



HITZ

Hizkuntza Teknologiako
Zentroa
Basque Center
for Language Technology



LATXA

Chatbot
abierto
en euskera

latxa.hitz.eus

Investigación Avanzada en Inteligencia Artificial Generativa

Informe anual
2025

ORGANIZACIÓN

HiTZ es un centro de investigación multidisciplinar líder en Tecnología del Lenguaje e Inteligencia Artificial Generativa, que reúne a expertos de siete departamentos de la Universidad del País Vasco. Su misión principal es investigar en tecnologías del lenguaje y del habla, con un fuerte enfoque hacia la transferencia de conocimiento y tecnología al sector empresarial. HiTZ está compuesto por dos grupos de investigación, IXA y Aholab, los cuales cuentan con una amplia experiencia que se remonta a 1993. Ambos han sido reconocidos como grupos de investigación de alto rendimiento por la agencia científica del Gobierno Vasco en su última evaluación. A lo largo de su trayectoria, los dos grupos han participado en más de 200 proyectos de investigación, desde iniciativas regionales hasta europeas, y han realizado más de 100 contratos con empresas de diversos sectores. HiTZ es miembro de BDVA y, a través de esa asociación, participa en la Alianza Europea de Inteligencia Artificial, Datos y Robótica.

HiTZ cuenta con más de 100 miembros, incluyendo informáticos, lingüistas y cinco técnicos de investigación, reconocidos internacionalmente en sus respectivas áreas científicas. En los últimos cinco años, los investigadores que actualmente forman parte del centro han publicado más de 200 artículos científicos. El grupo es líder en la aplicación de técnicas de aprendizaje profundo e inteligencia artificial generativa a texto, habla y modelos de lenguaje-visual. Su trabajo reciente en esta área ha sido citado más de 4.000 veces en los últimos dos años.

HiTZ también coordina la infraestructura de investigación distribuida CLARIAH-ES, el consorcio que administra la participación de España en CLARIN y DARIAH, ambas parte del Foro Estratégico Europeo sobre Infraestructuras de Investigación (ESFRI). Nuestro subdirector, German Rigau, lidera el consorcio CLARIAH-ES y es el coordinador nacional de España para CLARIN y DARIAH.

El impacto del centro en la sociedad se refuerza a través de la Cátedra de Inteligencia Artificial y Tecnología del Lenguaje (IA&TL), así como de dos másteres coordinados por el centro: el Máster Europeo Erasmus Mundus+ en Tecnologías del Lenguaje y la Comunicación (LCT) y el Máster en Análisis y Procesamiento del Lenguaje. Estos programas de máster, junto con un grado en IA y cuatro cursos de formación continua para profesionales, están diseñados para responder a las demandas de la industria y la investigación en este campo en rápida expansión. Además, HiTZ ofrece un Programa de Doctorado en Análisis y Procesamiento del Lenguaje.



Eneko Agirre

Director



German Rigau

Subdirector



Suna Şeyma Uçar

Miembro de la Junta



Aritz Farwell

Miembro de la Junta



Maite Oronoz

Miembro de la Junta



Inma Hernaez

Miembro de la Junta



Esther Miranda

Miembro de la Junta



Aitor Soroa

Miembro de la Junta

MENSAJE DEL DIRECTOR



Eneko Agirre
Director de HiTZ

La Inteligencia Artificial Generativa (IAG o GenAI) ha experimentado avances notables en los últimos años, especialmente en la Tecnología del Lenguaje (TL). Su rápido crecimiento se refleja en los Grandes Modelos de Lenguaje (GML o LLM), como muestran los progresos anuales de sistemas como GPT o Gemini. Una vez demostrada su capacidad para comunicarse con las personas, el pasado año dieron el salto más allá del lenguaje hacia la programación y las matemáticas. Han sido clave para ello nuevas técnicas como el *Reinforcement Learning with Verifiable Rewards* (RLVR), basado en valores verificables, y el *Test-Time Compute*. Al mismo tiempo, se ha dotado a los GML de la capacidad de utilizar herramientas y, gracias a su autonomía y a la coordinación con otros GML, se ha producido una explosión de agentes. La ampliación de las capacidades de los GML ha ido más allá del texto, incorporando el trabajo transparente con imágenes, vídeo y voz, lo que ha permitido su extensión a nuevas aplicaciones, como la interacción con lo que vemos en la pantalla. Resulta difícil creer que todo esto haya ocurrido en menos de un año.

En este contexto, las herramientas basadas en la IAG tienen un impacto profundo en muchos aspectos de la tecnología y de la sociedad. Esta era transformadora ofrece nuevas oportunidades a quienes sean capaces de aprovechar la tecnología de IAG más avanzada. No obstante, la irrupción repentina de la IAG ha generado importantes debates en torno a su desarrollo y despliegue responsables, como la transformación de numerosas profesiones, la desinformación, los sesgos perjudiciales, las elevadas necesidades energéticas y la ampliación de la brecha digital entre lenguas minoritarias y lenguas dominantes.

Para hacer frente a esta nueva era y aprovechar las oportunidades que ofrece, es fundamental fomentar la investigación abierta y pública en Tecnología del Lenguaje e IAG: sin investigación local, nuestra sociedad puede convertirse en un mero consumidor de tecnologías desarrolladas en otros lugares. Más allá de la TL, la IAG se está extendiendo a otras tecnologías, como la visión artificial o la robótica, y está impulsando una revolución en otras disciplinas científicas; por ejemplo, el Premio Nobel de Química de 2024 fue concedido a los creadores de Alpha-

Fold, una tecnología basada en la IAG, y los modelos basados en IAG se han impuesto en la predicción meteorológica.

En HiTZ asumimos plenamente esta nueva visión y hemos estructurado nuestra actividad en dos grandes líneas: IAG y TL por un lado, e IAG para las humanidades y las ciencias por otro. Muchos de los principales expertos en IAG y TL de España son miembros del centro HiTZ, y nos encontramos en la vanguardia de la investigación en este ámbito. Al mismo tiempo, nuestro centro multidisciplinar no deja de crecer. Como resultado, hemos publicado un número significativo de artículos en conferencias y revistas de primer nivel y contamos con más doctorandos que nunca. Lideramos redes de investigación en inteligencia artificial a nivel estatal y proyectos transformadores plurianuales. Un ejemplo claro de ello es Latxa, que desde la publicación de su primera versión en 2024 se ha convertido en el principal modelo abierto en euskera. Este logro obtuvo el premio al mejor artículo en ACL, la conferencia más prestigiosa del área. En 2025 lo adaptamos para que pudiera conversar con personas, alcanzando un nivel comparable al de GPT en evaluaciones públicas humanas. Latxa está actualmente disponible para pruebas, junto con los mejores modelos abiertos para el procesamiento del habla, bajo licencias abiertas.

Por último, la Cátedra HiTZ de Inteligencia Artificial y Tecnología del Lenguaje nos permite formar profesionales y transferir conocimiento y tecnología a la sociedad. Asimismo, la infraestructura de investigación CLARIAH-ES, coordinada por HiTZ, ha reforzado nuestras redes de colaboración nacionales e internacionales al proporcionar herramientas y recursos digitales para las ciencias sociales y las humanidades. Para concluir, trabajando activamente en la expansión de los beneficios de la IAG a nuevos dominios, hemos desarrollado modelos pioneros de lenguajes visuales y, junto con neurocientíficos, estamos participando en un ambicioso proyecto para estudiar la intersección entre el habla y el cerebro. Todos estos proyectos exitosos constituyen una base sólida para que HiTZ afronte con ilusión los retos futuros y, al mismo tiempo, sitúan a nuestro país como un referente en la investigación en Tecnología del Lenguaje e Inteligencia Artificial Generativa.

CLARIAH

HiTZ lidera CLARIAH-ES, la infraestructura distribuida de investigación digital que coordina la participación de España en los Consorcios Europeos de Infraestructuras de Investigación CLARIN (Infraestructura Común de Recursos y Tecnología Lingüísticos) y DARIAH (Infraestructura Digital de Investigación para las Artes y las Humanidades). CLARIAH-ES aspira a potenciar la investigación habilitada digitalmente y proporcionar a los investigadores acceso a recursos, herramientas y servicios digitales de vanguardia. Su enfoque multi- e interdisciplinar está diseñado para impulsar la investigación española en ciencias sociales y humanidades, fomentar la comunidad española de humanidades digitales y ayudar a posicionar estratégicamente a los investigadores de España en proyectos y programas internacionales, en particular, dentro del Espacio Europeo de Investigación. German Rigau, subdirector de HiTZ, es el coordinador nacional de España para CLARIAH-ES, CLARIN y DARIAH.

En la actualidad, CLARIAH-ES reúne a socios de doce instituciones y grupos de investigación líderes, de toda España, dedicados a impulsar las capacidades digitales en el ámbito de las ciencias sociales y las humanidades. En 2025, la infraestructura ha consolidado aún más su red de expertos y ha ampliado su proyección hacia la comunidad académica y el público general. En reconocimiento a su labor en este ámbito, CLARIAH-ES recibió el premio a la mejor infraestructura en el 7.º Congreso de la Sociedad Internacional de Humanidades Digitales Hispánicas.

HiTZ también lidera el nodo CLARIAH-EUS, que engloba a diversas instituciones y grupos dedicados a la investigación en y para la lengua vasca en el ámbito de las ciencias sociales y las humanidades. CLARIAH-EUS recibe financiación y apoyo del Gobierno Vasco, la Diputación Foral de Gipuzkoa, la Universidad del País Vasco (EHU) e HiTZ.

Más información aquí:

www.clariah.es y www.clariah.eus



German Rigau

Subdirector de HiTZ
Coordinador nacional en España
para CLARIAH-ES, CLARIN y DARIAH



Xabier Arregi

Coordinador de CLARIAH-EUS

 **CLARIAH-ES**

 **CLARIAH-EUS**

CÁTEDRA

En 2025 hemos consolidado la Cátedra HiTZ de Inteligencia Artificial y Tecnología del Lenguaje (IA&TL), puesta en marcha en 2024 en el marco de la Estrategia Nacional de Inteligencia Artificial (ENIA). La Cátedra es la única cátedra financiada del País Vasco y la única cátedra en España dedicada a la Tecnología del Lenguaje (más información en www.hitzeus/katedra).

En esta fase de consolidación, la Cátedra ha profundizado en sus tres pilares (docencia, investigación y transferencia del conocimiento) para maximizar el impacto positivo de IA&TL en la sociedad y en el sector productivo, fortalecer las actividades de investigación y afianzar la base de conocimiento que sustenta los programas académicos y profesionales. Este trabajo se lleva a cabo junto con nuestras entidades colaboradoras: Avature SL, la fundación Elhuyar, Ikerlan SCL, Euskaltel-MasOrange, Multiverse Computing y la fundación Tecnalía Research and Innovation. La relevancia, el número y la excelencia de estas organizaciones ponen de manifiesto la confianza del sector productivo en el potencial de la Cátedra para generar beneficios para la sociedad.

La Cátedra sigue ofreciendo cuatro cursos de especialización en IA Generativa, Aprendizaje Profundo y Tecnología del Lenguaje—actualizados con los últimos avances y dirigidos a profesionales, investigadores y estudiantes—, impartidos por expertos de primer nivel (hitzeus/training).



HiTZ

Adimen Artifizial eta Hizkuntza
Teknologiako Katedra
Cátedra de Inteligencia Artificial
y Tecnología de la Lengua

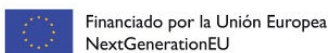


Aitor Soroa

Director de la Cátedra HiTZ de Inteligencia Artificial y Tecnología del Lenguaje (IA&TL)



Funded by:



Partner companies:



HiTZ EN NÚMEROS

Investigación y transferencia

31

proyectos de investigación

9

proyectos de transferencia de conocimiento

4

Otros proyectos

7

tesis doctorales defendidas
(3 internacionales)

23

artículos de revistas
(13 Q1)

41

artículos en congresos
(15 A o A+)

3

capítulos de libro

1

libro

TRANSFERENCIA

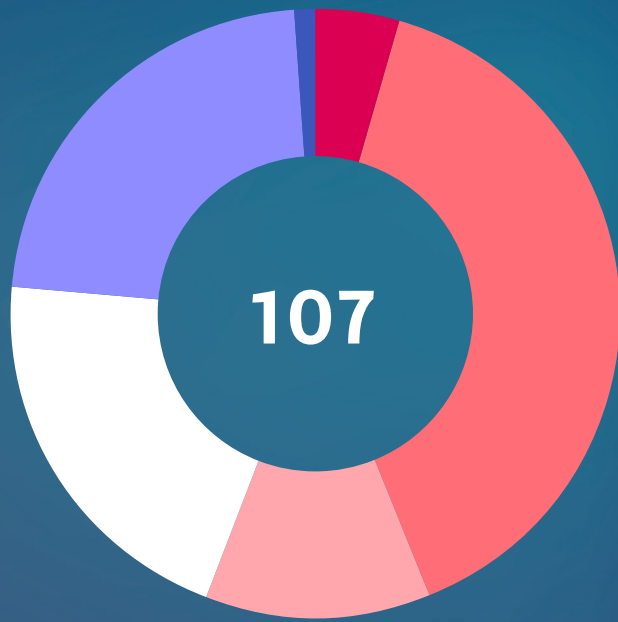
Todos los modelos, códigos y datos que hemos desarrollado para el euskera en el centro HiTZ están disponibles públicamente. Entre ellos se encuentran los mejores chatbots abiertos, modelos de reconocimiento de voz y modelos de síntesis de voz para el euskera, que permiten su uso comercial.

hitz.eus/es/resultados

Personas

Miembros

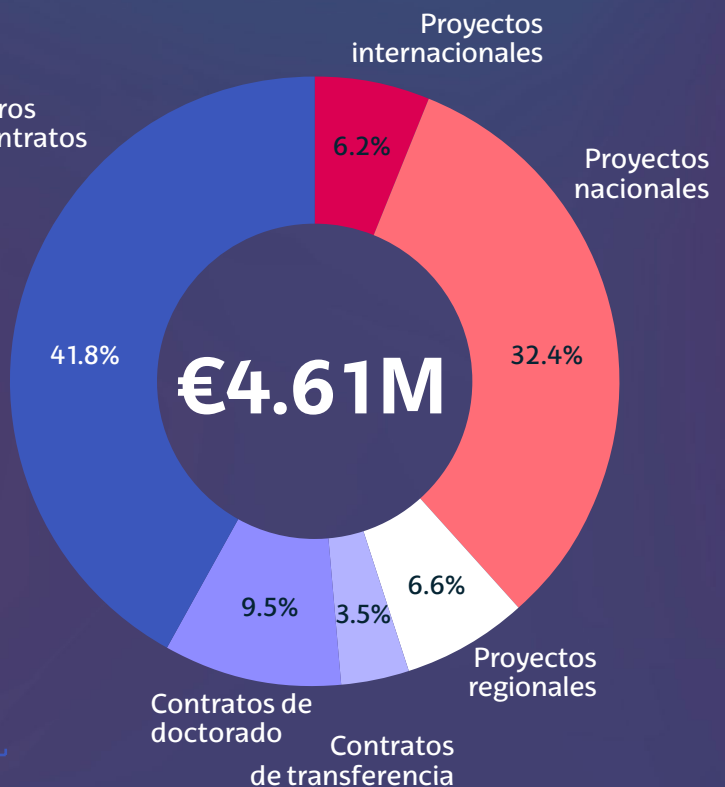
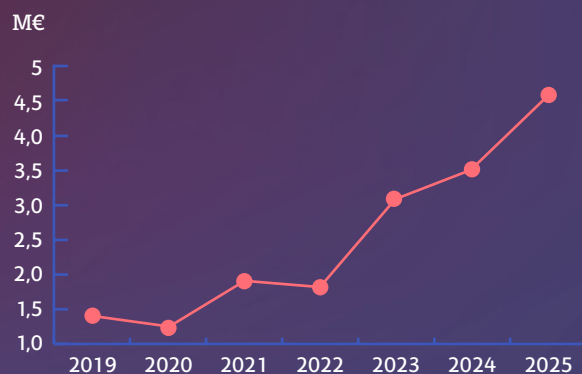
	Gestión y administración	5
	Profesores	42
	Investigadores postdoctorales	13
	Otras investigadoras	22
	Investigadores con becas predoctorales	24
	Colaborador académico	1



Alumnado

	Estudiantes de master	68
	Estudiantes de doctorado	46

Presupuesto



ÁREAS DE INVESTIGACIÓN



Investigador principal:

Extracción y recuperación de información



Aitor Soroa



Investigador principal:

Traducción automática



Gorka Labaka



Investigador principal:

Interacción persona-computadora



Gorka Azkune



Investigadora principal:

Recursos del habla y el lenguaje



Ainara Estarrona

INFRAESTRUCTURA

18

multiprocesador GNU/Linux

1

servidor SPARC Solaris

8

servidores H200 141GB vram GPU

36

GPU cluster A100 80GB vram

56

GPU cluster A100 80 GB vram (compartido / DIPC)

14

L40 GPUs con 48GB vram

26

GPU varios modelos

1.2

de capacidad de almacenamiento bruto en red

PB



Análisis de texto

Investigador principal:



Rodrigo Agerri



Tecnologías de Voz

Investigadora principal:



Inma Hernaez



Dominios médico y legal

Investigadora principal:



Arantza Casillas



Humanidades digitales y educación

Investigador principal:



Mikel Iruskietia

1

sala aislada acústicamente

1

Interfaz de audio Midi 4x4 Behringer

1

Quiet PC Sentinel Fanless i10

1

Delsys Trigno™ Wireless EMG System

Acceso a 900.000 horas de GPU (valorado en 2,2 millones de euros) en el Superordenador EuroHPC para la investigación de grandes modelos para lenguas europeas con pocos recursos.



PRÁCTICAS

68

estudiantes
en másters

46

estudiantes en
el programa
de doctorado

12

TFMs de
EMLCT
finalizadas

21

TFMs de
HAP/LAP
finalizadas

95

estudiantes en
cursos de formación
continua sobre LLM

3

Ikasiker

7

prácticas internas
y externas



ACTIVIDADES

31
seminarios



7
webinarios

3
talleres

4
premios

Proyecto ganador del III Certamen de Seminarios de la Fundación Duques de Soria de Hispanismo Internacional: "Humanidades digitales en acción: herramientas para el análisis de prensa histórica en español" (desarrollado por CLARIAH-ES)

Premio a la "Mejor herramienta, recurso e infraestructura desarrollada" para CLARIAH-ES, en la VII Edición de los Premios de la Sociedad Internacional de Humanidades Digitales Hispánicas (HDH)

Premio al Mejor Artículo o Presentación en el área de Ingeniería y Arquitectura, en la VI edición del congreso IkerGazte (Nazioarteko Ikerketa Euskaraz): Jaione Bengoetxea Azurmendi por el trabajo "Euskara eta gaztelaniazko kontra-narratiben sorkuntza: datuen sorrera eta ebaluazioa"

Premios TFG-TFM MONDRAGON Sariak 2025. Premio Eika al mejor proyecto en lengua vasca: Irune Palacios Orueta por el trabajo "Euskarazko Medikuntzaren Domeinuko Lehenengo Corpus Sintetikoaren Sorrera: Metodologia eta Hastapenak"



