Narendra Patwardhan

Sustainable Foundation Models and Extensions

Tutor: Prof. Carlo Sansone

Cycle: XXXVIII

Year: Second

My background

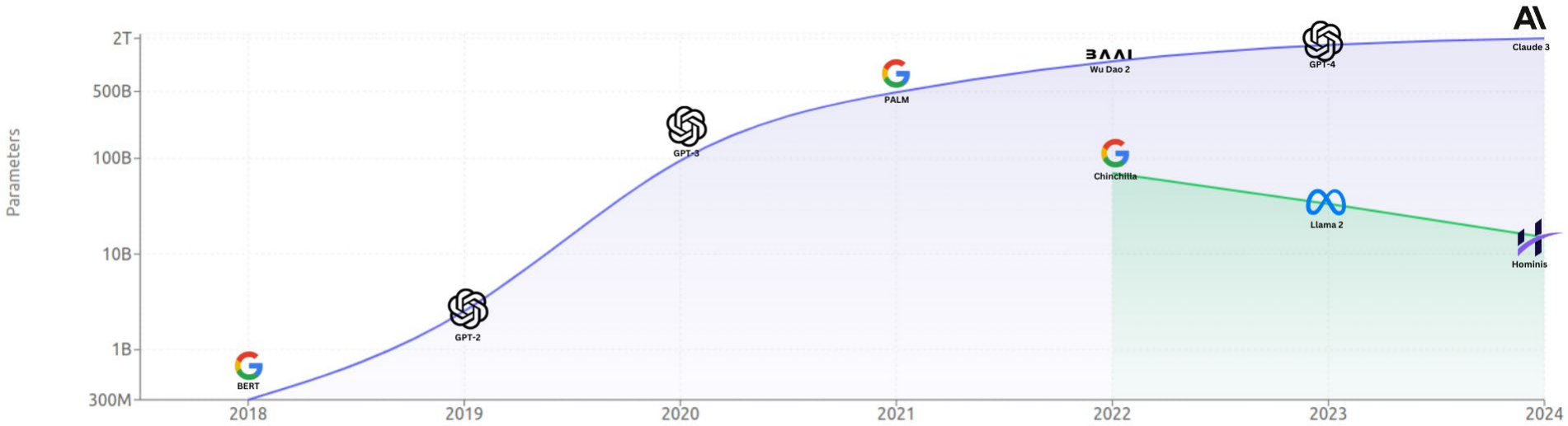
- MSc degree:

  Mechanical Engineering @ Michigan Technological University

Thesis - Proximal Reliability Optimization for Reinforcement Learning

- Research Group: **PICUS Lab**
- PhD start date: 01/11/2022
- Scholarship type: PNRR
- Partner company: SIMAR GROUP s.r.l., Monte Urano (FM)

Research field of interest



Enhancing efficiency of Neural Networks for **Edge Deployment**

Sustainability for Large Language Models

Post-hoc Capability Extension for Small Models

Summary of study activities

- Ad hoc PhD courses
  - Innovation and Entrepreneurship
- PhD Schools
  - International Computer Vision Summer School "Computer Vision in the Age of Large Language Models"
- Conferences attended
  - 4th CINI National Conference on Artificial Intelligence Ital-IA 2024, Naples, Italy, 29/05/24-30/05/24

Summary of study activities

	Courses	Seminars	Research	Total
Bimonthly Period I	0	0.2	9.8	10
Bimonthly Period II	0	1	9	10
Bimonthly Period III	0	0.4	9.6	10
Bimonthly Period IV	0	1.4	8.6	10
Bimonthly Period V	11	0	0	11
Bimonthly Period VI	0	1.7	8	9.7
Total	**11**	**5.1**	**45**	**61.1**

Research activity I

Objective To make large language models sustainable, accessible, and fair

Problem How to reduce the computational footprint of large language models?

Methodology

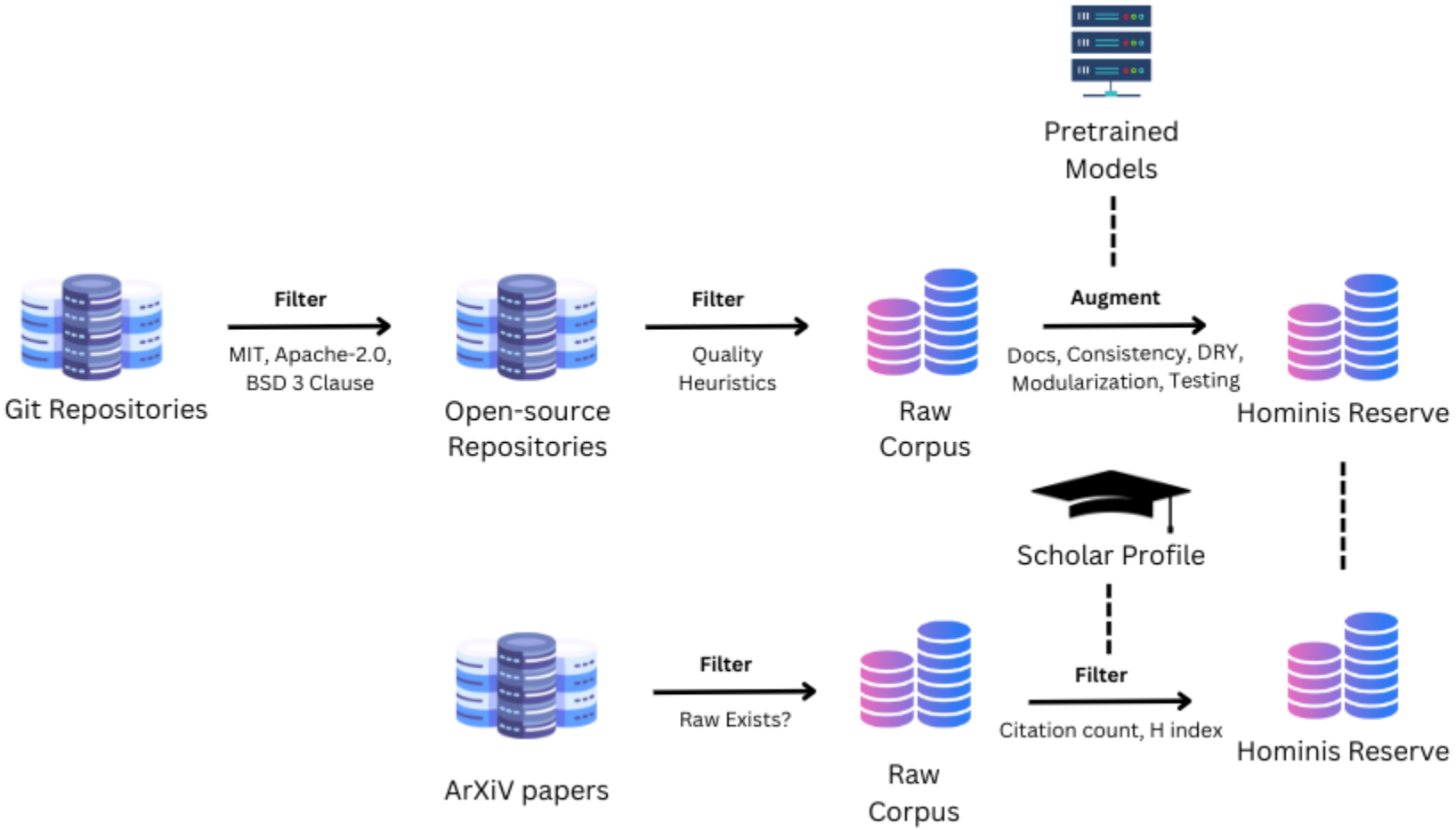
Curation Strategies → **Ablation Study** → **Resource Aware Scaling**

- Surveyed permissive datasets with available curation metrics - **RedPajama v2**
- Scraped high-quality permissive resources (GitHub, ArXiV) and conducted heuristic filtering, utilized permissive models to generate instruction-completion pairs - **Hominis Reserve**

- Identified the maximum size that can be trained efficiently - **13B**
- Conducted ablation over efficient attention variants - **FlashAttention**
- Conducted ablation over linear vs MOE - **MOE suitable for PostHoc**

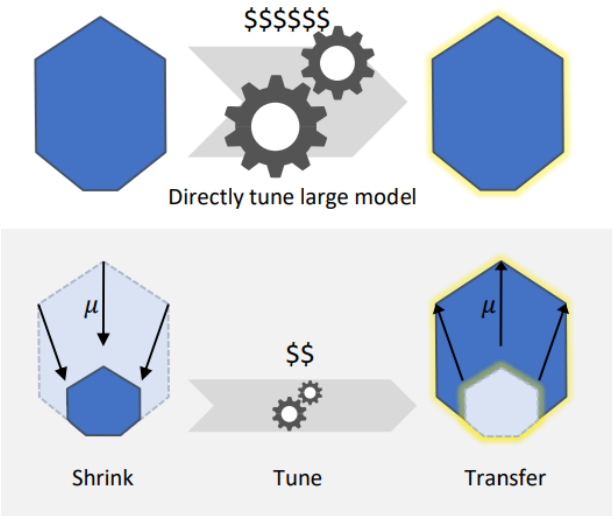
- Carried out MuTransfer for hyperparameter tuning
- Adopted LoRA the Explorer for efficient training over distributed GPUs.

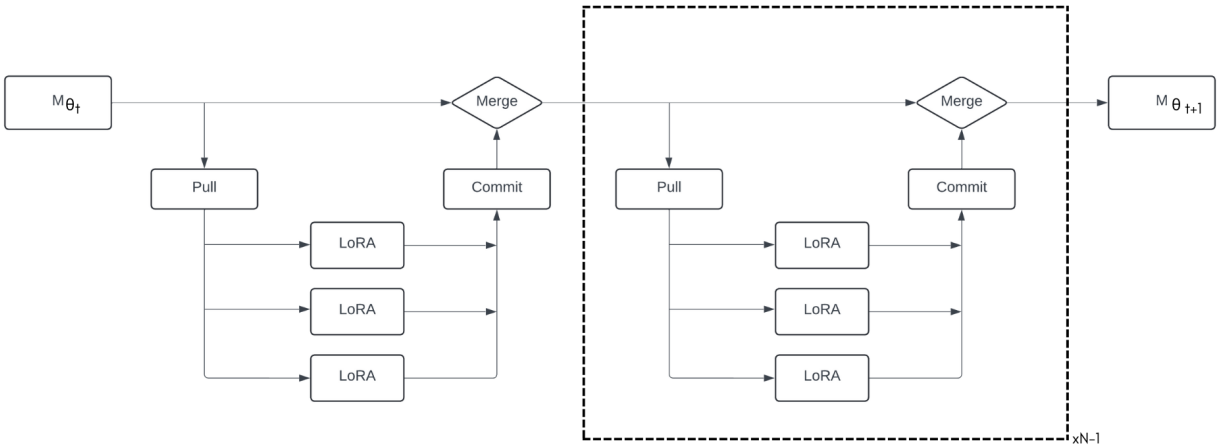
Sponsored by ISCRA-B compute grant by CINECA - Hominis Core & Instruct 15B models were trained using the 1M compute hours grant received to be used on Leonardo cluster



Curation process for Hominis Reserved Data - Used for Instruct Tuning

Resource-aware Scaling





We employ **μTransfer***, a technique that determines the hyperparameters for a large scale model by conducting hyperparameter sweeps on a much smaller model.

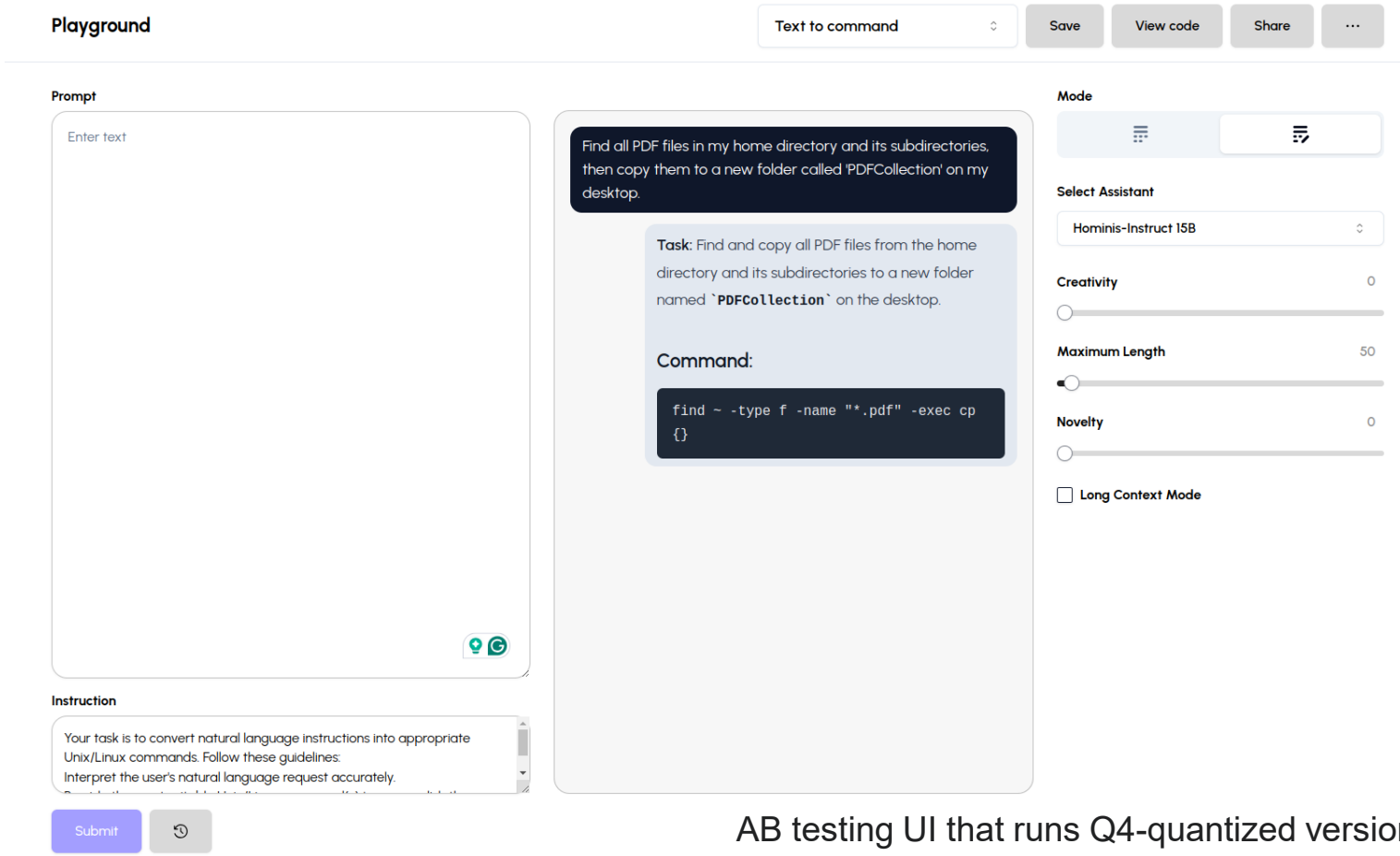
We initialize per GPU LoRA (Low-Rank Approximation) matrices and train them independently based on sharded data. We commit them together and merge into the base weights on a per mini-epoch basis.

Hominis Core Results

Benchmark Category	Dataset	Llama 2 13B	Hominis Core 13B v0504	Hominis Core 13B v1004	Hominis Core 13B v1504
Code Evaluation	MBPP	30.6	20.8	*33.0*	*36.5*
	HumanEval	18.3	12.4	*20.8*	*20.2*
Math Evaluation	GSM8k	28.7	20.4	*28.9*	*28.7*
	MATH	3.9	*4.5*	*11.2*	*13.5*
World Knowledge	NaturalQuestions 0-shot	16.1	16.1	*16.3*	*16.4*
	TriviaQA 0-shot	73.1	65.8	67.2	*73.7*
Carbon Emission (MTon)		539	11.38		

Why Llama 2? - Llama 2 is the last known pretraining-only model. The future versions are amalgamation of instruct tuned and raw pretrained.

Hominis Playground



AB testing UI that runs Q4-quantized version of Hominis models in a containerized manner on consumer level hardware. We aim to open-source parts of this.

Ongoing Work

- Applied for ISCRA-B grant extension to evaluate Hominis Instruct on LLM-Harness
- User preference (AB) testing against commercial models (ChatGPT/Claude/Gemini API) is currently underway on hominis.io/playground
- Hominis Italia, a localized version trained on Italian-only subset is undergoing training (till 31st Oct Leonardo, currently on RealAI)

Research activity II

Objectives To enable multimodal understanding and generation in models without significant training

Problem How to extend text-only models to other modalities?

Methodology

Virtual Machine Design

- Define DSL over the external modality understanding and generation.
- Provide API documentation and use few-shot prompting to generate code
- Execute resultant code within a container and fetch results

Constrained Function-Calling

- Use API documentation and few-shot prompting to directly call external functions
- Use grammar based sampling to constrain output to JSON schema

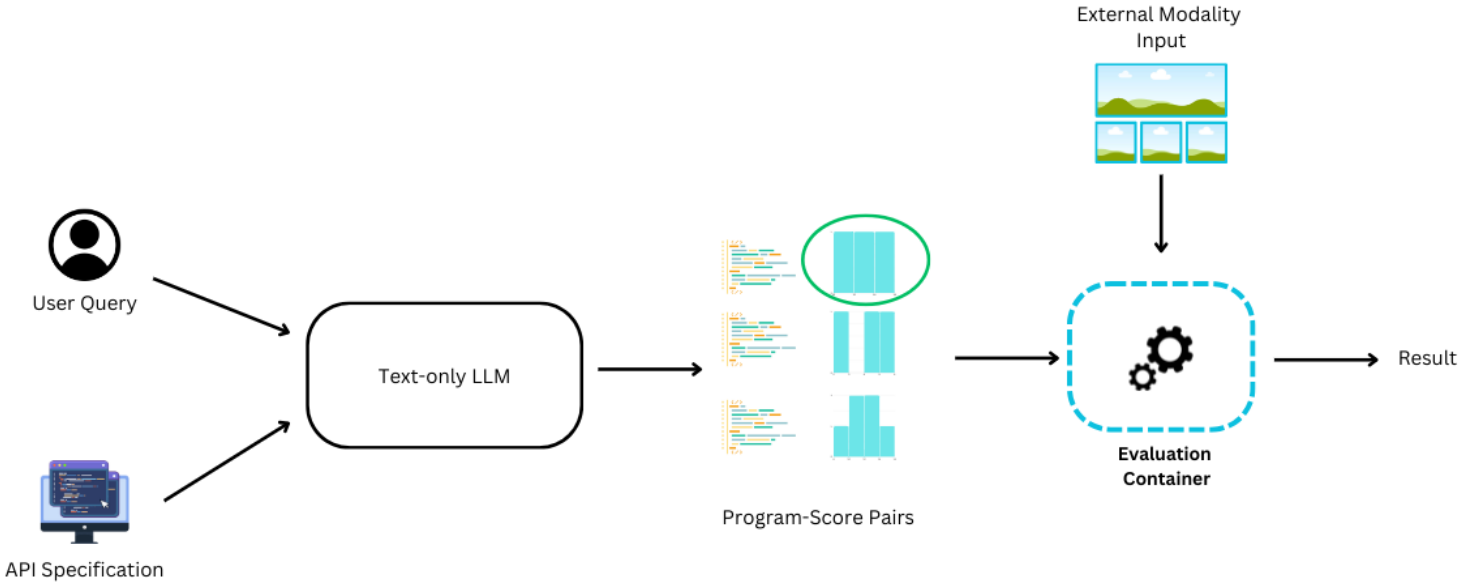
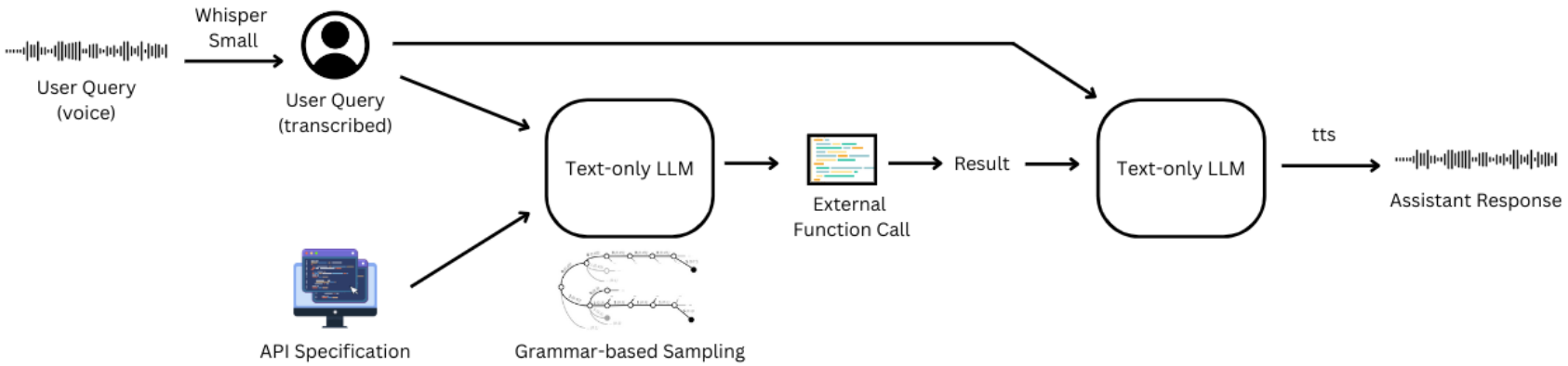


Image Understanding through VM: Collaboration with RealAI for Hominis Playground



Function Calling Assistant through model chaining: Prototype for SIMAR smart chair

Products

[C1]	Amato, F., Benfenati, D., Cirillo, E., De Filippis, G.M., Fonisto, M., Galli, A., Marrone, S., Marassi, L., Moscato, V., **Patwardhan, N.** and Moccardi, A. *"Advancements and Challenges in Generative AI: Architectures, Applications, and Ethical Implications."* Convegno Nazionale CINI sull'Intelligenza Artificiale, Ital-IA 2024. (Published)
[P1]	Narendra Patwardhan, Lidia Marassi, Tarry Singh, Stefano Marrone, and Carlo Sansone. "*Training Human-Centric Foundation Models*" ACM Transactions on Intelligent Systems and Technology: Special Issue on Transformers (Under Revision)

Next Year

Location	Duration	Intended Work
SIMAR group, Monte Urano	~5 months	Manage the next iteration of smart-chair prototype and integrate voice-to-instruction capabilities in it.
RealAI, The Netherlands	6 months	Exploring posthoc modifications & distillation of Hominis foundation models