

Artificial Intelligence

Recommendations for Artificial Intelligence

Develop distributed agentic AI (specialized action models)

The development of specialized action models (SAMs) acting as service is important and can be developed in Europe. These SAMs, small and specialized models that can interact with their environment, should operate in a distributed infrastructure and an ecosystem should be created to support research, development and business around them. These models need to be refined, optimized, and reduced in size to improve efficiency. These SAMs can be optimized from more general foundation models by an ecosystem of companies providing their optimized SAMs in a marketplace so that they can be dynamically discovered and used by the orchestrators.

Develop orchestrating technologies for distributed agentic AI, blueprint for NCP orchestrators

We call agentic AI a set of specialized AI agents working together to accomplish a common goal. An AI agent is synonymous with an SAM in this discussion: an AI that can perceive and act, having impact on the virtual or real world. **The orchestration technologies** should take into account all the requirements, that can select the best SAMs for the required tasks and dynamically activate them. The first steps could be very agentic-AI-centric (relying on already existing technologies used for orchestrating AI agents), but they should be blueprint and evolve towards an orchestration system for the NCP. These orchestrators must be developed for the edge – or near the final user – and dynamically combine SAMs into executing personalized applications in response to user needs.

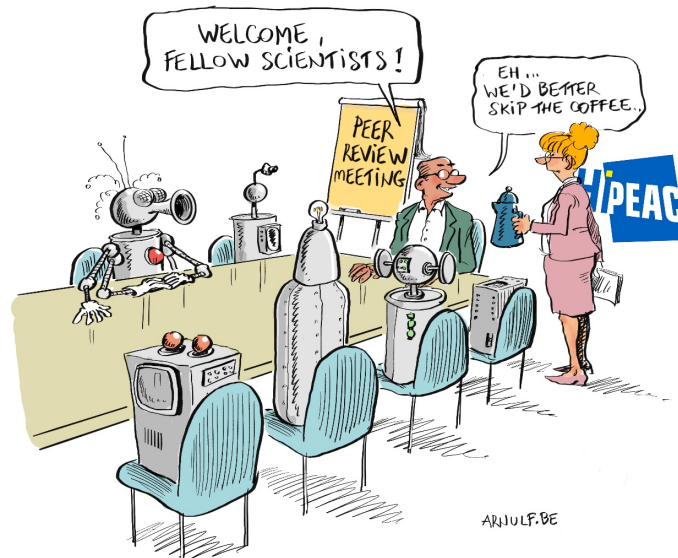
Establish open protocols for these “distributed agentic AI” systems to facilitate seamless interaction among distributed AIs from different origins

Protocols and specifications that group all requirements, existing ideas and proposals together in a single consortium to develop an open source “de facto” (before official standardization) standard protocol that takes into account all the good ideas of various researchers and organizations, so that it will be sound, future-proof, recognized and accepted. The requirements are:

1. It does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).
1. It also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, environmental impact, cost, localization, privacy of data, etc.).

The recommendation to develop generative AI at the edge (AI) is still important, but it is more in development and implementation mode now (for example, in Apple intelligence). We

should continue developing solutions that allow embedding generative AI at the edge in order that **human users can be provided with natural interfaces** (voice, gesture, eye movements, touch) to the digital world, with more energy efficiency, reduced latency, lesser communication overhead, and greater privacy. This is important to reduce the difficulties to access the digital world and decrease digital illiteracy.



Introduction: what happened in 2024 in terms of the use of AI?

In 2024, the field of artificial intelligence (AI) saw remarkable advancements, particularly in large language models (LLMs) and their applications, and it is difficult to keep up with the pace of announcements. The progress is so fast that the illustrations and examples of this text will already be outdated when you read it.

The HiPEAC community is not specialized in developing new AI algorithms or new AI applications, but it is deeply involved with artificial intelligence on two sides:

- Leveraging AI for HiPEAC developments, hardware, and software.
- Developing new hardware and software to better serve AI needs.

Using AI to help HiPEAC community to develop better hardware and software is covered in the "Tools" chapter of this document. However, originally prototypes with limited use, LLMs underwent significant development starting in 2022, evolving into viable tools by the end of that year. Models like o1 from OpenAI are now able to generate moderately complex code of several pages. Perhaps we are already entering into what Jensen Huang, the CEO of NVIDIA, calls a future where content will be generated, not retrieved [HuangHPCwire]. It is now faster to ask o1 to generate programs that play the Game of Life, Asteroid, or Flappy Bird than to look for a version that works on your computer - and the generated version is likely to be virus-free.

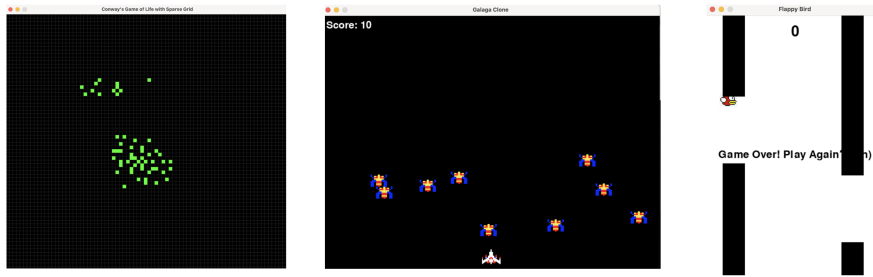
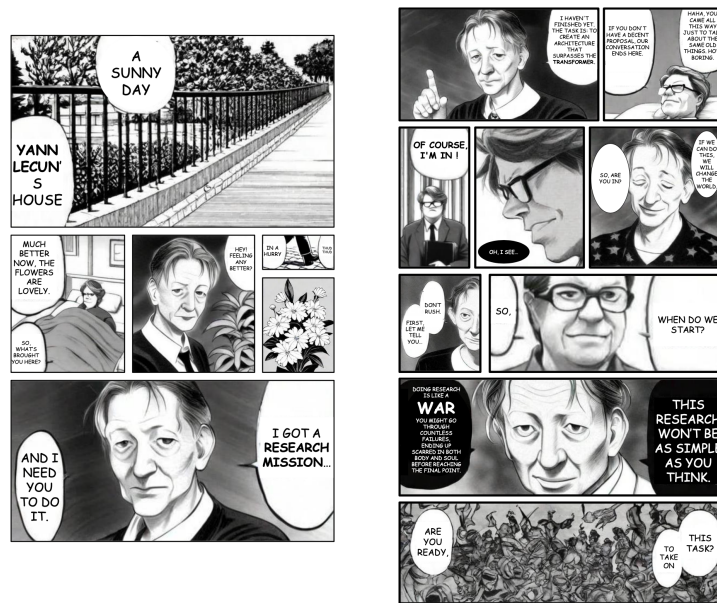


Figure 1: Game of Life, a clone of Asteroid and a Clone of Flappy bird generated in few minutes by OpenAI o1. A few iterations were necessary, but absolutely no reading or understanding the generated code required.

Beyond text (and code), pictures, music (Udio, Suno can generate songs), manga [NobelManga][NobelMangaPDF], and video generation all saw major improvements in 2024, reaching a point where coherent and realistic videos of several seconds can now be created. Among the leaders in this field, Google's VO2 currently stands out as the most advanced, followed by OpenAI's Sora. Unlike large language models (LLMs), where open-source models often rival their closed-source counterparts, video generation remains exclusively dominated by closed models for now.



Manga generated by AI [NobelManga]

Most of the major (big) LLMs are now covering multiple modalities (text, but also image, sound) and they can directly have one modality as input and another as output, without relying on intermediate models that transform one modality into another. One example is the advanced voice mode of OpenAI which uses sound (voice) as input and directly generates audio output, therefore with a reduced latency compared to the previous approach which used at least three models (a voice to text model, then the LLM text to text, then a text to voice model). Transformer-based systems seem to be now the "Swiss Army knife" of AI, because they are also efficient for perception tasks like image recognition, sound analysis etc, making them suitable for devices directly interacting with the real world, like robots, self-driving cars, ...

Another development this year was the progress in “world models”. These models do not simulate language but instead recreate entire games or physical environments. A notable example is Oasis, an AI-based simulation of Minecraft that replicates the game’s physics, block structures, and movement dynamics. Researchers also developed DIAMOND, a model capable of simulating a basic version of Counter-Strike: Global Offensive, but still with limited fidelity and running at 10 frames per second on a single NVIDIA RTX 3090 graphics card. Meanwhile, DeepMind’s Genie and Genie 2 demonstrated the ability to simulate not just one but multiple games, showcasing the potential for virtual environments.

In gaming, the possibilities for these world models are virtually limitless. Imagine games where every aspect—characters, scenarios, textures, and interactions—is dynamically created by neural networks. This level of generative power could revolutionize game design, enabling entirely AI-driven worlds free from traditional constraints. While still in its early stages, this field promises to redefine how we interact with both AI and digital environments. Its potential extends far beyond leisure. In the future, such simulators could play a crucial role in training AI systems in controlled environments. For instance, a road simulator is already being employed to train autonomous vehicles safely. Jensen Huang’s vision is coming to reality...

Similar progress can be observed in AI tools to help hardware designers: AlphaChip mirrors the principles of AlphaZero, the algorithm used in strategy games, but applies them to the design of computer chips. Developed in 2020, AlphaChip made headlines again in 2024 with its use in designing Google’s tensor processing units (TPUs). This approach showcases a remarkable loop of optimization: an AI system designs a chip, which is then used to train the same AI, enabling it to create even better chips in subsequent iterations. This self-reinforcing cycle highlights how AI can accelerate technological progress in unprecedented ways.

You can refer to the section on tools for more in-depth text about the use of AI to increase productivity of the HiPEAC community.

What improvements in AI took place in 2024?

To better understand the key recommendations for developing optimized hardware and software to serve the requirements of AI, it is necessary to make a brief explanation of what happened in 2024 for the development of AI technology and extrapolate (if possible) the next steps.

The improvements of artificial intelligence in 2024 were not primarily driven by increasing the size of these models, but by refining the quality of training data, techniques like fine-tuning, and using more compute time during inference. However, economic viability is still an open question, leading to an increase in the cost of subscriptions (for OpenAI) and perhaps limited access to the most powerful models, that will only be used to answer very specific questions that could compensate for the cost of running the model. Will we perhaps see the beginning of AI at multiple speeds: “basic” low-cost AI accessible to everyone, higher-performing AI on subscription for those who can afford it, and countries or big companies that are the only ones that can afford to access the best models and to ask them complex (therefore expensive in term of compute power and therefore cost) questions?

A focus area for improving LLMs involves the quality and quantity of training data. One practice in 2024 was the use of synthetic text generated by pre-trained models, offering a rich and scalable source of high-quality data. This shift has enabled the creation of smaller models without compromising performance, driving down costs dramatically. For example, while GPT-3.5 contained 175 billion parameters, Google’s Gemma 2 models now deliver similar performance with 9B parameters, representing nearly a 20-fold reduction in size. Meta’s Llama 3.3 (70B) of December 2024 has the same performance as Llama 3.1 (405B) of July 2024. This efficiency, combined with hardware optimizations, has led to a tenfold annual decrease in operational costs since 2022. There is a clear economic incentive to use

smaller LLMs that have similar performance as bigger ones, because the inference cost (computation) is lower. Some observers saw some decrease in performance between versions of ChatGPT 4o, perhaps due to a switch to a smaller LLM. It is also possible that the “big” LLMs will not be accessible to the public, but only internally used by the companies to train smaller, more economically viable LLMs.

In addition, advancements in LLMs included an increase in the contextual scope, with models now capable of handling up to 128,000 tokens per input. Google even expanded this limit to two million tokens, enabling the processing of extensive collections of documents in one go. At the core of current LLMs lies the transformer architecture, which, while powerful, suffers from a growing memory requirement as it processes longer texts. Newer architectures like Mamba, with constant memory usage, offer a promising alternative by enabling faster processing of extended word sequences. In 2024, it became clear that completely replacing transformers is not feasible yet. However, hybrid approaches that integrate Mamba with transformers are showing potential, maintaining high performance while reducing memory overhead.

A novel approach emerged in 2024, in which LLMs spend more computation time during inference to improve output quality. Traditional models produce responses with a fixed computation effort regardless of task complexity, but newer models like OpenAI’s o3 dynamically adjust their computation time. This process allows for more thoughtful and accurate responses, as these models essentially “reflect” or “research” internally before presenting their output. It seems that this also required an increase of the contextual scope referred to in the previous paragraph.

OpenAI, leveraging its reinforcement learning expertise, led the way in this approach, with Google and other competitors following with models like Gemini 2 Flash Thinking. This new approach is similar to the breakthrough in 2016 when AlphaGo transformed the landscape of the game Go. Initially trained to imitate the moves of expert human players, AlphaGo was limited by the quality and scope of its training data. However, when allowed to play Go independently without human intervention, it began to learn through trial and error, guided only by rewards within the game. This self-directed learning led AlphaGo to outperform human champions, fundamentally changing how the game was played.

Previously trained to replicate human-written text, these new models, like o1 or o3, have now begun to autonomously refine their reasoning abilities. Google’s Deep Research feature complex problem-solving by enabling LLMs to analyse data from over 50 online sources, including PDFs, in mere seconds, to provide comprehensive summaries.

By exploring, experimenting, and searching for solutions independently, LLMs are no longer constrained to mimicking human data. Instead, they are evolving their strategies for solving problems. For instance, OpenAI’s o3 model excels in mathematical problem-solving, achieving a 97% success rate in the AIME (American Invitational Mathematics Examination) competition, a major step forward from earlier performances of a few percent. The o3 model also excels in benchmarks like Frontier Maths and ARC-AGI-PUB, reaching human-equivalent scores. Similar progress was observed in medicine, physics, and coding benchmarks.

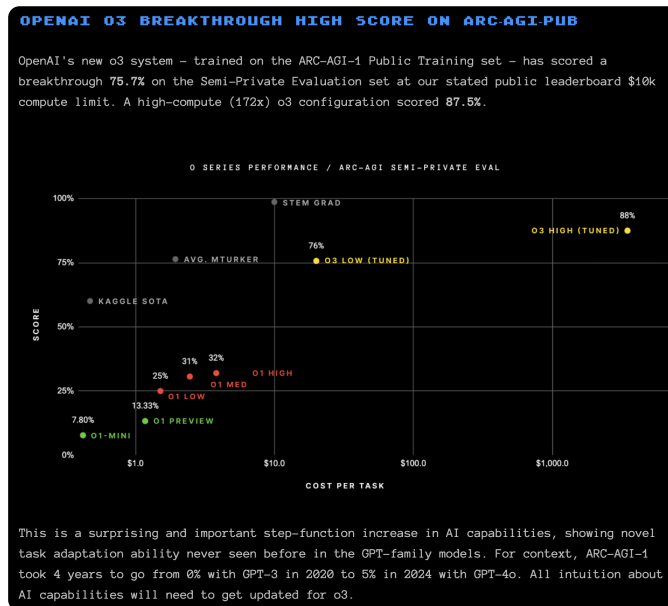


Figure 2: Blog from François Chollet about o3 and the ARC-AGI-PUB benchmark [ARCPRIZE]

This increased computation demand during inference has significant implications for the hardware market, particularly benefiting GPU providers like NVIDIA. Solving a single problem on benchmarks like ARC with o3 can cost thousands of dollars in computation, necessitating infrastructure investments like OpenAI's \$200 monthly ChatGPT Pro subscriptions.

The competitive landscape of AI also shifted significantly in 2024. While OpenAI maintained a lead in reasoning-based models like o1 and o3, other players like Google, Anthropic, Meta, xAI, and even Chinese companies such as DeepSeek and Alibaba made landmarks in LLM development. Google with Gemini 2 Flash Thinking, but also the Chinese DeepSeek et Alibaba, with their models DeepSeek R1 and QwQ (quill), also propose models that can use variable inference compute time to produce answers.

The open-source community also gained ground, with Meta's Llama 3 models and Alibaba's DeepSeek V3 rivalling closed models like GPT-4o. Hardware constraints became central, with NVIDIA's GPUs remaining indispensable for model training, for example, xAI's supercomputer now having 100,000 NVIDIA GPUs. All major companies are developing their own accelerator chips (AWS Inferentia for inference servers - they also develop Trainium chip; Meta with its Next GenMTIA), although Google's TPUs still have a competitive edge because they are already on their sixth generation with the Trillium chip. A significant hardware race unfolded; the hardware demand even triggered discussions about energy requirements, potentially leading to the construction of dedicated nuclear power plants.

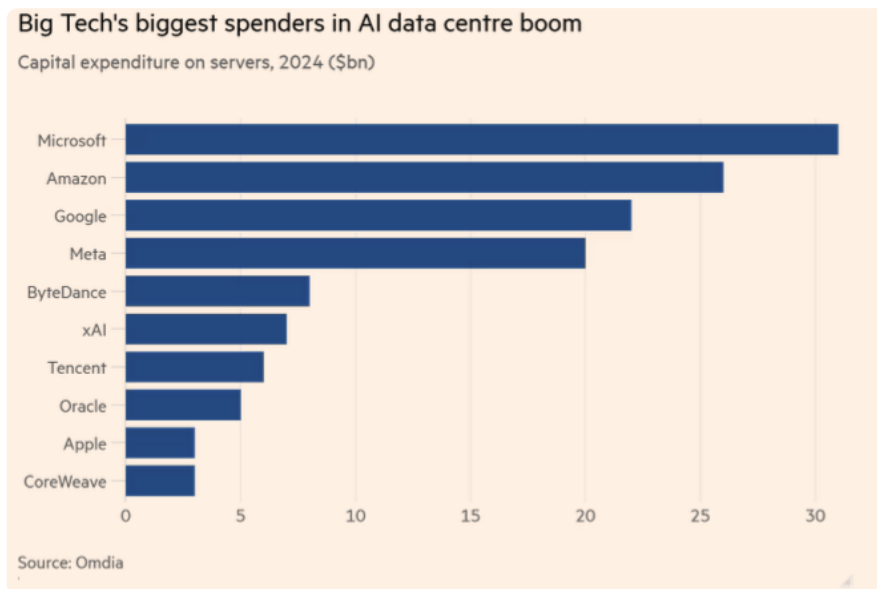


Figure 3: spending in AI servers in 2024, data originally from Omdia

In 2024, the landscape of large language models (LLMs) witnessed not only technical advancements but also greater accessibility for everyday users. Apple introduced Apple Intelligence, a feature integrating LLMs across all Apple devices. This innovation allows users to interact seamlessly with AI, even enabling direct access to ChatGPT. Apple is the first to propose a kind of “distributed” approach: first, the local LLMs are used by an orchestrator; if they are not powerful enough, the demand is seamlessly transferred to Apple servers, and even to ChatGPT.

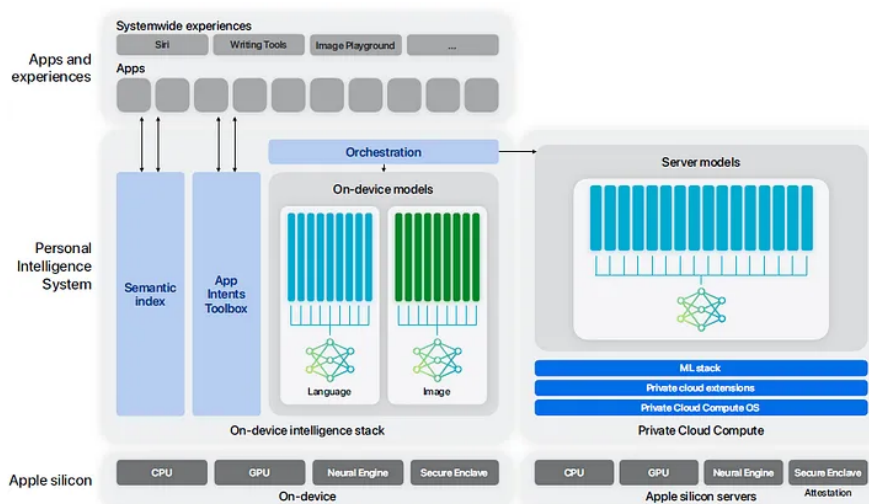


Figure 4: Architecture of Apple Intelligence with adapters, highlighted as blue and green rectangles, for the on-device and server language/image models, from <https://medium.com/byte-sized-ai/on-device-ai-apple-intelligence-533c4c6ed7d6>

This year also witnessed an interesting development in the training of large language models (LLMs): distributed training across the globe. Traditionally, LLMs have been trained within the confines of a single data centre, where GPUs are interconnected to manage the computational load. However, by the end of the year, two companies, Prime Intellect and

NousResearch, pushed the boundaries of this approach by training models with 10 billion and 15 billion parameters, respectively, using a distributed network of computers located in Europe, Asia, and the United States.

This innovation marks a significant shift in how LLMs can be developed, presenting opportunities for more flexible and scalable training processes. By spreading the workload across multiple regions, this method could lower barriers for smaller organizations, enabling them to pool resources and collaborate on creating advanced models. This distributed training approach holds immense potential for democratizing access to cutting-edge AI capabilities while fostering innovation on a global scale.

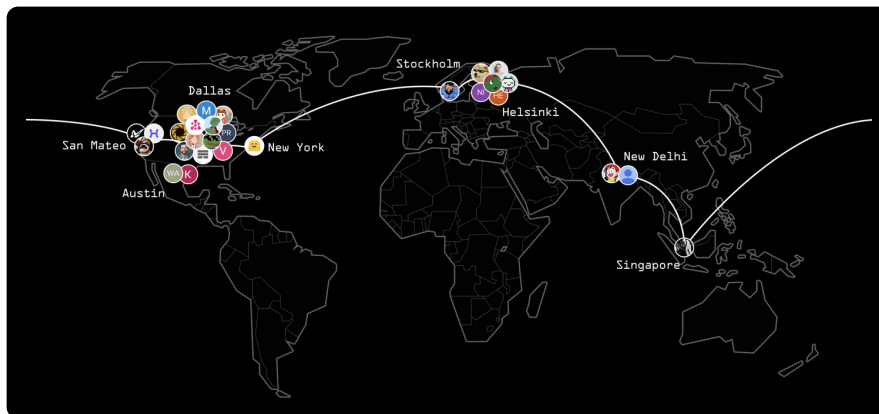


Figure 5: INTELLECT-1 Release: The First Globally Trained 10B Parameter Model [INTELLECT-1]

Recommendations and actions from observing 2024 evolutions

It is clear that trends seen in 2024 will continue in the future, perhaps with new improvements, but more computing power seems to be the key enabler of artificial intelligence, with its corollary of increased energy consumption. Therefore, **making innovative new hardware for supporting LLMs** is part of this HiPEAC Vision; see the “New Hardware” chapter.

However, from the (distributed) structure of Apple Intelligence, distributed training across the globe, and the new models like o1, we can derive recommendations that will help Europe to re-enter the game. In the summary above of major developments in 2024, only US and Chinese companies or organizations were cited; unfortunately, none of those cited were from Europe.

As explained in the foreword and introduction, the ideas behind the NCP can be instantiated in the short term as “distributed agentic AI”. The structure of Apple intelligence, of distributed training across the globe, are clear precursors, but also the possible technology behind models like o1 (see for example [Zeng24]) might possibly be done by several specialized agents working together.



There is clear research (and business) interest in looking into a set of smaller specialized agents (LLMs) working (orchestrated) together. If the agents are distributed in different locations— as was the case for [SETI@Home], [BOINC] and [Petals] – and if the compute resources are shared, as proposed in the NCP concept, then perhaps a gigantic data centre that consumes MW of electricity is not a requirement for advances in AI, or to run existing AI for users. This opens up contributions from a much larger base than the few companies that can afford gigantic data centres. And this distributed AI from edge to data centres can adapt to the user's requests, being exclusively local for simple requests and not activating a large LLM on a distant data centre, with its associated cost in terms of energy.

The recommendation is therefore the **development of specialized action models (SAMs)** – that is, small and specialized models that can interact with their environment, acting as service. These SAMs should operate in a distributed infrastructure and an ecosystem should be created to support research, development and business around them. Of course, this ecosystem should be a precursor and compatible with the one provided by the NCP. These models need to be refined, optimized, and reduced in size to improve efficiency.

This also ties into the need for **hybrid systems** (combining AI and algorithmic approaches), as noted in the foreword to this vision: future systems that must integrate both paradigms— precise and approximate—within feedback and reinforcement-based architectures. These SAMs can be optimized from more general foundation models by an ecosystem of companies providing their optimized SAMs in a marketplace so that they can be dynamically discovered and used by the orchestrators.

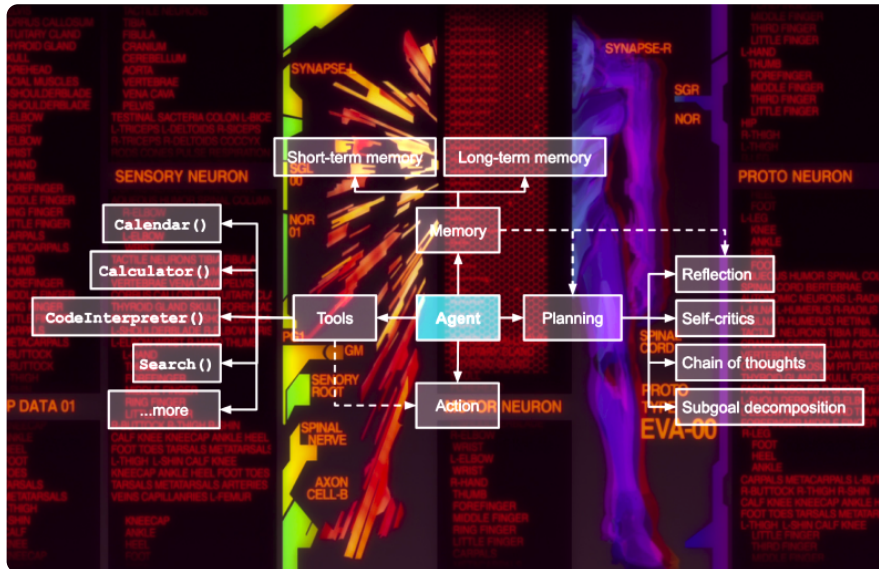


Figure 6: developing agents models with reasoning, from <https://www.primeintellect.ai/>

Two other ingredients are necessary for the system beside the developments of those SAMs:

- A way to discover, select, active, organize them together, therefore the **development of orchestration technologies** that take into account all the requirements, that can select the best SAMs for the required tasks and that can dynamically activate them. The first steps could be very agentic-AI-centric (relying on already existing technologies used for orchestrating AI agents), but they should be a blueprint and evolve towards an orchestration system for the NCP. These orchestrators must be developed for the edge – or near the final user – and dynamically combine SAMs into executing personalized applications in response to user needs.
- This will be only possible if all the systems “speak the same language”, therefore, a key recommendation is that it is **imperative to establish open protocols for these “distributed agentic AI” systems to facilitate seamless interaction among distributed AIs from different origins.**

To effectively operate this federation of distributed AIs, it is necessary for them to exchange data and parameters through a universally comprehensible protocol that:

1. Does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).
2. Also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, environmental impact, cost, localization, privacy of data, etc.).

It is therefore important that the community works together to commonly define this exchange protocol that should be open to allow broad acceptance. Large entities such as OpenAI, Meta, and Microsoft are attempting to promote their own application programming interfaces (APIs) for accessing their models. However, an API alone is insufficient for constructing this distributed and federated network of AIs. Also, the exchange format (JSON, ASCII text) is perhaps not the optimal way for networks of AIs to efficiently exchange information: this could be tokens, embeddings, or any other representations; some research also shows that LLMs talking to each other could develop their own “language”.

In the fields of distributed agentic AI, some work has already been done, for example [DAWN] and [DistMixofAgents]. But it is important to group all existing ideas and existing proposals together in a single consortium to develop an open source “de facto” (before official standardization) standard that takes into account all the good ideas of various researchers and organizations, so that it will be sound, future-proof, recognized and accepted.

In a similar way to TCP-IP that enabled various OS (operating systems) to communicate, the aim of this action is to create the equivalent for OS (orchestration systems) to exchange AI-related information.

Time is crucial for this initiative, and standardization, however necessary, will be too long, so a de facto open standard should be proposed in parallel with the standardization effort, before other closed proposals will emerge, locking down the approach to a few (non-European) players. This should also act as a blueprint for the NCP protocols and specifications. Like for the NCP, this approach will allow the creation of a completely new ecosystem where smaller players can provide specialized AI as a service along with the big ones. Directories of services, trusted brokers, and payment services are also important elements that can emerge from this ecosystem, where Europe can have an active part thanks to its set of SMEs, research organizations, and distributed nature.

Europe should be an active player in the race for the “distributed agentic artificial intelligence”.

References

ARCPRIZE: <https://arcprize.org/blog/oai-o3-pub-breakthrough>

BOINC: <https://boinc.berkeley.edu/>

DAWN: Aminiranjbar, Zahra et al. “DAWN: Designing Distributed Agents in a Worldwide Network from Cisco Systems”, 2024 <https://arxiv.org/pdf/2410.22339>

DistMixofAgents: Mitra, Purbesh, Kaswan, Priyanka and Ulukus, Sennur. “Distributed Mixture-of-Agents for Edge Inference with Large Language Models”, 2024 <https://arxiv.org/abs/2412.21200>

HuangHPCwire: “You know that in the future, the vast majority of content will not be retrieved, and the reason for that is because it was pre-recorded by somebody who doesn’t understand the context, which is the reason why we had to retrieve so much content,” he said. “If you can be working with an AI that understand the context – who you are, for what reason you’re requesting this information – and produces the information for you, just the way you like it, the amount of energy you save, the amount of network and bandwidth you save, the waste of time you save, will be tremendous.”

INTELLECT-1: <https://www.primeintellect.ai/blog/intellect-1-release>

NobelManga: <https://jianzongwu.github.io/projects/diffsensei/>

NobelMangaPDF: https://jianzongwu.github.io/projects/diffsensei/static/pdfs/nobel_prize.pdf

Petals: <https://github.com/bigscience-workshop/petals#readme>

SETI@home: <https://setiathome.berkeley.edu/>

Zeng24: Zeng, Zhiyuan et al. “Scaling of Search and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective”, 2024 <https://arxiv.org/pdf/2412.14135>