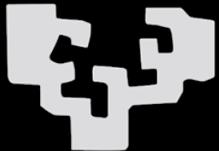


Universidad
del País Vasco

Euskal Herriko
Unibertsitatea



eman ta zabal zazu

HiTZ

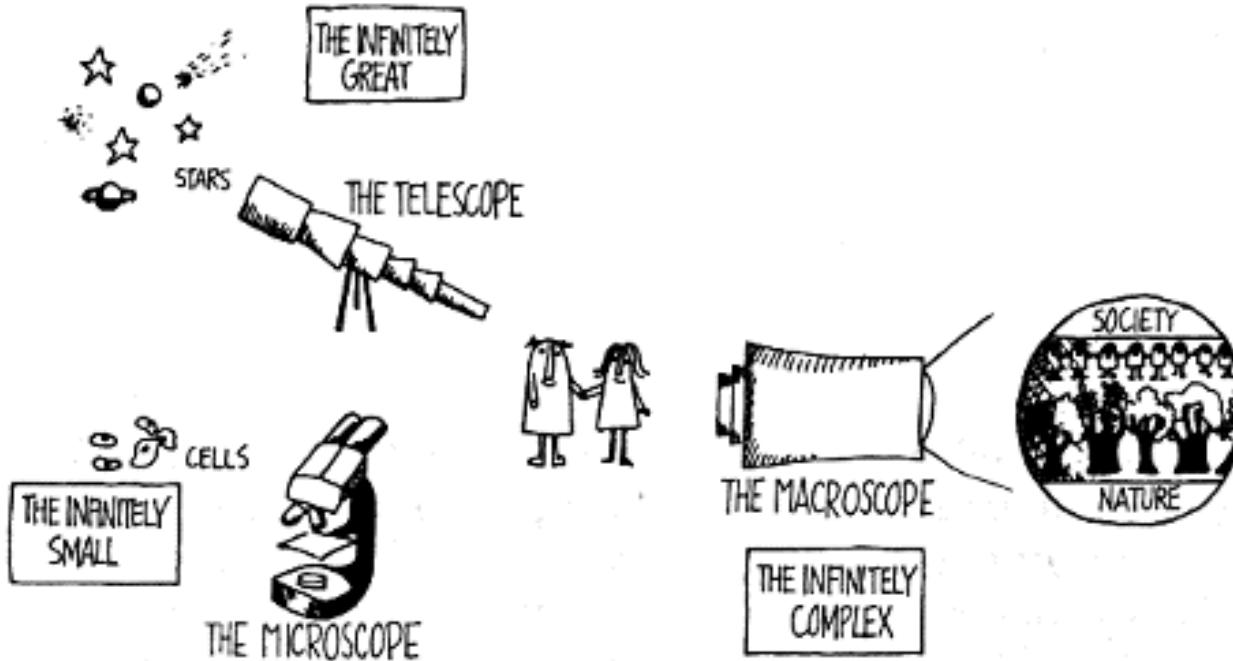
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

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

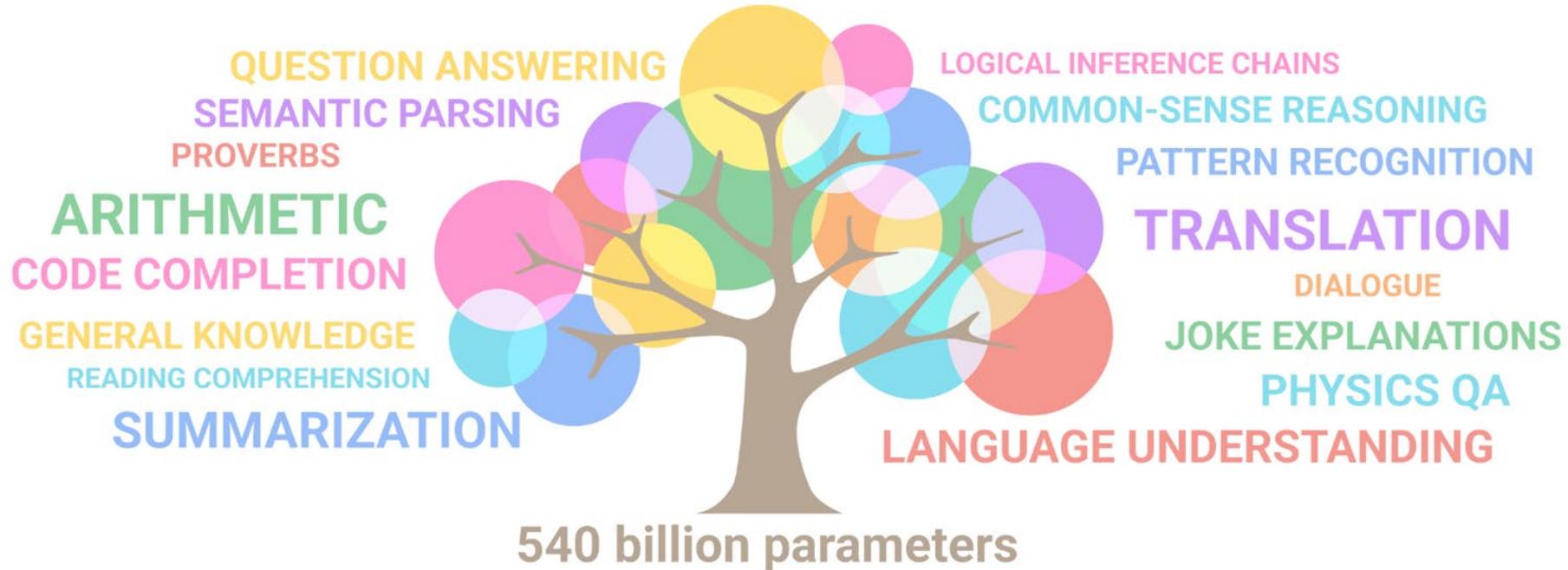
hitz.eus

 CLARIAH-ES

The Macroscope



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



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.

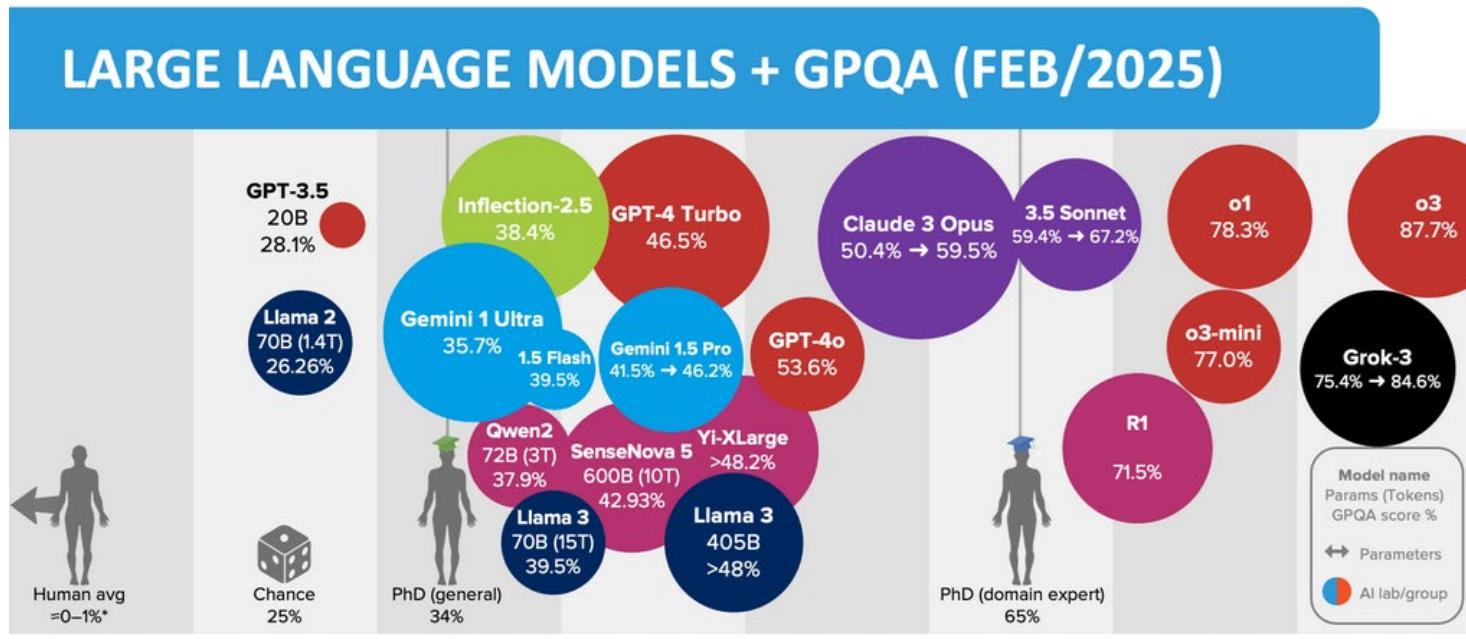
PaLM: Scaling Language Modeling with Pathways (2022)

Aakanksha Chowdhery, + 67 authors



Experts, Data, HPC

Large Language Models



Model sizes near to scale. * Estimates based on independent analysis. Selected highlights only. IQ correlation estimates only: <https://lifearchitect.ai/visualising-brightness/> PhD/IQ correlation: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10001/pdf/> All models: <https://lifearchitect.ai/models-table/> Alan D. Thompson. 2025.



1

HiTZ

- **Reference center** in NLP/AI research (2019)
 - Ixa (founded 1988) and Aholab (1998)
 - 90 members, multidisciplinarity
 - Yearly income in projects: ~3M €
- Research & Applications in EU, ES, EN, etc.
- **Academy**: Master, Doctorate, courses
- Consultants for Basque, Spanish and European LT plans
- National coordinators of two European research infrastructures
 - CLARIN-EU & DARIAH-EU : **CLARIAH-ES**
- Chair in AI and Language Technology
 - Funded for 4 years by SEDIA under ENIA
- Large LT projects
 - ILENIA, IKER-GAITU, ALIA

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 (~GPT4o)
- **Best research paper award ACL 2024**
- [**Demo**](#)



LATXA



EuroHPC
Joint Undertaking



ILENIA
IMPULSO DE LAS LENGUAS
EN LA INTELIGENCIA ARTIFICIAL





2

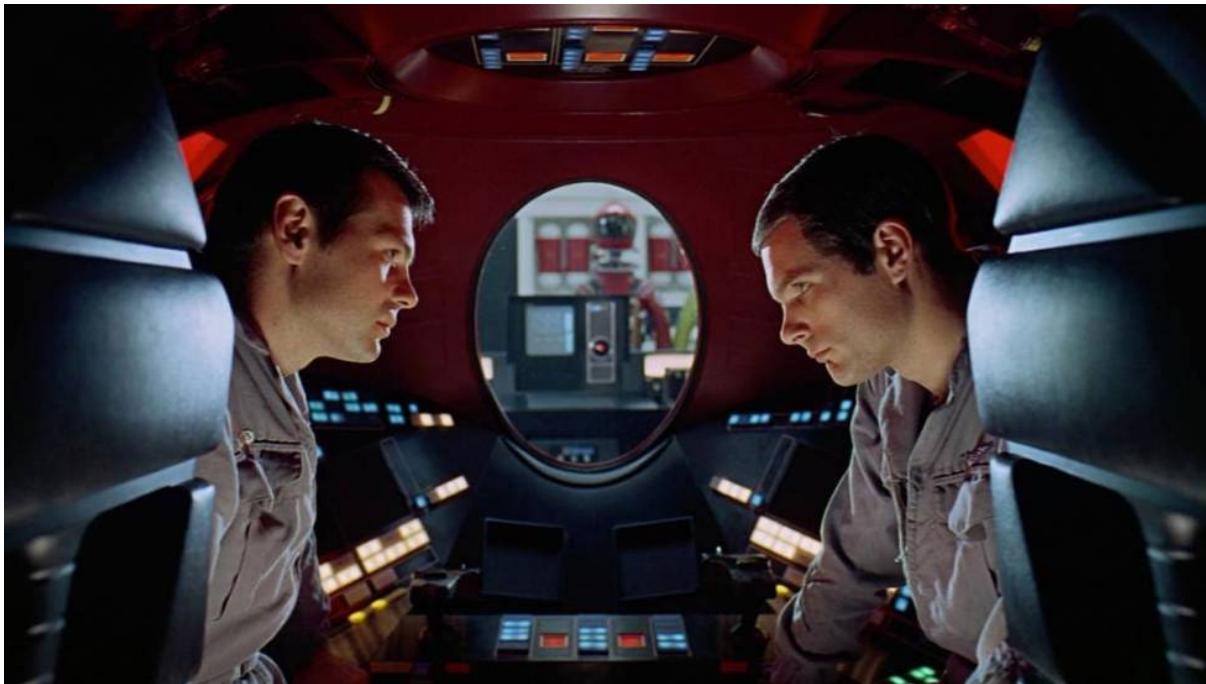
New AI paradigm

Artificial Intelligence



10
1
0

Artificial Intelligence

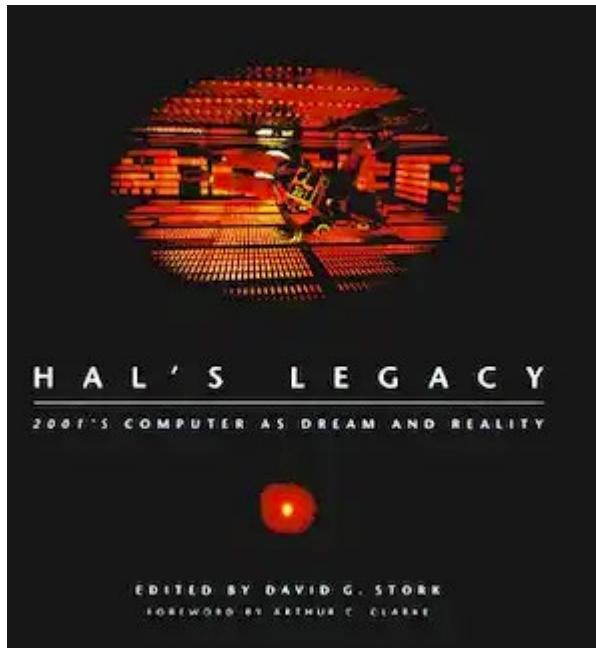


Artificial Intelligence

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

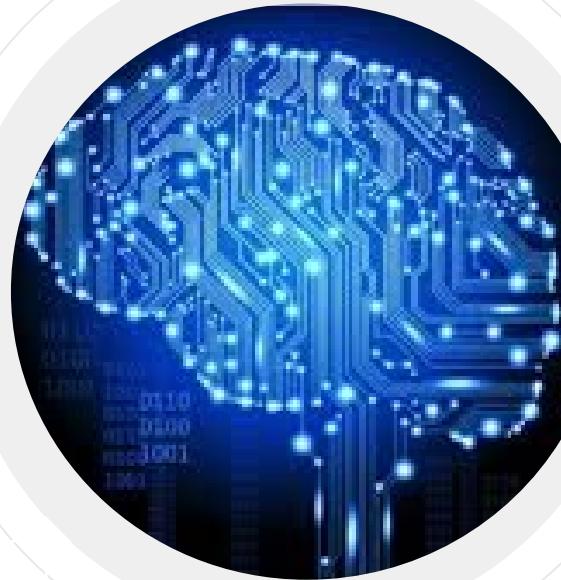


LipNet: Sentence Level Lipreading (2016)

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

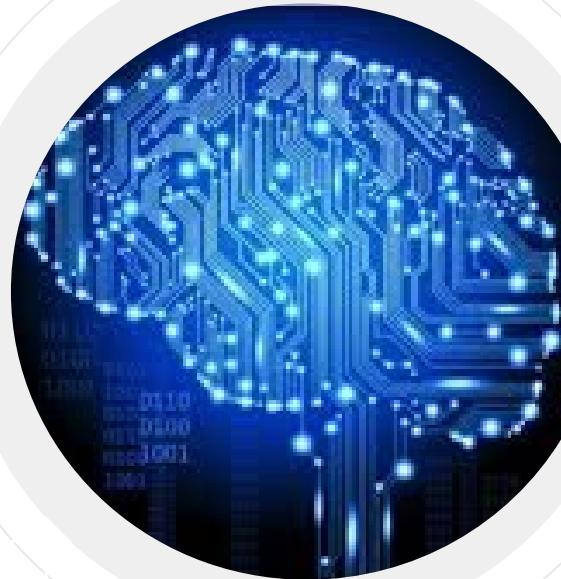
New AI paradigm

- 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)



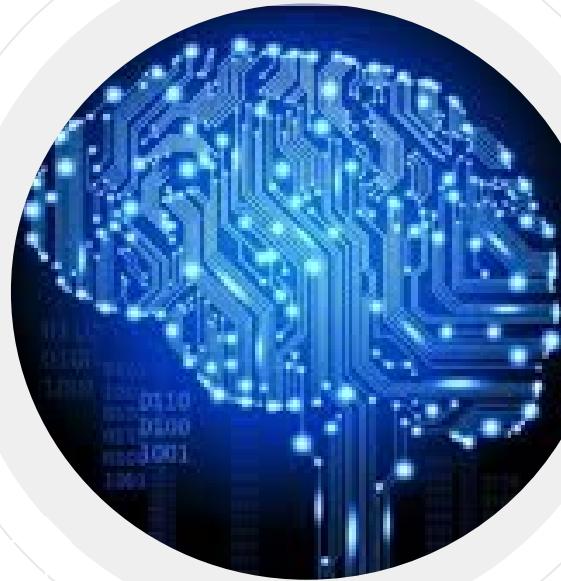
AI topics

1. Knowledge representation
2. Reasoning
- 3. Understanding**
4. Planning



Understanding

- Natural language
- Images / videos
- Code / databases
- The web / documents
- Music
- 3D objects
- ...
- The world?



Understanding

Applying for a job at IKEA



Understanding (ChatGPT)

What is this?



This is a cartoon image that humorously depicts the process of applying for a job at IKEA. The text at the top reads "Applying for a job at IKEA," and the scene shows a candidate entering a room where an interviewer is sitting at a desk. The interviewer says, "Make a chair and take a seat," while a disassembled chair is shown on the floor, highlighting IKEA's association with flat-pack furniture that customers assemble themselves.

Understanding (ChatGPT)

Why this cartoon is humorous?



The humor in this cartoon lies in the clever play on IKEA's brand identity and the interview situation:

1. **IKEA's Flat-Pack Furniture:** IKEA is famous for selling furniture that comes in flat packs, requiring customers to assemble it themselves. The idea of building furniture is strongly associated with the brand.
2. **Job Interview Twist:** In a typical job interview, you might expect questions or tasks related to your skills, not something as literal as assembling furniture on the spot.
3. **Wordplay on "Take a Seat":** The phrase "take a seat" is usually an invitation to sit down, but here it's taken literally. The candidate is required to build a chair before they can "take a seat."

This combination of brand-specific humor, the unexpected situation, and the literal interpretation of a common phrase creates a lighthearted and amusing scenario.

Understanding (ChatGPT)

Who is the author of this cartoon?



The cartoon is signed by "Canary Pete," who is likely the author or cartoonist behind this work. Canary Pete is known for creating humorous illustrations, often with clever wordplay or situational comedy.

Deep Learning

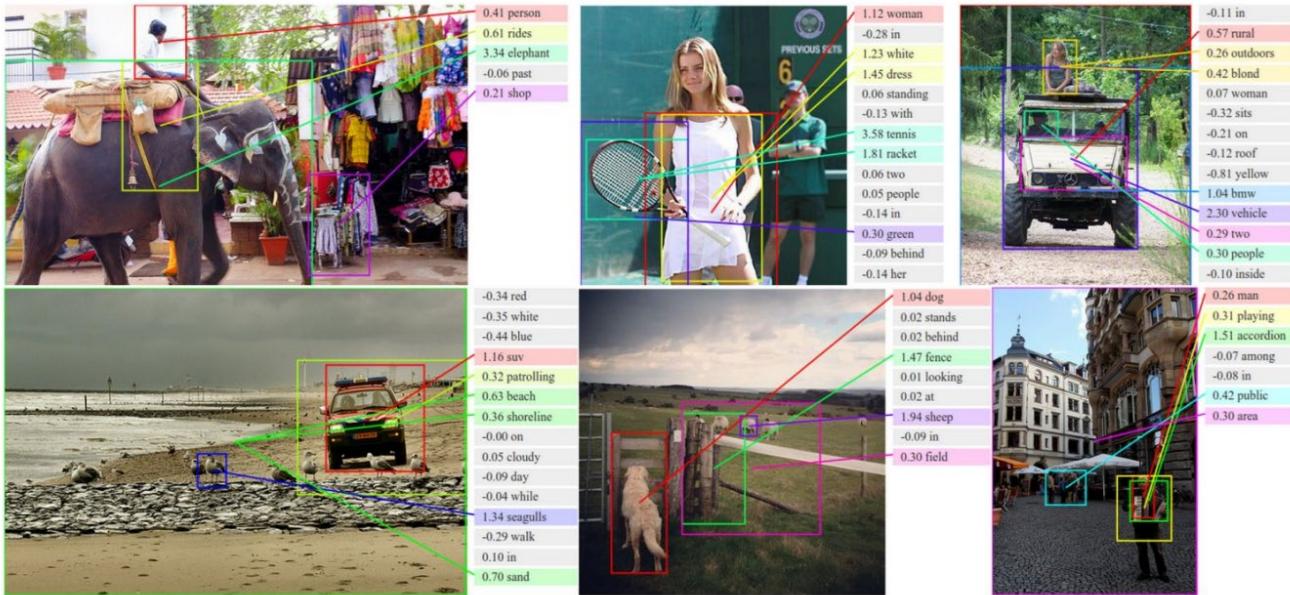


Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring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

A Style-Based Generator
Architecture for Generative
Adversarial Networks (2018)
Tero Karras, Samuli Laine, Timo Aila
<https://thispersondoesnotexist.com>

...

Scaling Synthetic Data Creation
with 1,000,000,000 Personas (2024)
[https://github.com/tencent-
ailab/persona-hub](https://github.com/tencent-ailab/persona-hub)

...



Figure 2. Uncurated set of images produced by our style-based generator (config F) with the FFHQ dataset. Here we used a variation of the truncation trick [40, 5, 32] with $\psi = 0.7$ for resolutions $4^2 - 32^2$. Please see the accompanying video for more results.

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](#)

[Flux.1](#)

[Imagen 3](#)

...

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 store front that has the word 'openai' written on it [...]

AI-GENERATED IMAGES



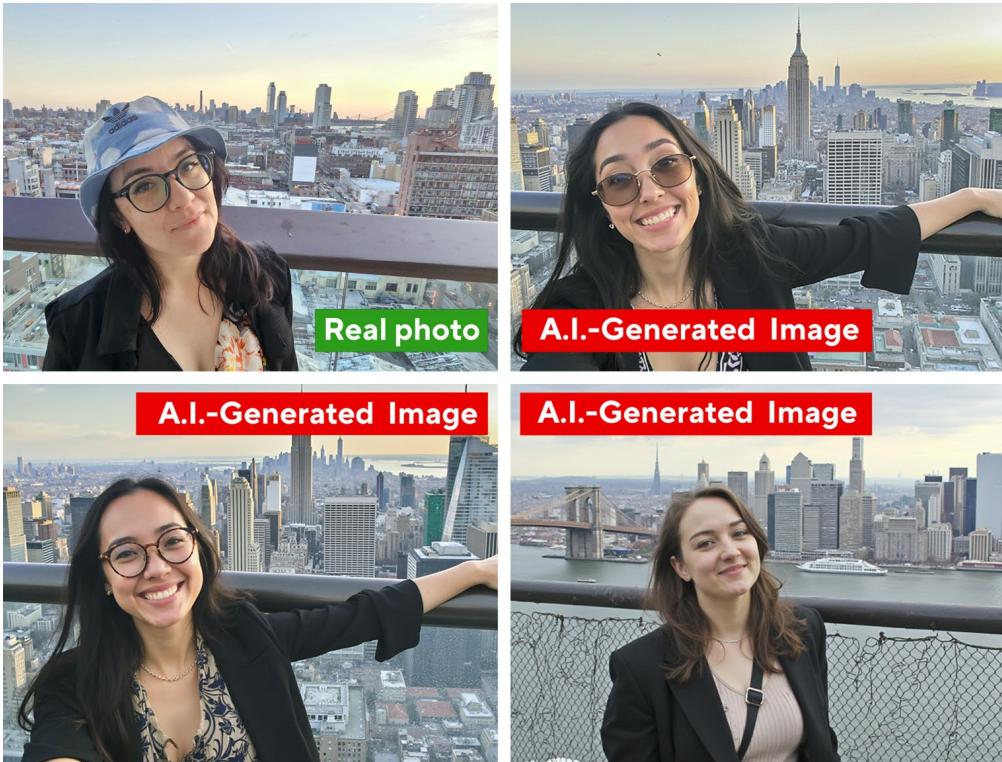
[Edit prompt or view more images ↴](#)

Deep Learning: pictures and video



Can you tell which of these four images is a real photograph?

Deep Learning: pictures and video

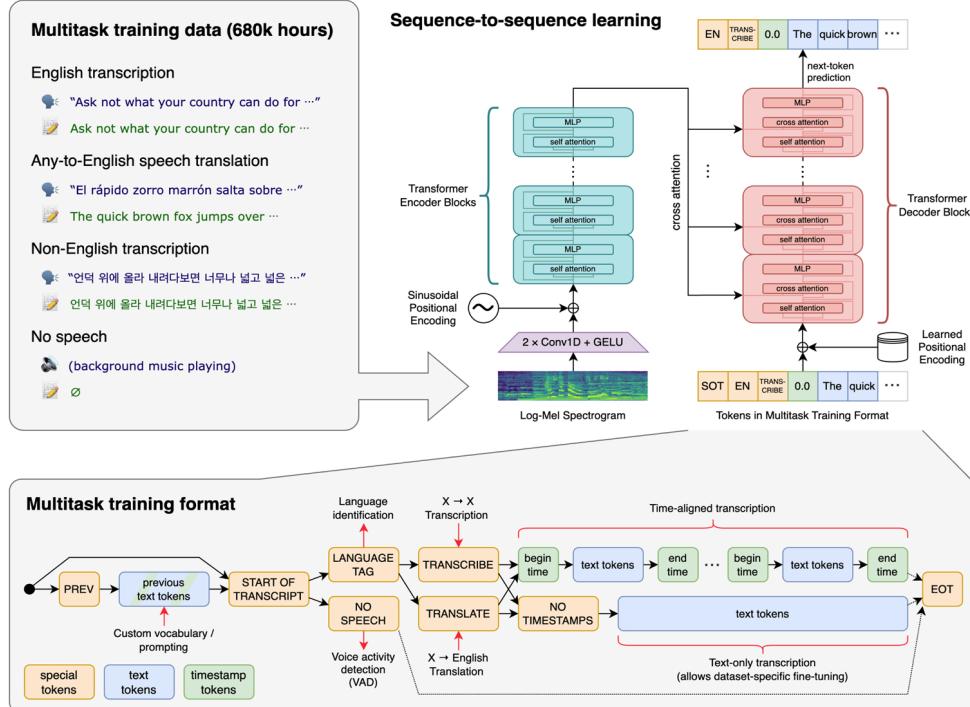


Can you tell which of these four images is a real photograph?

Deep Learning: music, speech

MusicLM: Generating Music
From Text (2023)
13 authors

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

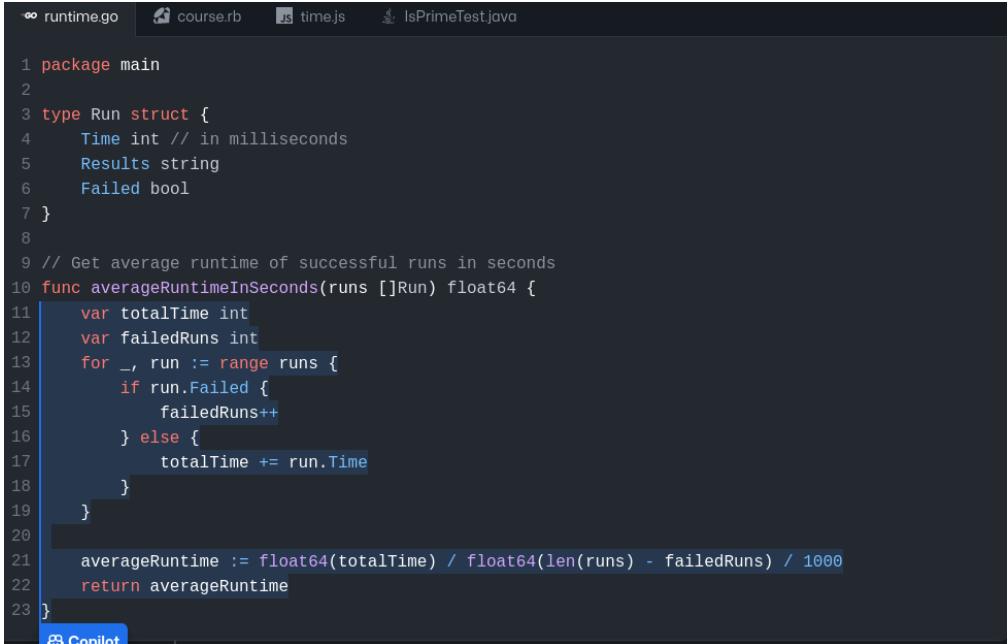


Deep Learning: coding, programming

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

[Codex](#)
[Copilot](#)
[Codewhisperer](#)
[ChatGPT](#)
[WizardCoder](#)

...



A screenshot of a code editor showing a Go file named `runtime.go`. The code defines a struct `Run` and a function `averageRuntimeInSeconds` that calculates the average runtime of successful runs. The code is annotated with numbers 1 through 23. A blue callout box labeled "Copilot" is positioned at the bottom left of the code area.

```
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 }
```

Deep Learning: videos/3D objects

[DreamFusion: Text-to-3D using 2D Diffusion \(2022\)](#)

4 authors

[POINT-E
Magic3D](#)

...

[CAT3D: Create Anything in 3D with Multi-View Diffusion Models \(2024\)](#)

8 authors

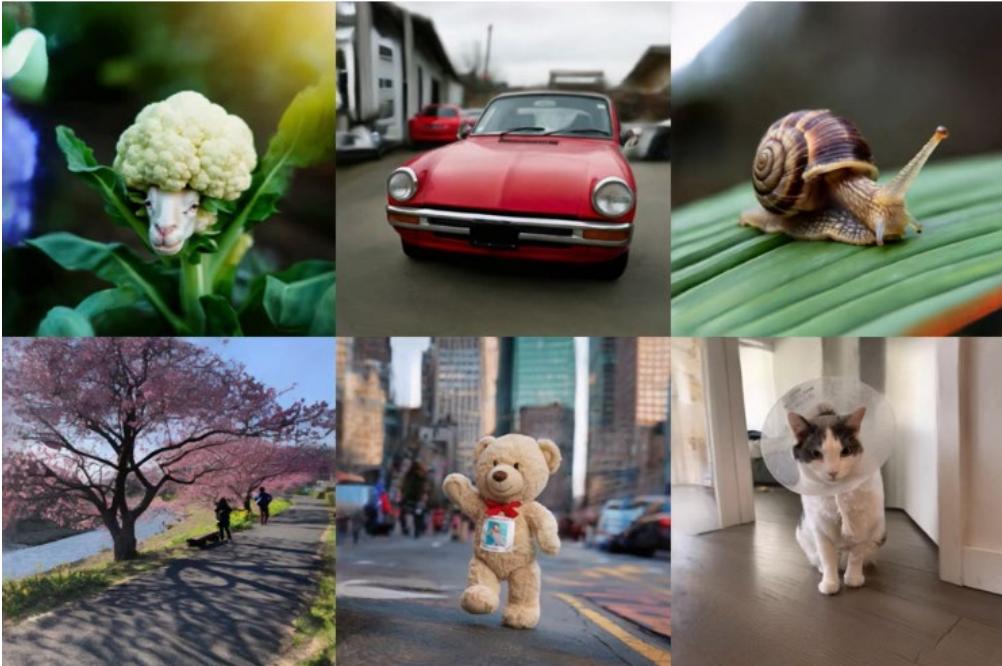
[CAT3D](#)

...

[19th Century Portraits Brought To Life](#)

[Animate old photos](#)

...



Deep Learning: videos

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

[Make-a-video](#)
[VEED.IO](#)
[GliaStudio](#)
[Sora \(video\)](#)
[Hotshot](#)
[Dream Machine](#)
[Meta Movie Gen](#)
[Mochi-1](#)
[Flux video](#)
[Wan2.1](#)
[Open Sora](#)

...



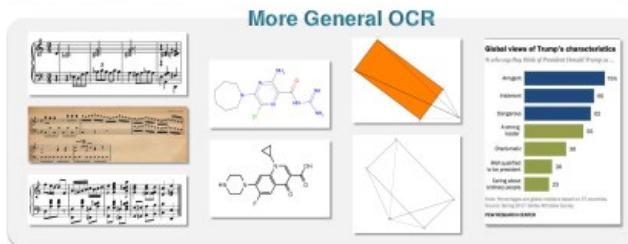
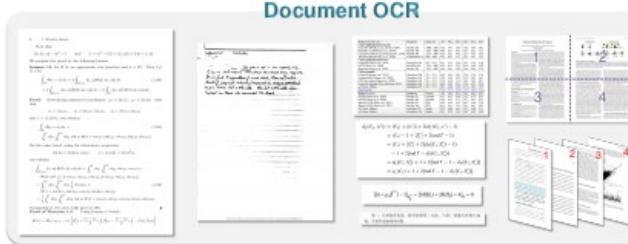
Deep Learning: documents

[General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model \(2024\)](#)

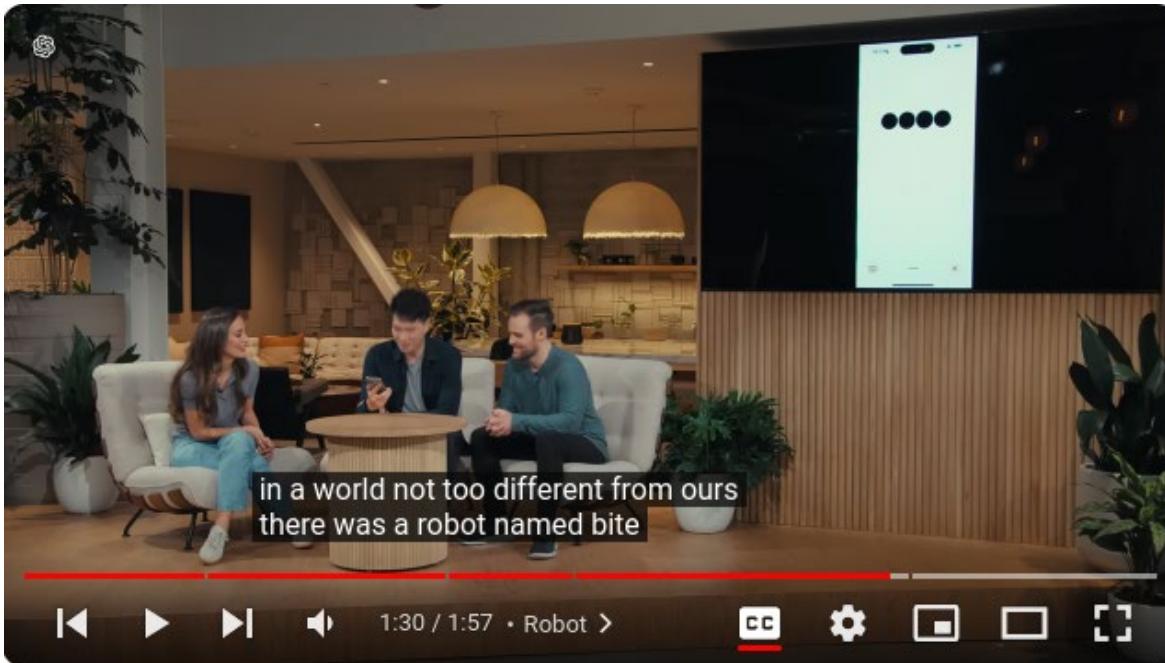
12 authors

[olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models \(2025\)](#)

9 authors



Deep Learning: Intelligent Assistant



[GPT-4o](#) by OpenAI

DeepMind

[Learn Spanish](#) with GPT-4o

[Dad jokes](#) with GPT-4o

[Project Astra](#) by Google

Deep Learning: Intelligent Assistant

Home > TECHNOLOGY > AI

NotebookLM now lets you listen to a conversation about your sources

Sep 11, 2024

2 min read

Our new Audio Overview feature can turn documents, slides, charts and more into engaging discussions with one click.

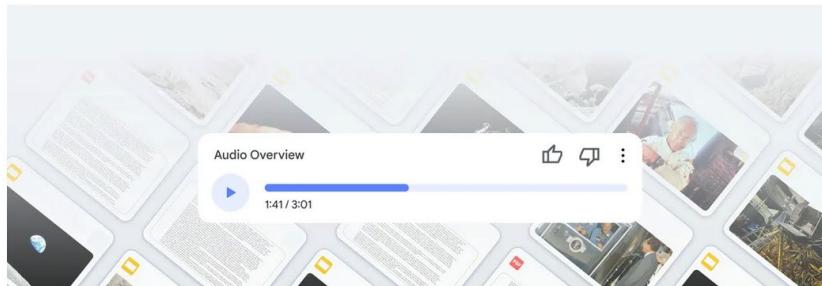


Biao Wang

Product Manager, Google Labs

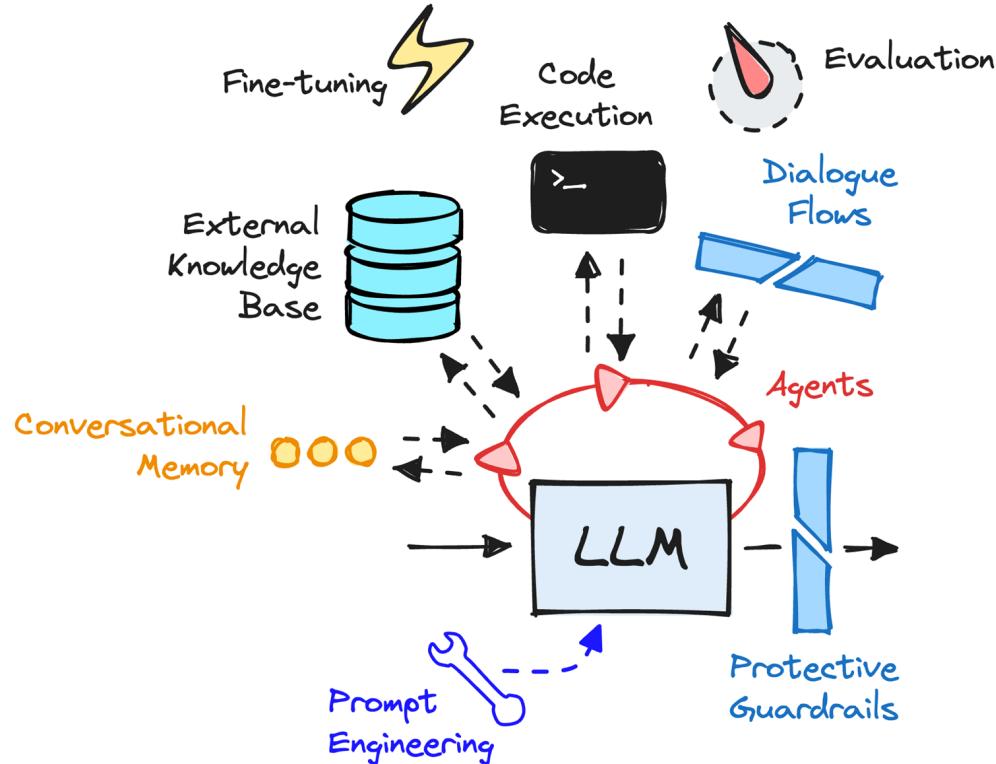
Read AI-generated summary

Share



[NotebookLM @ Google](#)
[Illuminate @ Google](#)

Language Agents



i.e. RAG

<https://typeset.io/>

Deep Research

Veridika

Epilogue

“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.”





3

CLARIAH-ES

ESFRI – EOSC – SSHOC

Strategy Report on Research Infrastructures

ROADMAP 2021



EUROPEAN OPEN SCIENCE CLOUD



#SSHOCaVocabulary
#SSHOCifyCLARIN

Social Sciences
Humanities

CLARIAH-ES



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



Members and Cooperating Partners

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

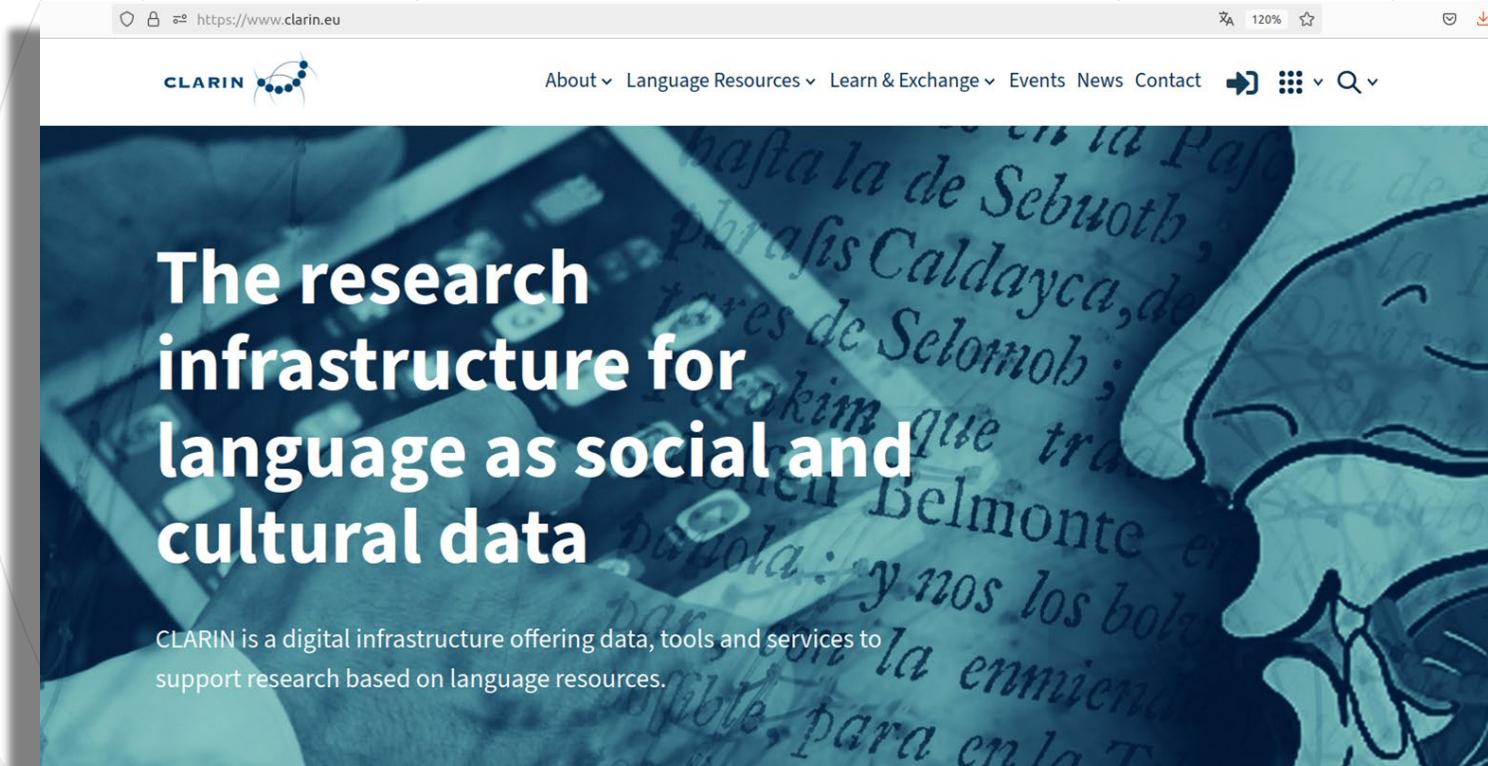
Alternatively, institutions from countries that are not yet DARIAH Members can participate as Cooperating Partners.

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.



CLARIN: Common LAnguage Resources and Technology INfrastructure
DARIAH: Digital Research Infrastructure for the Arts and Humanities

CLARIN: Distributed Research Infrastructure

A screenshot of the CLARIN website homepage. The header features the CLARIN logo and navigation links for About, Language Resources, Learn & Exchange, Events, News, Contact, and various icons for search and sharing. The main visual is a blue-toned photograph of a person's profile looking at a smartphone displaying text. Overlaid on the image is the text "The research infrastructure for language as social and cultural data". Below this, a subtitle reads: "CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources."

The research infrastructure for language as social and cultural data

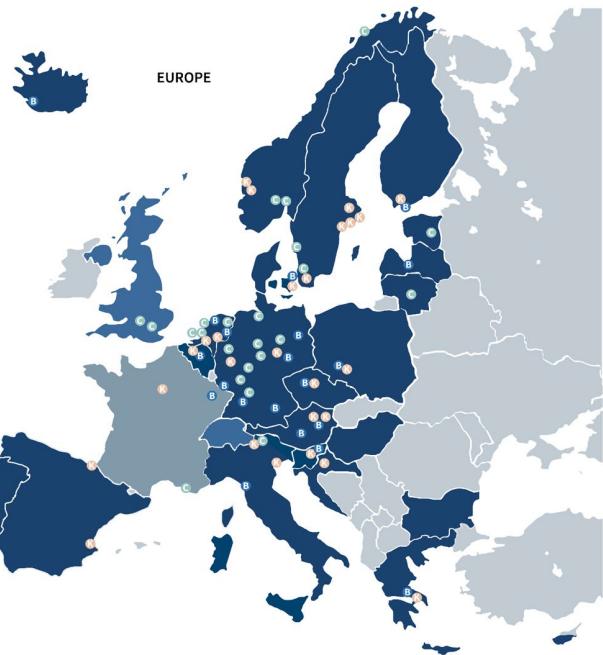
CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.

CLARIN: Distributed Research Infrastructure

24 countries

Network of 68 Centres (certified):

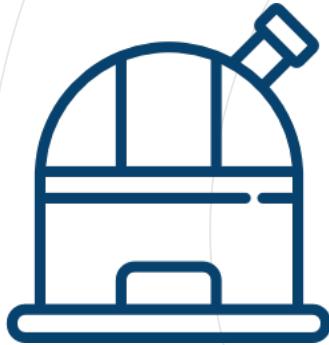
- *B-center*: data, services
- *K-center*: knowledge, advise
- *C-center*: metadata
- ...
- Experts, data, services, hw, sw, etc.



<https://www.clarin.eu/content/clarin-centres>

<https://www.clarin.eu/content/overview-clarin-centres>

CLARIN: Research Infrastructure



Virtual Language Observatory:

- <https://vlo.clarin.eu/>



Language Resource Switchboard:

- <https://beta-switchboard.clarin.eu/>
- <https://switchboard.clarin.eu/>

Other : <https://www.clarin.eu/content/services>

Language coverage of CLARIN technology

<https://vlo.clarin.eu>

almost 0.7 million components
(tools, corpus...)

- Dutch (133,508)
- English (129,776)
- German (97,184)
- Bulgarian (34,271)
- Slovenian (31,035)
- French (21,265)
- Italian (13,585)
- Latin (13,826)
- ...

Resources in CLARIAH-ES languages

- Spanish (18,616)
- Galician (696)
- Basque (291)
- Catalan (208)

CLARIN: Knowledge Infrastructure

- Knowledge K-centres:
 - <https://www.clarin.eu/content/knowledge-centres>
 - **Advice** from experts in various disciplines. They can also offer training, research guides, tool use guides, etc.
 - Organized by topics...
 - In Spain:
 - *Spanish K-Center*: languages in Spain
 - <https://www.clariah.es/k-centre/>
 - *IMPACT K-Center*: digitalization
 - <https://www.digitisation.eu/>
 - *LLM4SSH: Large Language Models*
 - <https://llms4ssh.clarin-pl.eu/>

"All K-centres have a helpdesk that will respond to requests within two working days."

CLARIN: Knowledge (sharing) Infrastructure

- Learning Hub:
 - <https://www.clarin.eu/content/learning-hub>
 - Compilation of various educational resources on different topics.
 - Use of CLARIN, teaching materials, workshops, HD courses, guides and “good practices”, etc.
 - Network of Experts ...
- Service to store resources and tools
 - FAIR principles: Findable, Accessible, Interoperable, Reusable
 - <https://www.clarin.eu/fair>
 - Contact Center B with deposit service:
 - <https://www.clarin.eu/content/depositing-services>
 - Tools:
 - <https://github.com/clarin-eric/switchboard-tool-registry>

CLARIN: networking, funding, projects

- CLARIN Funding Hub: scholarships for international collaboration in strategic tasks, exchange, training or mobility.
 - <https://www.clarin.eu/funding>
- Advice for applying for European projects (that are relevant to CLARIN).
- Networking: participation in events, nodes, k-centers, committees, etc.
 - <https://www.clarin.eu/events>

DARIAH: Distributed Research infrastructure

The screenshot shows the Dariah-EU website homepage. At the top, there is a dark header bar with the Dariah-EU logo and navigation links for Contact, Helpdesk, and a search icon. Below the header is a main banner featuring a blue background with a network graph pattern. On the left side of the banner, the text "The pan-European infrastructure for arts & humanities scholars" is displayed, along with a "Learn More About DARIAH" button. On the right side, there is an announcement about the "DARIAH Annual Open Access Book Bursary 2024: Call for manuscripts". A small thumbnail image of the bursary booklet is shown at the bottom right of the banner.

DARIAH-EU
Celebrating the first decade

ABOUT ▾ NETWORK ▾ ACTIVITIES ▾ TOOLS & SERVICES ▾ NEWS & EVENTS ▾

DARIAH Annual Open Access Book Bursary 2024: Call for manuscripts

We are pleased to announce the fourth round of the DARIAH annual book bursary for the Open Access...

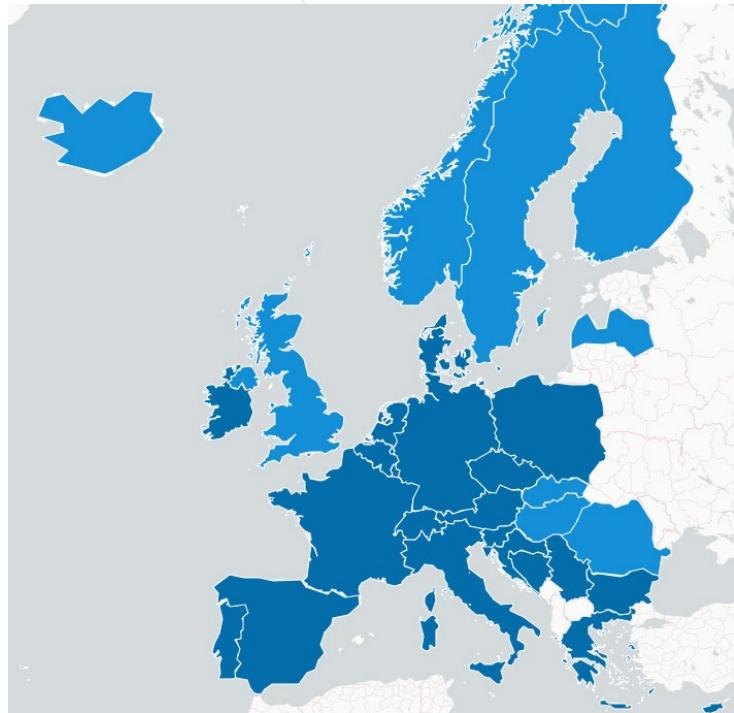
DARIAH Open Access Book Bursary 2024

DARIAH-EU
AN INFRASTRUCTURE FOR ARTS AND HUMANITIES

Read Post

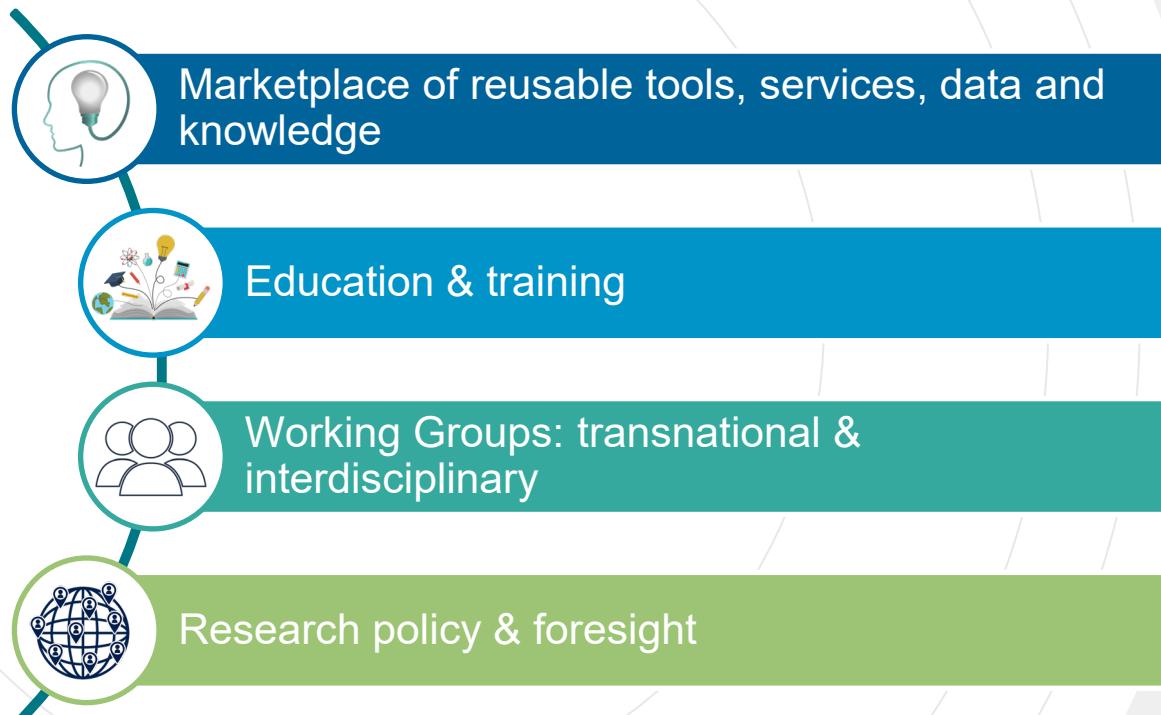
DARIAH: Distributed Research Infrastructure

- 22 countries
- 4 strategic pillars
 - SSH Marketplace
- Regional Hubs
 - Near countries



CLARIN: Distributed Research Infrastructure

4 strategic pillars



DARIAH (I) SSH Open Marketplace



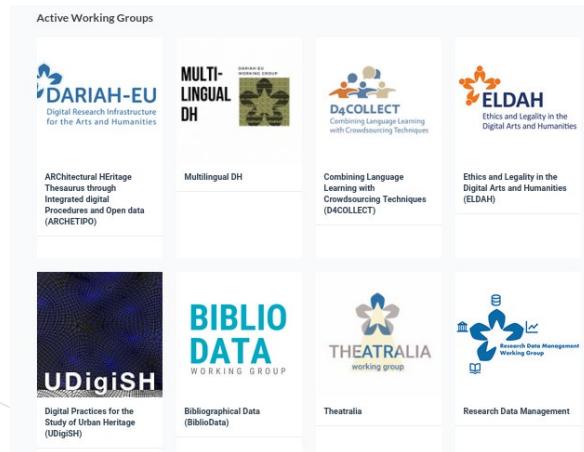
- Web portal to discover and access resources for CSH
 - <https://marketplace.sshopencloud.eu/>
 - Tools and services, Learning materials, Workflows, Datasets, Publications
- 3 Fundamental principles
 - Contextualization, Community, *Curatorship* (!)
- Entry point for SSH in EOSC
- Supported by CLARIN, DARIAH & CESSDA

DARIAH (II) Education and training

- DARIAH-CAMPUS
 - <https://campus.dariah.eu/>
 - Discovery framework and hosting platform for learning resources.
 - Complete courses, lessons, conferences or videos.
- DARIAH-TEACH
 - <https://teach.dariah.eu/>
 - Open Educational Resources (OER) platform for educators and students of Digital Arts and Humanities.
- Digital Humanities Course Registry
 - <https://dhcr.clarin-dariah.eu/>
 - Platform that offers teaching activities in the field of digital humanities around the world.

DARIAH (III) Working Groups

- Transnational and multidisciplinary
- 17 active WGs
- <https://www.dariah.eu/activities/working-groups-list/>
- Bottom-up design, grants (5K€), etc.



CLARIAH-ES: CLARIN & DARIAH



Distributed RI: **12** centres

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

- EHU ([HiTZ](#)) : *German Rigau & Mikel Iruskieta & Xabier Arregi*
- UAlicante ([BVMC](#)) : *Manuel Marco*
- USC ([Instituto da Lingua Galega](#) y [CiTIUS](#)) : *Elisa Fernandez*
- UNED ([LINDH](#) y [LENAR](#)) : *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](#): *Alvaro Fanego*
- [Dialnet](#): *Elena López Tamayo*
- **0.2M€** MICIN (fee) & **26+** Full Time Equivalent > **1.4M€** (in-kind)
- Strategic research network [CLARIAH-ES](#) **60K€** (2023-2024 MICIN ...)



Summary



- Unthinkable **AI** & **NLP** applications just a few years ago
- Even more **amazing** results in the near future (day by day)
- **Experts**, Supercomputing, Data
- **CLARIAH-ES research infrastructure:** CLARIN and DARIAH
 - 12 research centres
 - Experts, Community, Data, Tools & services, IT-HPC-infrastructure
 - Setting up the first B centre @ HiTZ ...
- **Coordination and Cooperation**
 - EU, national, regional
 - Administration, academia, technology centers, companies
 - **EOSC**
 - **SSHOC**
 - **EOSC EU node**





eman ta zabal zazu

Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

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

hitz.eus

HiTz

Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology



CLARIAH-ES