

eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

# HiTZ

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology

German Rigau Claramunt  
HiTZ Deputy Director  
National Coordinator of CLARIN and DARIAH

**hitz.eus**

# Artificial Intelligence



# Artificial Intelligence



# Artificial Intelligence

## HAL's Legacy 2001's Computer as Dream and Reality

Edited by [David G. Stork](#)

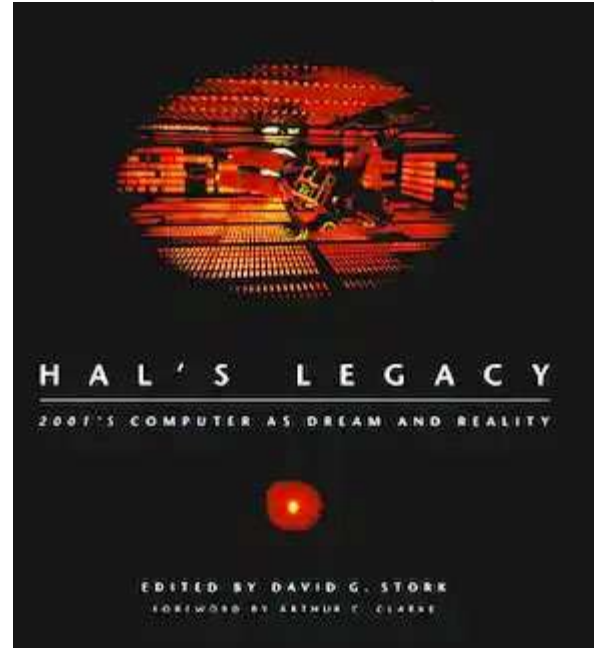
Foreword by [Arthur C. Clarke](#)

384 pp., 8 x 9 in, Paperback

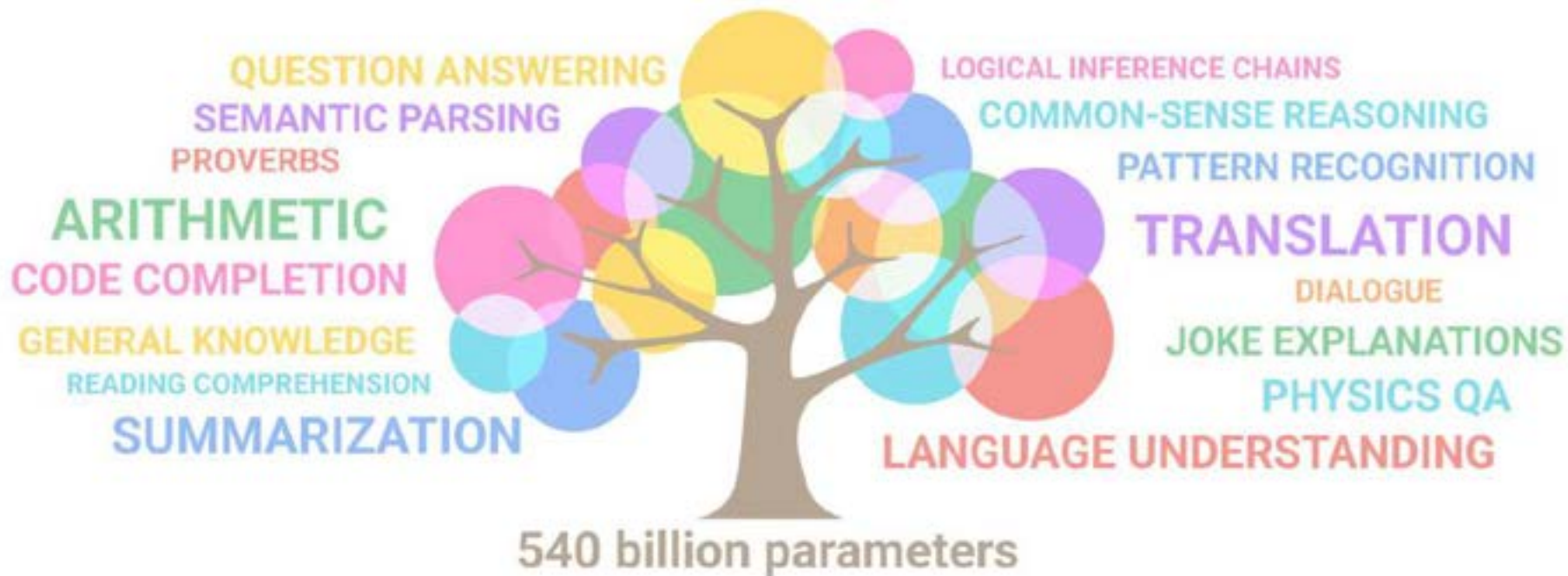
ISBN 9780262692113

Published: March 2, 1998

Publisher: The MIT Press



# Artificial Intelligence



**PaLM: Scaling Language Modeling with Pathways (2022)**  
Aakanksha Chowdhery, + 67 authors

# Preface

“Cuando creíamos que teníamos todas las respuestas, de pronto, cambiaron todas las preguntas.”

- Mario Benedetti

*“Just when we thought we had all the answers, suddenly all the questions changed.”*





1

**HiTZ**

# HiTZ Basque Center for Language Technology

- [Ixa](#) (founded 1988) and [Aholab](#) (1998)
- 88 members, multidisciplinary
- Yearly income in projects: ~3M €
- **Chair in Artificial Technology and Language Technology**

## Reference center in NLP/AI research

- Research & Applications
  - on Basque, English, Spanish and other languages
- [Academy](#): Master, Doctorate, [courses](#)
- Consultants for [Basque](#), [Spanish](#) and [European](#) LT plans
- Large LT projects
  - [ILENIA](#), [IKER-GAITU](#)
- National coordinators of two European research infrastructures
  - [CLARIAH-ES](#): [CLARIN](#) & [DARIAH](#)

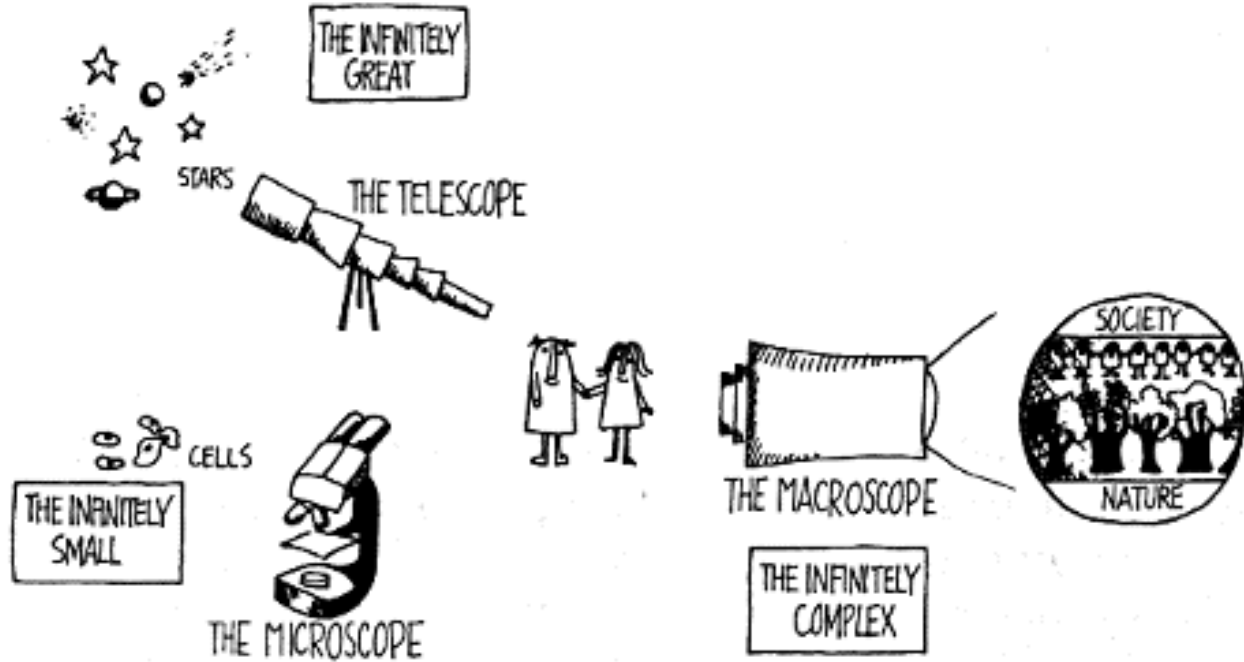




**2**

**The  
research  
infrastructure**

# The Macroscope



De Rosnay, J. [The macroscope](#), Harper & Row, NY, 1979

# ESFRI – EOSC – SSHOC



**EUROPEAN OPEN  
SCIENCE CLOUD**



**#SSHOCaVocabulary**  
**#SSHOCifyCLARIN**

○ Social Sciences  
○ Humanities

[CESSDA](#) ERIC - Consortium of European Social Science Data Archives

[CLARIN](#) ERIC - Common Language Resources and Technology Infrastructure

[DARIAH](#) ERIC - Digital Research Infrastructure for the Arts and Humanities

[ESS](#) ERIC - European Social Survey

[SHARE](#) ERIC - Survey of Health, Aging and Retirement in Europe

[E-RIHS](#) - European RI for Heritage Science - [CSIC](#), [CENIEH](#)

[OPERAS](#) - European RI for the open scholarly communication in the SSH.

...

# 2 CLARIN K-Centres and DESIR



## Spanish K-Centre

- UPF, UNED, UVigo, UPV/EHU

## IMPACT

- UA

## DESIR

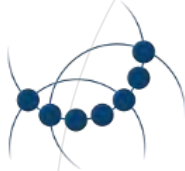
- UNED



# CLARIN-ERIC & DARIAH-ERIC



CLARIN



**CLARIN**: Common **L**anguage **R**esources and Technology **I**nfrasturcture  
**DARIAH**: Digital **R**esearch **I**nfrasturcture for the **A**rts and **H**umanities

# Language coverage of CLARIN technology

<https://vlo.clarin.eu>

more than 1,3 million components  
(tools, corpus...)

- English (177,273)
- Dutch (121,241)
- German (100,273)
- Bulgarian (34,658)
- Slovenian (31,022)
- French (26,728)
- Italian (14,546)
- Latin (12,883)
- ...

Resources in CLARIAH-ES languages

- Spanish (27,098)
- Catalan (1,801)
- Galician (859)
- Basque (498)



# Strategic network INTELE



- **INTELE**: (01-2020 to 09-2022) strategic network to promote the full membership in the CLARIN and DARIAH European research infrastructures.
  - **Promote** research in the humanities and social sciences
  - **Promote** international projects and programs.



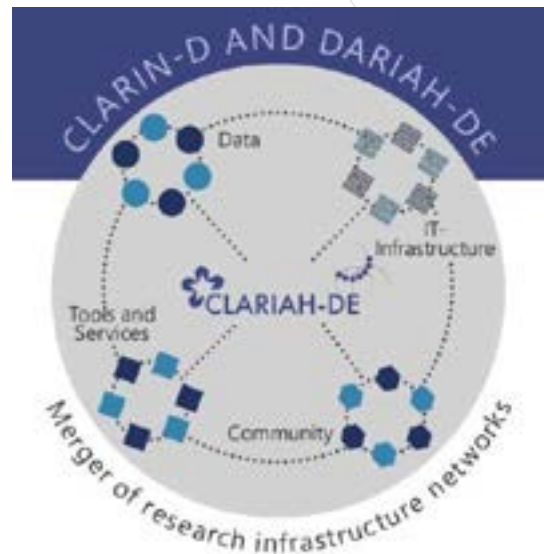
# Strategic network INTELE



## THE NETWORK IN EUROPE



The CLARIN, DARIAH and CLARIAH research infrastructures are active on a national and the European level. CLARIN and DARIAH both have the status of a European Research Infrastructure Consortium (ERIC).



The content-related and technological foundations created by CLARIN-D and DARIAH-DE will be aligned, integrated, further developed and jointly maintained in CLARIAH-DE.

# Strategic network INTELE



- 8 universities and research centers
  - EHU ([HiTZ](#)), UPF, UVigo, UNED ([LINDH](#)), UAlicante ([BVC](#)), UJAEN ([CEATIC](#)), UCM
  - USC ([Instituto da Lingua Galega](#))
  - CSIC ([Centro de Ciencias Humanas y Sociales](#))
  - [BSC-CNS](#)
- Support from:
  - CRUE, RAE, BNE, Instituto Cervantes, research networks and associations, etc.
  - More than **150** research groups from **35** universities
  - More than **750** researchers from **65** institutions signed the [INTELE Manifest](#)

# Strategic network INTELE



- Activities:
  - 2020-2022 **100+** Meetings
  - 2020 Workshop online: more than **130+** participants
  - 2021 **5** international webinars
  - 2021 Online workshop with experts from **6** research areas
  - 2022 Final workshop face-to-face with **120+** participants



# CLARIAH-ES proposal



- CLARIAH model ([Holand](#), [Germany](#), [Czech](#), [Finland](#), [Belgium](#), [Austria](#))
  - Union of CLARIN y DARIAH at a national level
  - sufficient *critical mass* with all relevant actors
  - Data, ICTs, HPC, tools and services, community
- Organization and financing (model from [Germany](#)):
  - CLARIAH-ES: CLARIN, DARIAH and integration
  - Contribution (all): centres, universities, regions, ministries, etc.
  - **Fee 0.2M€**/year during **5** years (MCIN)
  - In-kind **1.4M€**/year during **5** years
  - Surveillance: funding from strategic networks, etc.

# Requirements



- CLARIN
  - Fee: ~0.1M€/year
  - In-kind contribution: 0.5M-0.8M€/year
  - at least 1 [Certified B-centre](#)
- DARIAH
  - Fee: ~0.1M€/year
  - In-kind contribution: 0.6M€/year
- **Total: 1.4M-1.7M€/year during 5 years**
  - Total Fee: ~0.2M€/year
  - In-kind contribution: 1.1M-1.4M€/year
  - Operating expenses: trips, meetings, workshops, etc.
- MCIN signs two international agreements
- **All** Spanish researchers have access to both e-infrastructures

# Roadmap



- Prepare an agreed proposal: CLARIAH-ES
- Present the proposal to the institutions
  - Research centers and universities
  - Regional institutions and departments
  - MCIN, MUniversidades, PERTE
- Obtain the necessary support and commitments from all institutions
- Complete the design of the CLARIAH-ES consortium
- Sign the [protocolo general de actuación](#) CLARIAH-ES
- Sign the two international agreements (September 2023)
  - [CLARIN @ BOE](#)
  - [DARIAH @ BOE](#)

# Strategic network INTELE



# Strategic network INTELE





# CLARIAH-ES



- ERC members
- Observers
- Countries with participating centres
- Centre Providing Data
- Centre Providing Metadata
- Knowledge Centre



USA



SOUTH AFRICA



## Members and Cooperating Partners

DARIAH Members or Observers can be countries (EU Member States and Associated Countries) or intergovernmental organisations.

Alternatively, institutions from countries that are not yet DARIAH Members can participate as Cooperating Partners.

Currently, DARIAH has 22 Members and 19 Cooperating Partners in 11 non-Member countries: Egypt, Finland, Hungary, Iceland, Latvia, Norway, Romania, Slovakia, Sweden, the United Kingdom, and the United States of America.



# CLARIAH-ES: CLARIN & DARIAH



Distributed RI: **10** centres

- Community
- Data
- Tools & services
- IT-HPC-infrastructure

# CLARIAH-ES

- EHU ([HiTZ](#)) : *German Rigau & Mikel Iruskieta*
- UAlicante ([BVMC](#)) : *Manuel Marco*
- USC ([Instituto da Lingua Galega](#)) : *Elisa Fernandez*
- UNED ([LINDH](#)) : *Salvador Ros*
- UJAEN ([CEATIC](#)) : *Arturo Montejo-Ráez*
- UCM/MAD : *Dolores Romero + 5 public universities CAM*
- CSIC ([Centro de Ciencias Humanas y Sociales](#)) : *Daniel Riaño*
- [BSC-CNS](#) : *Maite Melero*
- [BNE](#): *Elena Sánchez*
- ULPGC ([IATEXT](#)): *Francisco Carreras*
- ... *Scayle, Dialnet?*
- **0.2M€** MICIN (fee) & **24+** Full Time Equivalent > **1.4M€** (in-kind)
- Strategic research network [CLARIAH-ES](#) **60K€** (2023-2024 MICIN)

# CLARIAH-ES office @ [HiTZ](#)

- Coordination with CLARIN-EU
- Coordination with DARIAH-EU
- CLARIAH-ES coordination
- CLARIAH-EUS coordination
- Work force
  - NC CLARIN and NC DARIAH - *German Rigau*
  - 1 FTE HiTZ - *4+ Technical Staff part time*
    - *Esther, Kike, Amaia,... also Aritz, Begoña, etc.*
  - 1 FTE UPV/EHU - *Mikel Iruskieta & Xabier Arregi half time each*
  - 1 FTE Basque Government - *Ainara Estarrona*
  - 1 FTE Diputación Gipuzkoa - *Xabier Goenaga*
  - ...

# CLARIAH-ES

- Infraestructura distribuida (estructural): Centros vs. Nodos
  - UPV/EHU => [CLARIAH-EUS](#)
  - UJaen => [CLARIAH Andalucía](#)
  - UCM => [CLARIAH-CM](#)
  - USC => CLARIAH-GAL
  - BSC => CLARIAH-CAT
  - UA => CLARIAH-VAL
  - UNED => CLARIAH-UNED
  - CSIC => CLARIAH-CSIC
  - BNE => CLARIAH-BNE
  - ULPGC => CLARIAH-Canarias

# Activities CLARIAH-ES 2023

- [Workshop CLARIAH-CSIC](#) Madrid, 11 May
- [DARIAH annual meeting](#) Budapest 6-9 June
- [CLARIN Centre Meeting 2023](#) Utrecht 13-14 June
- [Kick-off CLARIAH-ES](#) Alicante 15-16 June
- [EOSC tripartite event](#) Madrid, 19 September
- CLARIAH-ES office with [IATEXT](#) (ULPGC), Las Palmas Gran Canaria, 11 October
- [CLARIN annual meeting](#) Leuven 16-17-18 October
- [HDH Logroño](#) 18-20 de Octubre
- [Bibliotecas, datos, inteligencia artificial](#), BNE, Madrid 7-8 November
- [Workshop DARIAH-day](#) - Madrid 7 November
- DARIAH-EU - [NCC & GA](#) - Ljubljana 22-23 November
- [Workshop CLARIAH-EUS](#) Donostia 23 November
- [CLARIAH-ES presentation at UJA](#) - Jaén 14 December

# CLARIAH-ES (Alicante 2023)



# Activities CLARIAH-ES 2024

- [Workshop in Applied Digital Humanities](#) January 16 and 23, CCHS-CSIC, Madrid
- [CLARIAH-ES](#) project meeting - January 25-26, HiTZ, San Sebastián
- Meeting [Dialnet](#) - March 7 Logroño
- Participation in [Data Mining & IPR workshop](#) April 23th in Madrid.
- CLARIAH-ES meeting May 2nd in [Santiago de Compostela](#) (Galicia). 
- CLARIAH-ES meeting May 30th and 31th in [Madrid](#).
- Participation in [Dialnet Global Conference](#) May 29 to 31 in Logroño (La Rioja).
- ...
- [DARIAH annual event](#) Lisbon June 2024
- [CLARIN annual conference](#) Barcelona October 2024



# CLARIAH-ES (Donostia 2024)





3

# **New AI paradigm**

# New AI paradigm

- **Accelerated** discovery cycle
- Impressive progress
  - **Superior** to humans in many tasks
- **Deep** and **Reinforcement Learning**
- Application in **Language**, Vision, Robotics
- **LLMs** require **Experts**, **Supercomputing**, **Data**
  - **LLaMA-2** (70B):
    - 2 trillion tokens of training data (~20 million books)
    - 1.7M GPU hours A100-80Gb (~ 1000 A100 71 days)



# Deep Learning



Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ( $v_i^T \delta_i$ ). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

**Deep visual-semantic alignments for generating image descriptions (2014)**  
A Karpathy, L Fei-Fei

# Deep Learning



## **LipNet: Sentence Level Lipreading (2016)**

Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas

# Deep Learning

## Zero-Shot Text-to-Image Generation

(2021) Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever

<https://openai.com/blog/dall-e/>

[DALL-E 3](#)

[Midjourney](#)

[Stable Diffusion](#)

...

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



Edit prompt or view more images >

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



Edit prompt or view more images >

TEXT PROMPT

a storefront that has the word 'openai' written on it [...]

AI-GENERATED IMAGES



Edit prompt or view more images >

# Deep Learning: pictures and video



**v3 August 22**

**v4 November 22**

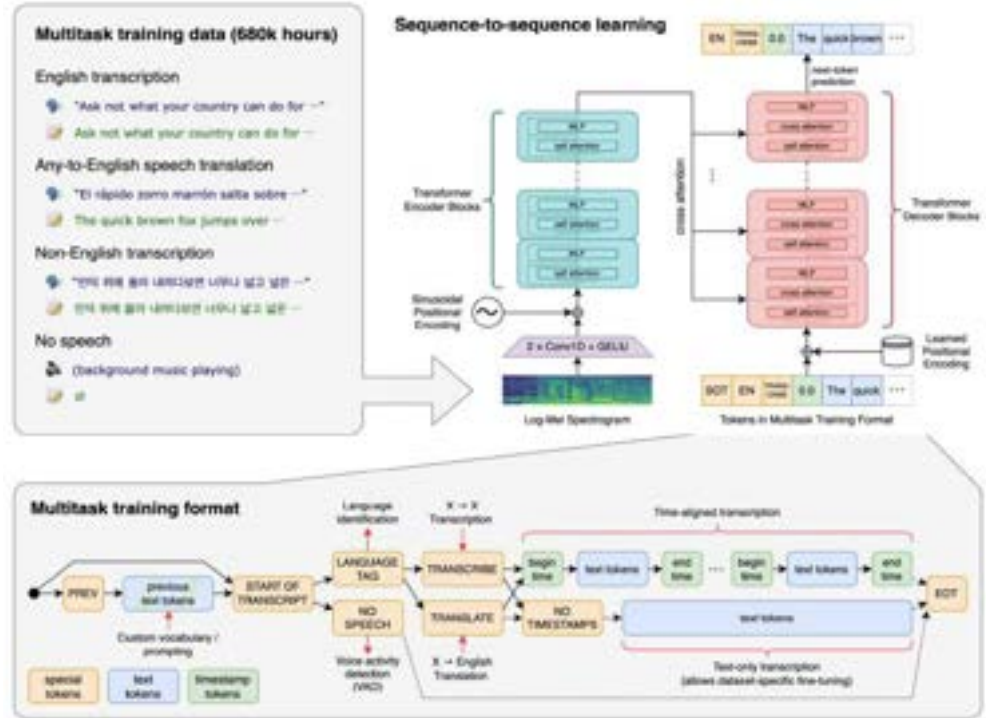
**v5 March 23**

<https://arstechnica.com/information-technology/2023/03/ai-imager-midjourney-v5-stuns-with-photorealistic-images-and-5-fingered-hands/>

# Deep Learning: music, speech

**MusicLM: Generating Music From Text (2023)**  
13 authors

[MusicLM](#)  
[UniAudio](#)  
[Elevenlabs](#)  
[Whisper](#)  
[PromptTT2](#)





# Deep Learning: videos

## [Make-A-Video: Text-to-Video Generation without Text-Video Data \(2022\)](#)

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman

[Make-a-video](#)  
[VEED.IO](#)  
[GliaStudio](#)  
[Sora \(video\)](#)

...



# Deep Learning: coding, programming

## [Evaluating Large Language Models Trained on Code \(2021\)](#)

58 authors

[Codex](#)

[Copilot](#)

[Codewhisperer](#)

[ChatGPT](#)

...

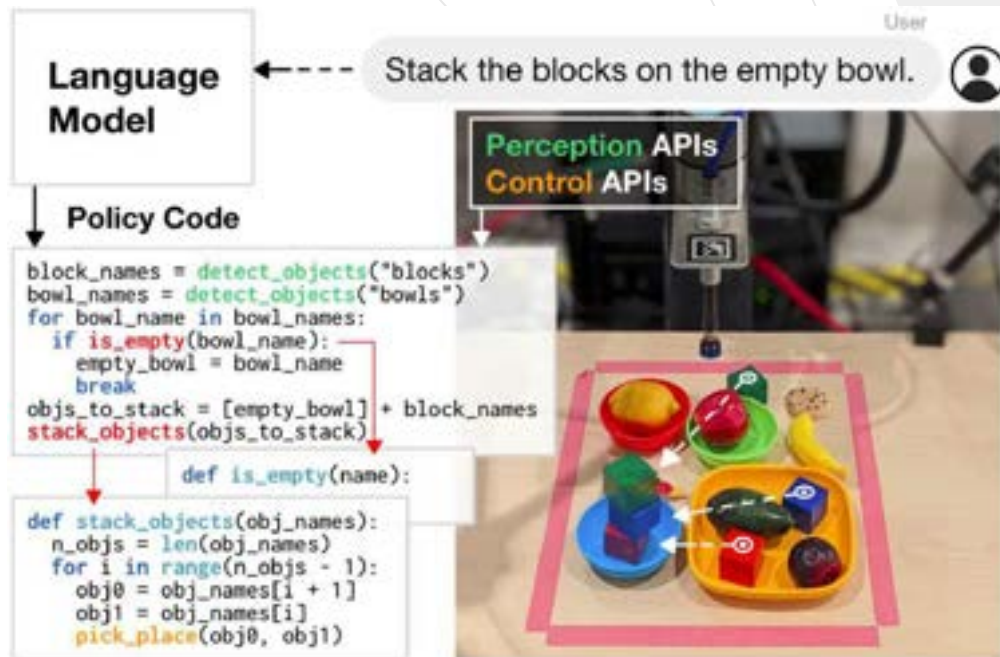
```
run:meso  course:  src:  |PrimeTest.java
1 package main
2
3 type Run struct {
4     Time int // in milliseconds
5     Results string
6     Failed bool
7 }
8
9 // Get average runtime of successful runs in seconds
10 func averageRuntimeInSeconds(runs []Run) float64 {
11     var totalTime int
12     var failedRuns int
13     for _, run := range runs {
14         if run.Failed {
15             failedRuns++
16         } else {
17             totalTime += run.Time
18         }
19     }
20
21     averageRuntime := float64(totalTime) / float64(len(runs) - failedRuns) / 1000
22     return averageRuntime
23 }
24 }
```

# Deep Learning: coding, programming

## Code as Policies: Language Model Programs for Embodied Control

(2022)

8 authors



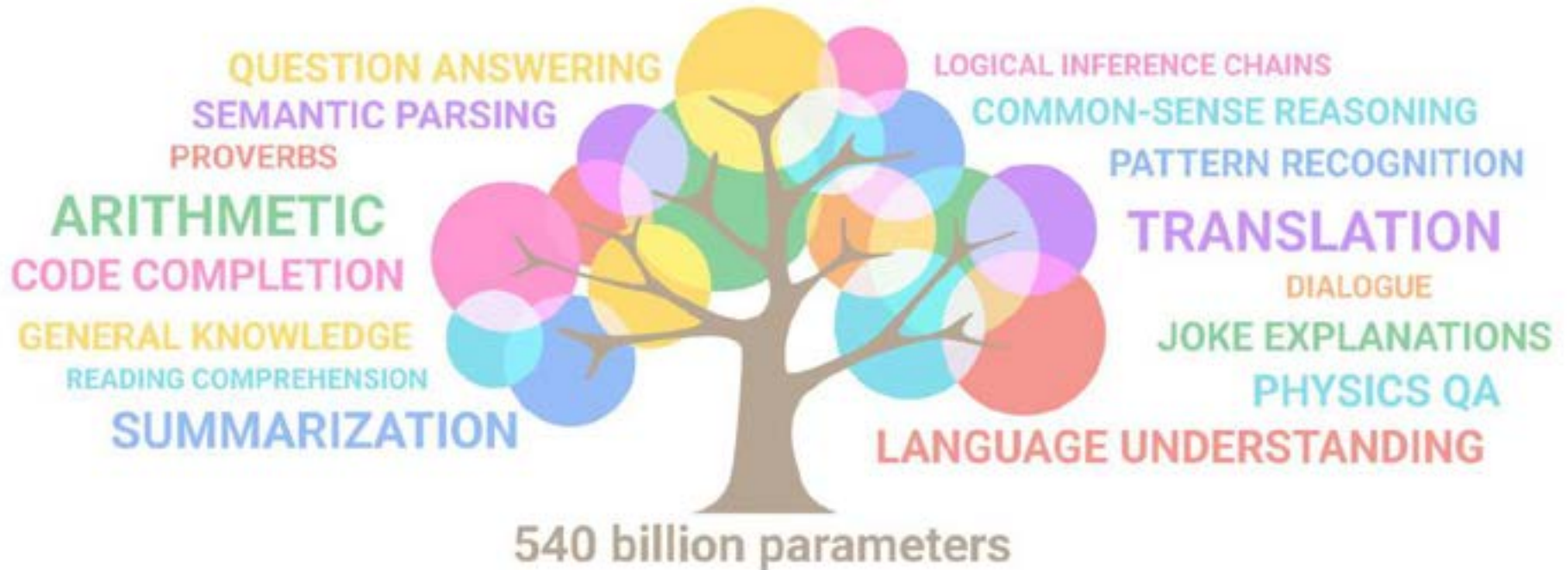
# Deep Learning: dubbing



[Example of translating, dubbing and changing facial expressions](#) by HeyGen

[VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time](#) by MS

# Large Language Models



**PaLM: Scaling Language Modeling with Pathways (2022)**

Aakanksha Chowdhery, + 67 authors



**Experts, Data, HPC**

# Large Language Models

## Explaining a joke

### Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

### Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

PaLM explains an original joke with two-shot prompts.

**PaLM: Scaling Language Modeling with Pathways (2022)**

Aakanksha Chowdhery, + 67 authors



**Experts, Data, HPC**

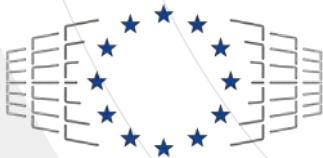


4

**An example: Latxa**

# Latxa: LLM for Basque

- Generative LLM for Basque
- Open ([LLaMA-2 license](#))
- **Largest Basque LLM** built to day
  - 7B, 13B, 70B
- Largest LLM trained in Spain
- Obtains state-of-the-art results
  
- Our research solutions apply to most languages (other than the largest)



**EuroHPC**  
Joint Undertaking



**ILENIA**   
IMPULSO DE LAS LENGUAS  
EN LA INTELIGENCIA ARTIFICIAL



# Latxa: LLM for Basque

- **Research problem**: publicly available text in Basque is scarce
  - Basque is 50th in Internet text volume
  - Large majority of language share the same limitation
- Collect a big corpora
  - Using diverse sources
  - Total, **4.3M documents** and **1.2B words**



# Latxa: LLM for Basque

- **Problem**: Still small corpus :-(
  - Solution: **Teach Basque** to an already pre-trained LLM in English
- Latxa is based on the [LLaMA-2](#) LLM
  - Pre-trained with 2 trillion tokens
- Also much less HPC-AI needed
  - Much less carbon footprint (!)

arXiv:2307.09288v2 [cs.CL] 19 Jul 2023

## LLAMA 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron<sup>1</sup>, Louis Martin<sup>1</sup>, Kevin Stone<sup>1</sup>

Pieter Albert, Amjad Al-Jarrah, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra,  
Pragati Bhargava, Shouni Bhowmik, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen,  
Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,  
Cythra Gao, Vladimir Gimenez, Nanang Gezal, Anthony Hartshorn, Sejjad Hossaini, Rui Hua,  
Hakan Inan, Marvin Kardes, Viktor Kerkov, Madian Khabsa, Isabel Kloumann, Artem Korneev,  
Parit Singh Koura, Mariia-Agneta Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovych,  
Yinghai Lu, Yuning Mao, Xavier Martinet, Taylor Mihaylov, Prudhvir Mishra,  
Igor Molybog, Yixin Nie, Andrew Noulton, Jeremy Reizenstein, Rishi Rungta, Kalyan Saluja,  
Alan Schelten, Ruan Silva, Eric Michael Smith, Ronjan Subramanian, Xiaoqing Tan, Binh Tang,  
Ross Taylor, Adina Williams, Jian Xiang Kuan, Pooja Xu, Zheng Yan, Bryan Zaionz, Huchen Zhang,  
Angela Fan, Melanie Kambhathur, Shantanu Narang, Aaryaman Rodriguez, Robert Stojic,  
Sergiy Edunov, Thomas Scialom<sup>2</sup>

GenAI, Meta

Abstract

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called LLaMA 2-Chat, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of LLaMA 2-Chat in order to enable the community to build on our work and contribute to the responsible development of LLMs.

# Latxa: Teaching Basque to LLaMa2

- Total grand: EuroHPC 1,4 million hours ~ 4,8 M€



**EuroHPC**  
Joint Undertaking

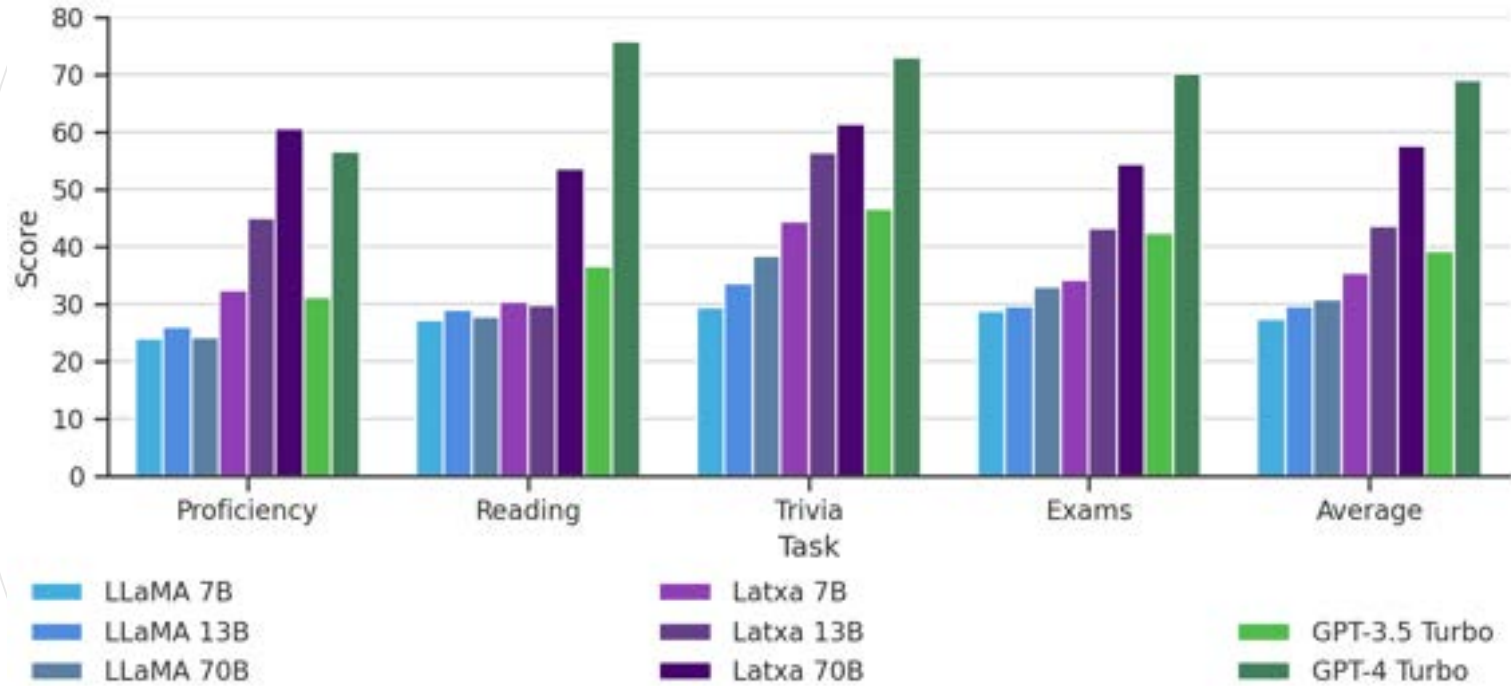


# Latxa: Evaluation

- **Problem**: How to test the model?
- Collect an evaluation dataset with Basque related tasks:
  - **EusProficiency** (language proficiency)
  - **EusReading** (reading comprehension)
  - **EusTrivia** (trivia questions from 5 areas)
  - **EusExams** (public examinations)



# Latxa: Results



# Latxa: LLM for Basque

© Chatbot

Kaixo

¿En que puedo ayudarte hoy?

Kaixo! Zertan lagundu zaitzaket gaur?

Necesito ayuda para desarrollar una función de Python que sume dos números, ¿puedes ayudarme?

Laguntza behar dut bi zenbaki batzen dituen python funtzio bat garatzen, lagunduko?

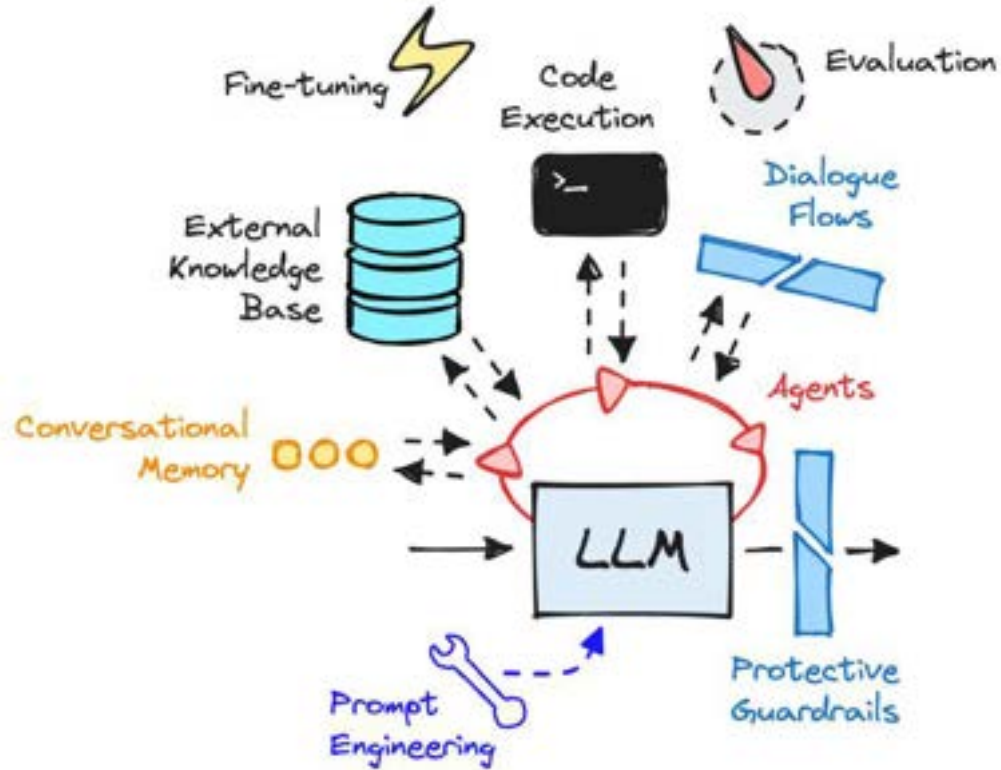
Noski! Hona hemen bi zenbaki batzen dituen python funtzioa:\n\npython\ndef add(num1, num2):\n return num1 + num2\n\nFuntzio honek num1 eta num2 parametroak hartzen ditu eta haien batura itzultzen du. Adibidez, 5 eta 7 zenbakiak batuz gero, emaitza 12 izango litzateke.

¡Por supuesto! Aquí hay una función de python que suma dos números.

```
python def add(num1, num2): return num1 + num2
```

Esta función toma los parámetros num1 y num2 y devuelve su suma. Por ejemplo, si sumas los números 5 y 7, el resultado sería 12.

# Language Agents



[LLM ecosystem components.](#)

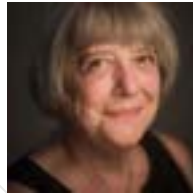
# Summary

- **Unthinkable** AI applications just a few years ago
- Even more **amazing** results in the near future (day by day)
  
- **CLARIAH-ES**: CLARIN and DARIAH
- **Coordination and Cooperation**



# Acknowledgements

- Ministerio de Ciencia e Innovación
- CLARIN - ERIC
- DARIAH - ERIC
- INTELE consortium & manifesto supporters
- Ministeries, C.C.A.A., PERTE, Oficina Español en el Mundo, FECYT, etc.
- Universities, la CRUE, el CSIC
- RAE, BNE, Instituto Cervantes, research networks and associations, etc.
  
- Bente Maegaard
- Steven Krauer
- Núria Bel
- Mario Tascón
- Xavier Gómez-Guinovart ...



# Epílogo

“No hemos llegado muy lejos  
pero hemos avanzado mucho.”

- Mikel Iruskietta

*“We haven't come very far  
but we've made a lot of progress.”*



# Future of CLARIAH-ES



eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

# HiTZ

Hizkuntza Teknologiako Zentroa  
Basque Center for Language Technology

German Rigau Claramunt

HiTZ Deputy Director

National Coordinator of CLARIN and DARIAH

[hitz.eus](http://hitz.eus)