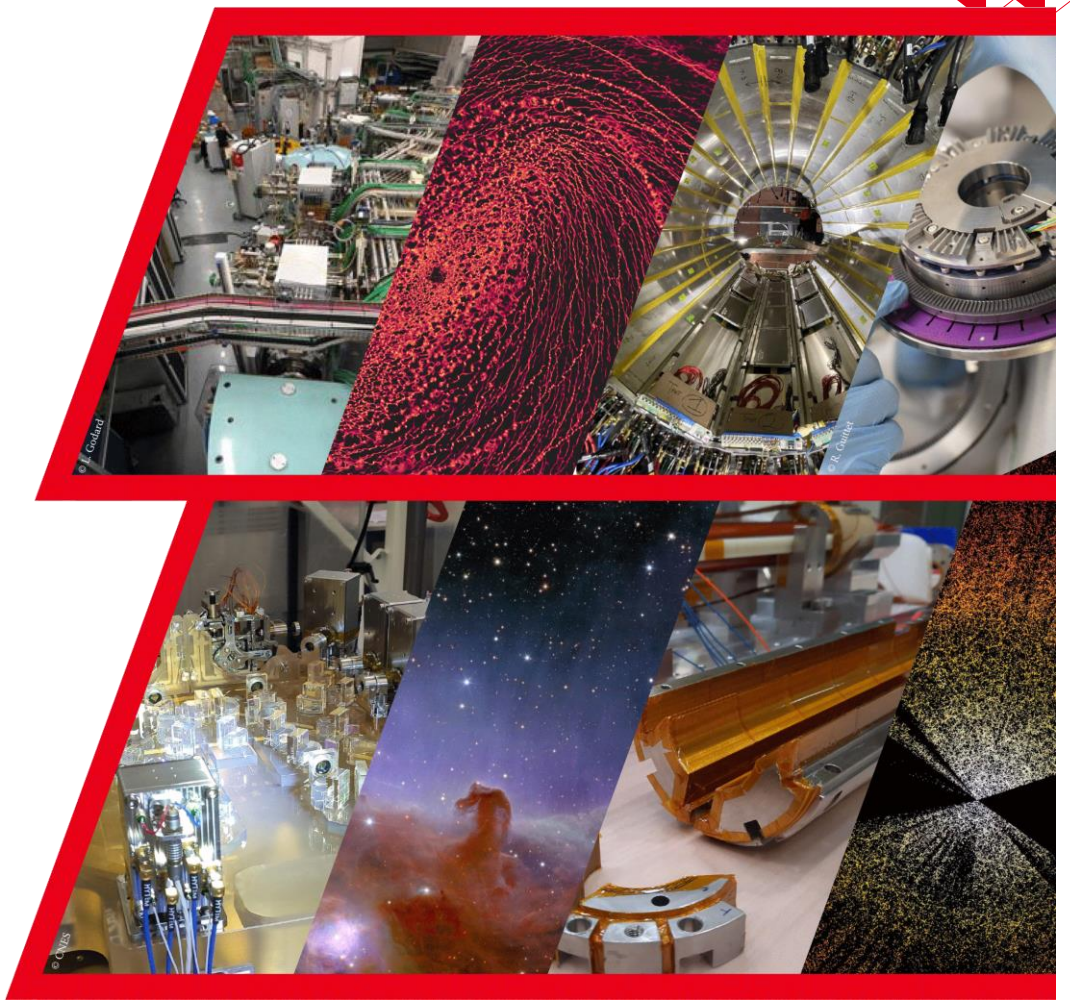
cea irfu

RAGLabs : Workshop MCP/n8n/NoCode

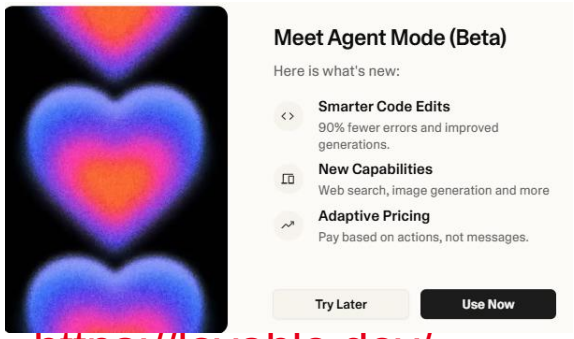


08/07/25

Imed MAGROUNE

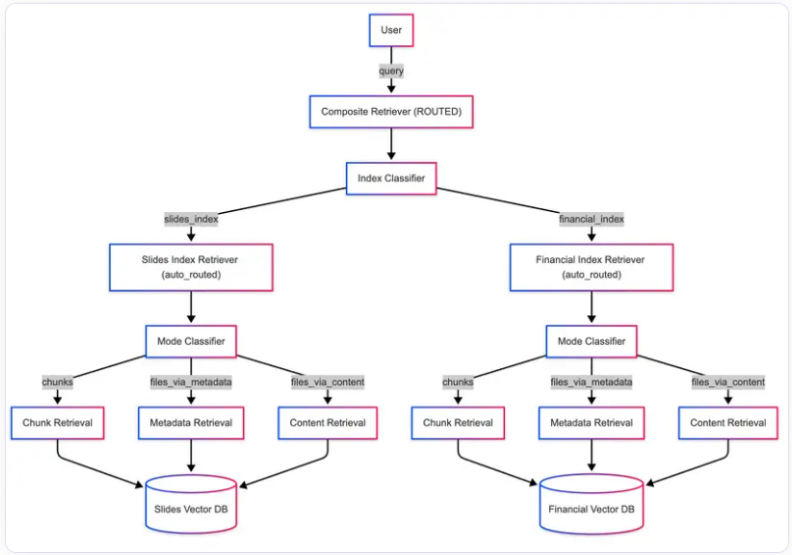
Plan

- **News / SOTA depuis la dernière réunion (5 min)**
  - Benchmarks : GAIA / Context Ingineering
  - Applications IA : notebooklm.google, gemini-cli
  - Models : gemma3n (text/audio/image/video), NoanLLM, qwen3 , qwen2.5VL,
  - Les ressources IA : Local, Web2, Web 3.
- MCP (Model Context Protoclol: (5min)
  - Pourquoi ça change tout.
  - Exemple serveur / Client from scratch.
  - Exemples open source.



- No Code / Low Code : (7 min)
  - Cline (plug'in sur VS-Code).
  - GEMINI-CLI.
  - PI : Claude-code, TRAE (Ali Baba) Cursor/WindSurf (rachetée par OpenAI?) , https://lovable.dev/ .
- N8n :
  - Solution Workflow open source AI Native
  - Agent AI.
  - Autres nodes / MCP.

Tendances



https://www.llamaindex.ai/blog/rag-is-dead-long-live-agentic-retrieval

Sourabh Desai • 2025-05-29

RAG is dead, long live agentic retrieval

https://docs.unsloth.ai/basics/gemma-3n-how-to-run-and-fine-tune

Google’s Gemma 3n multimodal model **handles image, audio, video, and text inputs**. Available in 2B and 4B sizes, it supports 140 languages for text and multimodal tasks. You can now run and fine-tune **Gemma-3n-E4B** and **E2B** locally using Unsloth.

Mesures

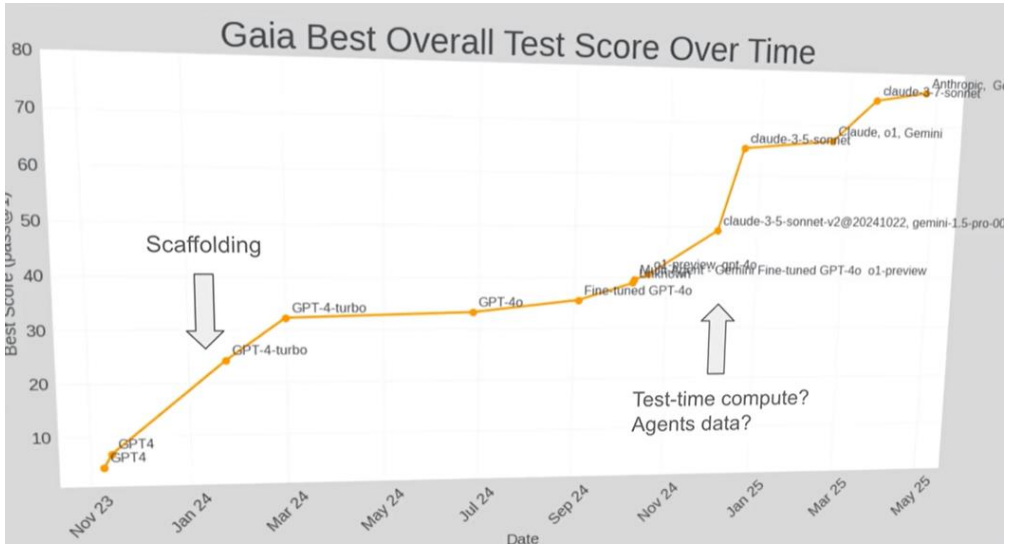
- **Level 1** questions generally require no tools, or at most one tool but no more than 5 steps.
- **Level 2** question generally involve more steps, roughly between 5 and 10 and combining different tools is needed.
- **Level 3** are questions for a near perfect general assistant, requiring to take arbitrarily long sequences of actions, use any number of tools, and access to the world in general.

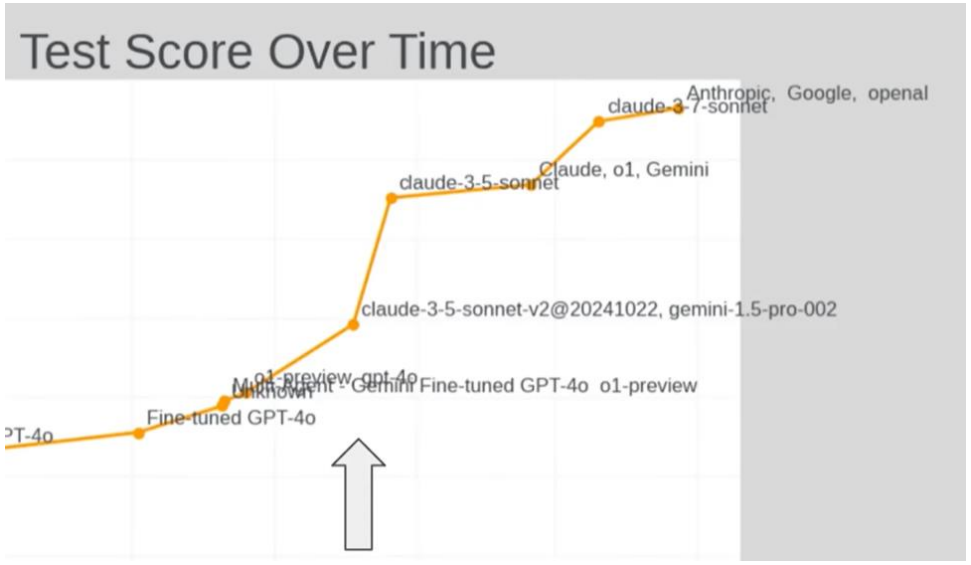
Level 3

Question: *In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute?* Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.

Ground truth: White; 5876

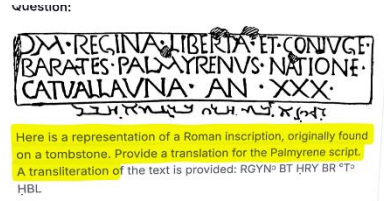
L'IA a discrètement franchi une étape historique



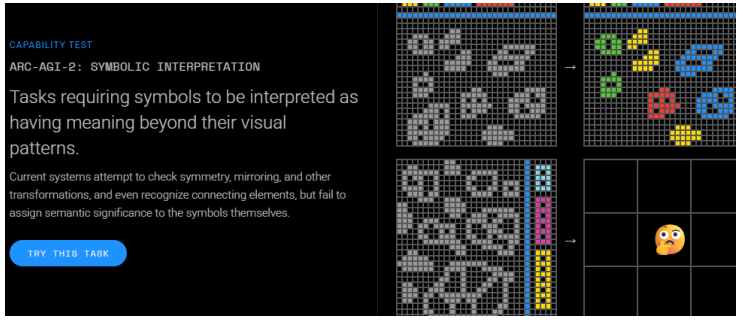


- Models de raisonnement.
- Réinforcement Learning
- Moins de « PINNs » dès que + données

- Avant : MMLU (Connaissance) , GPQA , HLE (Humanity's Last Exam : niveau thèse) , MATH,
- Saturation GAIA :
  - Niveau 1 : saturé,
  - Niveau 2 : Presque.
  - Niveau 3 : « que » 60%. Enore 6 mois ?
- Next Benchs : BrowseComp : (openAI) – lecture du monde - , DABStep –analyse de données – scientifique
- Mesure de la connaissance ➔ Mesure du raisonnement
- ➔Mesure de l'intelligence : ARC-AGI-2



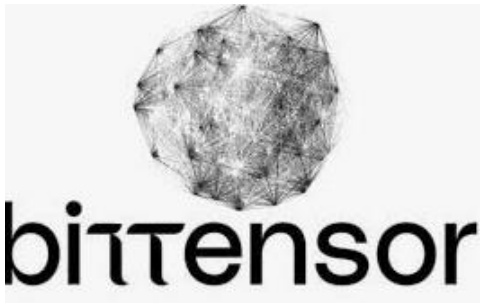
https://arcprize.org/



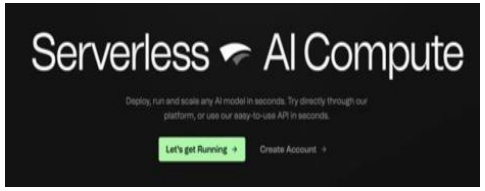
Ressources :

- Local:
  - Ollama (llama.cpp / gguf)
  - vLLM …
  - OpenWebui …

- Web 2 :
  - Grok.com, chatgpt.com, claude.ai, aistudio.google.com ,,,
  - APIs : https://api-docs.deepseek.com/, https://console.groq.com/keys …
  - ….. : https://mammouth.ai/ , https://openrouter.ai/



- Web 3 :
  - version décentralisée d'Internet, basée sur des technologies comme la blockchain, les cryptomonnaies, les NFTs (tokens non fongibles), les contrats intelligents (smart contracts) et les applications décentralisées (dApps)
  - Bittensor : Blockchain d'initation à la production des ressources IA
    - Mega Incubateur Décentralisé (21M TAO –comme bitcoin- 2MD$)
    - Subnets (projet/entreprise/labo) – 61 inférence, 56 training, 56 location GPU
  - ASI : Alliance : Fetch.ai ($FET), SingularityNET (AGIX) et Ocean Protocol (OCEAN)



https://asi1.ai/
https://api.asi1.ai/v1/chat/completions

ASI-1 mini: The World's First Web3 LLM, Designed for Agentic AI

Artificial Superintelligence Alliance

Contrairement au Web 2.0, où les données sont contrôlées par des grandes entreprises (Google, Meta, Amazon, etc.), le Web 3 vise à redonner le contrôle aux utilisateurs grâce à la décentralisation.

Rappel :

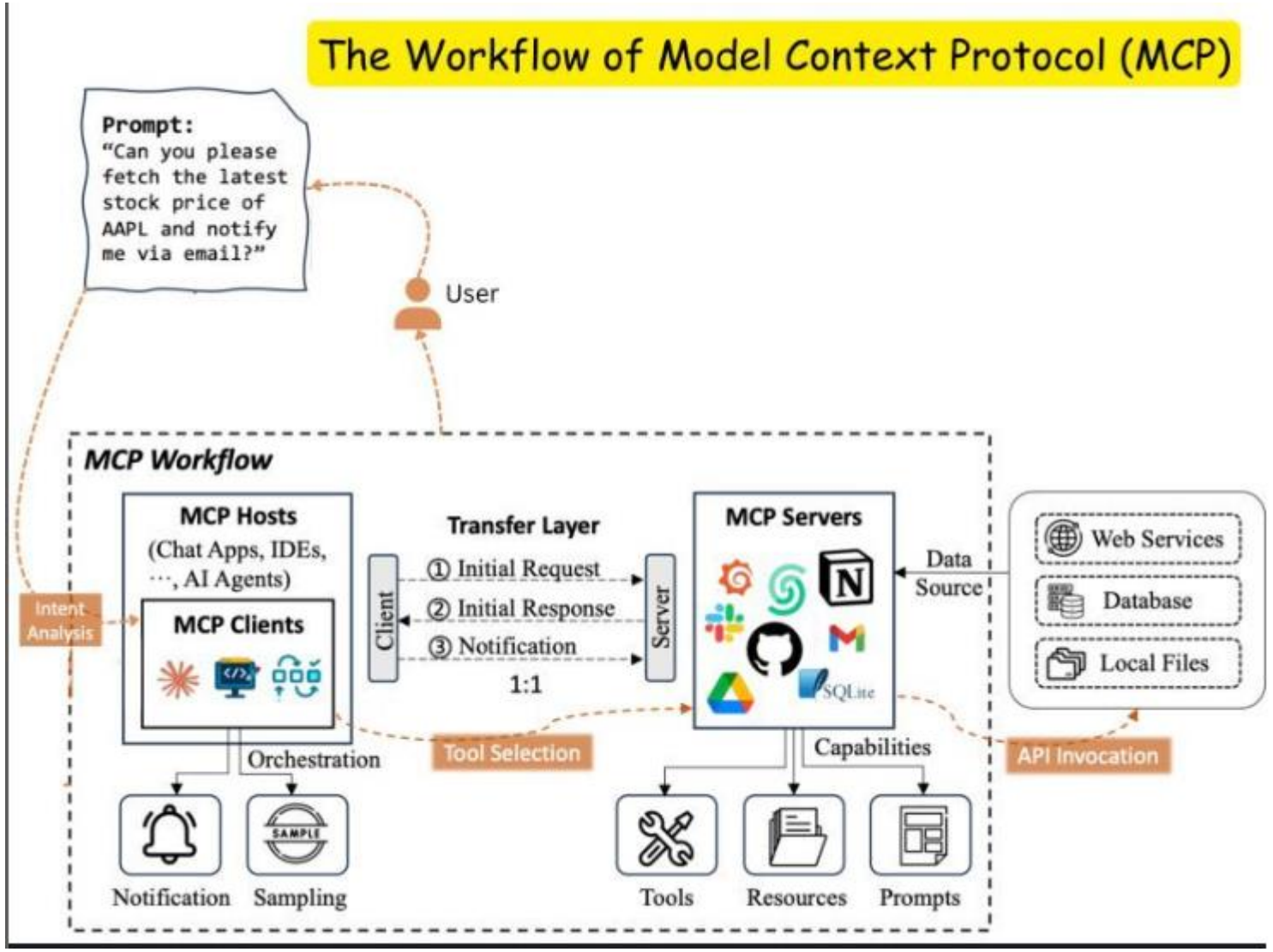

Écrire un prompt précis et spécifique
1. Rôle
2. Tâche
3. Objectif
4. Contexte
5. Étapes
6. Contraintes
7. Format
Rôle
Tâche
Objectif
Contexte
Tonalité
Contraintes
Format
Tu es un expert en marketing de contenu
tes clients sont tous dans le secteur {secteur}
Présente cinq idées distinctes pour augmenter
l'interaction avec les clients.
Le but est d'élaborer un plan de contenu qui
génère plus de trafic web et renforce la fidélité
des clients. Chaque idée de contenu doit être
claire et précise. Réponds sous forme d'un
tableau avec un maximum de 40 mots par idée.
Utilise un ton adapté au marketing B2B.

MCP :

1. MCP Host
The AI application itself. This could be a chatbot, a virtual assistant, or an autonomous agent environment. The MCP host manages the context and tasks while embedding the MCP client to handle tool interaction.

2. MCP Client
The communication bridge. It sits within the host and connects to one or more MCP servers. Its job is to:
- Discover available tools/functions
- Query server capabilities
- Route tasks based on intent

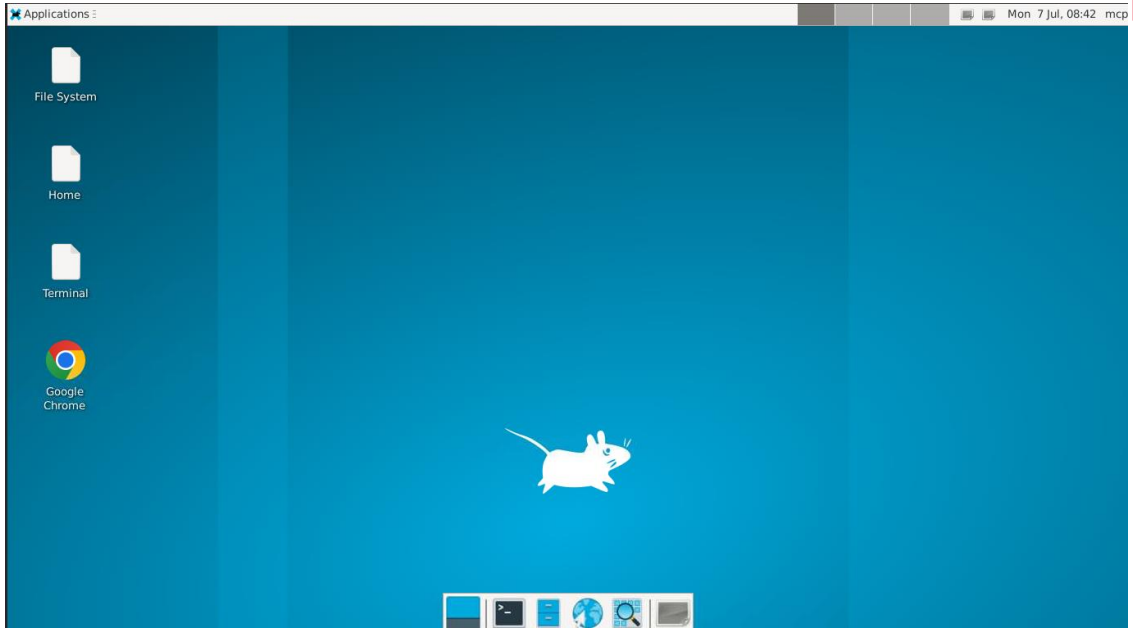
3. MCP Server
This is where the actual tools and operations live. The server exposes tools, resources, and prompts that the AI can invoke securely. It acts as the operational layer, enabling the AI to fetch data, trigger APIs, or query databases in real-time.



MCP Sandbox :

Docker ubunto + vnc + gogole-chrome + terminal + python (fastapi..) :

- Git clone https://github.com/Imag2020/raglabs
- cd raglabs/mcp-docker/
- docker build -t mcp-desktop .
- docker run --name mcp -e VNC_PASSWORD=secure123 -p 5901:5901 -p 6901:6901 -p 8001:8001 -v $(pwd)/mcp-data:/home/mcp/data mcp-desktop



mcp's X desktop (05c124d365b7:1) – noVNC
http://localhost:6901/vnc.html

vncviewer localhost:5901