Foundational Vision-LLM for AI Linkage and Orchestration

KHAN Zaid, KUMAR B G Vijay, SCHULTER Samuel, CHANDRAKER Manmohan

Abstract

We propose a vision-LLM framework for automating development and deployment of computer vision solutions for pre-defined or custom-defined tasks. A foundational layer is proposed with a code-LLM AI orchestrator self-trained with reinforcement learning to create Python code based on its understanding of a novel user-defined task, together with APIs, documentation and usage notes of existing task-specific AI models. Zero-shot abilities in specific domains are obtained through foundational vision-language models trained at a low compute expense leveraging existing computer vision models and datasets. An engine layer is proposed which comprises of several task-specific vision-language engines which can be compositionally utilized. An application-specific layer is proposed to improve performance in customer-specific scenarios, using novel LLM-guided data augmentation and question decomposition, besides standard fine-tuning tools. We demonstrate a range of applications including visual AI assistance, visual conversation, law enforcement, mobility, medical image reasoning and remote sensing.

Keywords

Computer vision, foundational model, agentic LLM, orchestration, AI linkage, reinforced self-training, visual assistance

1. Introduction

Computer vision is a key technology for NEC in a wide range of applications across health, finance, retail, mobility, remote sensing and safety. Our broad aim is to enable a dual strategy for NEC:

- To build defensive moats around AI businesses through strategically important foundational models.
- To build an aggressive toolkit that accelerates and diversifies impact in target application domains. This article outlines a foundational vision-LLM architecture

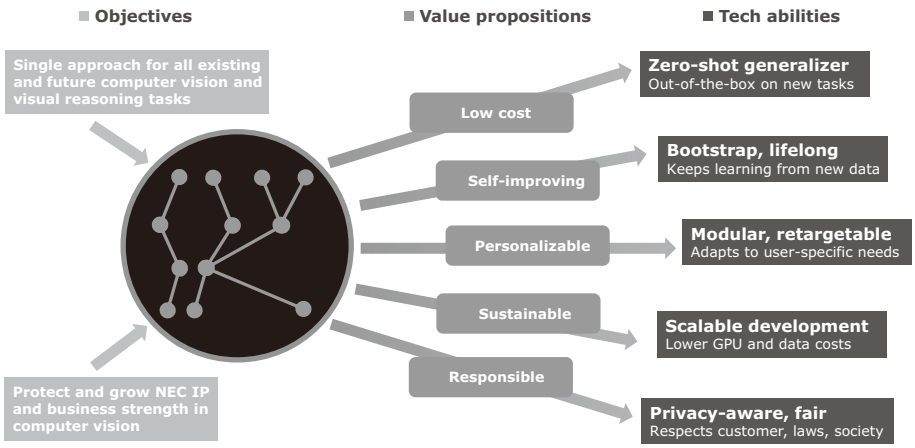

Fig. 1 Mission statements for creating our foundational vision-LLM.

to realize those aims.

Consider a typical computer vision solution, which requires customized effort through a workflow resembling the following:

(1) Customer explains their need in natural language or with examples.
(2) An engineer writes code based on available models, libraries and literature to solve the task.
(3) The deployment team tunes the solution to customer environment. We propose a vision-LLM that acts as an agent who understands new tasks and generates appropriate code, which then invokes existing engines and APIs to solve the given task. This design thereby unifies and leverages all NEC know-how on computer vision to solve any pre- or custom-defined visual task.

A vision-LLM requires different considerations from a traditional LLM, since visual data is not amenable to long-range reasoning, self-supervision is challenging and alignment with language is non-trivial. Our goal is to develop a low-cost, self-improving, personalizable, sustainable and responsible vision-LLM (**Fig. 1**). A key philosophy is for our vision-LLM to achieve a high level of physical grounding at minimal training cost, which we realize through design choices such as the agent-engine layers, as well as the use of code-LLMs and pretrained computer vision models. Our vision-LLM is distinct from a multimodal LLM like GPT-4V[1)], where our layered approach is more modular and efficient.

2. Summary of Architecture

A summary of our architecture is shown in **Fig. 2**. Our framework is comprised of three layers. First, a foundational layer with an LLM orchestrator that plans based on available code and documentation how to solve a new task. This layer also consists of domain-specific foundational models that are trained on very large datasets with zero-shot vision-language (VL) generalization ability. Second, an engine layer with VL models for specific tasks like image retrieval or object detection. Third, an application layer where tools like augmentation, decomposition, prompting and fine-tuning are available to adapt to customer-specific data or usage. We posit that such an approach allows for both competitive differentiation and market penetration.

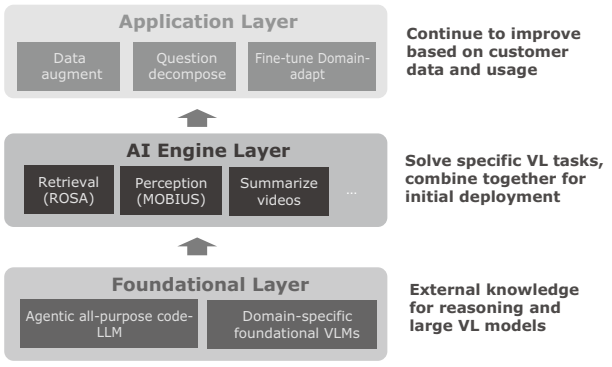

Fig. 2 Architecture of our vision-LLM.

3. Tech Details and Results

We outline how to realize the above architecture and key benchmark results.

3.1 Foundational layer

Our framework is based on an LLM orchestrator that generates a plan to accomplish new tasks using available tools, as well as large pre-trained domain-specific foundational models.

3.1.1 Agentic Vision-LLM Orchestrator

Solving a complex, novel task requires interleaving multiple steps of reasoning and perception, which must be composed with planning, backtracking, and sequential decisions. Such tasks are the next frontier challenge for intelligent systems, which we approach through foundational model-driven *AI agents*. We propose an LLM orchestrator that is given access to a battery of available tools and pretrained models, much like a human engineer. This allows the agent to solve tasks beyond the ability of the underlying LLM. For example, while LLMs may be prone to logical inconsistencies and poor arithmetic, an LLM-driven agent can delegate logical reasoning to a logic engine and arithmetic to a calculator. By allowing an agent to synthesize programs and invoke APIs,

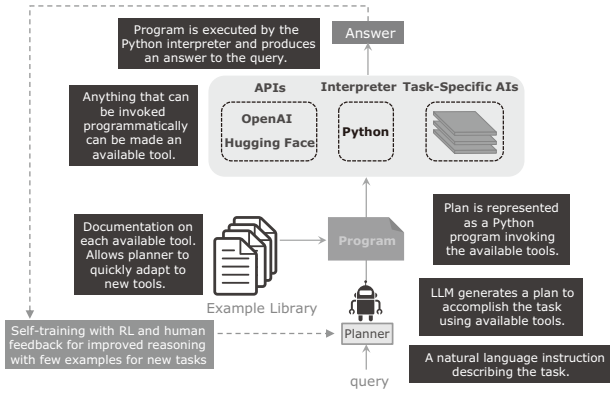

Fig. 3 Agentic Vision-LLM Orchestrator.

we can combine arbitrary tools to solve novel tasks.

At the core of the agentic architecture (**Fig. 3**) is an LLM that acts as a planner. Given a natural language instruction, the goal of the planner is to write a plan that accomplishes the task using available tools. The plan is represented as a formal program that invokes the available tools. To understand what tools are available and how it can use those tools, the planner consults a library of documentation and examples. This allows the planner to quickly adapt when new tools are added by reading their documentation. In principle, anything that can be programmatically invoked can be used as a tool by the planner using the code interpreter. As a starting point, we provide an environment that has access to task-specific AIs and third-party APIs. The plan (represented as a program) is then executed in the environment to produce an answer to the query.

A problem with frozen off-the-shelf LLMs as planners is that they lack experience with writing plans and can fail to understand nuances of tool use from documentation alone. But training an LLM to act as a planner requires training data, but no large-scale training data is available for writing programs that solve visual tasks. Our key insight is learning from feedback using reinforcement learning. We first design an environment in which the planner can write and execute programs. We provide the planner with an API through which it can invoke state of the art task-specific models. We then apply iterated reinforced self-training by using existing annotations for a vision-language task (**Fig. 4**). For example, given a dataset of image v, ground truth y and query q, we feed q into the planner, then run the generated program p on the image v. We compare the result of executing the program $\hat{y}$ with the ground-truth y to obtain a coarse reward signal, then apply a reward-weighted behavioral cloning loss. The trained planner outperforms a frozen planner based on ChatGPT by as much as 10%, 4%, and 10% on compositional variants of question answering, object detection, and image-text matching.

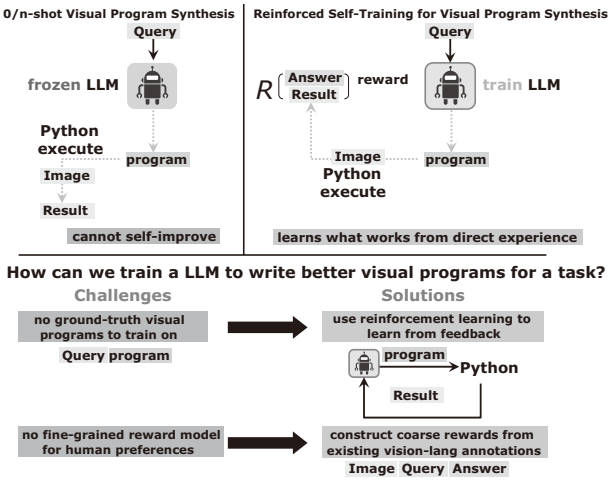

Fig. 4 More capable LLM planners who applied enhanced self-training.

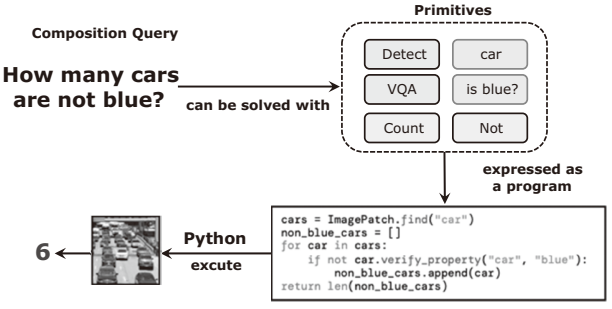

Fig. 5 An example of a generated program to solve a complex, novel visual task.

As an example for computer vision, consider a novel visual task that is difficult for end-to-end systems (**Fig. 5**). It can be solved by decomposition into primitive visual tasks (object detection, image-text matching) and logic, for which task-specific engines exist. The planner writes a Python program that controls the task-specific AIs through an API we provide to obtain necessary intermediate information about the image, then combines the acquired information with logic expressed in code to arrive at an answer.

Our AI orchestrator has many benefits:

- Solves existing or novel task specifications as composition of available vision modules, with automatically generated Python code.
- Planner trained with reinforcement learning for improved understanding of how to use available tools relative to a frozen planner.
- Human feedback for self-training for improved task reasoning, even with very few program examples.
- Efficient use of code-LLM with fewer parameters than general LLM.
- Parameter-tuning and data augmentation can be automatically handled.

3.1.2 Domain Foundational VL-Model

Our architecture can incorporate any existing vision model, but sometimes one may not exist in our library. Thus, we also propose foundational VL models (FVLM) in specific domains trained with a large amount of data to easily generalize to new tasks. Our key insight is that many task-specific models and datasets already exist,

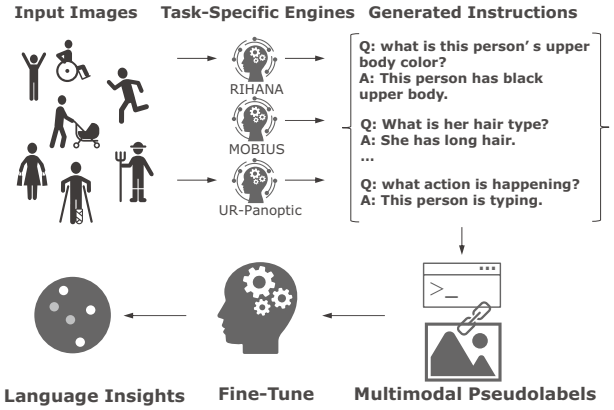

Fig. 6 We utilize existing vision data and models to develop domain-specific FVLM at low cost.

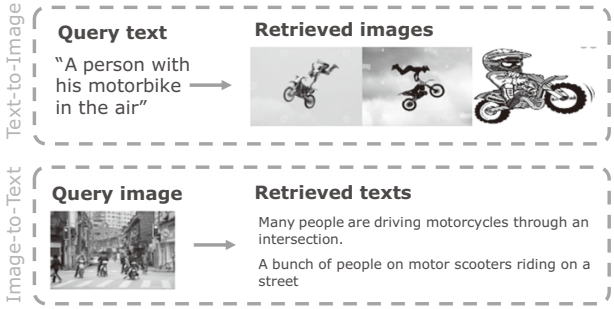

Fig. 7 The tight image-text alignment in ROSA enables accurate multi-modal retrieval.

which can be leveraged to train FVLM at low compute expense.

Examples of domains where we develop an FVLM are mobility and human analysis. Our mobility FVLM is developed using several large-scale autonomous driving datasets, along with the outputs of several object detection, segmentation, captioning and other models applied to them. Our human FVLM is trained using a collection of datasets and models for human attribute analysis, action recognition and human-object interaction. **Fig. 6** shows the overall pipeline for training our domain FVLM. Despite being significantly smaller in size (7B parameters), this new model improves performance by 0.5% compared to an existing closed model with 175B parameters.

3.2 Engine layer

The engine layer is comprised of a large zoo of AI models, standard tools such as finetuning and domain adaptation, together with documentation and usage examples, which have been developed for specific tasks

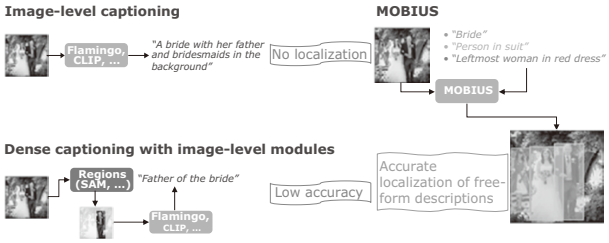

Fig. 8 Our open vocabulary detector MOBIUS provides accurate localization of rare object categories, with language descriptions.

like image retrieval, object detection, medical imaging, or remote sensing.

3.2.1 Vision-language retrieval

Our proposed model ROSA[2] is a data-efficient neural network that effectively aligns image and text modalities, which enables accurate image-to-text and text-to-image retrieval. In a zero-shot evaluation benchmark (**Fig. 7**), our model outperforms the state-of-the-art by 3% in Rank-1 text retrieval performance although competing models use 30 times more compute and require 100 times larger training datasets.

3.2.2 Open-world scene understanding

Another engine is our open-vocabulary object detector MOBIUS[3)4)] (**Fig. 8**), which can localize rare categories and objects described by free-form text descriptions. On a public open-vocabulary benchmark, where detectors are tasked to detect unseen categories without box annotations during training, MOBIUS outperforms the competition by 4.3% average precision (AP) points.

3.3 Application layer

While our foundational and engine layers already enable deployment of solutions in customer domains, the application layer will provide tools to achieve personalized solutions based on target data and usage. Specifically, we propose an approach to leverage a small amount of target data and an approach to decompose application-specific usage into easier to reason atomic segments.

3.3.1 Data augmentation

There is often insufficient data available for specialized tasks or domains. While collecting more annotations can be challenging, unlabeled images are often available. We

propose SelTDA[5] (Self-Taught Data Augmentation), a strategy which uses a generative vision-language model to produce synthetic data. Training with SelTDA results in improvements of up to 9.87%, 6.38%, and 29.81% on robustness, generalization, and reasoning respectively.

3.3.2 Question decomposition

Domain-specific tasks or those with specialized reasoning patterns are challenging for generalist models, especially in realistic data-scarce settings. We propose selective question decomposition[6] to improve generalist models on such tasks through dialogue that surfaces relevant context without extra data. We achieve improvements of up to 26% on medical datasets and correct up to 20% of prediction errors using decompositions.

4. Applications

We now discuss several applications for our foundational vision-LLM.

AI Visual Assistant: The vision-LLM yields outputs grounded in images or videos, while the orchestration through code allows a transparent step-by-step reasoning (**Fig. 9**). This can be used to aid the visually impaired, allowing them to find objects, navigate safely or perform other tasks with a language interface. This application is being validated with an NEC-X startup.

Visual Conversation: The vision-LLM enables joint reasoning with images and text along with external knowledge, which allows question-answering or conversation in multimodal data. Our application layer tools allow improvement of *13%* on the public OK-VQA benchmark for external knowledge-based language reasoning in images.

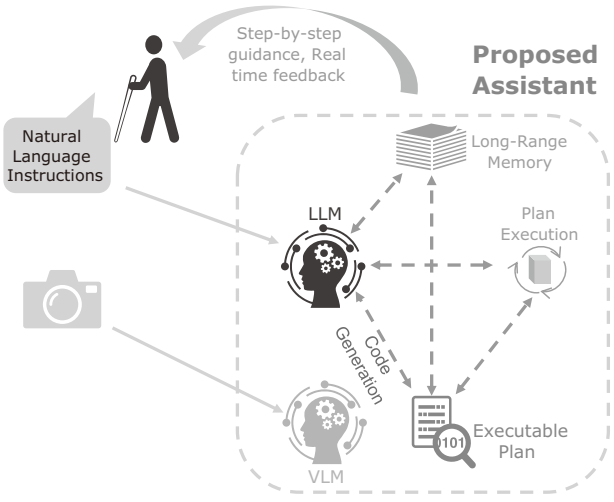

Fig. 9 Proposed visual AI assistant for NEC-X.

An example use-case is an NEC-X startup which seeks to improve the engagement of social media influencers with their audience by automating some interactions (**Fig. 10**).

Law enforcement: The VL-retrieval engine has been used for tattoo recognition for law enforcement, which allows a human operator to interpret tattoos and find matches based on semantic meaning beyond visual similarity (**Fig. 11**).

Mobility: A domain-specific FVLM for mobility scenarios has been demonstrated for automated insurance insights and summarization together with the VIR Lab at OH 2023 (**Fig. 12**).

Remote Sensing: The data augmentation methods in our application layer allow a foundational VLM to answer questions in satellite images even with a small amount

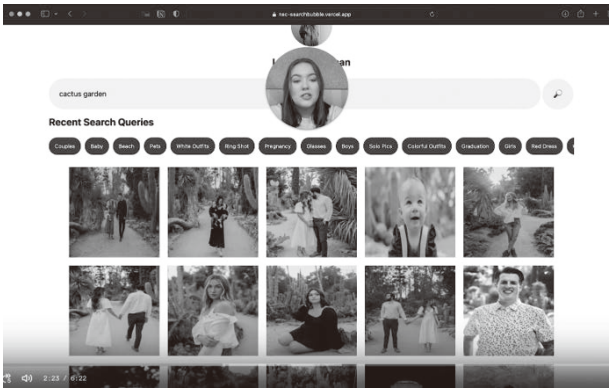

Fig. 10 Usage in an NEC-X VL-search app.

Fig. 11 Language-based tattoo analysis system.

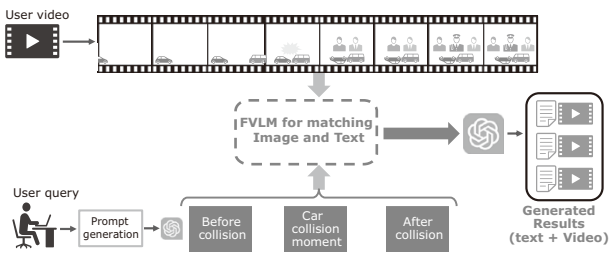

Fig. 12 Our domain-specific FVLM used for video summarization for a VIR Lab application.

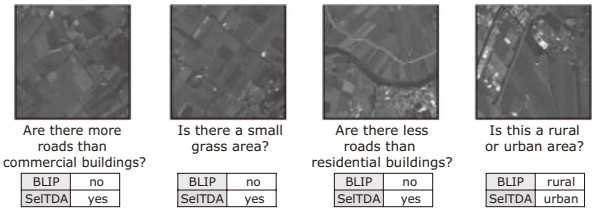

Fig. 13 Examples for remote sensing.

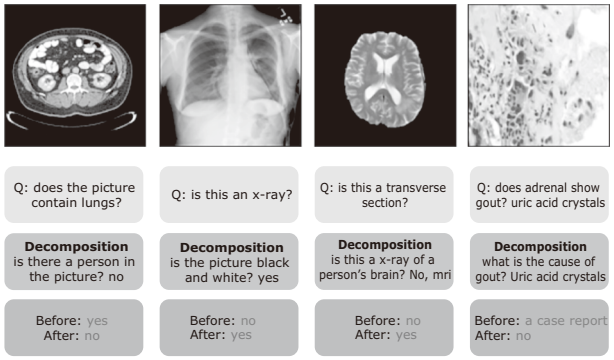

Fig. 14 Examples for medical image reasoning.

of data. We improve over BLIP on the RS-VQA benchmark (**Fig. 13**).

Medical Imaging: The question decomposition strategy in our application layer improves a generalist VLLM on QA over medical images, a data-scarce, domain-specific specialized task. In the public PathVQA, SLAKE and VQA-Rad benchmarks, we obtain improvements of 22%, 10% and 26% (**Fig. 14**).

5. Conclusion and Next Steps

We showcased an architecture for a foundational vision-LLM that will automate development and deployment of computer vision solutions by understanding customer tasks, then developing code to solve them using external knowledge and available resources. This will be supported by developing new FVLM to solve tasks in specific domains, as well as tools to rapidly customize in specific applications. Several next steps are being developed, including: (a) automatic tuning of parameters for deployment, (b) self-training to update task-specific models based on application rewards, (c) reducing hallucination and biases.

* ChatGPT is a trademark of OpenAI, Inc. in the United States.

* All other company names and product names that appear in this paper are trademarks or registered trademarks of their respective companies.

References

1) OpenAI: GPT-4 Technical Report,2023
https://arxiv.org/abs/2303.08774
2) Zaid Khan, Vijay Kumar BG, Xiang Yu, Samuel Schulter, Manmohan Chandraker and Yun Fu: Single-Stream Multi-Level Alignment for Vision-Language Pretraining”., European Conference on Computer Vision, 2022
https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136960725.pdf
3) Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Vijay Kumar B.G, Yumin Suh, Manmohan Chandraker and Dimitris N. Metaxas: Taming Self-Training for Open-Vocabulary Object Detection, 2023
https://arxiv.org/abs/2308.06412
4) Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B.G, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas: Exploiting Unlabeled Data with Vision and Language Models for Object Detection, 2022, European Conference on Computer Vision
https://arxiv.org/abs/2207.08954
5) Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu and Manmohan Chandraker: Q: How to Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images!, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.15005-15015, 2023
https://www.computer.org/csdl/proceedings-article/cvpr/2023/012900p5005/1POQzQuyw1O
6) Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker and Yun Fu: Exploring Question Decomposition for Zero-Shot VQA,” Conference on Neural Information Processing Systems, 2023
https://arxiv.org/abs/2310.17050

Authors' Profiles

KHAN Zaid
Northeastern University

KUMAR B G Vijay
Researcher
NEC Laboratories America

SCHULTER Samuel
Senior Researcher
NEC Laboratories America

CHANDRAKER Manmohan
Department Head
NEC Laboratories America
ProfessorUniversity of California San Diego

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.17 No.2 Special Issue on Revolutionizing Business Practices with Generative AI
– Advancing the Societal Adoption of AI with the Support of Generative AI Technologies

Remarks for Special Issue on Revolutionizing Business Practices with Generative AI
Approaches to Generative AI Technology: From Foundational Technologies to Application Development and Guideline Creation

Vol.17 No.2
June 2024

Special Issue TOP

Papers for Special Issue

Market Application of Rapidly Spreading Generative AI
NEC Innovation Day 2023: NEC's Generative AI Initiatives
Streamlining Doctors' Work by Assisting with Medical Recording and Documentation
Using Video Recognition AI x LLM to Automate the Creation of Reports
Understanding of Behaviors in Real World through Video Analysis and Generative AI
Automated Generation of Cyber Threat Intelligence
NEC Generative AI Service (NGS) Promoting Internal Use of Generative AI
Utilization of Generative AI for Software and System Development
LLMs and MI Bring Innovation to Material Development Platforms
Disaster Damage Assessment Using LLMs and Image Analysis

Fundamental Technologies that Enhance the Potential of Generative AI
NEC's LLM with Superior Japanese Language Proficiency
NEC's AI Supercomputer: One of the Largest in Japan to Support Generative AI
Towards Safer Large Language Models (LLMs)
Federated Learning Technology that Enables Collaboration While Keeping Data Confidential and its Applicability to LLMs
Large Language Models (LLMs) Enable Few-Shot Clustering
Knowledge-enhanced Prompt Learning for Open-domain Commonsense Reasoning
Foundational Vision-LLM for AI Linkage and Orchestration
Optimizing LLM API usage costs with novel query-aware reduction of relevant enterprise data

For AI Technology to Penetrate Society
Movements in AI Standardization and Rule Making and NEC Initiatives
NEC's Initiatives on AI Governance toward Respecting Human Rights
Case Study of Human Resources Development for AI Risk Management Using RCModel

NEC Information

2023 C&C Prize Ceremony