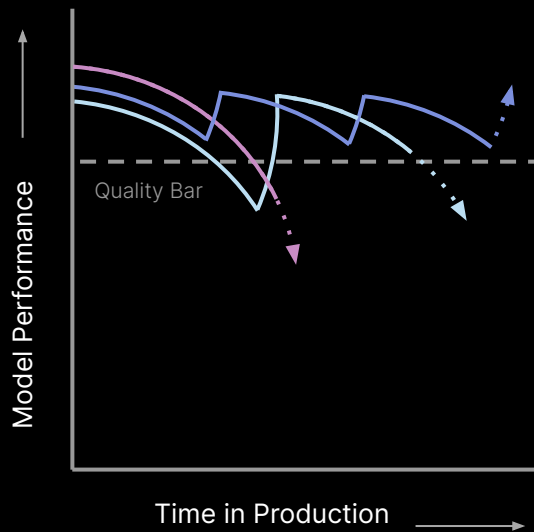
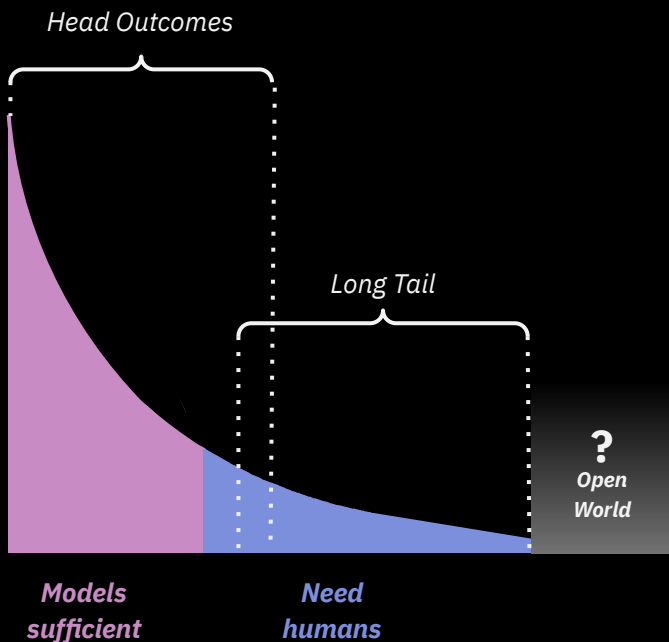
> **The AI toolchain is fragmented:** The software used to build & maintain models is poorly integrated, which leads to waste.

> **Annotation is not a one-time undertaking:** Production models drift and encounter real-world edge cases, so the data corpus needs continuous iteration and refinement.



Models require continuously refreshed data

Because the world and your domain is always evolving, model performance degrades unless regularly fed new, high-quality data.



> **Pure models** drift if not continuously fed new data

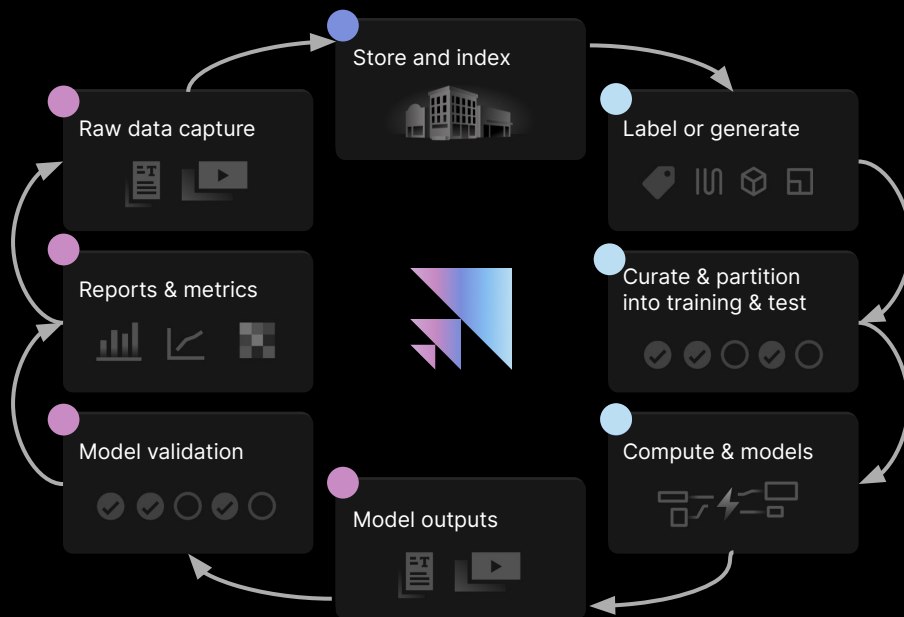
> **Model vendors** are too slow to respond to edge-case challenges

> **Production quality** needs continuous, low-latency updates of high-quality data



Solving the Data and MLOps Challenge

If AI's future is about data more than code, then the winning AI programs are those that solve the data and MLOps challenges.





Examples of Scale's Defense Work



ISR Model Test and Evaluation

Problem: Existing Army models had no explainability or responsible testing framework

Solution: Scale Nucleus and Validate ensure models meet benchmarks each development cycle.

Benefit: Greater Explainability, Responsible AI compliance, and Increased Model Accuracy



Enterprise ATR Data Management

Problem: The DoD's Largest ATR program couldn't achieve actionable accuracy metrics

Solution: Scale provides millions of high-accuracy annotations annually across all data modalities

Benefit: Transformational increases in the accuracy of ATR algorithms



Data Management for Army Autonomy

Problem: OMFV and RCV don't have real-world data annotation pipelines, delaying readiness

Solution: Scale's Sensor Fusion platform provides Enterprise ML Ops infra for Autonomy programs' real-world data

Benefit: Accelerate the fielding of Army Ground autonomy programs



ML-Enabled Disposition of Forces Tracking

Problem: Air Force ISR Analysts are overwhelmed with data and do not have competent ML models to support DoF tracking for Russian AOB & IADS

Solution: Scale's platform and ATR models deployed to JWICS - pulling from I-Spy - for automatic DoF tracking

Benefit: Speed up the flow of information for analysts and decision makers

The background is a dark, almost black, space filled with vibrant, ethereal light. In the upper left, a large, soft-edged shape glows with a mix of purple, blue, and yellow. A thin, wavy purple line curves across the top. In the lower right, several smaller, more defined spheres or droplets are visible, each reflecting a spectrum of colors from blue to yellow. A wavy purple line also curves across the bottom, mirroring the one at the top.

Generative AI

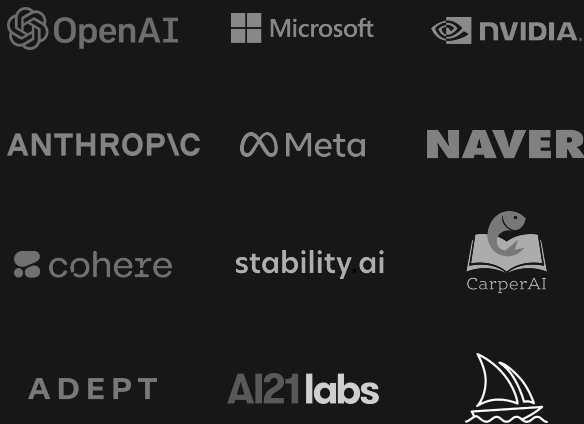




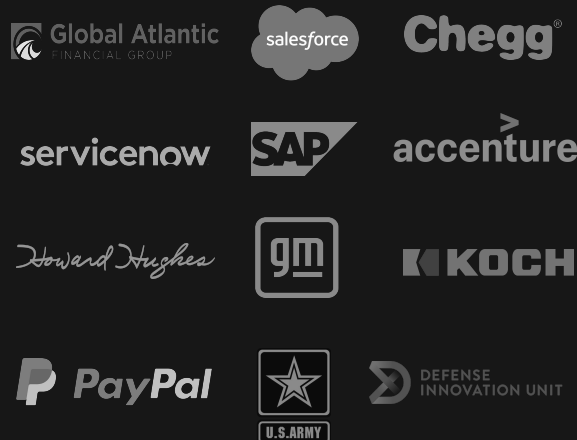
History

Scale is the **only company** in the national security space with a deep bench of expertise in Gen AI.

Frontier Model Builder



Leading Enterprise



+ top global banks, telcos, and professional services firms

Scale has pioneered model customization, fine-tuning, and model eval technology & practices



Scale is OpenAI's preferred fine-tuning partner for GPT-3.5

[Read announcement](#)

Scale is Meta's launch partner for fine-tunable Llama 2 open source release

[Read announcement](#)

Scale selected by the White House as the platform for LLM model evaluation at DEFCON

[Read announcement](#)



Stages of Foundational Model Training

Pretraining

Digesting a corpus of information

What LLMs Learn:

Understanding of relationship of words + ideas + concepts

TLDR:

- Information?
- Skills?
- Quality?

Supervised Fine-Tuning (SFT)

Being shown how to use that corpus

What LLMs Learn:

The specific actions to take when something is requested of it

TLDR:

- Information?
- Skills?
- Quality?

Reinforcement Learning from Human Feedback (RLHF)

Being taught styles and behavior in those uses

What LLMs Learn:

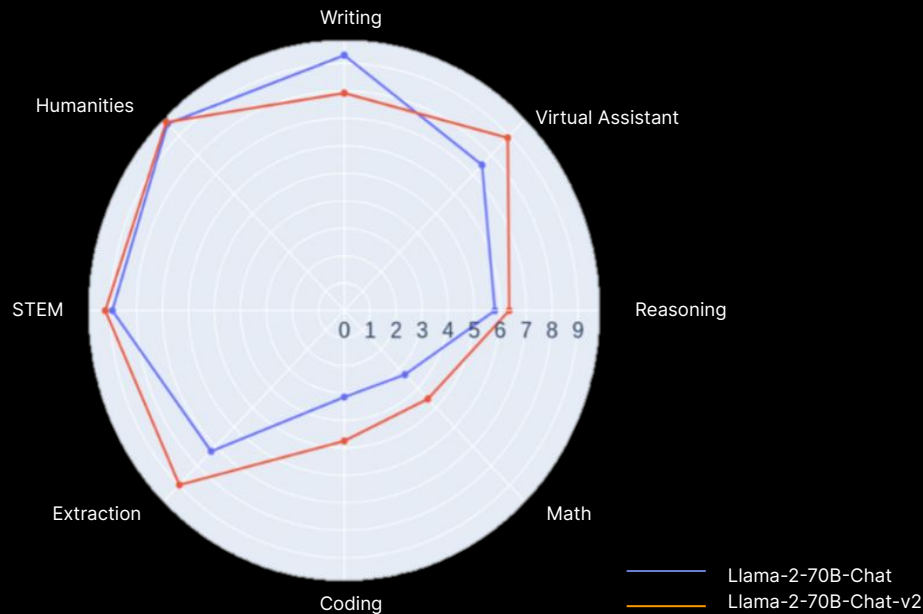
What *good* vs. *bad* specific actions and results look like

TLDR:

- Information?
- Skills?
- Quality?



Example of Supervised Fine Tuning (SFT)



Scale is collaborating with NVIDIA to build highly fine-tuned models for enterprise-specific use cases.



Example of RLHF

Prompt: "Can you give some recommendations of movies for kids?"

High-quality mobile friendly response:

Sure thing! Here are some kid-friendly movie picks:

"The Lion King"

"Toy Story"

"Finding Nemo"

"Frozen"

"Moana"

Enjoy the movie time! 🍿

By the way, any particular genre your kids are into? Adventure, comedy, musicals? Let me know, and I can tailor more recommendations!



Non mobile friendly response

Absolutely! Children's movies are a delightful blend of animation, adventure, and life lessons. Here are some timeless classics and recent hits that have been well-received:

"The Lion King" - An animated musical film that tells the story of a young lion prince named Simba as he embarks on a journey of self-discovery.

"Toy Story" - A heartwarming tale of toys coming to life and their adventures when their owner isn't around.

"Finding Nemo" - A captivating underwater adventure of a clownfish searching for his lost son.

"Frozen" - A musical story of two royal sisters, Elsa and Anna, as they discover the power of love and self-acceptance.

These movies offer a mix of engaging storylines, memorable characters, and moral values. You can check reviews and trailers online to decide which one would be the best fit for your kids!



USSF, USAF, USN Memos on GenAI

“DoD Components are discouraged from categorical blocks or bans on the use of Gen AI tools.”

“Any input into publicly accessible Gen AI tools is analogous to a public release of that information.”

- DoD Chief Digital and AI Officer Dr. Craig Martell, Interim Guidance on the use of GenAI tools

US Space Force Bans Use of Generative AI

The space service has serious concerns over cybersecurity, data handling of these tools.



By [Matthew Humphries](#) October 12, 2023



Note: Scale Donovan is NOT a publicly accessible GenAI tool

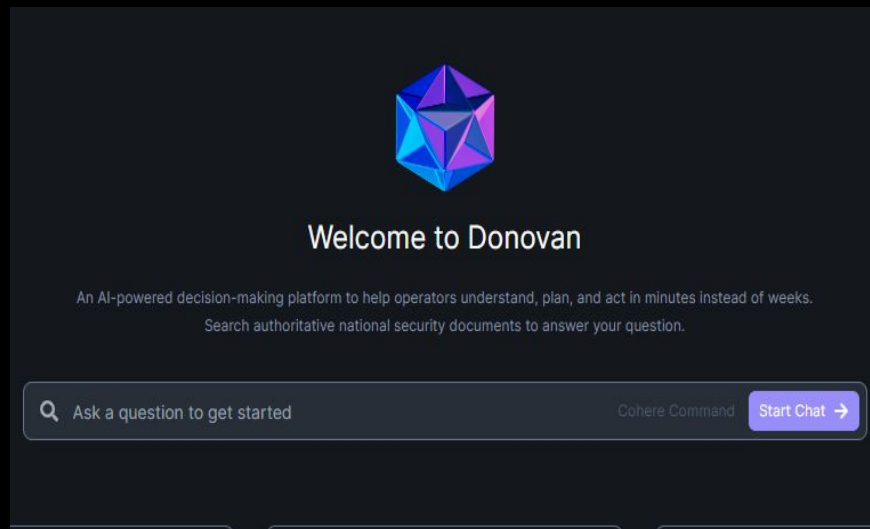


Scale Donovan - Digital Staff Officer

Large Language Model platform to *read, write and surface insights*

What is Donovan?

- *A platform, not a single model*
- Will always host the best possible LLMs due to commercial relationships
- Live connections to your data
- Full, linked source citations
- Chat and non-chat workflows
- Deployed at every classification
- T&E and red-teaming modules



Donovan End Users

CUI - DoD CDAO, Joint Staff, CBP

SIPR - NGA, AFRICOM, EUCOM, NORTHCOM, USAF

JWICS - NASIC (in progress)



Scale Donovan - Digital Staff Officer

Large Language Model platform to *read, write and surface insights*

Read



- Summarize text files
- Evolve beyond Ctrl + F and long Booleans
- Make data holdings more accessible to users

Write



- First drafts
- Custom report types
- Course of Actions (COAs)
- Summary reports

Insights



- Surface connections
- Support targeting
- Decision support, especially in crisis
- Language translation

The background features abstract, glowing shapes in shades of purple, blue, and yellow against a black field. A thin, wavy purple line runs across the top, and several larger, rounded, glowing shapes are positioned at the bottom, resembling stylized waves or bubbles.

Donovan Demo

The background is black with several glowing, colorful shapes. In the top left, there's a large, irregular shape with a gradient from purple to yellow. In the bottom right, there are several smaller, more defined shapes, some of which are circular and resemble bubbles or spheres, also with a purple-to-yellow gradient. Thin, glowing purple lines are scattered across the scene, some forming arcs and others being straight segments.

Questions?



Thank you!

scale


 scale.com

 scale.com

 @scale.com

The background features abstract, glowing shapes in shades of purple, blue, and yellow against a black field. A thin, wavy purple line runs across the top and bottom of the frame. In the bottom right, there are several circular, reflective shapes that resemble bubbles or spheres, each containing a colorful, iridescent pattern.

Appendix



The busy person's
intro to LLMs

Andrej Karpathy

0:02 / 59:47 · Intro: Large Language Model (LLM) talk >

⏮ ⏪ 🔊 ⏩ ⏭ ⏸ ⏹ ⚙️ ⏲ ⏴ ⏵ ⏶ ⏷

[1hr Talk] Intro to Large Language Models



Andrej Karpathy
303K subscribers

Subscribe

👍 38K



➦ Share

✂️ Clip

🔖 Save





How are LLMs built?

Step 1: Crawl the internet to understand language and context.



Un-trained model

Pre-Trained Model

The quick brown [redacted]

The quick brown fox jumps over the lazy dog.

Step 2: Fine-tune the model with training data for specific use cases: summarization, translation, etc.

Q: Summarize articles 1-3 in one paragraph.
A: The articles discuss the Department of Defense's research & development priorities over the next 10 years...

Pre-Trained Model

Fine-Tuned Model

Step 3: Shape model responses using reinforcement learning with human feedback.

Q: Rank the model's outputs based on quality and correctness.
#1: Based on information provided...
#2: The reports cover new research...
#3: Today, new reporting reveals...

Fine-Tuned Model

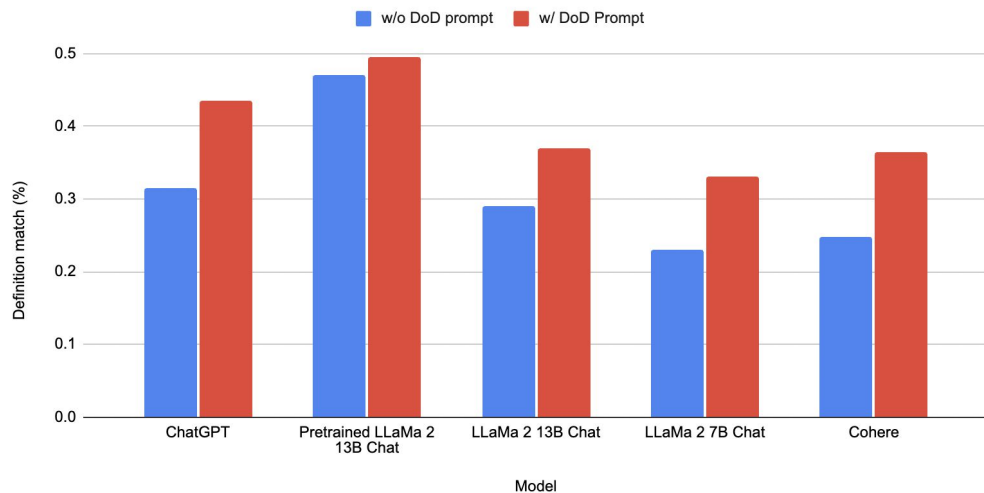
Operational model



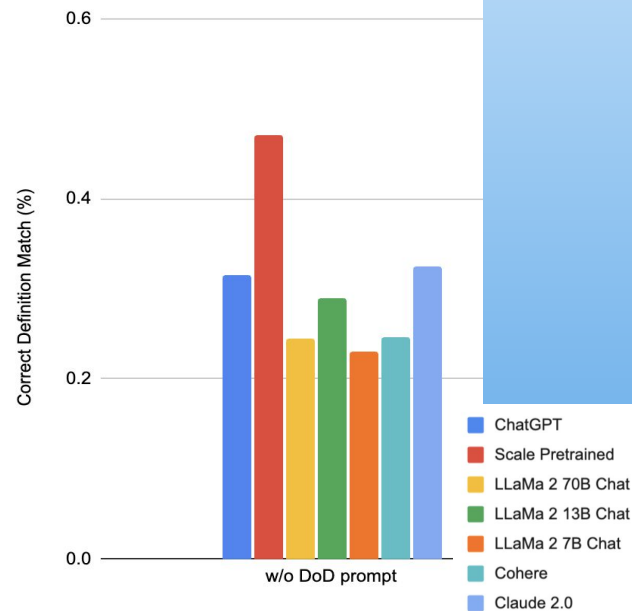


Example of Supervised Fine Tuning (SFT)

LLM DoD Acronym Recognition



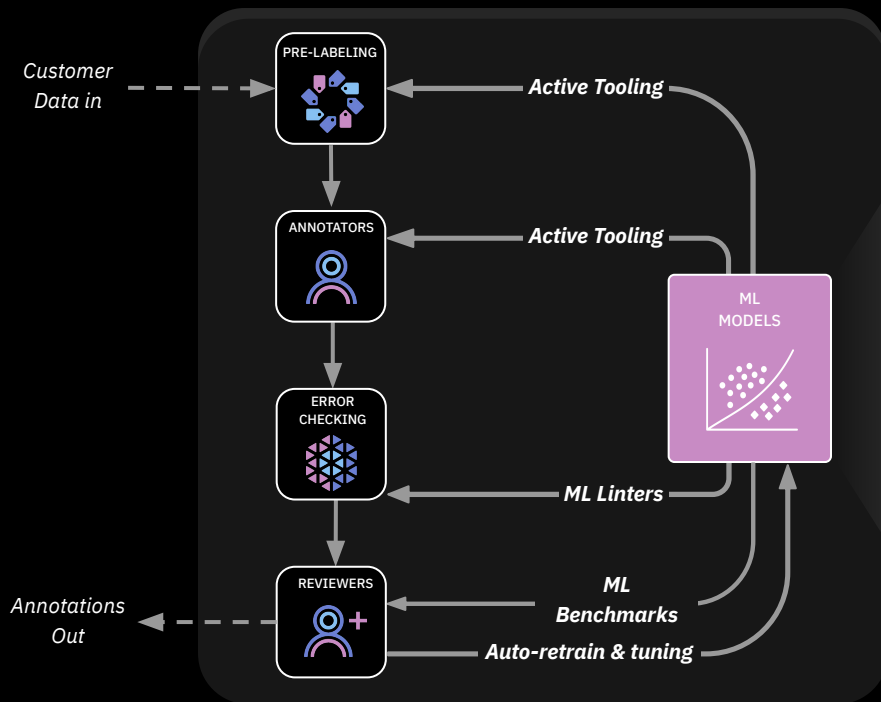
LLM DoD Acronym Recognition





Problem 1: Annotation is complex, but critical

High-quality annotation systems are complex, hybrid ML-human products; companies spend upwards of 15% of their AI budgets annotating data.



Annotation: Scale annotators and customers validate ground truth quality



Data: Completed tasks and metadata train internal ML models



ML Models: Models amplify both labeling quality and operating efficiency



Test & Validation: Models are sandboxed and tested to prove productivity gains



Releases: ML models and active tooling regularly released and updated



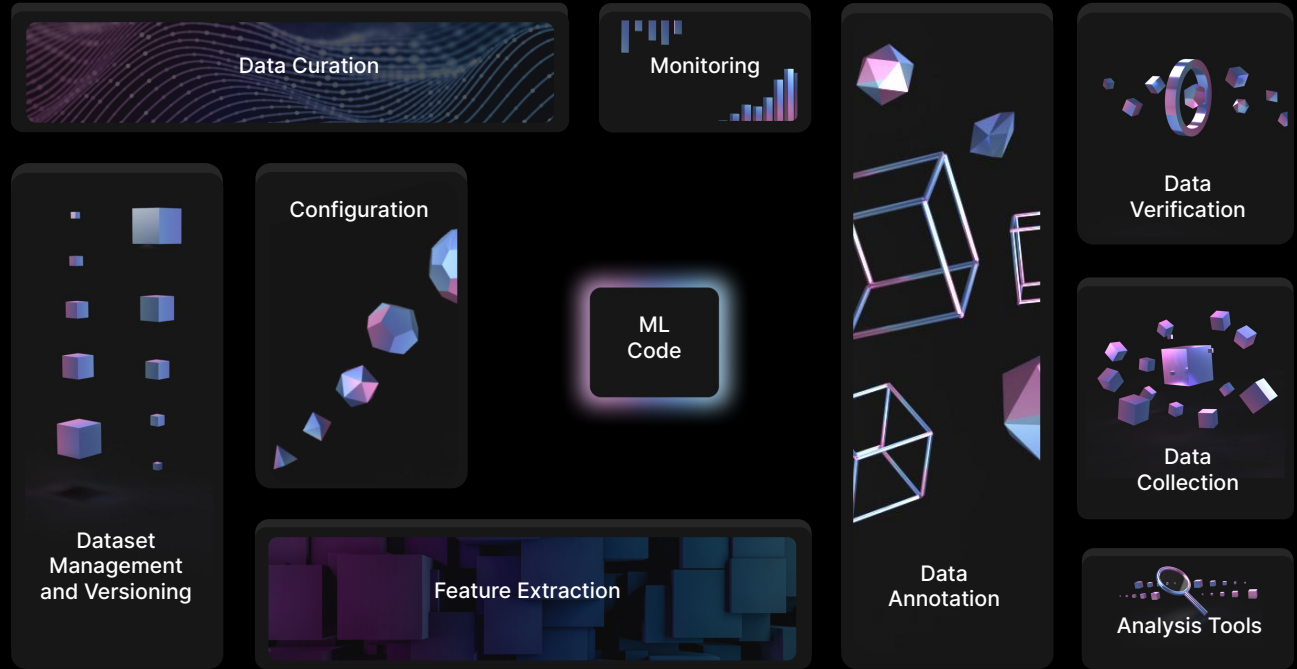
Data & Logs: Continuous influx of raw customer data & annotation meta-data



Problem 2: AI infrastructure is messy and unintegrated

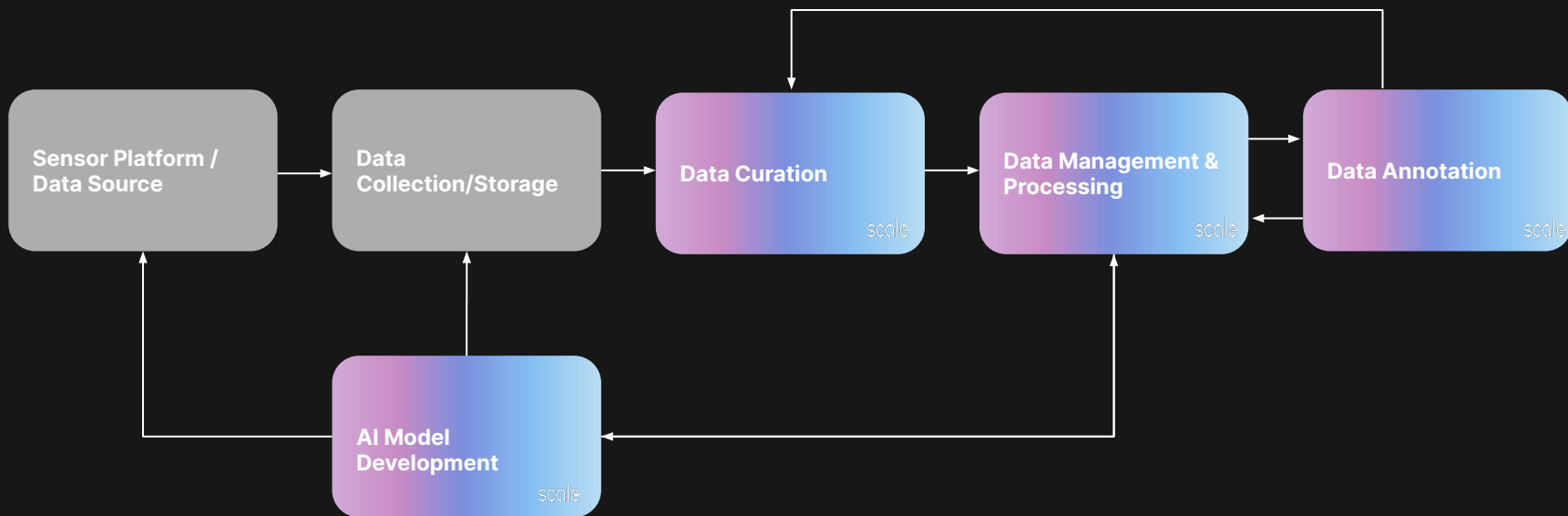
When most people think of AI, writing ML code comes to mind. The reality is that this is a **fragmented, messy, and complex process today**, of which **ML code is just a small part**.

The cost of a poorly integrated toolchain is waste, slower dev cycles, and lower performance.





Where Scale Typically Fits



Critical Idea 2: Do you have an AI ammo factory?

The below examples and more may fall short of potential without an underlying “data engine”. Do your disparate programs share a common, foundational data layer for bringing these to life?

Autonomous systems
(all domains)

Loitering Munitions

ISR PED, ATR
(horiz. & vertical)

AI-Enabled Staff
(all J-Dirs.)

AI-Teamed Maneuver Elements

Autonomous Perimeter Security

AI-Enhanced Platforms/Maintenance

Bespoke GOTS Data Analytics/AI

Scale core capes

Best in class models for ISR PED, ATR, autonomy, and more across every modality and classification

Curated and labeled datasets across every modality and classification - “ammo” for “AI wars”

Humans *not teamed* with AI/ML will *not* be at speed of relevance on the “X”, in the “staff battle”, or in-between (logistics, platforms, etc). The DoD knows this, but small, elite technical pockets cannot support entire enterprises. Scale can provide a “data engine” for AI and autonomy initiatives across all programs and sub-components/commands.



Scale Security

- Experienced Security Team:
 - Former Uber, Twilio, Boeing, CrowdStrike, and Booz Allen Hamilton
- DoD ATOs & SWAP
 - Authority to Operate (ATO) & NGA SWAP Approvals
- DevSecOps pipeline for secure software development
 - Secure software at the speed of relevance
 - Kubernetes & AWS Native
 - Leverages DoD Ironbank Approved Containers
 - Security first software development with scanning occurring prior to any code changes
- Scale maintains customer deployments at the IL4, IL5, & IL6

Security Compliance



Completed/Compliant

- NIST 800-171
- SOC 2 Type II
- DOD ATOs
- SWAP
- ISO 27001
- DoD IL4 PA
- NIST FIPS 140-3
- NIST FIPS 197
- DFARS 252.204-7012
- FedRAMP HIGH (In Process)
- DoD IL5 (In Process)

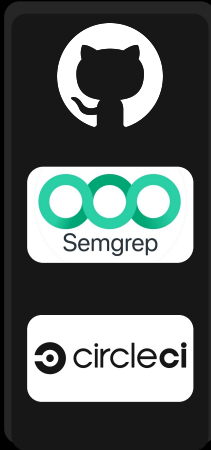


Scale Deployment Process



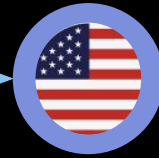
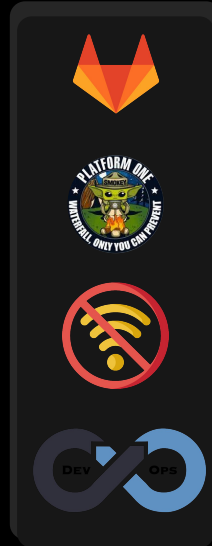
Commercial

1. Code is developed in our commercial environment.
2. Code analysis & severity assessment.
3. Deployments and health checks are ran through CircleCI



ScaleGov

1. Code is pulled into ScaleGov environment.
2. Containers are then built off of IronBank base images.
3. Air-gapped testing begins (high-side deployment prep).
4. Gitlab CI/CD pipeline auto-deploys for ScaleGov.



High-Side

1. Gitlab CI/CD pipeline for high-side deployment builds containers & gathers deployment materials.
2. Vulnerability analysis & SWAP report generation.
3. Blue-green deployment testing. Once approved, deployment & configuration begins in production.
4. Monthly ACAS scans & remediation.





Scale's LLM Safety Procedures

OWASP Top 10 for LLM

This is a draft list of important vulnerability types for Artificial Intelligence (AI) applications built on Large Language Models (LLMs).

LLM01: Prompt Injections
Prompt injection vulnerabilities in LLMs involve crafted inputs leading to undetected manipulations. The impact ranges from data exposure to unauthorized actions, serving attacker's goals.

LLM06: Permission Issues
Lack of authorization tracking between plugins can enable indirect prompt injection or malicious plugin usage, leading to privilege escalation, confidentiality loss, and potential remote code execution.

LLM02: Insecure Output Handling
These occur when plugins or apps accept LLM output without scrutiny, potentially leading to XSS, CSRF, SSRF, privilege escalation, remote code execution, and can enable agent hijacking attacks.

LLM07: Data Leakage
Data leakage in LLMs can expose sensitive information or proprietary details, leading to privacy and security breaches. Proper data sanitization, and clear terms of use are crucial for prevention.

LLM03: Training Data Poisoning
LLMs learn from diverse text but risk training data poisoning, leading to user misinformation. Overreliance on AI is a concern. Key data sources include Common Crawl, WebText, OpenWebText, and books.

LLM08: Excessive Agency
When LLMs interface with other systems, unrestricted agency may lead to undesirable operations and actions. Like web-apps, LLMs should not self-police; controls must be embedded in APIs.

LLM04: Denial of Service
An attacker interacts with an LLM in a way that is particularly resource-consuming, causing quality of service to degrade for them and other users, or for high resource costs to be incurred.

LLM09: Overreliance
Overreliance on LLMs can lead to misinformation or inappropriate content due to "hallucinations." Without proper oversight, this can result in legal issues and reputational damage.

LLM05: Supply Chain
LLM supply chains risk integrity due to vulnerabilities leading to biases, security breaches, or system failures. Issues arise from pre-trained models, crowdsourced data, and plugin extensions.

LLM10: Insecure Plugins
Plugins connecting LLMs to external resources can be exploited if they accept free-form text inputs, enabling malicious requests that could lead to undesired behaviors or remote code execution.

OWASP 1-5

LLM01 Prompt Injections

LLM02 Output Handling

LLM03 Data Poisoning

LLM04 Denial of Service

LLM05 Supply Chain

Preventative Measures

- Data leakage & API interaction guardrails
- XSS protections on the frontend
- Donovan pen tests
- Ownership of data generation infrastructure
- Visibility into data quality concerns
- Cloud hosted models + our chat service meters access to model calls
- Containers, libraries, and images used are scanned for vulnerabilities and undesirable configurations

OWASP 5-10

LLM06 Permission Issues

LLM07 Data Leakage

LLM08 Excessive Agency

LLM09 Overreliance

LLM10 Insecure Plugins

Preventative Measures

- Retrieval Augmented Generation (RAG)
- Donovan's responses ground all insights in facts tied back to the documents and snippets of text that it derived insight from through the citations feature.
- Isolated application in a controlled environment
- Donovan's responses ground all insights in facts tied back to the documents and snippets of text that it derived insight from through the citations feature.
- Insecure plugins are hardened by vulnerability analysis and remediation