

National Aeronautics and
Space Administration

AI at NASA Ames Supercomputing

Shubha Ranjan

NASA Advanced Supercomputing

Ames Research Center

April 27, 2025

Hugging Face

Pretrained Open-source models are available to access at NAS. These models are the basis for foundation models, allowing for faster development and testing of ML community use cases.

Purpose	Model Availability
Masked Language Modelling	google-bert/bert-base-uncased google-bert/bert-base-cased FacebookAI/roberta-base microsoft/deberta-v3-base distilbert/distilbert-base-uncased distilbert/distilbert-base-cased
Translation	google-t5/t5-base
Text Generation	openai-community/gpt2 meta-llama/Meta-Llama-3-8B meta-llama/Meta-Llama-3-70B meta-llama/Meta-Llama-3-70B-Instruct google/gemma-7b
Speech Recognition	openai/whisper-large-v3

AI/LLM@NAS

- Grace Hopper systems 350 – Nvidia H100
- Ensure relevant device drivers/CUDA support
- Build custom and provide conda environments for all architectures on the floor
- Provide help porting code to the multi-node clusters
- Open-Source Hugging Face Models Available in air gapped environment
- Hugging Face is a platform for Machine Learning comprised of multiple Python packages that can help train, deploy, and host open-source and licensed machine learning machine learning models and datasets.
- 13 open-source machine learning models are available to use at NAS through the Data Analytics Conda environments with Hugging Face.
- Models for masked language modeling, translation, text generation, and speech recognition are available to use.
- The Conda environments provide access to users for large open-source models like Llama 3 and GPT-2 locally without users having to download the models to their home directory.
- Acts as a central repository for common open-source models
- Allows for on-prem training and inference with hosted models
- New models are added on user request
- Local Hugging Face repository allows for easy model sharing between users.
- Models can take advantage of multiple GPUs, allowing for inference to run on multiple GPUs and training to run across multiple GPU nodes.