VILLANOVA

Modelli Linguistici su Misura per l'Europa

Andrea Zugarini, Senior Researcher

Leonardo Rigutini, Scientific Advisor

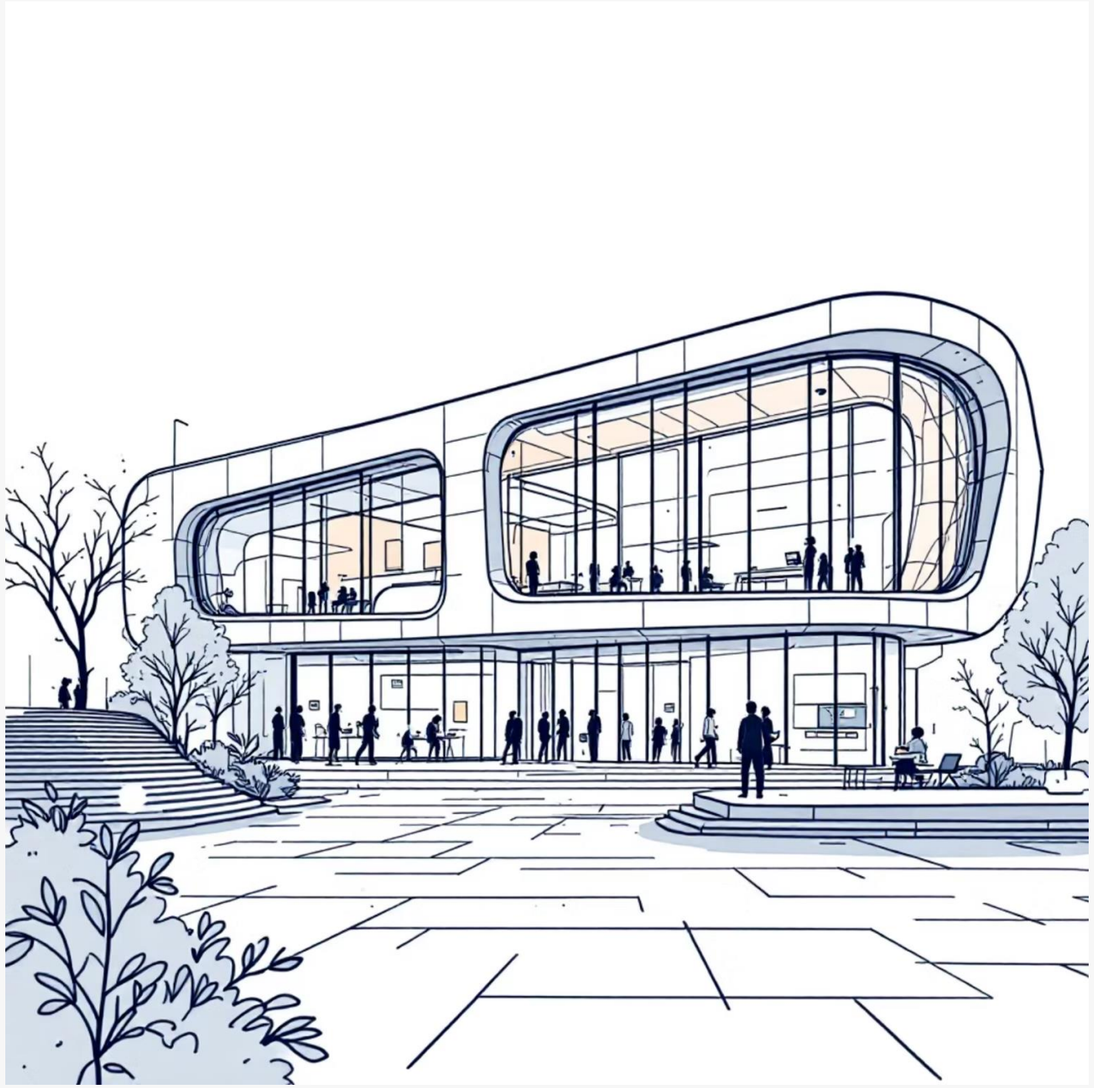
expert.ai

Vision

European AI

The Villanova project represents an ambitious initiative that aims to revolutionize the field of European artificial intelligence through the research, development, testing, and implementation of innovative solutions specifically designed for the European context.

The main goal is to create multimodal generative AI models that can produce high-quality text and multimedia content in real time, integrated with advanced time-series forecasting capabilities.



Two Work Packages

WP1: Multimodal Generative AI Models

Development of Large Language Models (LLMs) specifically trained for five European languages and five domains, integrated with time-series forecasting algorithms for composable smart applications based on cloud-edge architecture.

WP2: Open-Source Low-Code/No-Code Framework

Create a user-friendly, Generative AI-Driven open-source framework to facilitate collaboration between humans and AI, simplifying the creation and maintenance of advanced applications.

By providing reusable components, this approach democratizes AI, making it accessible to a broader range of users, including non-technical individuals and small businesses with limited budgets.

Use Cases

Five use cases designed to deploy the developed technologies of the project in **heterogeneous sectors and scenarios**.

TOURISM

Personalized Smart Tourist Guide

to enhance the tourist experience with tailored recommendations

LEGAL

AI-Assisted Legal Document Generation

to simplify drafting processes for legal professionals

AGRICULTURE

Cloud Platform for Farmers

to optimize crop yield and quality through data-driven insights

Public Administration

Digital Experience Platform (DXP) for Citizens

to streamline access to public administration services

Big Data Analytics and Knowledge Extraction

Data Analysis System for Public Administration

to improve decision-making and delivery of public services

WP1 - Objectives

A family of **multilingual**, **multimodal LLMs** built according to the principles of the AI act.

- Focus on five European languages: 
- Not just text, also support for multimodal inputs and outputs. Modalities: **images**, **audio** and **video**
- Integration of LLMs with time-series forecasting models

A **federated fine-tuning pipeline** to adapt models to specific domains and tasks. The framework should:

- Allow collaborative learning across different resources
- Maintain data privacy
- Favor the specialization of smaller, more **efficient** models to deploy applications on edge nodes

Application of Villanova models and fine-tuning pipeline for the **five use-cases**.

WP1 Development Roadmap

1

Distributed Data Collection

Collection and pre-processing of multimodal datasets (textual, multimedia, time series) with a distributed architecture

2

General Training

Developing general-purpose multimodal and multilingual models using advanced Transformer

3

Federated Fine-Tuning

Pipeline implementation for vertical model adaptation while maintaining data privacy through collaborative learning

4

Component Integration

Developing reusable components for composable applications with user-friendly interfaces and integrated ethical considerations

WP1 - Ongoing Activities

Data collection, pre-processing and annotation

Pre-training. Collection of ***4-8T tokens*** from web corpora (**FineWeb**, **FineWeb-2**,..) already filtered and selected based on content quality. Multilingual distribution skewed toward the five languages. Selection of curated data **0.4T-1T tokens** of high quality corpora for latest pre-training stages.

Post-training. Curated selection of Supervised-Fine-Tuning (SFT) and Preference Tuning data. Content generated by proprietary LLMs (such as chatGPT) is **excluded**. **Synthetic generation** of **multilingual** content with apache 2.0 licensed models, with human in the loop to validate and curate examples.

Multimodal data. Curated selection of existing data. Currently focussing on **images**. Most of the existing datasets are in English. We are creating synthetic multilingual and multimodal corpora.

Domain-specific content. Topic classification of documents from web corpora to identify content for the five use-cases for LLMs steering/fine-tuning. Definition of **vertical tasks** and data collection for each domain.

Evaluation and benchmarking. Multilingual evaluation of each training phase. Existing benchmarks extended for multiple languages. Dedicated benchmarks to assess domain-specific knowledge and instruction-following capabilities.

WP1 - Ongoing Activities

LLMs Trainings

Pre-training. Release of **base models** with different sizes: **2B, 7B, >13B**.

Post-training. Development of aligned versions of the base models via *instruction-tuning*, and *RL learning*.

Multimodal VLMs. Vision Language Models combining base models and pre-trained vision encoders.

Experimental activities on:

- Language and datasets distributions
- Pre-training stages and recipes
- Model architecture (tokenizers, MoEs)

Estimated **~1M GPU hours** for training, fine-tuning and experimental activities on LLMs.

WP1 - Ongoing Activities

Federated Fine-tuning Pipeline

Development of a pipeline to specialize models to specific **domains** and **tasks**.

Dedicated algorithms to allow **federated** and **distributed** learning.

Experimental activities on:

- **Continual Learning** and adaptation to vertical domains.
- Federated fine-tuning strategies and algorithms.

WP1 - Infrastructure

An HPC cluster with **8 physical nodes** hosted in Portugal.

Each node has:

- **8 x H100** with 80GB VRAM each
- 2TB RAM
- 3.2TB disk
- Infiniband connection

A shared storage of 130TB.

Cluster will increase up to 14 nodes in the next two months.

Another cluster is being hosted in Verona.

WP1 and AI Act

The AI Act influences much of the language model creation lifecycle.

Transparency

All data used for training must be formally certified according to European standards, guaranteeing origin, quality, and regulatory compliance.

Traceability

Implementation of end-to-end traceability systems for each dataset, allowing the complete reconstruction of the data path from source to end use.

Licenses and copyrights

Attention to licences and copyrights. Content generated by proprietary LLMs (such as chatGPT) is excluded in training.

Privacy and Security

Implementation of data protection measures compliant with GDPR and European privacy regulations, ensuring security and anonymization.

Progress status

- Infrastructure setup for pre-training
- Multi-node pre-training framework
- Synthetic data generation pipeline for SFT and multimodal examples
- Pre-training recipes tested on 2B LLMs
- Federated fine-tuning pipeline under development

Thank you for your attention