DELL Technologies

WORKSHOP

Want to learn how to Unlock your Inner Coder with AI - Assisted Programming? Join our interactive workshop at World Summit AI

Oct 10 , 11:30 – 12:30

Ronan Carey, Senior Director, Business Development, Dell Technologies
Maciej Mazur, Advisory Systems Engineer, Dell Technologies
Neil Bowden, Account Executive , Business Development – Sales, Dell Technologies
Olya Kozlova, Senior AI solutions architect, NVIDIA

NVIDIA.

https://dell.to/3N6gO6n

What are we running on

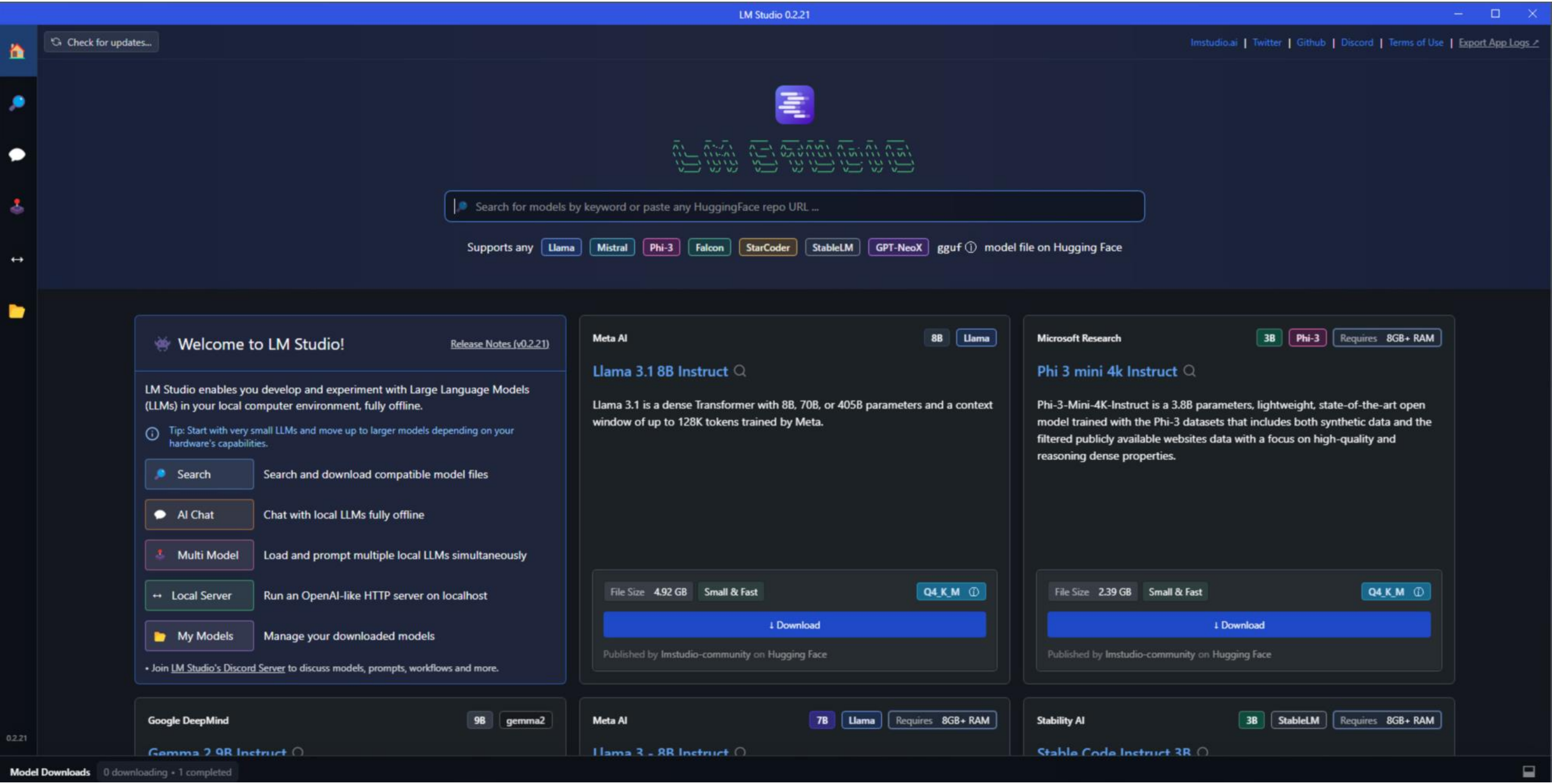
Windows

LM Studio

VS Code

DELL

NVIDIA



LM Studio 0.2.21
Check for updates...
lmstudio.ai | Twitter | Github | Discord | Terms of Use | Export App Logs ↗
0.2.21
Model Downloads
0 downloading • 1 completed

Search for models by keyword or paste any HuggingFace repo URL ...

Supports any Llama Mistral Phi-3 Falcon StarCoder StableLM GPT-NeoX gguf ⓘ model file on Hugging Face

Welcome to LM Studio!

Release Notes (v0.2.21)

LM Studio enables you develop and experiment with Large Language Models (LLMs) in your local computer environment, fully offline.

ⓘ Tip: Start with very small LLMs and move up to larger models depending on your hardware's capabilities.

Search	Search and download compatible model files
AI Chat	Chat with local LLMs fully offline
Multi Model	Load and prompt multiple local LLMs simultaneously
Local Server	Run an OpenAI-like HTTP server on localhost
My Models	Manage your downloaded models

• Join LM Studio's Discord Server to discuss models, prompts, workflows and more.

Meta AI — 8B — Llama

Llama 3.1 8B Instruct

Llama 3.1 is a dense Transformer with 8B, 70B, or 405B parameters and a context window of up to 128K tokens trained by Meta.

File Size 4.92 GB — Small & Fast — Q4_K_M ⓘ

↓ Download

Published by lmstudio-community on Hugging Face

Microsoft Research — 3B — Phi-3 — Requires 8GB+ RAM

Phi 3 mini 4k Instruct

Phi-3-Mini-4K-Instruct is a 3.8B parameters, lightweight, state-of-the-art open model trained with the Phi-3 datasets that includes both synthetic data and the filtered publicly available websites data with a focus on high-quality and reasoning dense properties.

File Size 2.39 GB — Small & Fast — Q4_K_M ⓘ

↓ Download

Published by lmstudio-community on Hugging Face

Google DeepMind — 9B — gemma2

Gemma 2 9B Instruct

Meta AI — 7B — Llama — Requires 8GB+ RAM

Llama 3 - 8B Instruct

Stability AI — 3B — StableLM — Requires 8GB+ RAM

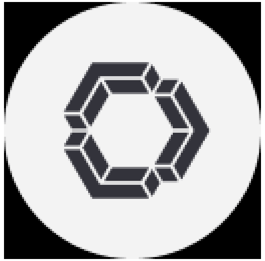
Stable Code Instruct 3B

Simple powershell

- how much free spece I have on disk
- open me a folder where drivers usually are on Windows
- change path to the Windows System32 drivers directory with drivers and then list all of them starting on letter „s” and having a number in the name

That’s easy but not super usefull, let’s download real programmer tools

vscode + python and continue extenstions



Basic web development

- index.html based website
- flask as web server
- generate a website with 2 tables with reports

Ok, now let’s create some data for those reports

jupyter

https://hackernoon.com/15-excel-datasets-for-data-analytics-beginners

How it works under the hood

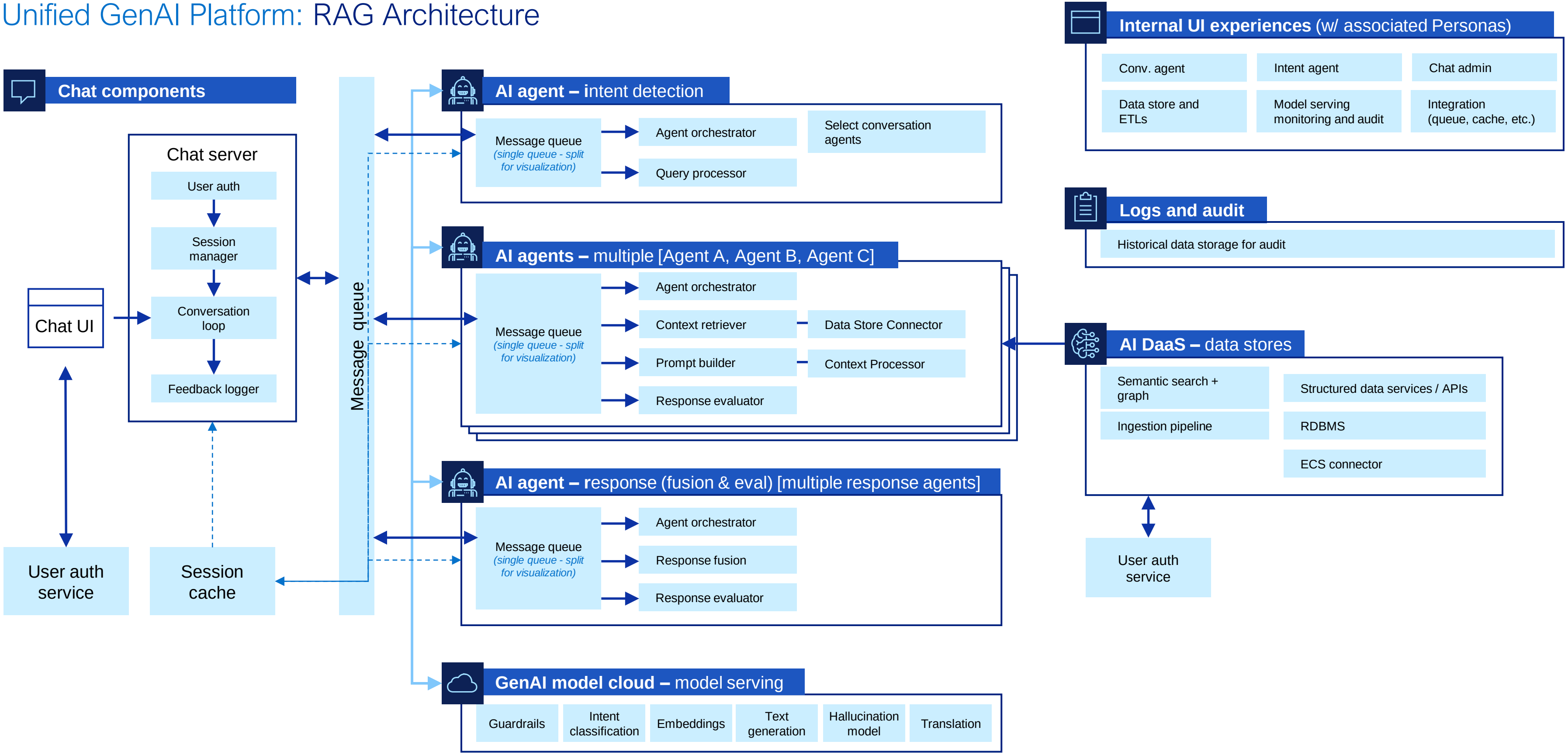
NVIDIA SLIDES HERE

Key take aways

- coding assistants can really work in multiple use-cases
- this is based on lot of input tokens and small amount of output tokens
- small models 7B-13B work best as latency is a key here
- proper use of caching is important (you opperate on the same code base all the time, no point in prompting with it all the time)
- proper prompt beats model and HW speed in terms of result
- use LLM’s also for learning , especially if you use Obsidian

Architecture details for RAG in general

Unified GenAI Platform: RAG Architecture



Chat components

Chat server
- User auth
- Session manager
- Conversation loop
- Feedback logger

Chat UI

User auth service

Session cache

Message queue

AI agent – intent detection
- Message queue *(single queue - split for visualization)*
- Agent orchestrator
- Query processor
- Select conversation agents

AI agents – multiple [Agent A, Agent B, Agent C]
- Message queue *(single queue - split for visualization)*
- Agent orchestrator
- Context retriever
- Data Store Connector
- Prompt builder
- Context Processor
- Response evaluator

AI agent – response (fusion & eval) [multiple response agents]
- Message queue *(single queue - split for visualization)*
- Agent orchestrator
- Response fusion
- Response evaluator

GenAI model cloud – model serving
- Guardrails
- Intent classification
- Embeddings
- Text generation
- Hallucination model
- Translation

Internal UI experiences (w/ associated Personas)
- Conv. agent
- Intent agent
- Chat admin
- Data store and ETLs
- Model serving monitoring and audit
- Integration (queue, cache, etc.)

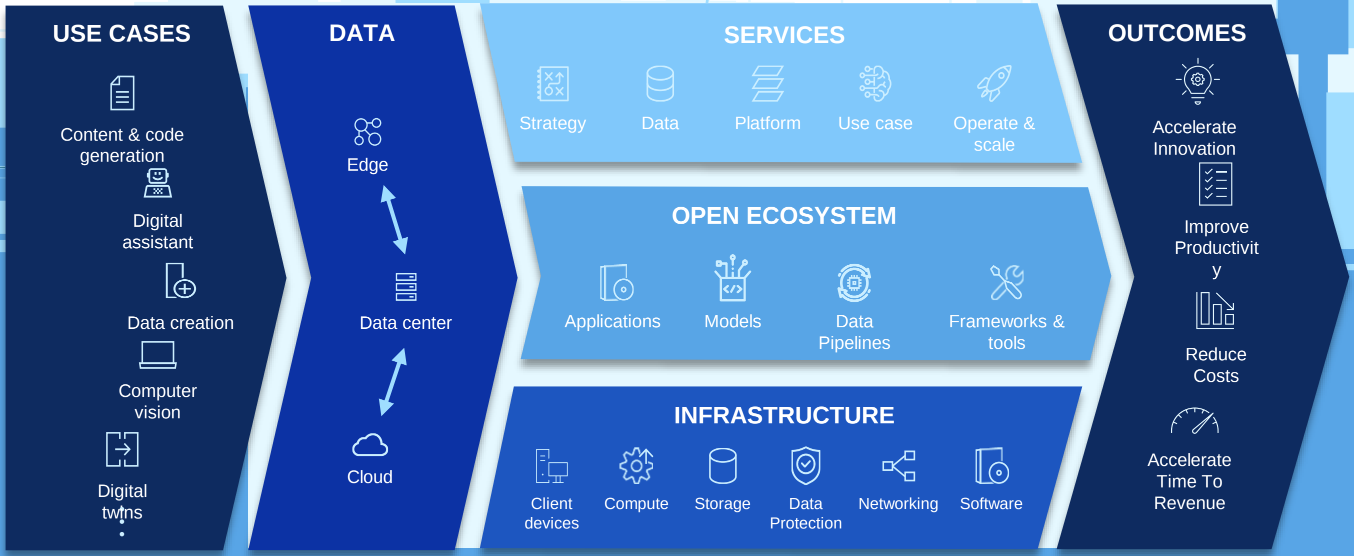
Logs and audit
- Historical data storage for audit

AI DaaS – data stores
- Semantic search + graph
- Ingestion pipeline
- Structured data services / APIs
- RDBMS
- ECS connector

User auth service

The Dell AI Factory

Dell’s approach to help customers embrace and implement AI



USE CASES

Content & code generation

Digital assistant

Data creation

Computer vision

Digital twins

DATA

Edge

Data center

Cloud

SERVICES

Strategy

Data

Platform

Use case

Operate & scale

OPEN ECOSYSTEM

Applications

Models

Data Pipelines

Frameworks & tools

INFRASTRUCTURE

Client devices

Compute

Storage

Data Protection

Networking

Software

OUTCOMES

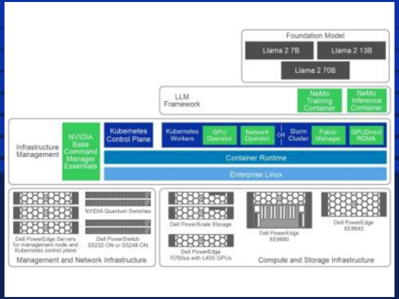
Accelerate Innovation

Improve Productivit y

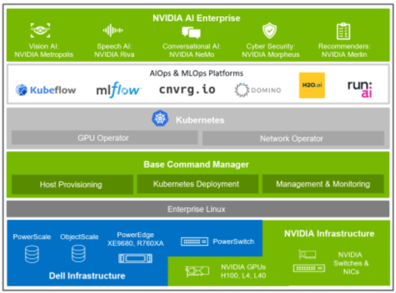
Reduce Costs

Accelerate Time To Revenue

Build Complex AI Systems with Ease

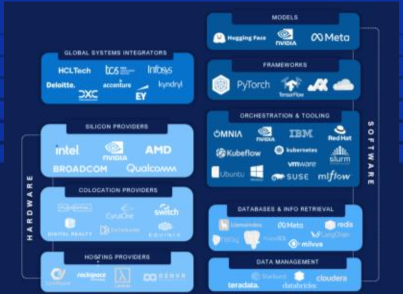


Access ready-made reference architectures and designs



Utilize Dell-validated, supported, and deployed hardware and software

Automate both deployment and day 2 operations

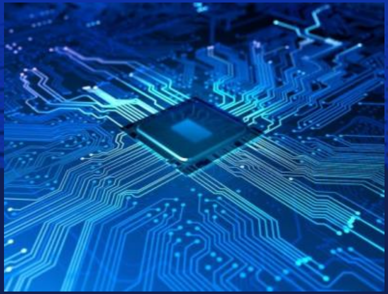


Leverage an ecosystem of partners led and vetted by Dell

Scale with Right-Sized AI Infrastructure

Increase GPU utilization with proper sharing and scheduling of resources

Leverage Dell experience in sizing for various industries and use-cases



Access heterogeneous accelerators from NVIDIA AMD intel

Ensure optimal performance by matching GPUs with appropriate storage, networking, and cooling

Accelerate AI Adoption and Deployment

Ensure the best availability and shortest delivery dates

Leverage our experience implementing AI projects internally at Dell

Deploy and manage with the support of Dell professional services in every geography

Benefit from worry-free legal, compliance, and financing processes