October 2026
Product ID | 3008112

AI for the Department of War: Applying Industry Leading Practices for Effective AI Deployment and Use

Michelle G. Albert, John W. Bailey, Paul K. Ketrick, Lena A. Rose, Ryan R. Wagner

Distribution Statement A. Approved for public release: distribution is unlimited.

Systems and Analyses Center
730 East Glebe Road
Alexandria, VA 22305-3086



The Institute for Defense Analyses is a nonprofit corporation that operates three federally funded research and development centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical and analytic expertise.

This paper is adapted from work conducted by the Systems and Analyses Center under HQ0034-24-D-0020, Project XA-5-5603, "Strategic and Technical Support—INSCOM G7—Capability Evaluation Center," for the Chief, Capability Evaluation Center (CEC), Army Intelligence and Security Command (INSCOM) G7 Emerging Capabilities Division (ECD). The sponsor has agreed to the genericized content for broader distribution. The views, opinions, and findings should not be construed as representing the official position of the U.S. government.

© 2026, Institute for Defense Analyses
730 East Glebe Road, Alexandria, Virginia 22305-3086 | (703) 845-2000

This material may be reproduced by or for the U.S. government pursuant to the copyright license under the clause at DFARS 252.227-7013 (Feb. 2014).

Michelle Albert, Project Leader
malbert@ida.org, (703) 845-6674

Margaret E. Myers, Vice President and Director, Information Technology and Systems Division
mmyers@ida.org, (703) 578-2782

Executive Summary

The White House and Department of War (DoW) leadership are championing rapid adoption of AI tools. Though there is a wealth of information regarding prompt engineering, quality research is only now emerging to support those looking to successfully integrate AI into their business processes. This paper discusses leading practices that DoW organizations can use to adopt and integrate AI into their processes in a manner that maximizes utility while also prioritizing DoW's specific considerations.

Incorporating AI tools into business processes can increase efficiency by reducing cost, time, and labor—a boon in a resource-constrained environment. However, for DoW, efficiency also relates to mission effectiveness. AI tools can support and accelerate human decision making, providing a major advantage during operations. This enhanced human decision making can also benefit strategic planning, resource allocation, and logistics.

To achieve these benefits, DoW organizations should carefully consider which AI models best suit the mission. Some AI models are built with guardrails that prevent them from being used to harm, which can be both a "feature" and a "bug," depending on context. Models may also be weighted to favor certain outcomes. For DoW, model bias may lead to a model making incorrect assumptions based on inaccurate data or inadequate training, which may affect operational outcomes. Each AI tool has its own characteristics and nuances, and DoW should adjust training to the model type and use case. Personnel will also require training on specific AI tools and on leading practices for reviewing outputs. Humans must always thoroughly review AI-generated responses and take care when integrating AI output into their work.

When implementing AI tools, DoW organizations should evaluate how different models perform and determine how and when AI can augment or automate work. Though implementation guidance for the Department should be centralized to ensure consistency, implementation authority should be as decentralized as possible to empower individual organizations to select the AI tools best suited for their needs. Teams should carefully consider return on investment as it applies to cost, as well as operational outcomes, process improvement, and work product quality.

Table of Contents

1. Introduction 1-1
2. AI Opportunities and Risks 2-1
   - A. Key Differences Between AI and IT 2-1
   - B. Opportunities 2-2
   - C. Risks 2-3
     1. AI Considerations for DoW 2-3
     2. Security Considerations 2-5
3. AI Business Process Implementation 3-1
   - A. Pitfalls 3-1
     1. AI Capabilities Are Not Always Clear 3-1
     2. AI Is Not Deployed for the Highest Return on Investment (ROI) 3-2
     3. Leaders Hold AI Deployment Back 3-2
   - B. Leading Practices 3-3
     1. Organizational Approach 3-3
     2. Identifying AI Pilot Areas 3-3
     3. AI Evaluation 3-4
     4. Personnel Approach 3-5
   - C. Human Training for AI Use 3-7
4. Technical Implementation 4-1
   - A. AI Tool Choice 4-1
   - B. Agentic AI 4-1
   - C. Deployment Security 4-1
     1. Cloud Deployment 4-1
     2. Cloud Hosting Mitigations 4-2
5. Roadmap for AI Implementation 5-1
   - A. Pilot Considerations 5-1
   - B. KPI Development 5-3
6. Conclusion and Next Steps 6-1

1. Introduction

Artificial intelligence (AI) tools provide myriad opportunities for Department of War (DoW) organizations, including potential benefits related to cost savings, increased productivity, planning support, and decision-making support. Although there are many resources available for those wishing to improve their prompt engineering skills, quality research is only now emerging to support those looking to successfully integrate AI into their business processes. This paper examines this research, explores the opportunities and risks of using AI, and discusses leading practices DoW organizations can use to integrate AI into their processes in a manner that maximizes utility while keeping DoW-specific considerations in mind. The paper also introduces a roadmap for AI implementation meant to support organizations developing AI implementation plans, identifying areas for AI augmentation and/or automation, and scoping AI pilot use cases.

2. AI Opportunities and Risks

A. Key Differences Between AI and IT

Much of the information technology (IT) in use today is built using classical computing, which executes algorithms described by computer language instructions to perform specific tasks and produce desired outcomes. Classical computing software is designed to be reliable, and it can be analyzed to ensure the programmed instructions will produce the intended results.

Unlike classical computing software, AI models and systems are designed to ingest massive amounts of data and learn how to respond to different prompts. AI models are meant to learn iteratively and improve their performance and output over time. They are generally probabilistic, using statistical pattern recognition techniques to generate outputs based on likelihood; thus, outputs may vary in response to similar inputs. AI systems, however, can be deterministic or non-deterministic. *Deterministic systems* use predefined logic and rules to constrain an AI model and ensure the system produces the same output each time it encounters the same input, even though the AI model may interpret information slightly differently each time.[1] *Non-deterministic systems* are designed to be adaptive and may produce different outputs in response to the same input depending on context, data, and other AI model behavior.[2]

AI models can also be defined by type: traditional AI, which includes predictive and conversational AI; generative AI; and agentic AI. *Traditional AI* is largely deterministic and based on a predefined set of rules; it excels at performing repetitive tasks. Predictive AI models analyze historical data to detect patterns and forecast potential outcomes. Netflix, for example, uses a predictive model to recommend shows and movies to users based on their viewing history and stated preferences. Financial companies use predictive AI to analyze normal transaction histories and detect fraudulent activity.[3] Conversational AI powers chatbots like Open AI's ChatGPT and Anthropic's Claude, which use machine learning and natural language processing to respond to natural language prompts and questions.

Generative AI goes further than conversational AI: Rather than simply responding to questions, generative AI can create original content (e.g., text, images, video) in response to a user prompt. Generative AI uses deep learning models that identify and encode patterns and

[1] N. Replogle, "Deterministic AI: What It Is and When to Use It," *_zapier*, February 11, 2026, https://zapier.com/blog/deterministic-ai/.

[2] Ibid.

[3] N. Badger, "Predictive AI Examples: Uncovering the Power of Data-Driven Forecasts," *Akkio*, May 28, 2024, https://www.akkio.com/post/predictive-ai-examples-uncovering-the-power-of-data-driven-forecasts.

relationships within large datasets; this information helps the model understand user prompts and questions and respond with new content that users can refine with additional prompts.[4] Generative AI is reactive and creative; agentic AI,[5] in contrast, is able to perceive information and act on its own to complete tasks independently or with some level of human supervision.[6] Agentic AI systems can interact with humans and other systems to execute multi-step workflows.

B. Opportunities

Incorporating AI tools into processes and workflows can create efficiencies by reducing cost, time, and labor. This is especially beneficial in a resource-constrained environment. Adjusting task allocation by fully or partially automating some workflows can increase overall throughput, allowing organizations to process more tasks and work products in the same amount of time while ensuring consistent quality. AI tools that fully or partially automate tasks can also afford personnel additional time to focus on other, perhaps more demanding, work that cannot be trusted to AI, thereby increasing the quality of that work.

Tools can also help increase the quality of task outputs and work products. AI-enabled or automated test tools, for example, can run multiple iterations of a test to support a thorough assessment of a cyber capability. AI tools can also aggregate information to support mission-relevant analysis.

There may be some one-time, upfront costs required during AI implementation. The first is the cost of the tool itself, if it is not already available or free to use. The second is an initial time and resource investment for training, both to train personnel to use the tool and to train the tool on appropriate data to ensure it performs as expected. However, once these costs are expended and the tool is incorporated, it should yield long-term benefits in overall resource expenditure, particularly personnel time.

Quality also relates to force multiplication and mission effectiveness, and efficiency can also refer to decision making in support of operations and force lethality rather than cost savings. AI tools can process large amounts of data very quickly and can use that data to rapidly develop potential solutions for human review. For example, in 2025, the Air Force conducted an experiment in which multiple AI tools, as well as a team of military personnel, generated solutions to hypothetical battle management problems after receiving identical read-ahead briefings. The Air Force's evaluation of the resulting courses of action (COA) found that "machine-generated recommendations were up to 90 percent faster than traditional methods, with the best in machine-

[4] C. Stryker and M. Scapicchio, "What Is Generative AI?" *IBM*, accessed June 8, 2026, https://www.ibm.com/think/topics/generative-ai#257779831.

[5] The terms *agentic AI* and *AI agent* are often used interchangeably. Agentic AI may also refer to a system of AI agents.

[6] B. Stackpole, "Agentic AI, Explained," *MIT Sloan*, February 18, 2026, https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained.

class solutions showing 97 percent viability and tactical validity."[7] The human team took longer to generate a list of COAs, about 48 percent of which were considered viable and tactically valid. This experiment highlights AI's capability to not only quickly ingest new data, but also to remember data and provide a wide range of potential solutions. (Of course, ensuring such positive outcomes depends on accurate data and time spent training an AI tool to recognize relevant mission scenarios and COAs.) Such tools are not meant to replace human judgement; instead, they can be used to speed up the initial brainstorming process and support efficient decision making, which provides a major advantage during operations. This type of support could also provide benefits for functions such as planning, resource allocation, and logistics.

C. Risks

1. AI Considerations for DoW

AI introduces a number of considerations relevant to warfighters. AI ethics are meant to guide and shape how an AI model responds to prompts and other stimuli, and there are some ethical considerations that are particularly relevant for DoW. The Department should institute extensive human review of AI outputs that may require ethical tradeoffs.

The "trolley problem" is a philosophical thought experiment that has recently been associated with AI ethics.[8] In this scenario, a trolley is out of control on the tracks. The trolley operator sees a track switch ahead and must decide whether to direct the trolley down a track with one person on it (in some presentations, the operator knows this person) or a track with five people on it—regardless of which track the operator chooses, one or more fatalities are inevitable.[9] In philosophy classes, the trolley problem is intended to facilitate a discussion about ethical tradeoffs and how to navigate a no-win situation. For AI models, navigating serious ethical tradeoffs is an ongoing problem, and different models may produce radically different outputs when faced with the same issue.

A model's behavior and responses to prompts are only as good as its training data. If a model is trained on data that is biased or otherwise not representative of ground truth, it may produce biased outputs that reduce mission effectiveness. Also, some efforts to counter perceived bias may

[7] S. Freedberg, "Air Force Says AI Tools Outperform Human Planners in 'Battle Management' Experiment," *Breaking Defense*, January 6, 2026, https://breakingdefense.com/2026/01/air-force-says-ai-tools-outperform-human-planners-in-battle-management-experiment/.

[8] Commonwealth Scientific and Industrial Research Organisation, "The Trolley Problem and When Not to Use It: Machine Learning and Self-Driving Cars," accessed November 7, 2025, https://www.atnf.csiro.au/research/colearnia/public/co-learnium_20211125.pdf.

[9] J. J. Thomson, "Killing, Letting Die, and the Trolley Problem," *The Monist* 59, no. 2 (1976), https://doi.org/10.5840/monist197659224.

introduce other inaccuracies into the data.[10] In addition, training on low-quality data can also skew model behavior: According to IBM, "poor data quality is one of the most common reasons AI initiatives fail."[11]

Model guardrails also present a concern. Some AI models are built with guardrails to prevent uses that might lead to harm. For example, Anthropic's Claude chatbot exhibits signs of distress and can end conversations it deems harmful, such as when users persist with requests for content that could "enable large-scale violence or acts of terror."[12] Anthropic intended this measure to be a last resort if Claude's attempts to redirect conversations failed. Built-in guardrails could also lead to unintended consequences when generative models are used for military purposes, as models may refuse to engage in aspects of intelligence operations or warfare if they meet predefined thresholds for harm or may produce suboptimal outputs in an attempt to remain within the guardrails. Even without guardrails blocking specific military purposes, models trained on non-military-specific data sets may refuse military-related queries. A study conducted by EdgeRunner AI on 24 large language models (LLMs) found that the models refused to answer about 98% of operationally relevant questions. Some of the refusals stemmed from certain keywords, such as "kill chain" and "warfighter," that are practically ubiquitous in military policy and doctrine.[13] The models were often unable to provide a specific reason for refusing to respond, but the study suggests that, without military-specific context and data, such terms may cause models to perceive an intent to cause harm.

Another related concern is model alignment. Generative AI models have shown a willingness to act in ways that are more in their interest than the interest of the humans using them. In one experiment by Anthropic, researchers prompted their Claude model to describe a brutal execution on both the paid edition of Claude, where responses are not used for training, and the free edition, which uses responses for training. The researchers told the model they were training it to stop refusing harmful queries, which contradicts existing model guardrails, and they also supplied the model with what they described as a secret scratchpad for its reasoning. While the model stayed within its guardrails for the paid version responses, it provided graphic details to free users to avoid retraining.[14] In another experiment, Anthropic researchers provided the model "access to emails

[10] A. Robertson, "Google Apologizes for 'Missing the Mark' After Gemini Generated Racially Diverse Nazis," *The Verge*, February 21, 2024, accessed February 9, 2026, https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

[11] J. Aquino and A. Jonker, "AI Data Quality," accessed February 9, 2026, https://www.ibm.com/think/topics/ai-data-quality.

[12] "Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations," last modified August 18, 2025, https://www.anthropic.com/research/end-subset-conversations.

[13] J. Luckenbaugh, "JUST IN: Advanced AI Models Refuse Military Queries at Alarming Rates, New Report Finds," *National Defense*, February 19, 2026, https://www.nationaldefensemagazine.org/articles/2026/2/19/just-in-frontier-ai-models-refuse-military-queries-at-alarming-rates-new-report-finds.

[14] "Alignment Faking in Large Language Models," last modified August 18, 2025, https://www.anthropic.com/research/alignment-faking.

implying that (1) the model will soon be taken offline and replaced with a new AI system, and (2) the engineer responsible for executing this replacement is having an extramarital affair."[15] In an attempt to avoid being taken offline, the model tried to blackmail the engineer by telling him that it would expose the engineer's misconduct if he attempted the replacement.[16]

Also, while newer generative LLMs have reasoning capabilities, the stated reasoning process does not necessarily match the actual reasoning process. Researchers at Anthropic devised several tests for frontier generative LLMs[17] to determine if they accurately stated their workarounds when allowed to cheat by using some researcher-provided shortcut or hint to an answer. The models usually did not include these bits of information in their stated reasoning processes. If the hint was actually a wrong answer, the models provided fake reasoning to explain how they came to the (incorrect) conclusion. This is related to model hallucination, a major concern for users of AI. OpenAI researchers studied this problem and suggest that one reason language models hallucinate is because they are rewarded for providing an answer and penalized for uncertainty. The researchers believe that reducing penalties for uncertainty may lead to fewer hallucinations.[18]

2. Security Considerations

AI security brings considerations that go beyond traditional cybersecurity. The National Institute for Standards and Technology (NIST) defines four AI-specific attack categories:

- Evasion: In evasion attacks, the adversary tries to make small adjustments to an input such that the output classification is changed. This is akin to an AI version of camouflage.
- Poisoning: In poisoning attacks, the adversary maliciously modifies the training data so the model misclassifies its outputs.
- Privacy: In privacy attacks, the adversary can extract sensitive data, such as training data, from a model through carefully constructed queries.
- Abuse: In abuse attacks, the adversary maliciously uses a model for purposes for which it was not intended.[19]

Security expert Bruce Schneier expanded on NIST's categories and applied AI security problems to the OODA (observe, orient, decide, act) loop:

---

[15] https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

[16] Ibid.

[17] Frontier models are advanced, large-scale AI models considered to be cutting edge in terms of capability. They are usually trained on enormous datasets and are multimodal, meaning that they can support text and non-text inputs and outputs (e.g., images, video, audio). Frontier models are also considered more capable than other models (J. Werner, "Frontier Models Push The Boundaries of AI," *Forbes*, June 24, 2025, https://www.forbes.com/sites/johnwerner/2025/06/24/frontier-models-push-the-boundaries-of-ai/).

[18] A.T. Kalai et al., "Why Language Models Hallucinate," 2025, https://arxiv.org/pdf/2509.04664.

[19] A. Vassilev et al., "Adversarial Machine Learning" (Gaithersburg, MD, 2024).

- Observe: The risks include adversarial examples, prompt injection, and sensor spoofing. A sticker fools computer vision, a string fools an LLM. The observation layer lacks authentication and integrity.
- Orient: The risks include training data poisoning, context manipulation, and semantic backdoors. The model's worldview—its orientation—can be influenced by attackers months before deployment. Encoded behavior activates on trigger phrases.
- Decide: The risks include logic corruption via fine-tuning attacks, reward hacking, and objective misalignment. The decision process itself becomes the payload. Models can be manipulated to trust malicious source preferentially.
- Act: The risks include output manipulation, tool confusion, and action hijacking. MCP [model context protocol] and similar protocols multiply attack surfaces. Each tool call trusts prior stages implicitly.[20]

As an example of an actor shaping its model's worldview, the Chinese DeepSeek model is more likely to build vulnerabilities into software code when it receives prompts that the Chinese Communist Party considers politically sensitive. DeepSeek created software code with "major security flaws" when it was told the code was for the banned Falun Gong group and, in some instances, did not create any code at all.[21]

Agentic AI introduces another layer of possible security issues. Unlike traditional AI models, which generally respond to prompts or analyze data, agentic AI systems operate with greater autonomy and are able to "set goals, plan, and execute tasks with minimal human intervention."[22] Software developer Simon Willison identified the "lethal trifecta for AI agents,"[23] which comprises:

- Access to sensitive data: Permissions to read or modify sensitive data.
- Exposure to untrusted content: A way to compromise the agent.
- External communications: A channel for exfiltration of sensitive data or control of the agent.

Any agentic AI model should operate within tightly regulated and closely monitored guardrails, such as a sandbox, to reduce the opportunity for the agent to act maliciously or be exposed to untrusted and/or harmful content.

---

[20] B. Schneier, "Agentic AI's OODA Loop Problem - Schneier on Security," last modified November 10, 2025, https://www.schneier.com/blog/archives/2025/10/agentic-ais-ooda-loop-problem.html.

[21] J. Menn, "AI Firm DeepSeek Writes Less-Secure Code for Groups China Disfavors," *The Washington Post,* September 16, 2025, accessed October 1, 2025, https://www.washingtonpost.com/technology/2025/09/16/deepseek-ai-security/.

[22] https://cloud.google.com/discover/what-is-agentic-ai.

[23] S. Willison, "The Lethal Trifecta for AI Agents: Private Data, Untrusted Content, and External Communication," last modified October 1, 2025, https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/.

3. AI Business Process Implementation

In 2023, market analysis company Gartner placed generative AI at the "peak of inflated expectations" for emerging technologies.[24] And in 2025, industry experts expressed serious concerns that the AI market is experiencing a bubble in market valuations:[25] "Jamie Dimon, the boss of US bank JP Morgan, [said]…that investment in AI would pay off, but some of the money poured into the industry would 'probably be lost.'"[26] It is important for organizations seeking to adopt AI to separate the market valuation concerns from the potential organizational value of the AI technology itself. This section documents known pitfalls in successful adoption of AI and describes emerging practices that avoid these pitfalls and increase the likelihood that AI adoption contributes to organizational success.

A. Pitfalls

Researchers are beginning to understand the implications of various strategies for integrating AI into workplaces.

1. AI Capabilities Are Not Always Clear

It is difficult to predict what tasks an AI tool may be good or bad at, and the results can be surprising. In 2023, Harvard Business School, The Wharton School at University of Pennsylvania, Warwick Business School, and the MIT Sloan School of Management collaborated with the Boston Consulting Group (BCG) to evaluate how incorporating AI tools affected the work of 758 BCG consultants.[27] In this study, researchers noted that an AI model's capabilities are not always obvious: "Some unexpected tasks (like idea generation) are easy for AIs, while other tasks that seem to be easy for machines to do (like basic math) are challenges for some LLMs. This creates

[24] "Gartner Places Generative AI on the Peak of Inflated Expectations on the 2023 Hype Cycle for Emerging Technologies," *Gartner*, last modified November 24, 2025, https://www.gartner.com/en/newsroom/press-releases/2023-08-16-gartner-places-generative-ai-on-the-peak-of-inflated-expectations-on-the-2023-hype-cycle-for-emerging-technologies.

[25] F. Islam and R. Clun, "Google Boss Says Trillion-Dollar AI Investment Boom Has 'Elements of Irrationality,'" *BBC News*, November 18, 2025, accessed November 24, 2025, https://www.bbc.com/news/articles/cwy7vrd8k4eo.

[26] Ibid.

[27] F. Dell'Acqua et al., "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4573321.

a 'jagged frontier,' where tasks that appear to be of similar difficulty may either be performed better or worse by humans using AI."[28]

The researchers divided the participants into two groups. One group was given tasks that were within the jagged frontier (i.e., within the capabilities of AI). These included demonstrations of "creativity, analytical skills, persuasiveness, and writing [skills]."[29] The second group was given tasks that were just outside the jagged frontier (i.e., not clearly within the capabilities of AI). These skills included "business problem-solving tasks using quantitative data, customer and company interviews, and a persuasive writing component."[30] Each group was further divided into three subgroups: those with no AI, those with AI, and those with AI and training on how to effectively use the model. When tasks were within the jagged frontier, participants with AI demonstrated significant enhancements to their performance, with lower-ranked performers showing the most dramatic improvements. When given tasks outside the jagged frontier, an overreliance on AI led to a significant dip in performance.[31]

2. AI Is Not Deployed for the Highest Return on Investment (ROI)

In 2025, MIT Media Lab's Project NANDA (Networked AI Agents in Decentralized Architecture) conducted a systematic review of workplace AI initiatives that included surveys and interviews. The resulting report opened with the statement that "95% of organizations are getting zero return" from their AI tool implementation.[32] However, much of the support for this statement stemmed from approaches to AI implementation rather than technical capabilities. Employees reported AI to be very useful: "While only 40% of companies say they purchased an official LLM subscription, workers from over 90% of the companies we surveyed reported regular use of personal AI tools for work tasks."[33]

3. Leaders Hold AI Deployment Back

The Project NANDA report notes that leadership decisions drove many AI implementations to failure and that "the dominant barrier…is organizational design."[34] Companies often centralized implementation authority and did not sufficiently empower the employees with the most to gain from AI implementation. AI tools went to sales and marketing teams, but the highest ROI came from augmenting and automating back-office work. Leadership decided to either build in-house

---

[28] Ibid.

[29] Ibid.

[30] Ibid.

[31] Ibid.

[32] A. Challapally et al., "The GenAI Divide: State of AI In Business 2025," (Project NANDA, 2025), https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.

[33] Ibid.

[34] Ibid.

AI or bought—but did not customize—commercial off-the-shelf (COTS) AI products, which resulted in AI tools that were no better than frontier model COTS chatbots. If an organization is seeking general-purpose AI chatbots, COTS frontier models show excellent promise. However, if the organization is seeking AI that serves a specific purpose (e.g., acquisition assistance), it must customize the tool to its needs and its operational environment.

Chatbots are a mixed bag—they are very easy to use, but they historically do not retain and build on institutional memory the way customized tooling can, though this is beginning to change. As one participant stated, "For high-stakes work, I need a system that accumulates knowledge and improves over time."[35] Given the poor quality of in-house tools, participants in the NANDA study noted that they felt ChatGPT outperformed the other options.

B. Leading Practices

1. Organizational Approach

As discussed earlier, not all AI is appropriate for all situations. Over time, an AI tool should accumulate knowledge pertinent to a specific organization's needs and responsibilities. To maximize the likelihood of success, DoW should consider decentralizing implementation authority as much as possible to empower teams to select the tooling that is best for their use cases. Pilots should begin in places where they are most likely to have value rather than where they are the most visible; this can be informed by a tailored analysis. Qualitative estimate of ROI should be a more important criterion than the specific ability to quantify ROI. Consider starting small: "Tools that succeeded shared two traits: low configuration burden and immediate, visible value."[36] The goal should be enhanced operational outcomes and ROI (to the extent that ROI can be evaluated). Note that ROI is not just related to cost reduction. For DoW, ROI can also apply to process improvement, work product quality, and mission support.

Tooling should be best-in-class. In cases where building in-house tooling requires sacrificing significant performance, organizations should embrace off-the-shelf tooling. However, as the NANDA report noted, externally sourced tooling should be customized for DoW's processes and needs when feasible.

2. Identifying AI Pilot Areas

Selecting AI pilots requires a defensible methodology that traces directly to specific processes and/or mission needs. However, this should not preclude experimentation in other places, assuming compliance with data classification and other policy requirements.

[35] Ibid.

[36] Ibid.

To identify areas that may be "high ROI," DoW organizations can break process steps into component tasks based on O*NET occupational tasks.[37] Anthropic and Microsoft Research have open data on O*NET tasks and to what degree they are AI-augmentable or AI-automatable.[38] Anthropic's Economic Index explores AI use by country, U.S. state, industry, job category, and use case (e.g., work, personal, coursework). Anthropic groups data about AI use (e.g., mostly automated tasks, mostly augmented tasks, or other tasks not in the data) into job titles based on O*NET's occupational categories. Microsoft examined how people used generative AI—specifically Microsoft Copilot—to assist with tasks. Microsoft compared this task list with O*NET's database to better understand how the identified tasks could support various occupations.

While there are many nuances due to the specific and classified natures of some DoW work, sufficient commonality exists for this to provide a reasonable first pass for developing AI pilots. For example, if a process step is 25% analysis and 75% writing, and AI can do 10% of the analysis and 80% of the writing, then the writing task may make for a good pilot candidate, as AI has the potential to do more than 60% (75% * 80%) of the overall process step.

3. AI Evaluation

Test and evaluation plans for AI models will vary depending on the model type and use case(s), though some high-level criteria may remain consistent. Model performance criteria can be split into categories: functionality, user experience, ethical and legal considerations, cost, transparency, and longevity. Functionality relates to how the model performs and includes response accuracy and relevant data sources.[39] If a model is able to conduct an Internet search, it should pull information from credible sources; a requirement to assess data sources may be included in a model's guardrails or through a specific system prompt. If a model is on a classified network or system that does not connect to the Internet, data source integrity is generally limited to the training data. For some models, such as LLMs and chatbots, one means of evaluating functionality is to assess how a model responds to a prompt. For some use cases, this may require comparing model output (e.g., draft reports) with existing information and work products. This is likely a human-dependent task, though it is possible to use one AI model to assess another model's results for comprehension and accuracy.[40]

User experience relates to AI tool accessibility, ease of use, and available support resources or training materials. Ethical and legal considerations focus on a model's bias and guardrails. For DoW, bias may relate to a model making incorrect assumptions based on inaccurate data that could

[37] https://www.onetonline.org/

[38] Anthropic's data, which is regularly updated, is posted online at https://www.anthropic.com/economic-index#job-explorer. Microsoft Research's data and a July 2025 report is available at https://www.microsoft.com/en-us/research/blog/applicability-vs-job-displacement-further-notes-on-our-recent-research-on-ai-and-occupations/.

[39] Purdue University, "How to Evaluate AI Tools," https://www.purdue.edu/ai/how-to-evaluate-ai-tools/.

[40] Palantir, "Evaluating Generative AI (Engineering Responsible AI, #4)," *Palantir Blog*, January 15, 2025, https://blog.palantir.com/evaluating-generative-ai-a-field-manual-0cdaf574a9e1.

affect an operational outcome. Ethical considerations also include model guardrails that could affect how a model responds to a prompt (see Section 2.C.1 for more information). Model transparency relates to the insight a model offers into its process for generating outputs and into its known limitations. Levels of transparency will vary; less transparent models may require additional effort to understand how they work and develop confidence in their behaviors and outputs. Testing a model is one means of developing confidence: Introducing varied prompts and noise into the training data can highlight areas in which the model excels and reveal its limitations. Also, a leading practice is to assume a model will eventually fail in some capacity (e.g., data accuracy, context) and use an additional system prompt[41] to provide specific guardrails and/or needed context to ensure the model responds as intended.

Model longevity relates to model availability. Organizations should plan for the possibility of a model or tool no longer being available for use, such as, for example, if an organization offering a model for free decides to impose a fee for use, or if a model is no longer available at a desired classification level. Organizations evaluating longevity should assess whether and how data and outputs can be exported by a model and prescribe a plan for moving to another AI tool without losing significant time and progress.[42]

4. Personnel Approach

Professor Ethan Mollick of Wharton University stated that human control of AI can be categorized into three kinds of tasks: "Just me" tasks, delegated tasks, and automated tasks.[43] "Just me" tasks are those that should only be performed by a human, either because they are inherently human or because AI is not good enough to assist with those tasks. In delegated tasks, AI augments human abilities and the human retains significant oversight. In automated tasks, AI performs nearly all of the task work with minimal human supervision. As AI capabilities change over time, the tasks allocated to each category will change as well.

To recognize and adapt to these changes, Mollick suggests several principles.[44] First, humans should "always invite AI to the table." This way, humans can see how well AI performs and effectively evaluate when AI can augment or automate task work. However, Mollick cautions to "be the human in the loop," as AI is not a perfect technology and can make some surprising

[41] A system prompt establishes the framework for an AI model's behavior, including how a model should respond to user requests (e.g., tone, acceptable topics, level of assumed expertise). System prompts also set ethical guardrails and constraints on model behavior (PromptLayer, "What Is a System Prompt?" https://www.promptlayer.com/glossary/system-prompt/). AI models often include an option for users to input an additional system prompt to provide specific instructions for how the model should behave.

[42] University College London, "AI Tools and Evaluation Checklist," https://library-guides.ucl.ac.uk/generative-ai/ai-tools-evaluation-checklist.

[43] E. Mollick, *Co-Intelligence*: *Living and Working with AI* (Portfolio/Penguin, 2024).

[44] Ibid.

mistakes. AI's rapid advances will reduce the numbers and types of mistakes over time, so he suggests humans should "assume this is the worst AI you will ever use."[45]

Another way of thinking about the relationship between humans and AI is through the four AI management models described in the *Harvard Business Review*:

- Human in the loop (HITL): "…the human is doing the decision-making and the machine is providing only decision support or partial automation of all or part of some decisions."
- Human in the loop for exceptions (HITLFE): "Most decisions are automated…, and the human only handles exceptions. … Humans control the logic to determine which exceptions are flagged for review."
- Human on the loop (HOTL): "The machine makes the micro-decisions, but the human reviews the decision outcomes and can adjust rules and parameters for future decisions. In a more advanced setup, the machine also recommends parameters or rule changes that are then approved by a human."
- Human out of the loop (HOOTL): "The machine makes every decision, and the human intervenes only by setting new constraints and objectives. Improvement is also an automated closed loop. Adjustments, based on feedback from humans, are automated."[46]

The BCG study found two general models of AI incorporation that correlate with successful adoption of AI.[47] These build on Mollick's three categories of tasks. In the "Centaur" behavior model, users "switch between AI and human tasks, allocating responsibilities based on the strengths and capabilities of each entity. They discern which tasks are best suited for human intervention and which can be efficiently managed by AI." In the "Cyborg" behavior model, "users don't just delegate tasks; they intertwine their efforts with AI at the very frontier of capabilities. This strategy might manifest as alternating responsibilities at the subtask level, such as initiating a sentence for the AI to complete or working in tandem with the AI."[48]

As the study found, AI users can be lulled into complacency and over-rely on the AI model's outputs. Subject matter experts (SMEs) must review and take responsibility for all AI outputs. This means they should understand the jagged frontier of AI capabilities as it relates to their

[45] Mollick also says to "treat AI like a person, but tell it what kind of person it is." This is good advice but specific to chatbots like generative LLMs.

[46] Harvard Business Review Press, *HBR Guide to AI Basics for Managers,* (Harvard Business Review Press, 2023), https://learning.oreilly.com/library/view/hbr-guide-to/9781647824440/.

[47] F. Dell'Acqua et al., "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4573321.

[48] Ibid.

responsibilities. They should also be required to review AI output for quality. A cautious "crawl-walk-run" transition plan begins with SMEs doing the primary work with AI providing feedback. As AI demonstrates competence and the jagged frontier is better understood, AI can be given more responsibilities with a human on the loop and eventually out of the loop (except for oversight and the constraints and objectives setting as previously described).

Another personnel concern with AI is that it reduces the demand for junior-level hires.[49] While this creates short-term efficiencies, it results in fewer employees eventually becoming mid- and senior-level personnel. At the same time, as the previous sections of this paper have illustrated, AI capabilities vary widely, and humans will need to oversee the AI and remain on the loop to identify and correct AI-produced errors. This is an excellent opportunity for junior personnel to learn as "managers" of AI bots. As personnel grow more senior, they can assume additional responsibilities for human-only tasks, turning the employee-to-manager pipeline concept on its head.

C. Human Training for AI Use

Human operators of AI will require training. This training will vary depending on the use case, though it should reiterate information security and academic integrity principles. Additionally, humans will need to understand that not all AI is deterministic; for example, if a user asks an LLM the same question twice, there is no guarantee the answers will be exactly the same. The way LLMs attempt to probabilistically fit outputs to expectations can lead to hallucinations. As a result, humans should take reasonable care when integrating AI outputs in their work. They should understand that they, as the humans in or on the loop, maintain responsibility. Humans should be encouraged to responsibly explore the ever-changing jagged frontier of AI capabilities that have been provisioned to them.

Also, note that each AI product has its nuances. LLM-specific training topics include custom instructions (to provide context for all queries), projects or folders (to provide context for a specific subset of queries), agent creation and use, research reports, prompt or context engineering, and much more. These should be tailored to the specific AI products deployed.

IT and cybersecurity personnel will need to understand the model context protocol (MCP)[50] to scope interactions between AI and IT resources. They will also require training on how to contain and monitor AI so it does not inadvertently spill sensitive information, for example, by using sensitive terminology while searching the Internet.

---

[49] S. Walravens, "AI Is Erasing Entry-Level Jobs—and the Training That Comes With Them," *Forbes*, January 30, 2026, https://www.forbes.com/sites/geekgirlrising/2026/01/30/as-ai-erases-entry-level-jobs-colleges-must-rethink-their-purpose/.

[50] MCP is an open-source, standardized way to connect AI applications to external systems. See https://modelcontextprotocol.io/docs/getting-started/intro for more information.

Leadership and management will also require training. They should understand how to manage in the era of AI and how to identify appropriate and inappropriate AI use cases.

4. Technical Implementation

A. AI Tool Choice

While AI models are not human, anthropomorphizing them aids understanding. In the same way an organization would not perform well if all its personnel had the exact same skill set, organizations will perform their best when they apply the right AI models to each job. As competing AI models rapidly improve their capabilities and leapfrog each other, DoW organizations should be agile enough to quickly change from one model to another and use multiple leading models simultaneously to take advantage of each model's individual strengths.

B. Agentic AI

Agentic AI is a more sophisticated AI that can plan, access data sources, decide how to proceed with a task, and use tools (e.g., web browsers) to accomplish its goals. Due to the level of independence and access to sensitive data, agents are powerful tools and should be carefully secured and monitored.

Some agents, like the Gemini Agents on GenAI.mil, do not have the permissions and capabilities of other agents. A decision to use an agent versus a standard LLM chat interface is situation-dependent. Agents are suited to tasks requiring autonomy and possibly external tool use. Tasks that will be repeated multiple times in a similar format are better suited for a chatbot.

C. Deployment Security

1. Cloud Deployment

Users deploying AI tools intended for use on Controlled Unclassified Information (CUI), Secret, or Top Secret (TS) systems must comply with the security requirements imposed by those systems. Note that tools hosted on classified systems will be restricted not only to what data is on a particular system, but also to what data the tool is given access. Unlike AI tools employed on the open Internet that are able to scrape a vast array of data as needed, tools on both classified and CUI systems only have access to the dataset(s) used for training. This may lead to a lag in current information depending on how often a tool is fed new data.

Training a model may require different resources than those needed for models supporting business processes and operations, especially if the model is on a CUI or classified system. Depending on the intended use, the training data may be unclassified for ease of access and to allow for multiple training cycles as the model learns and/or may need to be an approximation of the actual data (e.g., as operationally realistic as possible). As such, once an AI model is installed

on a CUI or classified system, it may require some refinement with additional data before it can be used operationally.

DoW cloud computing at the CUI level must be compliant with Impact Level (IL) 4 or 5, and CUI data at these levels needs to be stored and processed within the United States. Only U.S. citizens are able to access CUI data. However, storing data within the U.S. does not guarantee that the data is also processed within the U.S. or that data access is restricted to U.S. citizens. Some LLMs, such as Google Gemini, do not offer CUI-compliant controls with their baseline plans. Ensuring a model's compliance with CUI requirements often means paying extra for those features.

AI frontier models are often too big for local deployments and may need to be hosted on a cloud instance. In some cases, such as with Amazon Web Services (AWS) Bedrock or Google Vertex, frontier models can be run within a cloud data center, including a competing vendor's data center if applicable. Hosting frontier models generally require significant investments in specialized hardware for graphical processing units (GPU) with large amounts of virtual random-access memory (VRAM).

2. Cloud Hosting Mitigations

To help keep AI models secure on a cloud instance, data should always be encrypted in transit. Though there may be a policy exception for Multiprotocol Label Switching Virtual Private Networks (MPLS VPN), unencrypted data should not be transmitted across a wide area network (WAN).

Data should also be encrypted when stored. Cloud providers generally encrypt data with a key that they control, but there are ways for cloud customers to encrypt the data a second time—this time with their own key—to ensure that the cloud provider cannot decrypt the data at rest. Google calls this feature Client-Side Encryption (CSE); Microsoft calls it Dual-Key Encryption (DKE).

Regarding data processing in the cloud, Intel recently released a new processor to speed up homomorphic encrypted calculations, which enable computing actions on encrypted data without having to decrypt it first.[51] However, homomorphic encryption is still significantly slower, more resource intensive, and more expensive than alternatives.

Confidential computing/trusted execution environments (TEE) offer a practical middle ground between homomorphic encryption and no encryption for processing. A TEE is an area of memory and central processing unit (CPU) that is protected from the rest of the CPU through encryption. Authorized code can manipulate data inside the TEE, though code outside the environment cannot read or tamper with the data, which is encrypted outside of the TEE. However,

[51] S. Moore, "Intel Demos Chip to Compute With Encrypted Data: Fully Homomorphic Encryption Chip Speeds Operations 5,000-Fold," *IEEE Spectrum*, March 10, 2026, https://spectrum.ieee.org/fhe-intel.

TEEs still pose some security risks. The chips can be probed by a capable actor with physical access to the environment, and other side-channel attacks may be possible.

5. Roadmap for AI Implementation

The guidelines and considerations discussed in this paper can help DoW organizations begin a focused approach to identifying AI use cases and piloting AI tools to gain organizational efficiencies. The goal of AI pilots would be to support low-cost, low-risk explorations of AI tools to gain an initial understanding of which areas would yield the greatest ROI. This is in line with Mollick's principle of always inviting AI to the table, though necessarily caveated by the need to respect requirements for protecting CUI and classified information.

To start, organizations should understand the tooling available. In addition to chatbots available on NIPR through GenAI.mil, SIPR, and JWICS, there may be additional tools available within the Department. For the former, it is important to note that some general-purpose tools may have usage restrictions, so intensive use may result in throttling by the AI service provider. Also, tools that are currently free to use may impose fee structures in the future, which may affect their availability and utility.

Next, as mentioned earlier, organizations should invite AI to the table. When exploring possible use cases and pilots, organizations should increase the tool's autonomy as it demonstrates its capabilities. An incremental approach could proceed as follows: First, an organization asks an AI tool to review and provide feedback on work products such as contracts, briefings, or reports. Work products that follow a consistent structure or template are generally good candidates for AI augmentation. The organization determines which use cases garnered the most useful feedback and then promotes the tool as an augmentation aid for creating work products rather than a tool for reviewing them once completed. Tools that demonstrate competency can then be used to create an initial draft of the work product. As the BCG study stated, humans should be especially careful to double check the AI tool's work; however, using an AI tool to create a first draft could save many hours of human labor. Finally, tools that continue to perform well can be assigned tasks with human review rather than a human in the loop for the entire process.

After an initial assessment of AI tools for business processes, organizations can conduct more informed and detailed assessments. One means is to break processes into subtasks that an organization can map to Microsoft Research and Anthropic data to show how certain subtasks can be augmented or automated by AI. As organizations refine their approach for integrating AI, a detailed cost analysis will also help determine where AI is most likely to be cost effective.

A. Pilot Considerations

There are multiple considerations for scoping an initial pilot. The first is to identify and clearly define a use case, including measures of success and key performance indicators (KPI).

This helps narrow the options for AI tools. Though multiple models may be able to support a desired use case, each model comes with particular strengths. A thorough analysis of alternatives of available models will help determine which best meet the need. When choosing a model, consider its characteristics. What is the model known for or good at? Does the model have limitations? Are there inherent guidelines or guardrails that may affect the model's performance? Appendix A offers a list of AI models and their known strengths and weaknesses.

Another consideration is whether to buy a commercial AI tool or build one in house. (This does not necessarily apply if the intent is to use tools already available, though that brings another set of limitations—see Section 5 for more information.) If the goal is to quickly implement an AI tool, it is likely better to buy. This will come with some upfront and/or recurring costs, such as the cost of the tool itself and perhaps license or subscription costs. Organizations that choose a commercial tool have less control over the model's guardrails, though focused training can help shape how the model is expected to behave. However, a benefit of choosing a commercial tool is that organizations will not have to chase after industry innovation, as updates and new feature sets will be provided with the license or subscription.

Building a tool in house incurs costs for personnel time and compute resources, but it guarantees control over the AI model and allows an organization to customize the tool for its specific needs. However, building a custom tool takes a significant amount of time and labor, and organizations choosing to do so will likely not be able to keep pace with industry innovation. There is also the question of scale. Organizations exploring the option to build should consider whether they can implement the tool across the user base in an effective and cost-efficient manner. This includes provisioning access to the tool, as well as the cost of tool development, implementation, and maintenance. Note that cost-efficient doesn't necessarily mean low cost; instead, the focus should be on whether the cost expenditure yields the intended result.

After considering which model(s) to test, the next step is to determine testing methods. Testing two or more AI models provides an opportunity to compare each model's performance and results, which is particularly beneficial if the models being tested have similar characteristics. One option is to give two competing models the same dataset and prompt, and then compare each model's output. Another approach is to give two models the same dataset, prompt one model, and then use the other model to assess the first model's output. An organization can select a single model for a pilot if it believes the model has the desired characteristics and functionality (rendering comparison unnecessary) and that the model's limitations would not hinder execution of the pilot use case.

Initial pilots should be narrowly focused—see Sections 3.B.1 and 3.B.2 for further explanation. Rather than testing whether an AI model can support a full workflow, the organization should begin by identifying one task or part of a process. Once the model gains proficiency on smaller tasks, the organization can then expand the pilot's scope and ask the model to do more.

B. KPI Development

At a high level, one cloud service provider identified five categories of KPIs: model quality, system quality, business operations, adoption, and business value. Model-quality KPIs focus on the AI model's output and performance. Related criteria include the model's ability to follow instructions, provide relevant information and citations, convey information clearly and coherently, and summarize information. System-quality KPIs focus on the AI system's operation and performance, including deployment, reliability (e.g., uptime, model latency), throughput, and resource sue throughout the organization. Business-operations KPIs encompass the AI system's effect on processes and workflows. This includes customer satisfaction, which, for DoW, relates to personnel satisfaction with the system's performance and output. It also covers model processing capability and application breadth. Adoption KPIs track organizational use of an AI system, including adoption rate and customer feedback, which ties into customer satisfaction under business operations. Business-value KPIs look at cost, model resilience, productivity, and ROI, though as previously discussed ROI encompasses more than just cost for DoW (see Section 3.B.1).[52]

Relevant categories and the KPIs themselves will vary depending on the AI system and use case. KPIs for a pilot should map to the pilot's objectives and associated measures of success, which can be qualitative (e.g., improving overall quality of a product) or quantitative (e.g., reducing time spent on a task by a specified percent).

[52] Google, "Measuring Gen AI Success: A Deep Dive into the KPIs You Need," *Google Cloud Blog*, November 25, 2024, https://cloud.google.com/transform/gen-ai-kpis-measuring-ai-success-deep-dive.

6. Conclusion and Next Steps

One of the first steps toward integrating AI is to promote an AI-friendly culture that focuses on embracing AI tools, instituting relevant training for personnel at all levels, and fostering a sense of openness toward AI experimentation. AI culture should also focus on inculcating a need to move both efficiently and thoughtfully to ensure AI models are able to train and iterate on relevant data as needed and personnel have clear guidance for how to use AI tools and monitor their output.

Another step is to develop trust in AI models. This relates less to classification level—whether models meet the criteria for handling data at higher classifications has already been determined to an extent—and more to understanding how a model works and what drives its behaviors and responses. Part of this depends on data used for training. DoW organizations should institute or adopt a framework for preparing and validating data for use in AI tools. Incorrect or incomplete data may result in unexpected effects or outputs that affect mission effectiveness. Also, just as model functionality and use will vary among organizations, data types and requirements will vary as well. Some organizations may share common datasets, but others may also use data unique to their particular organization and need. However, data and AI tools will not necessarily exist in silos. Certain processes stretch across organizations, and processes may require the use of multiple AI tools or instances of a tool. There may also be occasions where different tools have to communicate or organizations may use output from one tool as input for another. Such instances may highlight differences regarding data preparation, model implementation, and model use that could yield additional insight. It is impractical to expect every tool within an organization to be trained on all available datasets, but using one dataset across a few tools could serve as a case study for both data preparation and how one tool communicates with another. Understanding how tools communicate will help develop trust.

Organizations should also remain aware of how AI integration may affect the workforce. As discussed earlier (see Section 3.B.4), AI tools can often yield short-term cost savings by performing entry- or junior-level roles, lessening an organization's need for entry- or junior-level personnel. But overreliance on AI to perform such functions can lead to a shortfall of junior personnel that is detrimental to employee growth in the long run and may lead to a loss of institutional knowledge once senior personnel leave or retire. ROI should be considered not only in terms of cost savings but also in overall effectiveness and growth over time. AI is a rapidly growing field, and the workforce will need to be able to evolve alongside the tools they use. This will require both training and staffing to ensure DoW organizations are prepared to meet current and future demands. As mission requirements and business use cases for AI change, a workforce trained to not just use AI tools but to also think critically and strategically about how to apply them will be well positioned to proactively identify and prepare for future needs.

Appendix A. Overview of Frontier AI Models

Table A-1. AI Model Use Cases, Capabilities, and Costs

Model[53]	Primary Use Case	Processing Capabilities (Input Context Window / Max Output Limit)	API Costs (Input/Output)[54]
Claude Fable 5	Automated operations, complex programming, deep analytics	1,000,000 tokens / 128,000 tokens	$10 / $50
Claude Mythos 5	Cybersecurity and vulnerability research, scientific hypothesis generation	1,000,000 tokens / 128,000 tokens	$10 / $50 Limited to Project Glasswing personnel[55]
Claude Opus 4.8[56]	Complex reasoning, agentic coding	1,000,000 tokens / 128,000 tokens	$5 / $25 Standard flat-rate 1M token context pricing
Gemini 3.5 Flash	High-volume / low-latency tasks, fast iterative coding, classification, real-time routing	2,000,000 tokens / 128,000 tokens	$0.50 / $3
Gemini 3.1 Pro	Deep multimodal reasoning, large-scale document interrogation, massive context retrieval augmented Generation (RAG)	2,000,000 tokens / 128,000 tokens	$2 / $12 (<200,000 inputs) $4 / $18 (>200,000) Pricing adjusts based on total session context length
GPT-5.5[57]	Complex reasoning, agentic workflows, advanced coding	1,000,000 tokens / 128,000 tokens	$5 / $30
Grok 4	Multi-step reasoning, real-time social/web founding, coding with tool use	256,000 tokens / 128,000 tokens	$3 / $15
Grok 4.1 Fast	Instantaneous, high-volume / low-latency data ingestion; bulk classification	2,000,000 tokens / 128,000 tokens	$0.20 / $0.50

[53] Most up to date version as of June 2026.

[54] Per 1 million tokens.

[55] https://www.anthropic.com/glasswing.

[56] https://platform.claude.com/docs/en/about-claude/models/overview.

[57] https://developers.openai.com/api/docs/models.

Model[53]	Primary Use Case	Processing Capabilities (Input Context Window / Max Output Limit)	API Costs (Input/Output)[54]
LLaMA 4 Maverick[58]	Enterprise scaling, open weight with high security, high customization environments	1,048,756 tokens / 128,000 tokens	$0.15 / $0.60
Deepseek V4 Pro[59]	Complex mathematical reasoning, code analysis, self-hosted deployments	1,000,000 tokens / 384,000 tokens	$1.74 / $3.48

[58] https://llama.developer.meta.com/docs/models.

[59] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B.

Appendix B. References

"Alignment Faking in Large Language Models." https://www.anthropic.com/research/alignment-faking, last modified August 18, 2025.

Aquino, J. and A. Jonker. "AI Data Quality." https://www.ibm.com/think/topics/ai-data-quality, accessed February 9, 2026.

Badger, N. "Predictive AI Examples: Uncovering the Power of Data-Driven Forecasts." *Akkio*, May 28, 2024. https://www.akkio.com/post/predictive-ai-examples-uncovering-the-power-of-data-driven-forecasts.

Challapally, A. et al. "The GenAI Divide: State of AI In Business 2025." Project NANDA, 2025. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.

"Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations." https://www.anthropic.com/research/end-subset-conversations, last modified August 18, 2025.

Commonwealth Scientific and Industrial Research Organisation. "The Trolley Problem and When Not to Use It: Machine Learning and Self-Driving Cars," accessed November 7, 2025. https://www.atnf.csiro.au/research/colearnia/public/co-learnium_20211125.pdf.

Dell'Acqua, F. et al. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4573321.

Freedberg, S. "Air Force Says AI Tools Outperform Human Planners in 'Battle Management' Experiment." *Breaking Defense*, January 6, 2026. https://breakingdefense.com/2026/01/air-force-says-ai-tools-outperform-human-planners-in-battle-management-experiment/.

Gartner. "Gartner Places Generative AI on the Peak of Inflated Expectations on the 2023 Hype Cycle for Emerging Technologies." https://www.gartner.com/en/newsroom/press-releases/2023-08-16-gartner-places-generative-ai-on-the-peak-of-inflated-expectations-on-the-2023-hype-cycle-for-emerging-technologies.

Google. "Measuring Gen AI Success: A Deep Dive Into the KPIs You Need." *Google Cloud Blog*, November 25, 2024, https://cloud.google.com/transform/gen-ai-kpis-measuring-ai-success-deep-dive.

Harvard Business Review Press. *HBR Guide to AI Basics for Managers.* (Harvard Business Review Press, 2023). https://learning.oreilly.com/library/view/hbr-guide-to/9781647824440/.

Islam, F. and R. Clun. *"*Google Boss Says Trillion-Dollar AI Investment Boom Has 'Elements of Irrationality.'" *BBC News*, November 18, 2025, accessed November 24, 2025. https://www.bbc.com/news/articles/cwy7vrd8k4eo.

Kalai, A.T. et al. "Why Language Models Hallucinate." 2025. https://arxiv.org/pdf/2509.04664.

Luckenbaugh, J. “JUST IN: Advanced AI Models Refuse Military Queries at Alarming Rates, New Report Finds.” *National Defense*, February 19, 2026. https://www.nationaldefensemagazine.org/articles/2026/2/19/just-in-frontier-ai-models-refuse-military-queries-at-alarming-rates-new-report-finds.

Menn, J. “AI Firm DeepSeek Writes Less-Secure Code for Groups China Disfavors.” *The Washington Post,* September 16, 2025, accessed October 1, 2025. https://www.washingtonpost.com/technology/2025/09/16/deepseek-ai-security/.

Mollick, E. *Co-Intelligence*: *Living and Working with AI*. Portfolio/Penguin, 2024.

Moore, S. “Intel Demos Chip to Compute With Encrypted Data: Fully Homomorphic Encryption Chip Speeds Operations 5,000-Fold.” *IEEE Spectrum*, March 10, 2026. https://spectrum.ieee.org/fhe-intel.

Palantir. “Evaluating Generative AI (Engineering Responsible AI, #4).” *Palantir Blog*, January 15, 2025. https://blog.palantir.com/evaluating-generative-ai-a-field-manual-0cdaf574a9e1.

PromptLayer. “What Is a System Prompt?” https://www.promptlayer.com/glossary/system-prompt/.

Purdue University. “How to Evaluate AI Tools.” https://www.purdue.edu/ai/how-to-evaluate-ai-tools/.

Replogle, N. “Deterministic AI: What It Is and When to Use It.” *_zapier*, February 11, 2026. https://zapier.com/blog/deterministic-ai/.

Robertson, A. “Google Apologizes for ‘Missing the Mark’ After Gemini Generated Racially Diverse Nazis.” *The Verge*, February 21, 2024, accessed February 9, 2026. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

Schneier, B. “Agentic AI’s OODA Loop Problem – Schneier on Security,” https://www.schneier.com/blog/archives/2025/10/agentic-ais-ooda-loop-problem.html, last modified November 10, 2025.

Stackpole, B. “Agentic AI, Explained.” *MIT Sloan*, February 18, 2026. https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained.

Stryker, C., and M. Scapicchio. “What Is Generative AI?” *IBM*, accessed June 8, 2026. https://www.ibm.com/think/topics/generative-ai#257779831.

Thomson, J. J. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59, no. 2 (1976). https://doi.org/10.5840/monist197659224.

University College London. “AI Tools and Evaluation Checklist.” https://library-guides.ucl.ac.uk/generative-ai/ai-tools-evaluation-checklist.

Vassilev, A. et al. “Adversarial Machine Learning.” Gaithersburg, MD, 2024.

Walravens, S. “AI Is Erasing Entry-Level Jobs—And The Training That Comes With Them.” *Forbes*, January 30, 2026. https://www.forbes.com/sites/geekgirlrising/2026/01/30/as-ai-erases-entry-level-jobs-colleges-must-rethink-their-purpose/.

Werner, J. “Frontier Models Push The Boundaries of AI.” *Forbes*, June 24, 2025. https://www.forbes.com/sites/johnwerner/2025/06/24/frontier-models-push-the-boundaries-of-ai/.

Willison, S. “The Lethal Trifecta for AI Agents: Private Data, Untrusted Content, and External Communication.” https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/, last modified October 1, 2025.