CreateAI's Available LLM Models (includes MyAI Builder, ASU GPT, Model Comparison)

Empowering Productivity and Innovation with Secure AI Models

Our platform provides a diverse range of cutting-edge language models in ASU's secure, walled off environment approved for us with FERPA, non-sensitive data.

- **Diverse Model Selection:** With a variety of models, you can choose the best fit for specific tasks, from drafting documents and summarizing data to creating intelligent chatbots and beyond.
- **Enhanced Security:** All models on our platform adhere to strict security and privacy standards, ensuring your data remains protected while you innovate with confidence.
- **Accelerated Productivity:** By automating repetitive tasks and delivering precise outputs, our AI models free up valuable time, allowing you to focus on strategic, high-value activities.
- **Driving Innovation:** The flexibility and reliability of our models open new possibilities for creativity and problem-solving, helping you stay ahead in a competitive landscape.

Whether you're optimizing workflows, fostering collaboration, or building transformative solutions, our secure AI models serve as the foundation for unlocking your full potential.

For most up to dates available in CreateAI, MyAI Builder, ASU GPT and Model Comparison: https://ai.asu.edu/createai-platform-available-llm-models

LLM Details

Claude 2

- **Definition**: Claude 2 is a conversational AI model developed by Anthropic, designed for tasks that require a mix of reasoning and creativity.
- **Strengths**: Excellent for casual conversations, simple task automation, and creative content generation.
- **Weaknesses**: Limited in handling highly technical or complex tasks compared to more advanced models.
- **Who Should Use It**: Ideal for users who need a reliable assistant for everyday tasks like scheduling, casual writing, or brainstorming.

Claude 2.1

- **Definition**: An improved version of Claude 2, offering enhanced reasoning and conversational abilities.
- **Strengths**: Better at following complex instructions and maintaining coherent, long-form discussions.
- **Weaknesses**: Still not as fast or efficient with large datasets compared to models optimized for data processing.
- **Who Should Use It**: Users needing help with slightly more complex workflows, such as creating reports, organizing projects, or answering detailed queries.

Claude 3 Haiku

- **Definition**: Claude 3 Haiku is designed to emphasize creativity and language generation in a poetic or concise format.
- **Strengths**: Outstanding for creative writing, poetry, and tasks requiring eloquence and expressiveness.
- **Weaknesses**: Not ideal for tasks requiring heavy reasoning or data-driven decision-making.
- **Who Should Use It**: Perfect for writers, marketers, or anyone looking to generate creative content quickly.

Claude 3 Opus

- **Definition**: Claude 3 Opus is a variant tailored for more structured writing and professional use cases.
- **Strengths**: Highly effective at producing structured documents, reports, and long-form content.
- **Weaknesses**: Less suitable for fast, casual conversation or rapid task completion.
- **Who Should Use It**: Professionals needing help drafting formal documents, reports, or longer articles.

Claude 3 Sonnet

- **Definition**: A model specialized in generating structured, poetic, and complex language forms, ideal for creative writing with a formal tone.
- **Strengths**: Great for generating structured content like sonnets, formal speeches, and complex storytelling.
- **Weaknesses**: Limited use outside of creative or structured writing tasks.
- **Who Should Use It**: Writers, poets, or anyone looking to create eloquent and formal written content.

Claude 3.5 Sonnet

- **Definition**: An upgraded version of Claude 3 Sonnet with enhanced understanding of complex language patterns.
- **Strengths**: Offers even better output quality for creative or structured writing tasks with advanced language capabilities.
- **Weaknesses**: May be slower in processing large amounts of data or technical content compared to other models.
- **Who Should Use It**: Users who need more polished, high-quality language generation for creative or formal purposes.

Claude 3.7 Sonnet

- **Definition:** An advanced version of Claude Sonnet with enhanced fluency and reasoning for both creative and structured content.
- **Strengths:** Delivers higher-quality output across longer and more complex prompts, with improved language finesse and task comprehension.
- **Weaknesses:** Still not optimized for highly technical, code-heavy, or data-intensive tasks.
- **Who Should Use It:** Ideal for users needing polished, articulate content—great for writers, marketers, and professionals working on reports or content creation.

Claude Instant

- **Definition**: A faster, lightweight version of Claude designed for speed in simple conversational tasks.
- **Strengths**: Extremely quick and responsive, great for real-time conversations or task completion.
- **Weaknesses**: Not as good at handling complex instructions or generating nuanced responses.
- **Who Should Use It**: Users looking for fast, efficient help with basic tasks like answering questions, quick writing, or chat interactions.

Geminia Flash 1.5

- **Definition**: A fast and efficient model designed for productivity and basic analytics.
- **Strengths**: Great for handling business-related tasks like simple data analysis, summaries, and basic decision-making processes.
- **Weaknesses**: Lacks the deep analytical capabilities of more advanced models when dealing with large datasets or complex scenarios.
- **Who Should Use It**: Professionals and business users who need quick, reliable results in everyday tasks like summarizing reports, handling data, or making simple business decisions..

Gemini Pro 1.5

- **Definition**: An upgraded version of Gemini Pro with enhanced processing power and ability to handle larger datasets.
- **Strengths**: Superior performance for long-form content generation, complex decision-making, and deep data analysis.
- **Weaknesses**: Can be slower in processing due to its heavy computational load, especially in simpler tasks.
- **Who Should Use It**: Ideal for users in business or academia who need comprehensive, data-driven reports, technical documentation, or decision support tools.

Gemini 2 Flash

- **Definition:** A fast, responsive model built for productivity, communication, and rapid task completion.
- **Strengths:** Great at summarizing, answering questions, and light data tasks at high speed.

- **Weaknesses:** Not designed for deep analysis or creative content generation.
- **Who Should Use It:** Ideal for professionals needing reliable and fast AI assistance in day-to-day workflows.

Gemini 2 Flash Lite

- **Definition:** A lightweight version of Gemini 2 Flash optimized for instant responses in low-resource environments.
- **Strengths:** Ultra-fast, low-latency outputs for simple tasks and embedded applications.
- **Weaknesses:** Limited depth; not suitable for complex prompts or nuanced conversations.
- **Who Should Use It:** Great for chatbots, mobile assistants, or time-sensitive responses.

Gemini 2 Flash Pro

- **Definition:** A premium version of Gemini 2 Flash with enhanced reasoning and data capabilities.
- **Strengths:** Balances speed with deeper task understanding, suitable for business intelligence and advanced communication.
- **Weaknesses:** Slightly slower than Lite, and less capable than full-scale models in massive data tasks.
- **Who Should Use It:** Business users and teams needing quick but intelligent AI assistance for decision-making and reporting.

GPT-3.5 (16K)

- **Definition**: GPT-3.5 (16K) is an upgraded version of GPT-3 with a higher memory capacity for handling longer conversations or more complex tasks.
- **Strengths**: Excellent for tasks that require context retention over extended conversations, such as customer support or detailed writing.
- **Weaknesses**: Not as advanced as GPT-4 in terms of reasoning, creativity, and general language understanding.
- **Who Should Use It**: Ideal for users who need assistance with slightly longer tasks, such as drafting emails, writing reports, or engaging in multi-turn conversations.

GPT-4

- **Definition**: GPT-4 is one of the most advanced language models developed by OpenAI, known for its superior reasoning and language understanding.
- **Strengths**: Great for almost any task, from creative writing to complex data analysis and programming help.
- **Weaknesses**: Can be slower compared to lighter models like GPT-4 (Turbo) and might require more computational power.
- **Who Should Use It**: Professionals, creatives, and anyone who needs high-quality content, coding assistance, or advanced language capabilities.

GPT-4 (32K)

- **Definition**: A version of GPT-4 with expanded memory, capable of handling very large documents or complex conversations over extended periods.
- **Strengths**: Ideal for processing long documents, in-depth reports, and detailed conversations without losing context.
- **Weaknesses**: Slower and more resource-intensive compared to standard versions of GPT-4, making it less suited for casual tasks.
- **Who Should Use It**: Users who work with long-form content, legal documents, or complex multi-turn conversations where memory is critical.

GPT-4 (Turbo)

- **Definition**: A faster and more optimized version of GPT-4, designed to deliver quicker responses while maintaining the model's robust capabilities.
- **Strengths**: Combines the advanced reasoning of GPT-4 with much faster response times, making it great for real-time use cases.
- **Weaknesses**: May not handle extremely large or complex tasks as well as the full GPT-4.
- **Who Should Use It**: Users who need quick, high-quality responses in real-time, such as customer service, live support, or on-the-fly content generation.

GPT-4o

- **Definition**: GPT-4o is a streamlined variant of GPT-4, optimized for everyday tasks that don’t require the full computational power of GPT-4.
- **Strengths**: Ideal for daily use, balancing performance and resource efficiency for users who want good results without overloading their systems.
- **Weaknesses**: Less effective than GPT-4 for highly complex tasks or in-depth analysis.
- **Who Should Use It**: Perfect for users who need a reliable AI for day-to-day work, like drafting emails, generating summaries, or answering questions.

GPT-4o Mini

- **Definition**: A more compact and lightweight version of GPT-4o, focused on speed and efficiency for quick tasks.
- **Strengths**: Very fast and responsive, ideal for shorter tasks and instant feedback.
- **Weaknesses**: Limited when it comes to handling complex or detailed tasks compared to larger models.
- **Who Should Use It**: Users looking for a lightweight model to quickly complete simple tasks such as answering questions or generating short content.

Llama3 40 5b

- **Definition**: Llama3 40 5b is a model developed by Meta, optimized for large-scale, data-driven tasks and deep learning processes.
- **Strengths**: Great for handling complex datasets, research, and analytical tasks with high accuracy.

- **Weaknesses**: Requires more technical expertise to use effectively; not ideal for casual or everyday tasks.
- **Who Should Use It**: Researchers, data scientists, and professionals who need a robust AI for data analysis, machine learning, or advanced research tasks.

Llama3 70b

- **Definition**: The most powerful variant in the Llama3 series, designed for deep learning tasks with large datasets and heavy computational needs.
- **Strengths**: Unmatched in terms of processing power for large-scale machine learning tasks, excellent for scientific research or complex decision-making.
- **Weaknesses**: Overkill for everyday tasks or small-scale use cases; requires significant computational resources.
- **Who Should Use It**: Ideal for experts in AI, machine learning, and data analysis who need to run large models or handle massive datasets.

Llama3 40 8b

- **Definition**: A middle-ground model in the Llama3 series, balancing performance and resource efficiency.
- **Strengths**: Provides strong performance for mid-tier tasks without requiring as many resources as the Llama3 70b.
- **Weaknesses**: Still not as user-friendly as smaller models for non-technical tasks.
- **Who Should Use It**: Users with intermediate-level needs in AI, machine learning, or data processing who want a balance between performance and resource consumption.

*Mistral 8 7b**

- **Definition**: Mistral 8* 7b is designed for tasks that require efficient computation while handling moderately complex data.
- **Strengths**: Offers a good balance between speed and computational efficiency for tasks like document processing and data analysis.
- **Weaknesses**: Not as powerful as larger models for handling highly complex datasets or in-depth decision-making processes.
- **Who Should Use It**: Business users or analysts looking for efficient data processing without needing extensive resources.

Mistral Large

- **Definition**: A larger variant of the Mistral series, designed for more demanding tasks and greater scalability.
- **Strengths**: Great for handling more complex, larger-scale tasks, such as advanced data analytics and large-scale automation.
- **Weaknesses**: Requires more resources and is slower compared to lighter models for simple tasks.
- **Who Should Use It**: Ideal for businesses or teams that need a robust AI for advanced data processing and decision-making workflows.

Nova Pro

- **Definition:** Nova Pro is a high-performance AI model designed to handle advanced tasks that require complex computations, scalability, and precision. It is ideal for use cases where accuracy and depth of analysis are critical.
- **Strengths:** Nova Pro offers superior computational power, making it capable of processing large-scale datasets and tackling highly complex tasks. Its exceptional accuracy and performance make it ideal for advanced applications, such as predictive modeling and generative AI.
- **Weaknesses:** Despite its strengths, Nova Pro has higher resource requirements, including significant memory and processing power. It may also have longer processing times compared to lighter models and can be more expensive to deploy and maintain.
- **Who Should Use It:** Nova Pro is best suited for data scientists, AI researchers, and enterprises working on mission-critical projects. Teams handling large datasets or requiring advanced machine learning capabilities will benefit the most from this model.

Nova Lite

- **Definition:** Nova Lite is a versatile AI model that balances performance and resource efficiency. It is tailored for mid-level tasks that require reliable yet moderately complex AI solutions.
- **Strengths:** This model offers an excellent trade-off between accuracy and resource consumption. It delivers faster response times compared to Nova Pro while being cost-effective and scalable, making it suitable for medium-complexity tasks.
- **Weaknesses:** While Nova Lite performs well for many tasks, it is not designed for extremely complex or large-scale projects. Its accuracy and capabilities are lower than those of Nova Pro, limiting its effectiveness in highly specialized applications.
- **Who Should Use It:** Nova Lite is ideal for mid-sized organizations seeking reliable AI solutions with manageable overhead. Professionals working on tasks like data classification or recommendation systems will find this model effective. It is also suitable for teams that prioritize a balance between speed and accuracy.

Nova Micro

- **Definition:** Nova Micro is a lightweight AI model optimized for quick, low-resource tasks. It is perfect for environments where efficiency and speed are more important than advanced computational capabilities.
- **Strengths:** Nova Micro's minimal computational requirements make it ideal for resource-constrained environments. Its high speed and efficiency allow it to handle simple, repetitive tasks effectively, and it integrates seamlessly into lightweight applications or devices.
- **Weaknesses:** Due to its lightweight design, Nova Micro has limited capabilities for handling complex or large datasets. Its accuracy and functionality are reduced compared to Nova Pro and Nova Lite, making it unsuitable for more demanding use cases.
- **Who Should Use It:** Nova Micro is best for startups, small teams, or applications that require fast, real-time responses. It is particularly useful for chatbots, simple data processing tasks, or use cases with lower accuracy requirements.

O1 :

For deep reasoning and deeper understanding

- **Definition**: O1 is a versatile AI model designed for a wide range of tasks, from simple conversational AI to more advanced content generation.
- **Strengths**: Flexible, able to handle both basic tasks and moderately complex content creation.
- **Weaknesses**: May not handle highly specialized or deeply technical tasks as efficiently as more focused models.
- **Who Should Use It**: Perfect for users who need an all-rounder that can handle a variety of tasks with ease, without requiring extensive expertise.

O1 Mini

For deep reasoning and deeper understanding and also for more technical tasks

- **Definition**: A lighter version of O1, optimized for faster, simpler tasks where quick output is prioritized over complexity.
- **Strengths**: Quick and easy to use for simple tasks like generating short content or answering queries.
- **Weaknesses**: Limited in terms of depth and complexity; not suitable for tasks requiring high levels of reasoning or data processing.
- **Who Should Use It**: Users who need fast, straightforward assistance with light tasks like email drafting or casual writing.

O3 Mini

- **Definition:** A compact, efficiency-focused model designed for streamlined conversational AI and task support.
- **Strengths:** Responsive, reliable, and well-suited for quick chats and light productivity tasks.
- **Weaknesses:** Lacks the sophistication needed for detailed, analytical, or creative projects.
- **Who Should Use It:** Best for users seeking a smooth, fast assistant for daily tasks like writing help, summaries, or support interactions.

Short Summary

1. Reasoning-Focused Models

- **Claude 3 Haiku, Claude 3 Opus, Claude 3 Sonnet, Claude 3.5 Sonnet, GPT-4**
  These models are generally well-suited for tasks involving natural language understanding, critical reasoning, and complex context analysis. They're adept at generating coherent, contextually relevant text, often used in reasoning-heavy tasks like summarization, dialogue, and customer support.

2. Math-Focused Models

- **GPT-4 (32K), Gemini Pro, Llama3 70b**
  These models perform well in mathematical reasoning and complex calculations due to their larger token limit and advanced training on mathematical datasets. They're suitable for applications requiring high accuracy in mathematical problem-solving, such as finance or quantitative analysis.

3. Coding-Focused Models

- **GPT-4 (Turbo), Claude Instant, Command, Command Light**
  These models are optimized for tasks that involve coding and software development, with strengths in understanding programming languages, debugging, and generating code snippets. They're used for code generation, code completion, and other software development tasks.

4. General Purpose / High Similarity Across Domains

- **GPT-3.5 (16K), GPT-4o, GPT-4o Mini, Titan Express (8K), Titan Lite (4K)**
  These models are versatile and can perform reasonably well across reasoning, math, and coding. They provide a balanced performance across tasks, making them suitable for general-purpose applications where a broad range of abilities is needed.

Creative Writing

- **Claude 3 Opus, GPT-4, GPT-4 (32K)**
  These models are strong in generating expressive and engaging content, ideal for tasks requiring a creative touch like storytelling, article writing, or creating persuasive text.

2. Mathematics

- **GPT-4 (32K), Gemini Pro, Llama3 70b**
  These models excel at handling complex math tasks, offering accuracy for scientific, quantitative, and problem-solving applications.

3. Summarization

- **Claude 3 Haiku, GPT-4 Turbo, Titan Express (8K)**
  These models are efficient and precise in summarizing long texts, making them suitable for creating concise overviews of reports, articles, or documents.

4. Cost-Efficiency

- **GPT-3.5 (16K), Titan Lite (4K), Llama3 40 5b,** Mistral 3B
  These models offer solid performance at a lower cost, making them ideal for high-usage applications that require a balance of quality and affordability.

5. Quality of Response

- **Claude 3.5 Sonnet, GPT-4 (32K), Gemini Pro 1.5**
  These models deliver highly accurate, contextually aware responses, suitable for tasks where quality and reliability of response are critical, such as customer service and knowledge-based queries.

6. Reasoning

- **Claude 3 Sonnet, GPT-4, Llama3 70b**
  Known for strong logical and contextual reasoning, these models are well-suited for complex problem-solving and understanding nuanced, layered queries.

7. Instruction Following

- **GPT-4 Turbo, Claude Instant, Command Light**
  These models excel at following instructions precisely, making them ideal for technical support, step-by-step processes, and other guided procedural tasks.

8. Mathematical Ability

- **GPT-4 (32K), Claude 3 Haiku, Llama3 70b**
  With high mathematical accuracy, these models are effective for academic or technical tasks involving rigorous math, from calculus to algebra.

9. Coding

- **GPT-4 Turbo, Claude Instant, Command, Command Light**
  These models are optimized for coding-related tasks, including code generation, debugging, and providing step-by-step code instructions, making them highly suitable for software development, code assistance, and programming education.

Deep Reasoning : O1 Models

Multilingual: Gemini and Llama and GPT models are multilingual