Grigori Fursin – Curriculum Vitae

gfursin@gmail.com cKnowledge.org/gfursin scholar.google.com/citations?user=IwcnpkwAAAAJ Paris, France

British national with permanent residency and full work rights in France. As Head of the R&D Lab at FlexAI, I use AI to co-design more efficient and cost-effective AI systems. I am also a Founder and Architect of cKnowledge.org and cTuning.org, Former Founder and Architect of a collaborative AI benchmarking and optimization platform acquired by OctoAI (now Nvidia), Former VP of MLOps at OctoAI, Former Co-Director of the Intel Exascale Lab, Former Senior Tenured Scientist at INRIA, and Former Adjunct Professor at the University of Paris-Saclay with a PhD in Self-Optimizing Compilers and Systems from the University of Edinburgh.

Biography and motivation

I am a forward-thinking and agile computer scientist, inventor, full AI/ML stack engineer, strategic advisor to startups, investors, and executive teams, educator, mentor, long-time open source contributor, and passionate advocate for open science. I hold a PhD in self-optimizing compilers and systems from the University of Edinburgh. My interdisciplinary background spans computer engineering (with expertise in co-designing the full hardware-software stack from the cloud to the edge), machine learning, AI systems, data analytics, workflow automation, knowledge management, and physics and electronics. I am passionate about building innovative solutions to real-world problems and about unifying and automating R&D processes to enhance efficiency and reduce costs.

Fascinated by the prospects of AI and robotics, I began my R&D career in the mid-1990s as an undergraduate, taking on a technical leadership role to develop Hopfield-based analog semiconductor neural networks from scratch. This included complete development and automation of software, hardware, models, and datasets for training, inference, electronic simulation, and prototyping—since nothing existed at the time.

This project took much longer than I originally expected and revealed numerous issues in R&D methodologies, tools, and inefficiencies in computer engineering. As a result, I decided to switch to computer science and pursue PhD research to address these challenges. This interdisciplinary foundation and experience enabled me to pioneer and champion visionary uses of machine learning, AI, crowd-tuning, and crowd-learning to co-design more efficient, cost-effective, and scalable computer systems—including compilers, runtimes, software, and hardware—during my PhD at the University of Edinburgh and postdoctoral research at Inria.

I led R&D efforts that addressed the growing complexity of modern systems and served as a precursor to self-optimizing and agentic systems, AutoML, workflow automation, agent-based optimization, federated learning, reproducible experimentation, and universal, efficient, technology-agnostic compute. It also enabled me to initiate and support open science and reproducibility initiatives starting in 2008, when I launched cTuning.org (followed by cKnowledge.org with my Collective Knowledge Technology aka CK in 2014) and released all my research code, data, models, and experiments for our ML-based self-optimizing compiler—considered the first of its kind (ACM TechTalk'21). I was honored to receive the ACM CGO Test of Time Award, multiple Best Paper Awards, the INRIA Award for Scientific Excellence, and the EU HiPEAC Technology Transfer Award for this research and open-source tools.

After serving as a senior tenured research scientist at INRIA, an adjunct professor at the University of Paris-Saclay, and co-director of the Intel Exascale Lab, I transitioned my research and open-source tools into industry. I first established a non-profit cTuning foundation and co-founded a successful engineering company to automatically benchmark and optimize deep learning across diverse software and hardware stacks, with a focus on mobile phones and edge devices. I helped bootstrap it as CTO and Chief Architect, quickly growing it to $1M+ in revenue with just 4 people, thanks to my CK automation technology. I then joined Entrepreneur First, a highly selective company-building program for scientists and technologists, where I learned to build lean startups and avoid common pitfalls. As a result, I founded and bootstrapped two startups in the fields of performance optimization, MLOps automation, and knowledge management—the latter of which was acquired by OctoAI (now part of NVIDIA).

During that time, I invented the Collective Mind automation language (CM/CMX), which was adopted by MLCommons—a consortium of over 100 AI and systems companies—to test and benchmark a wide range of AI models and datasets across diverse hardware and software platforms, from cloud to edge. My CM technology has automated thousands of MLPerf submissions and enabled the discovery of some of the most performance- and cost-efficient AI solutions using commodity servers, outperforming high-end systems from Nvidia. I am now developing the next generation of this automation.

At the same time, I remained actively involved in community service and open-source initiatives. I helped establish MLCommons and launch reproducibility efforts at ACM and IEEE conferences: cTuning.org/ae . I also introduced a unified artifact appendix, which has since been adopted by major conferences such as ASPLOS, CGO, PPoPP, SuperComputing and MICRO. Finally, I co-organized several successful Quantum Hackathons, including one at Ecole 42 in Paris, where we utilized my CK workflow automation and platform for collaborative benchmarking and optimization of Quantum workloads (Hackathon page and a list of my events).

Throughout my career, I've been honored to collaborate with and learn from brilliant minds across leading universities, non-profits, startups, and companies — including Google, Amazon, Meta, Arm, AMD, Intel, IBM, Qualcomm, NVIDIA, Raspberry Pi, OpenAI, Tesla, OctoAI, Neural Magic, Red Hat, Dell, HPE, Lenovo, Apple, INRIA, ACM, IEEE, HiPEAC, MLCommons, and the Linux Foundation: Acknowledgments (1), Acknowledgments (2), and Acknowledgments (3).

My passion lies in applying my knowledge, experience, and tools to accelerate the journey from deep tech research to real-world production—while building intelligent, self-optimizing systems. I regularly support startups, enterprises, universities, non-profits, researchers, students, and investors in rapidly prototyping novel ideas, launching innovative deep-tech projects, reducing time to market, and delivering tangible impact through collaborative, reproducible, interdisciplinary, quantifiable, and automated R&D methodologies.

While I actively prototype full-stack projects and contribute hands-on, I bring the most value in roles such as strategic advisor, technical manager, R&D lab head, educator, research scientist, or senior individual contributor. I focus on bridging research, engineering, and product teams—helping them navigate complex, rapidly evolving technology landscapes, manage project complexity, avoid common pitfalls, and achieve meaningful outcomes efficiently, even with with limited resources and time.

In 2024, I began prototyping the next generation of my Collective Knowledge platform, aimed at solving the complexity of AI systems. My goal is to develop a universal compute engine that simplifies running models on any hardware with any software in the most efficient and cost-effective way while significantly reducing all associated costs. This initiative builds on my existing work, including Collective Mind, virtualized MLOps, MLPerf, Collective Knowledge Playground, and reproducible optimization tournaments. For more details, see my white paper, and feel free to reach out if you would like to learn more. I also joined FlexAI as Head of their R&D Lab, where I am developing FlexBench to track state-of-the-art models like DeepSeek and to benchmark and optimize them across diverse software and hardware stacks. This work is based on the MLPerf methodology and my MLCommons CM workflow automation framework.

In my spare time, I enjoy spending time with my two children, reading, learning new skills, playing soccer (having competed semi-professionally), hiking, traveling, teaching, developing automations and platforms for collaborative and reproducible R&D, and brainstorming creative projects.

Employment History

2024-cur. FlexAI (France) ***Head of R&D Lab***

- I am setting up an R&D lab focused on leveraging AI to develop more efficient and cost-effective systems for inference and training, building on my interdisciplinary expertise in machine learning, software/hardware co-design, benchmarking, and data analytics.
- I lead the development of FlexBench, an open-source benchmarking framework for evaluating and optimizing HuggingFace LLMs—such as DeepSeek and LLaMA 3+—in terms of accuracy, performance, and cost across diverse software and hardware stacks. FlexBench leverages MLPerf methodology and workflow automation tools I built such as Collective Knowledge: see our MLPerf inference v5.0 submission with DeepSeek
- Core technologies used: HuggingFace models and datasets, vLLM, PyTorch, Triton, TensorRT, Nsight, MLPerf, OpenSearch, MLCommons CMX, FastAPI, Docker, Bayesian search, reinforcement learning and LLMs, Nvidia and AMD GPUs.

2023-2024 MLCommons (USA) ***Coordinator and developer of MLPerf automations***

- I bootstrapped the development of Collective Mind automation recipes for MLOps, MLPerf, and the ABTF (Automotive Benchmarking Task Force). This was part of a collaborative engineering effort to run MLPerf Inference benchmarks across a wide range of models, datasets, software stacks, and hardware platforms from various vendors — all in a unified, reproducible, and automated way:
    - https://access.cknowledge.org/playground/?action=scripts
- My MLCommons Collective Mind framework (CM) with MLPerf automations was successfully validated by automating ~90% of all MLPerf Inference v4.0 performance and power submissions. It also enabled identification of top-performing and cost-efficient software/hardware configurations for AI systems from different vendors.

2023-cur. cKnowledge.org (France) ***Founder and Chief Architect***

- I am prototyping an open platform to learn how to co-design more efficient and cost-effective ML and AI systems through community challenges and reproducible optimization tournaments—powered by my MLCommons CM automation framework and virtual MLOps automation recipes.
    - White paper: https://arxiv.org/abs/2406.16791
    - Project website: https://access.cKnowledge.org

2021-2023 OctoAI acquired by NVIDIA (USA) ***Vice President working remotely from Paris***

- I have developed the 2nd generation of the open-source Collective Knowledge workflow automation technology (aka Collective Mind with MLOps and MLPerf automation recipes) - it was adopted by @MLCommons to modularize MLPerf inference benchmarks and automate development, optimization and deployment of AI systems across diverse models, data sets, software and hardware:
    - o https://github.com/mlcommons/ck
    - o https://doi.org/10.5281/zenodo.8105338

2019-2021 cKnowledge.io (France) – acquired by OctoML.ai ***Founder and Chief Architect***

- I proposed and implemented an MVP of a virtual MLOps platform that enabled for the first time automatic and collaborative co-design, benchmarking, comparison and optimization of ML Systems from unified LEGO-like bricks for diverse ML models, frameworks, datasets and platforms from different vendors.
- With my automation, 3 MLPerf benchmark contributors submitted 3x more results than 12 companies combined.
- Partnered with MLCommons, ACM and the Linux Foundation AI to expose optimization problems in systems to ML researchers and practitioners via my platform during reproducible tournaments: cKnowledge.io/reproduced-results.
- Demonstrated how to combine my technology with predictive modeling to auto-generate adaptive libraries that accelerated convolution kernels in 8 popular CNNs by up to 4x over best vendor solutions on the latest high-end Nvidia GPUs and low-power Arm GPUs (research results published in ACM TACO'21).
- Honored to give a prestigious ACM TechTalk (Feb 2021): bit.ly/acm-tech-talk .

2015-cur. cTuning foundation (France) ***Founder and Director of Research and Development***

I established the cTuning foundation to support my open science initiatives and develop a common methodology and open-source tools for collaborative and reproducible Systems ML R&D: cTuning.org.

- **Collective Knowledge Framework (CK) – became an official MLCommons framework in November 2021**
    - Developed the open-source Collective Knowledge framework (CK) to automate and accelerate ML Systems research by converting ad-hoc research projects into a database of portable, customizable and modular workflows, reusable artifacts, common APIs and extensible meta descriptions: github.com/mlcommons/ck .

  - CK has been downloaded 270K+ times and successfully used by Fortune 50 companies, startups and universities to accelerate benchmarking, optimization and co-design of ML Systems by 10..50x : cKnowledge.org/partners.
- **Reproducibility initiatives and ML Systems hackathons**
  - Introduced artifact evaluation at 6 ACM/IEEE conferences including MLSys, ASPLOS, CGO and PPoPP to reproduce results from published papers: cTuning.org/ae. Half of accepted papers undergo this evaluation now.
  - As a founding member of the ACM taskforce on reproducibility, I introduced the first Reproducibility Checklist and co-authored the Artifact Review and Badging policy widely used across most ACM conferences now.
  - Co-organized 6 successful reproducible ML and quantum hackathons in collaboration with ACM, IBM, Arm, Intel, NVidia, Google, Microsoft, CERN, AMD and multiple universities: cKnowledge.org/request and cKnowledge.org/quantum. IBM hired the winners of our last hackathon in Paris out of 50 participants.

2015-2019 dividiti Ltd (UK) ***Co-founder and CTO***

- Co-founded an engineering company to validate my CK technology in production and led it to €1M+ revenue.
- Led a team of 5 researchers and engineers to develop CK-powered ML workflows for our Fortune 50 customers:

2018 Amazon (UK) ***R&D project partner (Research Project Manager and Tech Lead)***

- Successfully prototyped CK workflows connected with SageMaker to automatically scale deep learning on AWS using C5 instances with MXNet, TensorFlow and BigDL from the edge to the cloud: bit.ly/ck-aws.

2017-2018 General Motors (USA) ***R&D project subcontractor (Research Project Manager and Tech Lead)***

- With my CK technology, GM reduced the time needed to explore and select Pareto-optimal AI/SW/HW stacks from different vendors for self-driving cars from 3 months to a few days: youtu.be/1ldgVZ64hEI.

2015-2019 Arm (UK) ***R&D project subcontractor (Research Project Manager and Tech Lead)***

- Developed CK workflows for Arm to fully automate internal benchmarking and optimization of SOTA AI and ML workloads across diverse frameworks, models, compilers and platforms: cKnowledge.io/nn-components.

2010-2011 Intel Exascale Lab (France) ***Co-director and Head of application optimization group (on sabbatical)***

- Co-directed Intel Exascale Lab in France with 24 researchers and engineers and reported directly to the Lab's CTO.
- Proposed, designed and led the development of a framework and a web-based platform to fully automate the software/hardware co-design process for Exascale systems using autotuning workflows and ML: c-mind.org/repo.
- Promoted to Senior Scientist at INRIA and received the prestigious INRIA award of scientific excellence.

2006-2014 INRIA (France) ***Tenured research scientist***

- I was the first researcher in the world to demonstrate how to automatically learn optimization heuristics in production compilers, reduce optimization cost by an order of magnitude and get up to 4x speedups and 30% code size reductions on previously unseen workloads in comparison with the best manually tuned heuristics.
- Developed a compiler plugin framework and program feature extractor included to the mainline GCC in collaboration with Google and Mozilla to support ML-based testing and optimization in production compilers.
- Led the development of the world's first ML-based compiler across 20 researchers and engineers in 5 teams in the €2M MILEPOST project with IBM, ARC, U.Edinburgh, CAPS Entreprise and INRIA. This work was commercialized by IBM, Synopsys, Intel and STMicroelectronics: www-03.ibm.com/press/us/en/pressrelease/27874.wss.
- Co-advised 2 PhD students who successfully graduated from the University of Paris Saclay.
- Developed cBench and 2 data sets (MiDataSets and KDataSets) to create more realistic conditions for Systems ML research. They are now used by the leading universities and companies including Arm, Google and Facebook.
- Developed the cTuning.org platform considered the first in the world to crowdsource ML-based program autotuning across diverse platforms provided by volunteers similar to SETI@home: https://bit.ly/ctuning.

2005-2007 INRIA (France) ***Postdoc funded by the EU HiPEAC fellowship***

- Co-authored the successful €2M MILEPOST grant proposal to build the world's first self-optimizing compiler based on my PhD research and powered by crowd-tuning and Machine Learning.
- Was awarded a tenured research scientist position at INRIA to lead this project.

1999-2005 University of Edinburgh (UK) ***Research project manager (research associate)***

- Led 2 of 4 work packages in the €1M EU MHAOTEU project. This comprised 12 researchers and administrators.
- Built an open-source polyhedral compiler and autotuning infrastructure used by the HiPEAC network of excellence.
- Successfully delivered all objectives and won the EU HiPEAC postdoctoral fellowship (~5 out of 500 applicants).

Education

I am passionate about lifelong learning and regularly take Coursera and other online courses to acquire new skills or refresh existing knowledge: www.linkedin.com/in/grigorifursin/details/certifications

2019 Entrepreneurs First (France)

- I was selected for Entrepreneur First's second cohort in Paris where I learnt how to build deep tech startups and MVPs from scratch, while avoiding common pitfalls and minimizing risks.
- This experience enabled me to create the virtual MLOps platform acquired by OctoAI in 2021. OctoAI was acquired by Nvidia in 2024.

1999-2004 University of Edinburgh (UK) ***PhD in Computer Science*** **(advisor: Prof. Michael O'Boyle)**

- I demonstrated for the first time that it is possible to perform program autotuning in production and achieve up to 10x speed-ups on previously unseen programs in comparison with SOTA commercial compilers.
- Prepared the foundations for ML-based autotuning that enabled the world's first self-optimizing compiler in 2009.
- Won a highly selective EU HiPEAC postdoc fellowship to continue my research at INRIA (~5 out of 500 applications).

1997-1999 MIPT (Russia) ***MSc in Computer Engineering*** **(summa cum laude and golden medal, GPA=4.0)**
1993-1997 MIPT (Russia) ***BSc in Physics and Electronics*** **(summa cum laude, GPA=4.0)**

- Developed a prototype of an analog semiconductor Hopfield neural network to accelerate inference by 10x.
- Developed all related software and data sets for electronic circuit simulation, NN training and inference.
- Developed a web platform that enabled unified and remote access to diverse HPC resources for ML researchers.
- Won a prestigious Overseas Research Student Award (ORS) selected by the Committee of Vice Chancellors and Principals of the UK Universities to continue my research at the University of Edinburgh.

Licenses & certifications

- 2025: MCP: Build Rich-Context AI Apps with Anthropic (DeepLearning)
- 2025: AI Agents and Agentic AI with Python & Generative AI (Coursera)
- 2025: Foundations of Project Management (Coursera/Google)
- 2024: Generative AI with Large Language Models (Coursera)
- 2024: Efficiently Serving LLMs (DeepLearning)
- 2024: Intro to Federated Learning (DeepLearning)
- 2024: Quantization Fundamentals with Hugging Face (DeepLearning)
- 2023: Learning How to Learn (Coursera)
- 2021: Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization (Coursera)
- 2021: Structuring Machine Learning Projects (Coursera)
- 2021: Neural Networks and Deep Learning (Coursera)
- 2021: AI for everyone (Coursera)
- 2020: Machine Learning (Coursera)

Academic research (tenured research scientist at INRIA with PhD in CS from the University of Edinburgh)

- I prepared the foundations to combine machine learning, autotuning, knowledge sharing and federated learning to automate and accelerate the development of efficient software and hardware by several orders of magnitude (Google scholar);
- developed self-optimizing compiler with collective tuning considered to be the first in the world (MILEPOST GCC);
- developed Collective Knowledge and Collective Mind technology and started educational initiatives with ACM, IEEE, HiPEAC, Raspberry Pi foundation and MLCommons to bring my research and expertise to the real world to benefit everyone;
- prepared and tought M.S. course at the Paris-Saclay University on using ML to co-design efficient software and hardare (self-optimizing computing systems);
- gave 100+ invited talks about my R&D;

Project management and open-source development

2024-cur: leading the development of the next generation of my MLCommons Collective Mind and Collective Knowledge technology to help researchers, engineers and students co-design software and hardware for more efficient and cost-effective AI;
2023-2024: led the development of the Collective Knowledge playground to benchmark and optimize AI/ML Systems via reproducible optimization challenges and tournaments;
2022-2024: led the development of the Collective Mind automation framework (CM) to modularize AI/ML systems and make it easier to benchmark and optimize them across diverse and rapidly evolving models, data sets, software and hardware from different vendors - I donated CK and CM to MLCommons to benefit everyone and continue developing at as a community effort (see white paper);
2014-2019: developed the Collective Knowledge framework to automate and accelerate design space exploration of AI/ML/SW/HW stacks while balancing speed, accuracy, energy and costs;
2010-2011: led the development of the cTuning 2 automation framework to benchmark emerging workloads across diverse hardware at Intel Exascale Lab;
2007-2009: led the development of the ML-based compiler and the cTuning.org platform across 5 teams to automate and crowdsource optimization of computer systems - this technology is considered to be the first in the world;
2007-2009: led the development of the compiler plugin framework in collaboration with Google and Mozilla that was added to the mainline GCC powering all Linux-based computers and helped to convert production compilers into research toolsets for machine learning;

Community service (collaboration with MLCommons, ACM, IEEE, HiPEAC and other organization)

2025: Program Committee Member, ACM Conference on Reproducibility and Replicability 2025
2024: Co-organizer of the artifact evaluation at IEEE/ACM MICRO'24
2024: MLPerf liaison at the Student Cluster Competition at SuperComputing'24
2022-2024.: Founder of the MLCommons taskforce on automation and reproducibility - I donated my open-source Collective Mind technology to MLCommons to benefit everyone and continue developing it as a community effort to modularize and automate MLPerf benchmarks (see my ACM REP'23 keynote and white paper for more details).
2020-cur.: A founding member of MLCommons to accelerate ML and systems innovation.
2020-cur.: A founding member of the ACM SIG on reproducibility.
2019-cur.: An early member of the MLPerf.org.
2017-cur.: A founding member of the ACM taskforce on reproducibility.
2015-cur.: Author of the unified artifact appendix and reproducibility checklist now used and extended by many ACM and IEE conferences;
2015-cur.: co-organizer of many reproducible optimization tournaments and hackathons to co-design efficient AI and ML systems powered by my CK&CM technology;
2014-2016: co-author of the ACM artifact review and badging policy.
2014-cur.: helped to reproduce 150+ research papers from ML and systems conferences;
2014-cur.: Introduced reproducibility initiatives and artifact checklists at MLSys, ASPLOS, MICRO, CGO, PPoPP and other ML and systems conferences to validate results from published papers (see white paper and ACM TechTalk for more details);
2014-cur.: Founder and president of the cTuning foundation, France.
2008-cur.: Founder and the architect of cTuning.org.

Awards

I have been honored to receive 7 major research awards including the ACM/IEEE CGO’17 test of time award for my ML for Systems research (considered to be the most influential even 10 years later), 2 best papers awards at CASES and HiPEAC, 1 best presentation award at CGO, EU HiPEAC technology transfer award and postdoc fellowship and INRIA award for scientific excellence.

Key presentations and publications

- Invited ACM TechTalk'21: Reproducing 150 Research Papers and Testing Them in the Real World
- HPCA'25: MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from Microwatts to Megawatts for Sustainable AI
- ArXiv white paper'24: Enabling more efficient and cost-effective AI/ML systems with Collective Mind, virtualized MLOps, MLPerf, Collective Knowledge Playground and reproducible optimization tournaments,
- ACM REP'23 keynote: Collective Mind: toward a common language to facilitate reproducible research and technology transfer
- Nature Machine Intelligence'23: Federated benchmarking of medical artificial intelligence with MedPerf
- Philosophical Transactions of the Royal Society'21: Collective knowledge: organizing research projects as a database of reusable components and portable workflows with common interfaces
- Quantum Collective Knowledge Hackathon at École 42 (Paris): GitHub with code and photos
- Joint presentation with Amazon at O'Reilly AI conference'18: Scaling deep learning on AWS using C5 instances with MXNet, TensorFlow, and BigDL: From the edge to the cloud
- Presentation from General Motors about my Collective Knowledge technology: Collaboratively Benchmarking and Optimizing Deep Learning Implementations
- ArXiv preprint'18: A collective knowledge workflow for collaborative research into multi-objective autotuning and machine learning techniques
- ArXiv prepring'15: Towards Performance- and Cost-Aware Software Engineering as a Natural Science
- Report of Dagstuhl Perspectives Workshop'15: Artifact Evaluation for Publications
- ACM TRUST'14 at PLDI'14: Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering
- IJPP'11: Milepost gcc: Machine learning enabled self-tuning compiler
- PLDI'10: Evaluating Iterative Optimization Across 1000 Data Sets
- MICRO'09: Portable compiler optimisation across embedded programs and microarchitectures using machine learning
- HiPEAC'09: Predictive runtime code scheduling for heterogeneous architectures
- CGO'07: Rapidly selecting good compiler optimizations using performance counters
- CGO'06: Using machine learning to focus iterative optimization

Key software developments that I initiated and prototyped before handing them over to the community

- My Collective Knowledge (CK) platform and Collective Mind (CM) workflow automation technology (hosted on the MLCommons GitHub): https://github.com/mlcommons/ck , https://doi.org/10.5281/zenodo.8105338
- Virtual MLOps scripts and MLPerf automations:
  - https://github.com/mlcommons/ck/tree/master/cmx4mlops/repo/flex.task
  - https://github.com/mlcommons/ck/tree/master/cm4mlops/cm4mlops/repo/script
- Prototype of a next-generation virtual MLOps platform (universal AI compute): https://access.cKnowledge.org
  - https://access.cknowledge.org/playground/?action=challenges
  - https://access.cknowledge.org/playground/?action=contributors
- Prototype of FlexBench to benchmark vLLM across diverse hardware, software and HuggingFace models using the MLPerf methodology and CM automation:
  - https://github.com/mlcommons/inference_results_v5.0/tree/main/open/FlexAI/measurements/cmx-flexbench-cuda-1xH100-vllm-0.7.3-pytorch-2.5.1-huggingface-16d94432c8704c14/DeepSeek-R1-Distill-Llama-8B/Server
- Artifact Evaluation website and unified artifact appendix adopted and extended by major CS conferences:
  - https://cTuning.org/ae
  - https://github.com/ctuning/artifact-evaluation
  - https://github.com/ctuning/artifact-evaluation/blob/master/docs/checklist.md
- Discontinued Collective Knowledge portable ML/AI solutions (2019-2021): https://cknow.io/solution/demo-obj-detection-coco-tf-cpu-webcam-linux-azure
- Discontinued Collective Knowledge crowd-results (2015-2018): https://cKnowledge.org/repo.html
- Discontinued Collective Mind repository (2011-2014): https://c-mind.org/repo
- Discontinued Collective Tuning portal (2007-2011): https://cTuning.org/wiki