



Definition of the use cases for the 5 application domains: data and requirements

Deliverable D4.1, version 1.0

Project name	Large Language Models for the European Union	
Project acronym	LLMs4EU	
Grant agreement number	101198470	
Project web-site	www.llms4eu.eu	
Deliverable Lead	Expert.ai	Flavio Merenda
Contributors	Expert.ai	José Manuel Gómez Pérez
Internal reviewers	FBK ALT-EDIC	Alessio Brutti Maria Velasco

History of Changes

Version	Date	Changes
1.0	24 September 2026	First version

This document is a public report. The information herein is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.



**Co-funded by
the European Union**

*This project is funded by Digital Europe programme
under Grant Agreement no. 101198470*

Contents

1.	Executive summary.....	7
2.	Introduction.....	8
2.1.	Background.....	8
2.2.	Objectives and document organization.....	8
3.	Defining and prioritising LLMs4EU use cases.....	10
3.1.	Selection criteria.....	11
3.2.	Specified use cases (fully defined).....	12
3.3.	Other use cases under consideration (partially defined).....	23
4.	Use case data and requirements collection.....	24
4.1.	Collection methodology	24
4.2.	Collected data and requirements.....	28
5.	Use case statistics.....	42
5.1.	Language coverage	42
5.2.	Data overview.....	44
5.3.	Fine-tuning types.....	45
5.4.	Thematic areas coverage	47
6.	Conclusions.....	49
	References.....	50

Table of Abbreviations

Abbreviation	Extended
AI	Artificial Intelligence
ALT-EDIC	Alliance for Language Technologies - European Digital Infrastructure Consortium
API	Application Program Interface
ASR	Automatic Speech Recognition
B	Billion
CFT	Continual Fine-Tuning
CPT	Continual Pre-Training
CPU	Central Processing Unit
D4.1	Deliverable 4.1 Definition of the use cases for the 5 application domains: data and requirements
D6.2	Deliverable 6.2 Legal guidelines for data, services and models
DPO	Direct Preference Optimization
EU	European Union
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPU	Graphics Processing Unit
HPC	High Performance Computing
IE	Information Extraction
LLM	Large Language Model
LoRA	Low-Rank Adaptation
M	Month
NER	Named Entity Recognition
NLP	Natural Language Processing
O3	Objective 3: Remove technical and operational obstacles to the production of models covering all the official EU languages and regional languages of Alliance member countries, targeting low resourced languages and reducing the amount of data required (labelled as well as unlabelled)
OCR	Optical Character Recognition
PEFT	Parameter-Efficient Fine-Tuning
QA	Question Answering
Q&A	Questions and answers
QLoRA	Quantized Low-Rank Adaptation
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
RE	Relationship Extraction
R&D	Research & Development
SFT	Supervised Fine-Tuning
SME	Small and Medium Enterprise
T2.1	Task 2.1 Identification and Specifications for Language Data
T4.1	Task 4.1 Collection of adaptation/fine-tuning requirements for the use cases and for generic domain adaptation
T4.2	Task 4.2 Development and optimization of the fine-tuning adaptation strategies
T4.3	Task 4.3 Design and implementation of an API framework for multi-model fine tuning on cloud-edge infrastructures
T4.4	Task 4.4 Fine tuning of the pre-trained models for all EU languages on 5 industry-validated domains and release of the finetuned models

*D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU*

T4.5	Task 4.5 Adoption and assessment by the use-case owners
WP2	Work Package 2 Data Collection and Data Infrastructure
WP4	Work Package 4 Methodologies, algorithms and infrastructure for model fine-tuning
WP5	Work Package 5 Distributed Evaluation Centre for LLMs
WP6	Work Package 6 Legal and ethics compliance aspects
XML	eXtensible Markup Language

List of Tables

Table 1. Defined use cases.....	18
Table 2. Fine-tuning types.....	25
Table 3. Overview of the use cases in the Science domain.....	31
Table 4. Overview of the use cases in the Public Services domain.....	34
Table 5. Overview of the use cases in the Tourism domain.....	36
Table 6. Overview of the use cases in the Telecommunications domain.....	38
Table 7. Overview of the use case capabilities in the Energy domain.....	41
Table 8. Project-relevant languages categorized by group.....	42
Table 9. Overview of language support across use cases.....	43
Table 10. Overview of the nature of project data across use cases.....	45
Table 11. Overview of the fine-tuning techniques required across use cases.....	46
Table 12. Mapping between use cases and thematic areas.....	47

List of Figures

Figure 1. Fine-tuning types.....	26
Figure 2. SciDEV training workflow.....	29
Figure 3. ClaimAnalysis use case workflow	29
Figure 4. Research_SSH use case workflow.....	30
Figure 5. Energy domain model training workflow.....	39

1. Executive summary

This report presents the LLMs4EU project use cases, their definitions, their prioritisation, and the collection of requirements needed to execute the core activity of the project: model fine-tuning [1]. This document serves as the technical backbone of the project and provides a roadmap for the development and implementation of technological solutions designed jointly by use case providers and technology partners in the field of GenAI [2]. It characterizes the models in terms of innovation, technological impact, and potential limitations, with the objective of ensuring that European companies, and especially SMEs, have access to the tools and resources required to remain competitive in language technologies, particularly LLMs [3].

The introductory section situates the fine-tuning process to be carried out in WP4 within the context of the overall project. It explains the background and motivation for this work, defines its objectives, and describes the structure of this report.

The core of the document details the proposed use cases and the criteria applied for their proposal and prioritization. A substantial part of the report is dedicated to collecting and analysing the fine-tuning requirements needed to establish a clear development plan for GenAI technologies and to enable effective collaboration among all technical partners involved. It describes the methodology used to collect, validate, and consolidate training plans from use case owners and collaborators, addressing both functional and non-functional requirements. These include model capabilities and innovation aspects, supported languages, identified training data, compliance with legal and ethical standards, and technical aspects such as base model selection, fine-tuning methods, and computing resource requirements for both training and deployment.

Building on this methodology, the roadmap establishes a structured fine-tuning approach to guide all partners in training the project's defined use cases. It also serves to support an iterative process for continuous refinement and expansion to additional use cases under study.

Since the project aims to provide EU citizens and companies with competitive solutions in the field of GenAI, this document should be considered a living document. It may expand over time to include additional use cases, which will be fully evaluated and executed once each use case has been clearly defined in terms of fine-tuning, evaluation, and data planning. This document presents an overview of the use cases that have currently reached a mature stage of development. In addition, the methodology described in this document (Section 4.1) is intended not only for the specific use cases presented in this document (Sections 3.2 and 3.3) but also as a general framework that can be applied to future use cases beyond the scope of the project. This methodological approach may also serve as a reference for organizations and companies involved in model fine-tuning and evaluation activities.

2. Introduction

This section introduces the role of WP4 within the broader context of the project, outlines the objectives of this work package, and describes the organization and structure of the report.

2.1. Background

The LLMs4EU project, led by ALT-EDIC, aims to preserve Europe's linguistic and cultural diversity by supporting the development of GenAI for all EU languages¹. It brings together leading AI players to provide European companies, especially SMEs, with tools and resources for training and deploying LLMs, ensuring compliance with European regulations (AI Act², GDPR³, etc.) and promoting competitiveness in language technologies⁴.

In the context of the project, WP4 represents the backbone of the technical development of the project. It contributes to Objective O3 by fine-tuning language, speech, and multimodal foundation models [4] for the languages identified in T2.1 and for the tasks relevant to the five industry-validated application domains (see Section 3). This report, deliverable D4.1, details the work conducted under Task T4.1, which involves systematically gathering comprehensive requirements for model fine-tuning across the five application domains covered by the use cases. D4.1 addresses key aspects, including relevance, expected impact, technological challenges, data availability, and regulatory constraints. The outcomes of T4.1 guide the development of fine-tuned models in T4.4. Additionally, T4.1 supports T4.2 by guiding the development of methodologies for domain adaptation and fine-tuning, as well as the design of composable, distributed APIs for integration into the model catalogue in T4.3.

2.2. Objectives and document organization

The objective of this document is to describe the process adopted within the project to define the use cases, carried out in collaboration with stakeholders, data providers, and technological partners, as well as the methodology and roadmap developed to collect the necessary requirements for implementing the solutions associated with these use cases.

The information reported in this document reflects the collaborative effort of the various partners involved in each specific use case as summarized in the use case technical coordination table, a project-internal use-case registry that defines each use case and assigns leaders for each development area involved, including technical integration, HPC provisioning, data creation, evaluation, fine-tuning, and legal, ethics, and compliance. More specifically, this document focuses on the definition of the use cases in the technical coordination table and the plans developed to establish the training procedures. This is an ongoing process that prioritizes use

¹ https://european-union.europa.eu/principles-countries-history/languages_en

² <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

³ <https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr>

⁴ <https://digital-strategy.ec.europa.eu/en/policies/language-technologies>

cases with fully defined training plans, and the methodology adopted will be applied in the future to extend the process to use cases still under study. We will refer in this document to *fully defined* use cases as those that include a complete fine-tuning plan. In contrast, *partially defined* use cases are those that have been proposed and discussed and, in some cases, partially developed, but are currently missing a complete fine-tuning plan and have not been prioritized. Nevertheless, these use cases could be completed during the project, their fine-tuning plans finalized, and they could have a significant impact. The document also provides statistics on the languages covered in the use cases, the data utilized throughout the project, an overview of the fine-tuning techniques employed, and their connection to the thematic areas defined in task T4.2, which supports the theoretical development of these technologies and ensures that the most suitable techniques are applied to address the specific requirements of each use case.

The remainder of this document is organised as follows. Section 3 presents the use cases, both fully and partially defined, including the criteria used for their selection. Section 4 outlines the methodology employed to collect use case requirements and provides a detailed description of the requirements gathered for each domain. Section 5 presents statistical analyses derived from the collection process, and Section 6 summarises the conclusions.

3. Defining and prioritising LLMs4EU use cases

Open-source packages and APIs have expanded access to advanced AI, including LLMs, allowing European businesses to apply GenAI in translation, transcription, content generation, and data analytics. Their effectiveness varies across EU languages, mainly because some lack sufficient high-quality, domain-specific training data [5]. Building reliable models requires datasets that are large, relevant, and legally compliant.

Building on the three foundational elements common to all AI initiatives, which are *access to data and computational resources*, *clearly defined market opportunities*, and *socio-political priorities*, the ALT-EDIC, following a consultative process with its member states, identified five seed use cases. Each seed use case was proposed within a distinct application domain. During the first year of the project, these initial use cases were further refined, and their respective application domains were expanded with additional use cases.

The seed use case selection was guided by several criteria: (i) *the level of interest from member states*, (ii) *the engagement and availability of industrial partners*, (iii) *the accessibility of data for training and fine-tuning AI models*, and (iv) *the potential for replication across multiple European countries*. The selected use cases defining the five application domains are as follows:

1. **Science** (led by *CROSSLANG*): science-based assistant for researchers and policy makers.
2. **Public Services** (led by *Notilyze*): leveraging public administration data to develop AI-enhanced assistance to public services.
3. **Tourism** (led by *Arctur*): leveraging cultural heritage data to develop a Cultural Tourist AI Assistant promoting sustainable tourism development.
4. **Telecommunications** (led by *Orange*): NLP research on datasets of EU operators (Orange, Tiscali, KPN BV, etc.) for the deployment of AI services.
5. **Energy** (led by *EDF*): developing a specific multi-modal and multi-dimensional AI solution to ease European nuclear engineers mastering, use and creation of highly normative and regulatory requirements content.

Building on the five defined application domains, this section presents the use cases identified during the project's first year, beginning with the general criteria applied across all domains and then detailing the domain-specific cases. It also includes use cases that, although identified as relevant and promising, have not yet reached sufficient maturity to be considered fully defined. These cases are briefly described, highlighting the specific challenges encountered in their development and discussing possibilities for further refinement over the course of the project.

This is an ongoing process, and the document should be considered a living reference that will guide the inclusion of potential future use cases. It provides a snapshot of the status of defined use cases, with the aim of expanding the pool of relevant applications and equipping EU citizens and companies with competitive tools in the field of GenAI.

The information presented has been extracted from the technical coordination table, which tracks, across work packages, all relevant details for each use case as well as the partners involved in its development.

3.1. Selection criteria

Within the five strategic application domains (*Science, Public Services, Tourism, Telecom, and Energy*), seed use cases were selected for their relevance to European stakeholders and their potential to strengthen Europe's digital competitiveness and market positioning. During the first year of the project, the initial use cases were further refined, and additional use cases were introduced through contributions from partners in each application domain. For each domain, the designated domain leader coordinated the identification, development, and formal definition of both the original and newly proposed use cases. A variety of factors were considered in defining these cases, with particular attention to the project's core objective: providing European citizens and SMEs with tools designed to pragmatically address concrete tasks while respecting all European language diversities and values. The proposed use cases were developed through a rigorous methodology designed to define them in accordance with five core pillar criteria:

1. **Relevance:** Address the practical importance of the use case, including alignment with EU priorities such as digital sovereignty, inclusion of multilingual populations, and support for public services, research, and industry. Emphasize how the outcomes will benefit society, government, and SMEs.
2. **Impact:** Highlight expected outcomes and benefits, such as improved accessibility of AI technologies across languages, increased competitiveness of EU AI solutions, and promotion of linguistic diversity. Include societal, economic, and technological effects.
3. **Technological challenges:** Covering official EU languages and regional languages of member countries, targeting low resourced languages and reducing the amount of data required (labelled as well as unlabelled).
4. **Data availability:** Assess the existence, accessibility, and quality of linguistic datasets across different EU languages and domains. Include strategies for collecting, sharing, and curating data while addressing gaps in low-resource languages and ensuring sufficient coverage for training robust LLMs.
5. **Regulations:** Ensure compliance with EU existing and upcoming data and AI regulations as well as the fair and ethical treatment of data and AI processes. This includes GDPR, AI Act requirements, transparency, accountability, and data protection by design.

Taking these criteria into account, domain application leaders, in coordination with project partners, led the proposal, selection, involvement and prioritization of use cases during biweekly meetings throughout the first half of the project's first year. These meetings helped define the use cases within the context of the project. To track the progress of proposed use cases through their definition, ALT-EDIC proposed the creation and maintenance of a technical coordination table to capture all information related to each specific use case. This table includes all relevant stakeholders and their defined roles across work packages involved in the development process, including *data collection and data infrastructure (WP2), methodologies, algorithms and infrastructure for model fine-tuning (WP4), distributed evaluation centre for LLMs (WP5), and legal and ethics compliance aspects (WP6)*.

The following sections present the defined use cases, as well as those that, although proposed, have not reached sufficient maturity to allow for a complete definition of the required specifications needed to propose a fine-tuning plan yet.

3.2. Specified use cases (fully defined)

This section presents the use cases that provided a fully defined fine-tuning plan. We will briefly describe them according to the selection criteria applied, such as their innovative technological capabilities, the languages covered, the data involved, and specific potential legal considerations. All this information has been extracted from the use case technical coordination table, which tracks, across WPs, all relevant information related to each use case, as well as all the partners involved in its development. We will organize them by application domain and provide a final summary at the end.

Science

Research and innovation are central to a European knowledge-based economy and society. By establishing a European open infrastructure for science-oriented LLMs, research disparities can be reduced, multilingual content promoted, decision-making enhanced, and industry–research collaboration strengthened. Consequently, the Science domain has proposed three use cases focused on developing domain models and analytical tools to directly support and accelerate scientific research and innovation.

SciDev will develop a model specialized for the scientific domain [6]. Along with DomainAdapt, which targets the telecommunications domain, these constitute the two domain-specific adaptations in the project that require further pre-training to improve the models' domain expertise [7]. This model will assist developers in managing scientific terminology, domain-specific knowledge, and interactive user engagement, with the goal of functioning across multiple EU languages, including English, French, Finnish, Romanian, and Greek. The model will be trained on millions of publicly available scientific publications and validated in other science-related use cases (ClaimAnalysis and ResearchSSH), with its performance compared against generic models that are not specialized.

Building on the resulting model from SciDev, **ClaimAnalysis** will enable researchers, pharmaceutical companies, and patent attorneys to extract, verify, and explain key claims in scientific papers [8], particularly in the biomedical field. Its capabilities will include automatically identifying main claims and cross-referencing them against relevant literature to determine whether they are supported or refuted, leveraging structured datasets such as FEVER [9] and SciFact [10].

The **Research_SSH** use case will focus on advanced semantic search in the Social Sciences and Humanities. It will enable natural language queries over bibliographic databases, support contextualized searches, and generate synthetic, context-rich answers with precise references and publication excerpts. Its design will draw inspiration from NotebookLM [11] for RAG-based contextual QA [12], ResearchRabbit [13] for dynamic literature mapping, and Scite for citation-aware search.

All sciences use case initiatives will incorporate legal safeguards, ensuring data privacy and compliance with open-access and data governance requirements through careful dataset selection and anonymisation.

Public Services

The development of AI in the public sector holds significant potential to transform how governments interact with citizens and manage information [14]. By providing guidance on permits, certifications, and legal matters, AI can enhance accessibility and efficiency of public services. It supports experts in delivering accurate, timely, and consistent information, helping counter disinformation and improve decision-making. Furthermore, AI enables inclusivity through multilingual and multimodal capabilities, while increasing productivity by automating repetitive tasks, even in contexts with limited resources. The proposed use cases focus on developing AI tools to improve access to public administration information and citizen services, while addressing challenges and strategies for effective public sector implementation.

Efforts include **PADA** (Public Administration Document Assistant), an AI system designed to support public employees in handling administrative documents through QA, RAG, and IE [15]. The system will be trained and validated on datasets derived from thousands of Italian local public records, complemented by custom evaluation questions to ensure robustness, reliability, and strong domain alignment.

LuxHubAI is a use case focused on strengthening digital support for Luxembourgish, a low-resource language spoken daily by around 300 000 people [16, 17]. Although the Zenter fir d'Lëtzeburgerger Sprooch (ZLS) has promoted and formalized the language since 2018, linguistic datasets and language technologies remain fragmented across institutions. To address this, ZLS developed SproochHub, an open-source platform integrating speech-to-text, large text corpora, advanced search tools, and audio alignment features, but interoperability is still limited. LuxHubAI aims to federate SproochHub through LuxHub, aligned with European Language Data Space standards, and to develop an AI assistant that supports users in navigating both the platform and the shared data catalogue.

LuLoBots, coordinated by the Ministry for Digitalisation, will create a Luxembourgish-language AI assistant for government chatbots. It will let users ask questions in Luxembourgish and get accurate answers, even when the official knowledge is in French or English. The system works by linking Luxembourgish queries to existing expert information, helping integrate Luxembourgish into public services and making administrative tools more accessible to citizens.

The **AI4LC** (AI Assistant for Legal Consultations) use case aims to fine-tune LLMs for in-context QA in the legal domain [18]. The system will allow users to ask natural-language legal questions and receive answers explicitly grounded in legislative texts, such as laws, regulations, and directives. Trained on tens of thousands of human-written question-answer examples, it will retrieve relevant legislation via semantic search, interpret the user's query, and generate traceable, accurate answers.

Similarly, **ELISA** (Efficient Language Interface for Services Access) is intended to support citizens of Sibiu municipality in navigating local government services. It will aim to address common

challenges such as complex bureaucratic procedures, unclear documentation requirements, and difficulty accessing information across multiple official sources. The system will provide guidance on permits and authorizations, appointment scheduling, form completion, civic procedures, and social assistance applications. Its primary target users will be Romanian-speaking residents seeking information and assistance related to municipal services.

The **IRLT_LLM** (LLM driven Irish Language Transcription) use case has the objective of developing an AI system for transcribing Irish language audio and visual content [19, 20] from TG4, Ireland's primary Irish-language broadcaster. Much of TG4's archival material lacks transcription, making cataloguing and search difficult, and the scarcity of large-scale Irish datasets limits ASR capabilities. The project plans to generate automatic transcriptions of existing archives, create a model for future programming, train a large-scale Irish speech model, and assess transcription quality to identify gaps in current datasets. The strategies developed will serve as a case study for extending ASR solutions to other low-resource languages.

The **ARTE-PT** use case aims to enhance Portugal's public services portal⁵ chatbot by fine-tuning an LLM to improve accuracy, clarity, and usefulness across public services. While the current assistant is powered by a proprietary model, the new version will build on open models, contributing to Portugal's digital sovereignty. This upgraded assistant will help users including citizens, foreign residents, and newcomers navigate administrative rules, procedures, and required documents through natural language guidance, simplified explanations, and multi domain support. Fully integrated into the live portal, the system provides information and procedural guidance at national scale, ensuring immediate real-world impact and long-term sustainability.

The models that will be fine-tuned in the **GenAI4Med** use case will support healthcare experts by answering domain-specific questions, generating smart reports on healthcare R&D topics, and providing activity summaries for healthcare organizations. Key capabilities include answering expert queries with source-based information, reporting on the state-of-the-art in specific R&D domains, summarizing [21] organizational R&D activity, identifying emerging trends over time, and assessing an organization's positioning relative to current state-of-the-art research. The system will leverage both public and private healthcare data, with access to private data determined by user permissions.

Finally, the **Diavgeia** use case goal is to adapt LLMs to understand Greek and public administration terminology, enabling key IE from official decisions and documents on the Diavgeia platform. The fine-tuned models will serve two main purposes: simplifying the document-publishing process by automatically completing metadata fields and supporting public administrations in generating actionable insights through structured data extraction, trend identification, and anomaly detection. These capabilities aim to make Diavgeia a searchable, interpretable, and analysable knowledge base, enhancing transparency and data-driven decision-making in Greek public administration.

⁵ <https://www.gov.pt>

Across all initiatives, legal considerations are being carefully addressed, with training and evaluation data anonymised as needed and compliance with data privacy, national regulations, and open-access requirements ensured from the outset.

Tourism

Europe's rich cultural and natural heritage attracts many tourists, but sustainable tourism faces challenges like overcrowding and uneven visitor distribution. This work focuses on developing AI tools to improve access to cultural heritage information, offering interactive, on-demand guidance that covers popular sites while actively promoting lesser-known attractions [22].

HeritAI will provide an assistant capable of conversational interaction and QA, giving users easy access to factual data on cultural heritage sites tailored to their needs [23]. The system will operate across multiple languages, including Slovenian, Italian, Polish, Dutch, Portuguese, Greek, and English. Its capabilities will combine RAG with conversational QA, leveraging a dataset of at least 100 000 tokens per location and an estimated 1 000 questions for training and evaluation. Legal considerations will be addressed by relying on publicly accessible data and maintaining the research-oriented nature of the project, ensuring that all data usage complies with relevant regulations and privacy standards.

Telecommunications

The telecommunications sector includes both technical areas such as network operations, field services, and customer care, and regulatory areas governed by laws like the AI Act, Digital Markets Act⁶, and NIS2 Directive⁷. Technical documents in this sector often have specialized vocabulary, tables, graphs, API specifications, and code. Language technology tools can help streamline processes, support tasks, and improve customer services in many European languages. However, the sector's highly sensitive data makes it difficult to create realistic synthetic datasets. Taking this context into account, work in the Telecom domain will focus on developing AI tools to analyse customer interactions and support operational decision-making, while ensuring data privacy [24].

DomainAdapt, as the already described SciDev for the Science domain, will adapt a foundational model to acquire broad telecom knowledge, enabling generation for a wide range of tasks across English, French, Italian, Dutch, Slovenian, and Polish. The use case involves further pre-training and instruction fine-tuning on a large set of relevant documents and synthetic instructions covering telecom sub-domains, including technical, regulatory, and customer care. The resulting model is expected to acquire multi-language knowledge of these domains and serves as a foundation for a wide range of practical applications in the telecom sector.

ConvAnalytics will extract relevant information from customer-agent or customer-chatbot conversations, including summaries, call reason identification, proposed solutions, sentiment

⁶ <https://digital-markets-act.ec.europa.eu>

⁷ <https://digital-strategy.ec.europa.eu/en/policies/nis2-directive>

analysis, and estimated urgency levels. Its capabilities combine summarization and classification, trained initially on 250 annotated conversations in French, with a target of 500 annotated conversations for each supported language, including French, Italian, Dutch, Slovenian, and Polish.

PIID (Personal Identifiable Information Detection) will focus on developing a tool for effective anonymisation and pseudonymization of customer-related data [25]. The tool will include entity detection capabilities to ensure sensitive data is properly handled. This will allow datasets to be safely shared and used for model training, while ensuring full compliance with GDPR and other relevant data privacy regulations. Legal considerations will be carefully addressed.

Finally, **KGGen** (Knowledge Graph Generation) will deliver a tool for automatic domain-specific knowledge graph generation from unstructured text [26]. The system operates in two main stages. First, it performs NER [27] to identify and classify relevant entities within input documents; these entities become the nodes of the knowledge graph. Second, it performs relationship extraction [28] to detect semantic connections between entities and represent them as structured triples (subject–predicate–object), which define the edges of the graph. Relationship extraction will be carried out using an LLM to capture complex semantic relations. For the NER component, the use case will systematically compare two approaches: LLM-based extraction versus encoder-only models [29] optimised for sequence labelling. By benchmarking multiple encoders and LLMs, similar to the PIID use case, KGGen aims to determine the most accurate and efficient architecture for building reliable, structured knowledge graphs from domain texts.

Energy

The nuclear energy sector faces challenges in its digital transition due to a complex regulatory environment, inconsistent international practices, and diverse professional vocabularies. Requirements and recommendations are spread across multiple documents and repositories, complicating standardisation and knowledge management [30]. This application domain aims to develop specialised multi-modal, multi-dimensional LLMs for the nuclear industry, leveraging its ability to handle complex data dependencies, semantic and syntactic variation, and scarce data.

The proposed **Nuclear Requirement Engineering (NER)** use case focuses on developing AI tools to support requirements engineering and system analysis. Activities such as requirement engineering extraction, elicitation, and formalisation aim to capture, refine, and structure requirements across multiple languages, including English, French, Czech, Finnish, and Swedish. Building on these, the use case will enable users to input two lists of requirements along with a System Breakdown Structure and generate an AI-assisted allocation matrix that links requirements to system elements with justifications. Iterative refinement can be applied, and the resulting matrices can be exported for further use in engineering tools. Capabilities rely on specialised LLMs to analyse requirements and system architectures, and training data may include synthetically derived datasets [31] to cover diverse language combinations and matching scenarios. Legal considerations are addressed by adhering to relevant norms, such as AFNOR (Association française de normalisation) standards, which restrict AI usage in certain contexts, and ensuring compliance with applicable data governance and privacy regulations.

Overview and analysis of defined use cases

Table 1 provides an overview of all defined use cases, highlighting key aspects such as the languages covered, data availability, legal considerations, and the partners contributing to each use case. It also outlines those involved in data acquisition, planning, or evaluation, illustrating the extensive collaboration and diverse expertise mobilized across the LLMs4EU project for each use case. The table reports the languages with the two-letter ISO 639:2023⁸ codes.

⁸ <https://www.iso.org/standard/74575.html>

Table 1. Defined use cases

Domain	Name	Description	Capability	Languages	Data		Legal aspects		Involved partners
					Available amount of representative input data for testing	Estimated amount of data to be created (if any)	Legal issues already cleared	Potential other legal issues	
Science	SciDev	<i>Goal:</i> provide a science domain model; <i>Target:</i> developers	Handle scientific terminology, style, and user interactions	de, el, en, es, fi, fr, it, pl, pt, ro	Millions of publications are available for fine-tuning a base model (foundational model), originating from databases hosted by various partners in the Science group.	It is probably not necessary to create new data.	Check on licenses is ongoing. A subset of the data is likely to be open enough for usage on the short term. In order to be compliant with GDPR, automated anonymisation will be applied to remove personal data (such data is not necessary for domain-wide fine-tuning).	All metadata in ISTEX are open to everyone, but most full texts have restricted availability. Access requires individual partners to sign an NDA with CNRS, or CNRS to add information to Appendix 1 of CA; to be discussed.	Expert.ai, Pleias CNRS, BSC CNS, IMEC, LIGHT ON, ATHENA RC, OPENAIRE AM, ICIA, UTU, INRIA, ALT-EDIC
	ClaimAnalysis	<i>Goal:</i> validate claims in scientific papers (especially biomedical); <i>Target:</i> scientists, pharmaceutical companies, patent attorneys	Extract main claims from scientific texts and verify them against relevant literature, providing supporting or refuting explanations	de, en, es, fr, it	Datasets like FEVER or SciFact	A small subset (~5 000 data samples), automatically annotated from SciFact/FEVER or synthetically generated, designed to enhance the verification step			Expert.ai, Pleias, ICIA, UTU, BSC CNS, ATHENA RC, TU/e, Babelscape, INRIA, ALT-EDIC
	Research_SSH	<i>Goal:</i> obtain contextualized answers; <i>Target:</i> researchers, archivists, librarians	Query bibliographic databases	en, fr, it	Publications and metadata from databases of partners in the Science group, and testing data derived from those publications	Need for specific dataset for QA, with various question types (size TBD)	Using full texts (with personal data, and with various licenses) or metadata (also containing personal data) to test performance of a RAG system should not lead to legal issues, as it does not involve modifying the model.		CNRS, CNR-ILC, OPENAIRE, ICIA, BSC CNS, ATHENA RC, LIGHT ON, IMEC, LNE, LIU, CNR, Pleias, IMEC, INRIA, ALT-EDIC
Public Services	PADA	<i>Goal:</i> extract information and answer questions in natural language; <i>Target:</i> public administration	RAG, QA, IE	it	We are developing new benchmarks based on 1 469 deliberations and 4 264 determinations from local public records in Apulia, including a manually annotated pilot set for	An estimated set of 50+ custom test questions covering extractive and generative tasks	Public tests based on anonymised data		TALIA, ALT-EDIC, GISKARD AI, Babelscape, CNR



Co-funded by
the European Union

This project is funded by Digital Europe programme
under Grant Agreement no. 101198470

D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU

		employees			NER and over 20 custom test questions for both extractive and generative tasks. These resources will support fine-grained evaluation of document-level and entity-level information retrieval, summarization, and QA.				
	LuxHubAI	<i>Goal:</i> help navigate language and data hubs; <i>Target:</i> users of Luxembourgish SproochHub and LuxHub by LIST	RAG with satisfactory performance on Luxembourgish	de, en, fr, lb, pt					LIST, ULUX, ALT-EDIC
	LuLoBots	<i>Goal:</i> enable Luxembourgish language support on public service platforms; <i>Target:</i> users of Luxembourgish public services	Bidirectional translation en/lb and fr/lb	de, en, fr, lb					
	AI4LC	<i>Goal:</i> provide legally grounded answers based on legislation; <i>Target:</i> users seeking guidance on legal situations	RAG and domain-specific LLM fine-tuning	lv	Laws, Regulations, Directives	Up to 30-50 000 examples for in-context QA from human-written data sources; 50-100 000 synthetically generated examples for in-context QA		Personal / sensitive information in training data	TILDE, ALT-EDIC
	ELISA	<i>Goal:</i> guide citizens through e-government services; <i>Target:</i> users of Sibiu Municipality portal	Solve complex user problems by understanding natural language, performing multi-step reasoning, and using tools to interact with municipal systems	ro	The project will not use real citizen data for training. Instead, a large-scale synthetic dataset will be generated. An initial test set of 500+ queries will be developed to evaluate the pilot.	A large synthetic dataset will be created using a "teacher" LLM. This model will generate thousands of realistic, multi-turn conversational scenarios and tool-use examples based on Sibiu's official, publicly		A formal Legitimate Interest Assessment (LIA) under GDPR and the training data summary for the EU AI Act will have to be initiated before deployment.	CERTSIGN, ICIA, ALT-EDIC

D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU

			(e.g., checking taxes, submitting applications)			available documents and regulations			
IRLT_LLM	<i>Goal:</i> enable audio-to-text transformation and summarisation for search and retrieval; <i>Target:</i> users accessing audio-visual content	Provide high-quality transcription for content and moderate-quality transcription for metadata generation	en, ga	700 hours of audio-visual programming, 1 800 text files, 1 437 bilingual TV listing documents, Irish language cataloguing description of approx of 4 400 News bulletins	The priority is to convert 700 hours of audiovisual material to text offline with post-editing, then long-term batch processing for archives as a subsequent extension	Sharing agreement is needed - TG4 has ownership of some data while copyright may need to be resolved for other data for final production system.		DCU, TG4, FBK , BUT , Vocapia , LNE, ALT-EDIC	
ARTE-PT	<i>Goal:</i> provide accurate 24/7 answers; <i>Target:</i> users of Portuguese multilingual public service chatbot	Fine-tune LLMs for a chatbot supporting citizens with 2,300 public services	de, en, es, fr, it, pt, uk, bn, hi, ne, ru, ur, zh					AMA IP, TILDE, LIU, ALT-EDIC	
GenAI4Med	<i>Goal:</i> provide evidence-based updates on healthcare developments; <i>Target:</i> researchers and practitioners	Enable natural language queries with synthetic/reporting features, evidence-based answers, domain adaptation, hallucination-free responses, multilingual support, and ontology-based annotation	de, en, es, fr, it	Currently not available; a test set of questions needs to be developed that incorporates checking of new publications as well	Only test benchmark needs to be created, and ontology needs to be translated into the 35 EU languages			LIGHTON, TNO, GISKARD AI, LIU, MainAI, ALT-EDIC	
Diavgeia	<i>Goal:</i> answer questions based on government decisions and extracted relationships; <i>Target:</i> users of the Diavgeia portal	Train LLMs on program textual data and create a smart agent capable of answering complex questions and producing reliable results	el	The Diavgeia Program contains approx. 69 billion words in 69 million published documents (and growing). In the context of the project we will start with a sample of 1 000 documents with apprximenyl 1 000 word each (1 million tokens).	As a user agent executes by corresponding on user questions, no data is expected to be created in this manner. However, there should be capabilities to create reports therefore a small amount of data could be created by this means (a fraction of the input data	Documents contain PII and should be anonymised before used for fine-tuning. Otherwise use only documents containing company information (about 25M documents)	Diavgeia contains personal data and in this regard anonymisation is required. Data share is problematic since the use of the documents for other purposes apart from transparency is not allow. So we cannot provide the personal data e.g. open data. In	ATHENA RC, GRNET MoDG, ALT-EDIC	

D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU

					We will see the results and if required we can extend the sample.	nevertheless).		this regard we plan to use an infrastructure in the premises of Diavgeia and prepare an MOU with the organisations than will get access to the data. The MOU is expected to be ready the following two weeks.	
Tourism	HeritAI	<i>Goal:</i> enable conversation and QA; <i>Target:</i> general users (tourists across Europe)	Easy access to factual data on cultural heritage sites adapted to user	el, en, it, nl, pl, pt, sl		RAG dataset - at least 100 000 tokens per location, conversational Q/A dataset - 1 000 questions estimate			ARCTUR, PWR, TALIA, AMA, CINECA, LIGHT ON, Dataaction B, FBK, ATHENA RC, Blits.ai, JSI, Babelscape, ELDA
Telecom	DomainAdapt	<i>Goal:</i> foundation model adaptation to acquire general knowledge of the telecom domain; <i>Target:</i> developers in the telco domain	Generation for a wide range of telco tasks	en, fr, it, nl, pl, sl					ORANGE, BSC CNS, Pleais, TK
	ConvAnalytics	<i>Goal:</i> extract insights from customer-agent/chatbot conversations (summary, call reason, solutions, sentiment, urgency); <i>Target:</i> customer support teams	summarisation and classification	fr, it, nl, pl, sl	250 annotated data available for French	objective of 500 conversations annotated for each language		RGPD, data must be anonymised	ORANGE, TS, KPN, VILLANOVA, BSC CNS, JSI, IPI PAN, Babelscape, TNO, TK
	PIID	<i>Goal:</i> anonymise and pseudonymise customer data; <i>Target:</i> organisations handling	Entity detection and pseudonymisation	fr, it, nl, pl, sl					KPN, BSC CNS, Babelscape, JSI, LIGHT ON, ORANGE, TK

D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU

		customer information and developers in telco domain							
	KGGen	<i>Goal:</i> generate a knowledge graph from unstructured data for semantic-driven RAG chatbot; <i>Target:</i> users needing answers to complex queries		fr, nl					KPN, BSC CNS, Pleias, TK
Energy	Nuclear Requirement Engineering (NER)	<i>Goal:</i> Support the extraction, elicitation, and formalisation of nuclear requirements through AI-assisted allocation to system elements, generating a justified and refinable matrix that can be exported to Excel.	Specialised LLMs analyse requirements and system architectures, generating allocation matrices with justifications	cs, en, fi, fr, sv		Depending on the data that can be derived from the RE-elicitation use case, synthetic data may need to be derived to cover the different matching cases in more language combinations		Some norms (eg AFNOR) prevent usage with AI	EDF, CEA, ASSYSTEM E&OS, GENCI, LIGHT ON, LNE, TNO, ELDA, Pleias, APTUS.AI, EDF

3.3. Other use cases under consideration (partially defined)

This section briefly outlines use cases that are still under consideration. These proposed use cases have high potential, and their development could deliver significant additional benefits to the project. These may require further exploration, additional effort, or deeper analysis from a data or legal perspective. Consequently, priority has been given to fully defining the other use cases. Nevertheless, project partners are actively working to develop these ideas across all components and to integrate them into the overall project scope.

The **SciDa** use case, in the Science domain, aims to enable the discovery and exploration of scientific knowledge through natural language queries over the OpenAIRE Graph⁹. By leveraging interconnected metadata on publications, datasets, software, and projects, SciDa would provide evidence-based, retrieval-driven summaries with clear provenance and direct links. This use case emphasises trustworthy synthesis of scientific knowledge, delivering structured insights to support downstream research workflows. Partners involved in this use case are: OPENAIRE, LIGHT ON, ICIA, BSC CNS, ALT-EDIC, TU/e, IBL-BAS, Pleias, INRIA and ALT-EDIC.

Luxembourgish Speech-to-Text Model (LuxSTT), in the Public Services domain, is a use case focused on the comprehensive evaluation of existing Speech-to-Text models for Luxembourgish. LuxSTT uses spoken Luxembourgish as input and produces text as output, with the aim of identifying models suitable for further use by interested parties. The evaluation provides a basis for comparing existing solutions, assessing their applicability to Luxembourgish speech-to-text needs, and drawing possible improvement opportunities through future fine-tuning and adaptation. Partners involved in this use case are: ULUX, LIST, LNE.

⁹ <https://graph.openaire.eu>



4. Use case data and requirements collection

This section specifically describes the collection of requirements needed to effectively develop and implement the use cases listed in the previous section. Once the use cases were proposed within each application domain, expert.ai, as task leader, tracked their progress by attending domain meetings and holding regular discussions with domain leaders. During this work, expert.ai developed a methodology to collect fine-tuning requirements and refine initial proposals into well-defined development plans. A standardised template was used to translate each use case into a clear LLM fine-tuning plan, enabling models to be trained for targeted, practical solutions. This section also describes the methodology used to develop the fine-tuning requirements template and presents the requirements collected for each application domain.

4.1. Collection methodology

This section describes in detail the methodology adopted by expert.ai to collect the fine-tuning requirements for each specific use case within the five application domains. expert.ai developed a dedicated fine-tuning template to collect all the essential information needed to implement a fine-tuning process. The template was shared with project partners and subsequently refined and validated. For each use case, a fine-tuning template has been filled, completed by the partners who proposed the use case, and monitored by the corresponding application domain leader. This process supported the definition and development of the use cases themselves, creating an iterative loop with other work packages involved in the use case lifecycle, including WP2 for data acquisition, WP5 for evaluation, and WP6 for legal and ethical compliance. The following subsection describes in detail the structure of the fine-tuning template.

Fine-tuning template

The fine-tuning template was developed as a guideline to collect all the requirements for each specific use case, with the goal of producing a detailed plan to obtain a trained, specialized model that addresses the target task. The template is structured as follows:

1. Introduction
2. Fine-tuning types
3. Base model selection
4. Datasets
5. GPU infrastructure scheduling
6. Training
7. Expected impact of the fine-tuned model
8. Model deployment

Introduction

This section provides an overview of the use case, including its context, scope, and main objectives. It outlines the technologies and methodologies to be adopted and describes the designed training pipeline needed to achieve the intended outcomes.

Fine-tuning types

This section provides a detailed description of the different fine-tuning types required to achieve the expected model outputs. As illustrated in Figure 1, these types range from the broadest to the most specific tuning. They include further pretraining on domain data, supervised instruction tuning, domain alignment and enrichment, user preference alignment, and finally, downstream task fine-tuning. A model may require one or more fine-tuning types to achieve the desired outputs, depending on the selected base model and the specific requirements of the task. The template includes all these possible tuning types, with a detailed description of each, as outlined in Table 2. For each fine-tuning type, partners could indicate “yes” or “no” and provide an explanation. Completing these details ultimately defines the fine-tuning pipeline needed to generate the desired model.

Table 2. Fine-tuning types

Fine-tuning type
Further pre-training on domain data: Continuing to train a pre-trained LLM for the language modelling task on additional domain data to improve basic model skills within that domain by adding domain knowledge into the model. Typically, a large corpus of domain-specific documents is required for this. Training approach: self-supervised.
Domain-specific supervised instruction tuning: Fine-tuning a language model using examples of desired behaviour (inputs paired with correct outputs) to improve its ability to follow instructions in the domain context. Instruction tuning is key for conversational applications. Most pre-trained LLMs are already tuned for general instruction following. If the target domain is very specific and the model is unable to follow instructions relevant to the use case, gathering a dataset of instruction input and their outputs to fine-tune the model on it is probably necessary. Training approach: supervised fine-tuning.
Domain alignment and enrichment: Adapting model responses to better fit the context and knowledge of a specific domain, and supplementing its output with relevant, domain-specific insights. At this step, the LLM is trained to put the emphasis on its knowledge about the domain, increasing its truthfulness, reducing hallucination resulting from lack of domain awareness, and to prefer domain-specific rather than generic outputs. Training approach: supervised fine-tuning + reinforcement learning.
User preference alignment: Adjusting model outputs to align with user expectations around tone, safety, bias, and content appropriateness, ensuring it avoids toxic or harmful language and uses domain-appropriate terminology. Training approach: supervised fine-tuning + reinforcement learning
Downstream task fine-tuning: Adapting an LLM to perform a specific, narrower task, such as classification, named-entity recognition, sentiment analysis or question

answering, using task-specific data to achieve better performance on that task.

Training approach: supervised fine-tuning.

Other forms of learning, such as transfer learning, e.g., from similar languages or domains, distillation, continual learning, etc.

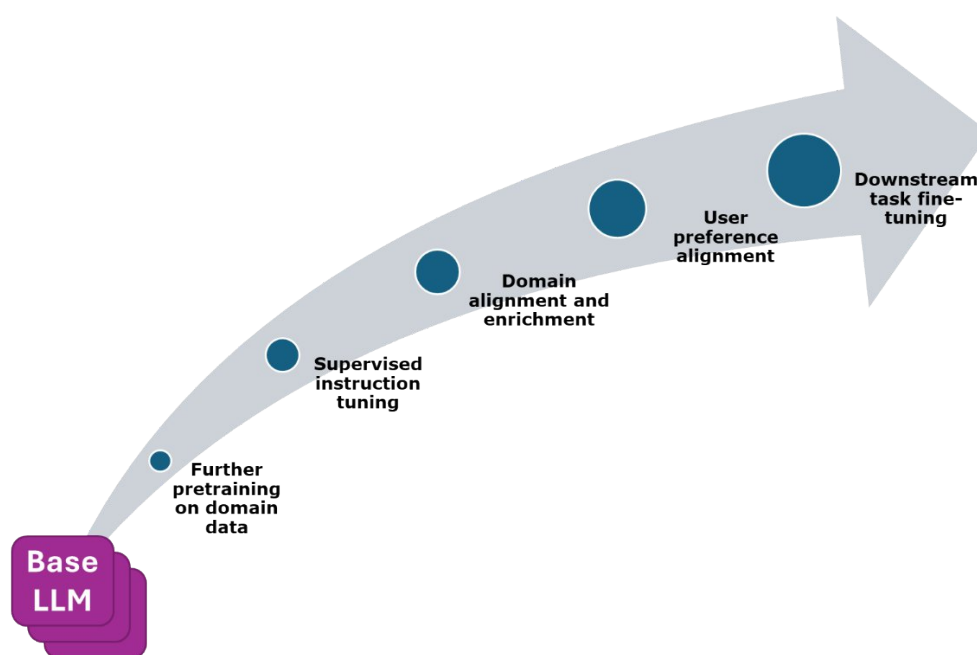


Figure 1. Fine-tuning types

Base model selection

This field focuses on evaluating and selecting candidate foundation models for fine-tuning by considering both technical characteristics and practical constraints. It examines each model's intended use, such as conversational abilities, reasoning focus, or general versus domain-specific capabilities, as well as the composition of its pre-training dataset and known limitations, including biases or language coverage restrictions. It also considers accessibility and governance, noting whether models are open source, have open weights, or are closed, and specifying licensing terms that govern use, distribution, and modification. Model assessment typically includes measuring pre-fine-tuning performance through benchmarks, user evaluations, or other metrics, while additional technical details, such as architecture type and parameter count, are systematically documented to support informed fine-tuning decisions.

In the context of LLMs4EU, model selection will prioritize compliance with all relevant EU standards and regulations, ensuring that chosen solutions align with ethical, legal, and technical requirements. Preference will be given to models with open weights and those trained on openly available datasets, supporting transparency, reproducibility, and collaborative development across the European research and innovation community. Representative models reflecting these selection priorities include OpenEuroLLM¹⁰, a major candidate currently in development

¹⁰ <https://openeurollm.eu>

within a sister parallel project, Salamandra [32], and EuroLLM [33]. This approach will maximize accessibility while fostering innovation, enabling the development of high-quality, trustworthy AI solutions that can be widely adopted across domains.

Datasets

This section focuses on the datasets planned for use in fine-tuning. For preexisting datasets, it provides references or links to their descriptions, while for newly created datasets, it outlines the creation plans and, where relevant, annotation guidelines, referring to the data creation plans developed within WP2. In all cases, the intended purpose of each dataset is specified, such as further pre-training, instruction tuning, domain alignment and enrichment, preference alignment, or downstream task fine-tuning.

As for model selection, dataset selection and creation will prioritize compliance with all relevant EU standards and regulations, ensuring ethical and legally compliant handling of data. Preference will be given to open and publicly available datasets to promote transparency, reproducibility, and collaborative use across the European research and innovation community. WP2, responsible for the data creation plan, and WP6, overseeing legal assessment, will guide and support the adaptation or creation of datasets, ensuring high-quality, trustworthy AI models while fostering accessibility and innovation across domains

GPU infrastructure scheduling

This section focuses on estimating the computational resources and training schedule required for fine-tuning. It assesses lead time and effort, including preliminary tasks such as annotation, labelling, and dataset assembly, expressed in person-hours and calendar duration. It specifies whether training is a one-time process or involves multiple model versions, summarizing the planned schedule and frequency. Additional considerations include the required model storage per GPU node and the estimated GPU hours for training, based on a standard configuration of NVIDIA¹¹ H100 GPUs with 80 GB of VRAM each.

Training

This field focuses on outlining possible training strategies for fine-tuning. It provides an overview of the training process and a justification for the chosen strategies, such as SFT [34] or RL [35]. It specifies whether PEFT [36] techniques, like LoRA [37], will be used or if full model training is required for the use case. The field also documents planned best practices, including checkpoint saving and evaluation, and provides estimates for key hyperparameters such as training epochs, learning rate, and batch size.

Expected impact of the fine-tuned model

This field describes the overall scope, usefulness, and benefits of the resulting fine-tuned model. It outlines the model's intended use and potential impact, including quantifiable improvements

¹¹ <https://www.nvidia.com>

where possible. It highlights benefits related to openness and transparency, support for multilingualism and cultural diversity, and adherence to best practices in data governance. It also emphasizes collaborative development and sharing to ensure the model contributes positively to the broader community.

Model deployment

This field outlines the deployment requirements for the fine-tuned model, focusing on the infrastructure needed to support expected usage. It describes the deployment strategy and schedule, infrastructure requirements based on projected demand, and the type and scale of hardware needed, such as GPU servers, CPUs, or cloud instances. It also specifies runtime environment details, including containers, orchestration tools, and APIs, along with scalability plans for varying workloads. Additional considerations include reliability, performance, and integration with existing systems or platforms.

4.2. Collected data and requirements

This section summarizes, for each application domain, the information collected from the use case fine-tuning templates: the types of fine-tuning to be performed, the selected base models, the datasets to be used for training, estimates of computational resources and techniques required, and the specifications for deploying the resulting tools.

Science

Within the Science domain, three use cases have been fully defined, covering a broad range of fine-tuning approaches that will enable models to acquire extensive capabilities within this domain.

Fine-tuning types

The **SciDev** use case aims to develop a domain-wide scientific model (Figure 2) using a four-stage approach that follows the first four fine-tuning techniques from the proposed schema (see Table 2 and Figure 1). First, the model will undergo further pre-training (CPT) on high-quality scientific data sources. Second, SFT will be applied using over 100 000 domain-specific instructions to refine model behaviour. Third, domain alignment and enrichment will be performed. Finally, user preference alignment will be carried out. This methodology is expected to achieve an impact comparable to other science-oriented LLMs, such as OmniScience [38]. The development of this domain model will prioritize open-weight architectures, reflecting LLMs4EU's commitment to openness.

*D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU*

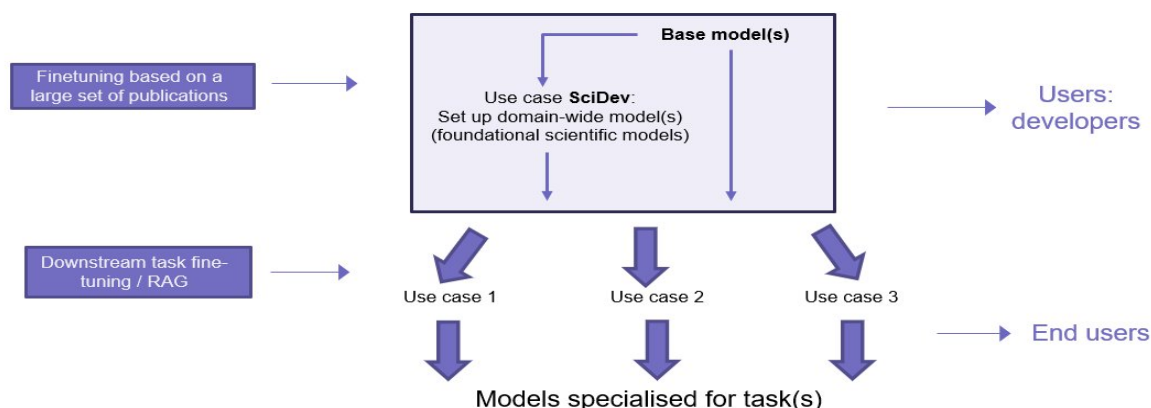


Figure 2. SciDEV training workflow

Building on this domain model, the **ClaimAnalysis** and **Research_SSH** use cases will focus on downstream fine-tuning tasks: ClaimAnalysis on claim extraction and verification (Figure 3) using SFT, and Research_SSH on RAG and QA capabilities (Figure 4) using PEFT. Together, these three use cases aim to provide versatile and highly relevant tools for the scientific domain.

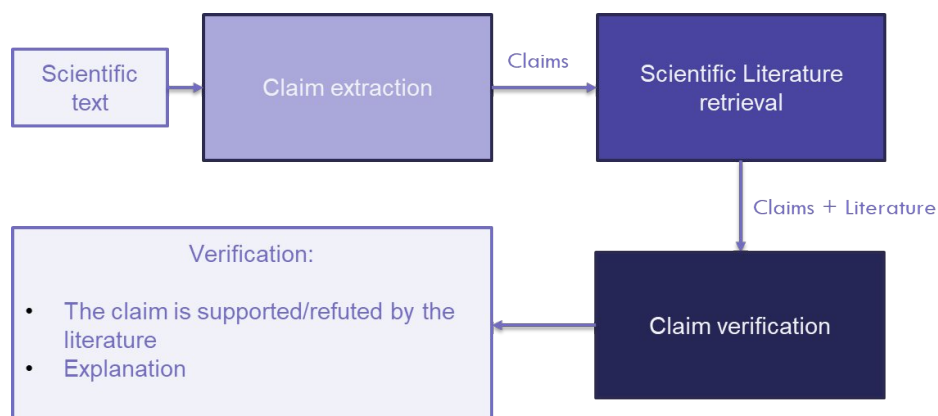


Figure 3. ClaimAnalysis use case workflow

*D4.1 Definition of the use cases for the 5 application domains:
data and requirements, LLMs4EU*

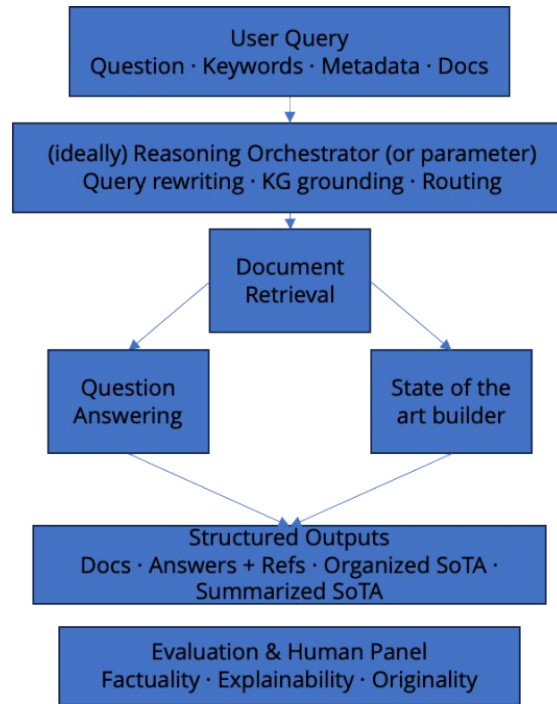


Figure 4. Research_SSH use case workflow

Base model selection

Given the importance of multilingualism and broad coverage of EU languages, a primary base model will be selected among EU open-source, multilingual candidates, such as the Salamandra suite (2B, 7B, and 40B), OpenEuroLLM and EuroLLM, selecting the most suitable model at the start of the training process. Partners expect this domain model to underpin the other use cases in the Science domain and serve as a key resource for model developers across Europe.

Data

Particular attention has been given to open data. A substantial volume of approximately 30–35B tokens of reliable, multilingual data will be used for additional pre-training, along with 100–250 000 domain-specific instructions for SFT. Sources include Pleias OpenScience, Open Research Europe, OpenAIRE, and others. The data will be carefully curated to ensure EU language coverage and scientific field diversity, with regional contributions (e.g., ISTEEX, University of Turku, Athena RC) enhancing multilinguality. The downstream tasks will involve approximately 150 000 documents across ClaimAnalysis and Research_SSH, including metadata, RDF, and XML formats required for the latter use case.

Computation demand, training techniques and deployment strategies

The training will be extensive, particularly for the domain-wide model, which spans different scales from 2B to 40B parameters. Training is expected to take approximately 110–130 days, with an additional ~40 days for downstream tasks. Extra computing resources have also been accounted for to evaluate each model developed for the various use cases.

Table 3. Overview of the use cases in the Science domain

Use case	Fine-tuning types	Base model family	Datasets	GPU scheduling (maximum estimation)	Training techniques	Deployment
SciDev	Further pre-training on domain data; Domain-specific supervised instruction tuning; Domain alignment and enrichment; User preference alignment	Salamandra, OpenEuroLLM, EuroLLM and Falcon	OpenScience subset from Pleias Common Corpus, arXiv, Semantic Scholar, Pubmed, Open Research Europe, OpenAire, Daring-Anteater dataset, others	130 days (8 x A100)	CPT, SFT	Server with dedicated GPUs
ClaimAnalysis	Downstream task fine-tuning	Scientific Domain Wide or OpenEuroLLM, EuroLLM and Falcon	FEVER, SciFact, RoSE, CLIMATE-FEVER, SciClaimHunt	25 days (8 x A100)	SFT	Server with 2 dedicated GPUs and vLLM
Research_SSH	Downstream task fine-tuning	Scientific Domain Wide or OpenEuroLLM, EuroLLM and Falcon	Hypotheses.org, Isidore, Nakala, AIUCD, CLIC, UD, IJCL, KG relevant data, others	15 days (8 x A100)	SFT, PEFT (LoRA)	sandbox version of the Isidore platform

Public Services

Nine distinct use cases have been proposed within the Public Services domain, representing the highest number across all domains. This prominence is reflected in the requirements collection phase, which highlights a wide range of operational needs.

Fine-tuning types

Three use cases, **AI4LC**, **IRLT_LLM**, and **Diavgeia**, will undergo CPT using specialized domain data. During this phase, **AI4LC** will focus on pretraining to better capture legal language patterns. **IRLT_LLM** will target gaps in open-weight foundation speech models for the project's target languages. **Diavgeia** will focus on pretraining to enhance understanding of Greek public sector discourse and administrative terminology.

A consistent number of use cases require domain-specific supervised instruction tuning. Use cases such as **PADA**, **LuLoBots**, **AI4LC**, **ELISA**, and **ARTE-PT** necessitate adaptation to specialized QA tasks in highly specific domains related to citizen services, often integrating external knowledge through RAG techniques. Meanwhile, the **GenAI4med** and **IRLT_LLM** use cases will adopt supervised tuning tailored to domain-specific retrieval and ASR.

LuxHubAI, **LuLoBots**, **AI4LC**, and **ARTE-PT** will undergo domain alignment and enrichment, using SFT and RL to improve documentation and language knowledge and reduce hallucinations. The **IRLT_LLM** use case will apply the same techniques to align the recognition of English utterances in broadcast television, accounting for factors such as noise, background music, speaking speed, and other domain-specific characteristics.

ELISA and **AI4LC**, for which safety, tone, policy guardrails, and appropriate disclaimers have been identified as key factors, will also incorporate user preference alignment tuning.

Finally, a wide variety of downstream task fine-tuning will be performed in use cases including **PADA**, **AI4LC**, **ELISA**, **GenAI4Med**, and **Diavgeia**, covering tasks such as IE, QA, RAG, FAQ-based interaction, summarization, and classification.

LuLoBots and **ELISA** will also explore additional learning paradigms, focusing on transfer learning and continual learning, respectively.

Base model selection

Considering the selected base models, this domain, like the scientific one, welcomed the recommendation to use open-weight and open-data models, including pretraining models such as EuroLLM and Salamandra. In addition to these, models specifically developed for the target languages of particular use cases will also be considered, such as Mistral [39] for French and Luxembourgish, TildeOpen [40] for Nordic and Eastern European languages, Gervasio [41] for Portuguese, and *Meltemi* [42] for Greek. Beyond these, various other models necessary for different steps in the use case pipelines will be considered to fully perform the tasks, including encoder models for retrieval and ASR, such as POWSM [43] and OWSM v4 [44].

Data

All use cases, except **GenAI4Med**, will leverage custom datasets provided by final users and curated in coordination with WP2 for dataset generation, curation, and management, and with WP6 for legal assessment.

PADA will rely on the Albo Pretorio dataset, which includes approximately 5 500 documents from the public healthcare district of Apulia.

LuxHubAI will use the ZSL Documentation and SproochHub manual to generate instruction data, while LuxIT already contains 220 000 instruction-answer pairs in Luxembourgish.

LuLoBots will source data with institutional and private sector partners, such as the National Library of Luxembourg, RTL Media, and the Commission nationale pour la protection des données (CNPd), which provides a database of roughly 2 000 Q&A pairs in French for training.

AI4LC will use a dataset of 35 000 real-world citizen legal questions and official answers from the E-Consultations section of the public Latvian portal¹², grounded in Latvian legislation with references from the legal acts of the Republic of Latvia portal¹³, and will further generate synthetic data to expand coverage and robustness.

Within the **ELISA** use case, a dataset of roughly 1 000 instruction-response pairs for the specific domain have already been created and curated by validators, along with approximately 400 data samples for DPO [45] and alignment of tone, helpfulness, and safety.

¹² <https://lvportals.lv>

¹³ <https://likumi.lv>

The datasets available to the **IRLT_LLM** use case comprise 700 hours of untranscribed TG4 TV speech, 15 hours of transcribed Irish from Common Voice (including occasional code-switching), orphaned Irish transcriptions without audio, and open monolingual Irish data such as Wikipedia (Vicipéid).

For **ARTE-PT**, fine-tuning will use a dataset of real-world dialogues collected from interactions between users and the chatbot on the Portuguese national public services portal¹⁴. Approximately 3 000 dialogues are currently available, with this number expected to grow to around 10 000 by the end of the project.

The **Diavgeia** dataset manages raw data from the Greek portal¹⁵, including government decisions and public administration documents, comprising around 45 million documents, each averaging 1 000 words.

Unlike the previous use cases, which rely on proprietary or curated datasets, the **GenAI4Med** use case will leverage a subset of multiple open datasets, including ~40 million PubMed¹⁶ articles, ~550 000 ClinicalTrials¹⁷ studies, ~4.8 million expert contacts from SO¹⁸, ~31 200 MeSH¹⁹ keywords, ~3.8 million EPO patents²⁰, ~420 000 grants from various sources, and ~38.1 million OpenAlex²¹ articles.

For more information on the datasets, especially proprietary or newly created ones, please refer to the data creation plans (D2.3) and the legal plans (D6.2), where they are described in more detail.

Computation demand, training techniques and deployment strategies

Each use case provided an estimate of the GPU requirements needed to train their models, as well as information for running them on hosting machines. Most use cases will employ PEFT strategies, mainly using LoRA and QLoRA, with a smaller number also performing CFT and fully SFT. The estimated GPU demand ranges from a couple of days up to 830 days, typically computed assuming 1 to 8 GPUs, although actual requirements may vary depending on the available hardware. These estimates are particularly important for a crowded domain like this one, to ensure computation resources are distributed according to each use case's specific needs.

¹⁴ <https://www.gov.pt>

¹⁵ <https://diavgeia.gov.gr>

¹⁶ <https://pubmed.ncbi.nlm.nih.gov>

¹⁷ <https://clinicaltrials.gov>

¹⁸ <http://www.sequenceontology.org>

¹⁹ <https://www.ncbi.nlm.nih.gov/mesh>

²⁰ <https://www.epo.org/en/searching-for-patents>

²¹ <https://openalex.org>

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

Most models will be compatible with widely adopted frameworks in the LLM community, including Python²², PyTorch²³, vLLM²⁴, as well as REST APIs, orchestration and agentic frameworks, LDS, and will require GPU computation to deliver their functionality.

Table 4. Overview of the use cases in the Public Services domain

Use case	Fine-tuning types	Base model	Datasets	GPU scheduling	Training techniques	Deployment
PADA	Domain-specific supervised instruction tuning; Downstream task fine-tuning	Gemma 3, EuroLLM and Salamandra	Albo Pretorio subset	7 days (1 x A100)	PEFT (QLoRA)	Server with a single high-memory GPU (e.g., NVIDIA L40S or H100), vLLM, Docker, REST API
LuxHubAI	Domain alignment and enrichment; Other forms of learning	Mistral, GML, Phi, Salamandra, OpenEuroLLM	ZSL Documentation, SproochHub manual, LuxIT	45 days (4 x A100)	PEFT (LoRA)	BESSER Agentic Framework, LDS
LuLoBots	Domain alignment and enrichment; Other forms of learning	OpenEuroLLM, Tower, OPUS-MT, Mistral, LuxMT, LuxT5	National Library of Luxembourg, RTL Media, Commission nationale pour la protection des données (CNPd)	33 days GPU compute hours per month on the Luxembourg supercomputer MeluXina	SFT	Integration in Guichet.lu platform
AI4LC	Further pre-training on domain data; Domain-specific supervised instruction tuning; Domain alignment and enrichment; User preference alignment; Downstream task fine-tuning	TildeOpen LLM	TildeOpen legislative subset, Ivportals.lv E-Consultations, Synthetic domain data	830 days (1 x H100)	CPT, SFT	Server with a single H100, H200, or two L40/L40s GPUs, vLLM, REST API
ELISA	Domain-specific supervised instruction tuning; Domain alignment and enrichment; User preference alignment; Downstream task fine-tuning; Other forms of learning	Salamandra, EuroLLM, Mistral	Instruction and preference custom data	2 days (1 x H100)	PEFT (QLoRA), DPO	REST API with retrieval backend, Docker, FAISS
IRLT_LLM	Further pre-training on domain data; Domain-specific supervised instruction tuning; Domain alignment and enrichment; Other forms of learning	Wav2Vec, Open Whisper, POWSM, OWSM v4	TG4, Common Voice subset, Vicipéid	580 days (1 x A100)	CPT, SFT	Server with GPU, Pytorch, Huggingface
ARTE-PT	Domain-specific supervised instruction tuning; Domain alignment and enrichment	Gervásio	Portuguese national public services portal	90 days (10 x 80G GPUs)	SFT	Server with 20 GPUs, REST API

²² <https://www.python.org>

²³ <https://pytorch.org>

²⁴ <https://vllm.ai>

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

GenAI4Med	Domain-specific supervised instruction tuning; Downstream task fine-tuning; Other forms of learning	Mistral, EuroLLM, Salamandra	PubMed, Clinical Trials, Contacts, MeSH, EPO Patents, Grants, OpenAlex	7 days (8 x A100)	PEFT (LoRA)	Server with a single high-memory GPU, Docker, vLLM, Transformers, REST API
Diavgeia	Further pre-training on domain data; Downstream task fine-tuning	Meltemi	Diavgeia governmental portal	30 days (1 x H100)	PEFT (LoRA)	Server with NVIDIA L4 or A100 GPUs

Tourism

The Tourism application domain has focused its efforts on a single use case, aiming to equip a model with the most important skills and capabilities needed for the domain.

Fine-tuning types

The tourism application domain proposes a single, high-impact use case, **HeritAI**, aimed at developing multilingual conversational AI assistants that provide tourists with accurate, location-aware information about European cultural heritage sites, museums, and destinations. This use case involves multiple tasks requiring the application of various fine-tuning strategies.

Domain-specific supervised instruction tuning has been identified to handle question-answer pairs about cultural heritage sites and instruction-following examples for itinerary planning. Domain alignment and enrichment will prioritize retrieved cultural heritage information over parametric knowledge, ensure citation of retrieved sources when appropriate, acknowledge when information is unavailable in the knowledge base, and reduce hallucinations by grounding responses in retrieved context.

In a second phase, user preference alignment fine-tuning may be applied to adjust model outputs according to expectations regarding tone, safety, bias mitigation, and content appropriateness.

Finally, downstream task fine-tuning will address extractive and abstractive QA over cultural heritage documents, multi-turn conversational capabilities for tourist interactions, and OCR optimization for improved character detection in low-resource languages. Additional downstream tasks include generating descriptions of cultural heritage artifacts from images, transcribing tourist voice queries, and producing natural-sounding audio guides.

Base model selection

Ideally, the use case will develop one or two specialized models for QA and RAG, supported by additional models for embeddings [46], speech processing [47], and vision tasks [48]. In line with the project's openness-by-design principle, generative foundation models from the OpenEuroLLM, EuroLLM, and Salamandra projects have been selected as base models. For OCR requirements, PaddleOCR [49], Kraken²⁵, and LightOnOCR-2 [50] may be considered. For speech

²⁵ <https://kraken.re>

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

processing, open-source ASR models, or language-specific systems like the Slovene Conformer²⁶, will be used to transcribe audio content.

Data

The first version of the assistant will support Slovenian, Polish, Italian, Portuguese, Greek, Dutch, and English, leveraging multiple curated data sources to build a RAG-based knowledge base. Core datasets include up to 50 000 data samples (up to 20M tokens) of the Europeana Cultural Heritage Corpus²⁷ (55M documents accessed via APIs), filtered for quality, translated where needed, and structured with embeddings and indexed metadata; publicly available documents from HeritAI pilot destinations (processed via OCR and RAG ingestion workflows); a Romanian Tourism QA dataset provided by ICIA; and potentially synthetic QA and RAG data generated from Wikipedia²⁸.

Computation demand, training techniques and deployment strategies

HeritAI partners estimated GPU requirements for supporting a multilingual, multi-capability bot, considering all fine-tuning strategies. On a node with 4 A100 GPUs, training would take ~200 days, reducible with more GPUs. Parameter-efficient techniques, specifically LoRA, will be employed to minimize computational demands.

For deployment, the system requires GPU-equipped servers, containerized management environments (K8s²⁹), efficient storage for RAG workflows (S3³⁰), and libraries like LiteLLM [51] and vLLM to optimize LLM serving and inference.

Table 5. Overview of the use cases in the Tourism domain

Use case	Fine-tuning types	Base model	Datasets	GPU scheduling	Training techniques	Deployment
HeritAI	Domain-specific supervised instruction tuning; Domain alignment and enrichment; User preference alignment; Downstream task fine-tuning	EuroLLM, OpenEuroLLM, Salamandra, PaddleOCR, Kraken, LightOnOCR-2, Open Whisper, Sloven Conformer	Europeana Cultural Heritage Corpus, various custom tourism QA datasets	200 days (4 × A100)	PEFT (LoRA)	Servers with NVIDIA GPUs (L40s or better), container management environment (k8s), S3 compatible Object storage, LiteLLM, vLLM

²⁶ <https://live.european-language-grid.eu/catalogue/tool-service/23852>

²⁷ <https://www.europeana.eu>

²⁸ <https://www.wikipedia.org>

²⁹ <https://kubernetes.io>

³⁰ <https://aws.amazon.com/s3>

Telecommunications

In the Telecommunications application domain, efforts have focused on four use cases that include different fine-tuning approaches, pipelines involving multiple model architectures, and the development of domain-specific datasets.

Fine-tuning types

The use cases defined in this domain require multiple fine-tuning approaches. The **DomainAdapt** use case, which can serve as a domain telecom model, and the **PIID** use case will undergo an initial phase of further pre-training. In the first case, this step will adapt the LLM to specific subdomains of the telecommunications ecosystem, namely *technical*, *regulatory*, and *customer care environment*. In the second case, it will improve the performance of an encoder-only model for NER.

DomainAdapt, **ConvAnalytics**, and **PIID** will subsequently undergo supervised instruction fine-tuning as well as domain alignment and enrichment, as these use cases must be trained to follow instructions and interact effectively with users, to perform domain-specific technical tasks while aligning model outputs with expected formatting.

DomainAdapt will also handle user preference alignment, a crucial step for a model that manages interactions with end users as a chatbot and that also serves as a telecom model to be used as the basis for the other use cases in this domain.

Finally, **ConvAnalytics**, **PIID** and **KGGen** will undergo downstream task fine-tuning to effectively perform their respective tasks: specific data analysis: specific data analysis, NER and RE.

Base model selection

As in other use cases, European models, preferably with open weights and datasets, have been prioritized in this case, such as OpenEuroLLM and Salamandra. In addition to decoder LLMs, the development of an encoder-only model for NER is also planned. For this purpose, a European alternative to widely used encoder models, mmBERT [52], has been considered.

Data

In the context of the project, the telecommunication application domain is one of the most sensitive in terms of data, as it involves customer-sensitive information. For this reason, the availability of open data is limited or largely irrelevant, and significant effort during the first months of the project has been dedicated to defining data requirements and plans. Specific attention in numerous telecommunication meetings has been given to data anonymisation practices, including a comprehensive review of the state of the art, which has been considered as a possible safe preprocessing approach, allowing telco partners to use parts of their internal data sources securely. Partners, particularly Orange and KPN, have contributed extensively to the discussion and development of anonymisation systems.

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

Extensive meetings and interactions with other work packages, especially WP2 and WP6, have led to the development of numerous data plans dedicated to generating synthetic data particularly relevant for the specific tasks at hand. With the exception of the PIID use case, which identified part of the required data in open datasets (Nemotron PII³¹, PII-masking-400k³²), all other data used for model training will be automatically generated as synthetic data. Further details can be found in the data plans developed within WP2.

Computation demand, training techniques and deployment strategies

The estimated GPU requirement for this application domain is considerable due to the need to generate a large amount of training data for both further pre-training and the other types of fine-tuning planned for model development. The estimate exceeds 3 000 days on a single H100 GPU with 80 Gb memory; however, this requirement will be reduced by accessing multiple nodes, which this domain application is coordinating with HPC partners in the context of T4.4. Given the specific needs of this use case, as well as other use cases with high computational requirements, such as Science, these considerations will be reflected in the execution of fine-tuning plans, and consistent computational resources will be provided accordingly.

Overall, each use case is planned to perform both full fine-tuning as well as the use of PEFT strategies. Deploying these models will require GPU servers, a container management environment, S3-compatible object storage, accounting, and model hosting capabilities.

Table 6. Overview of the use cases in the Telecommunications domain

Use case	Fine-tuning types	Base model	Datasets	GPU scheduling	Training techniques	Deployment
DomainAdapt	Further pre-training on domain data; Domain-specific supervised instruction tuning; Domain alignment and enrichment	OpenEuroLLM, Salamandra, Mistral	Custom synthetic instruction data	2083 days (1 x H100)	SFT, PEFT	Accessible through project catalogue
ConvAnalytics	Domain-specific supervised instruction tuning; Domain alignment and enrichment; Downstream task fine-tuning	Telco DomainAdapt model	Custom synthetic conversational data	416 days shared with PIID (1 x H100)	SFT, PEFT	GPU servers, container management environment, S3 compatible object storage, accounting and model hosting
PIID	Further pre-training on domain data; Domain-specific supervised instruction tuning; Domain alignment and enrichment; User preference alignment; Downstream task	DeBERTa, EuroLLM, Mistral, Salamandra	Nemotron PII, PII-masking-400k, custom PII data	416 days (1 x H100) shared with ConvAnalytics	SFT, PEFT	GPU servers, container management environment, S3 compatible object storage, accounting and model hosting

³¹ <https://huggingface.co/blog/nvidia/nemotron-pii>

³² <https://huggingface.co/datasets/ai4privacy/pii-masking-400k>

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

	fine-tuning					
KGGen	Downstream task fine-tuning	EuroLLM, Mistral, Salamandra, DeBERTa, mmBERT	Manually annotated and custom synthetic KG data	416 days (1 x H100)	SFT, PEFT	GPU servers, container management environment, S3 compatible object storage, accounting and model hosting

Energy

The Energy application domain aims to develop multilingual AI assistants for regulatory analysis in the nuclear energy sector, following a requirement engineering approach towards processing Nuclear Engineering Requirements (NER), including four main capabilities: requirement extraction, elicitation, formalization, and regulatory compliance analysis and comparison. Like the Telecommunications domain, the Energy domain is characterized by highly sensitive data and requirements, and this has been carefully considered in planning and development. The proposed fine-tuning strategy aims to develop a single, comprehensive model covering the entire domain, capable of handling most tasks, complemented by specialized multilingual or encoder models to manage complex pipelines.

Fine-tuning types

The proposed fine-tuning begins by adapting a generalist LLM to the energy domain, producing the *energy domain model*, which serves as the foundation for addressing the various capabilities across the use case (Figure 5).

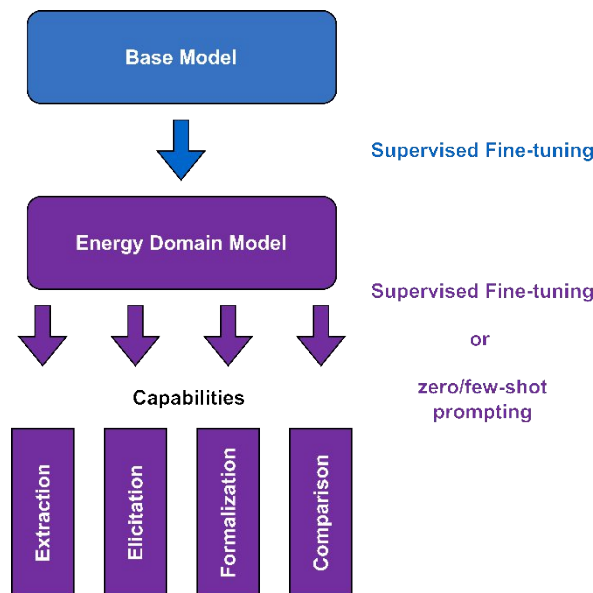


Figure 5. Energy domain model training workflow

The model will undergo a domain specific supervised instruction tuning using a large dataset generated from public documents related to nuclear plant operations and the nuclear regulatory domain.

Subsequently, it will receive domain alignment and enrichment fine-tuning to manage graph-RAG capabilities, enabling it to prioritize retrieved information over parametric knowledge, cite sources from retrieved documents when appropriate, acknowledge when information is not available in the knowledge base, and reduce hallucinations by grounding responses in the retrieved context.

Once the model has acquired the ability to manage domain-specific information, it will be fine-tuned for various downstream tasks corresponding to the core Energy capabilities. This will enable the model to perform requirement detection, categorization, rewriting, and consistency and compliance analysis. Additional capabilities may be addressed through specialized models, such as encoder-only models for retrieval tasks, or in cases where training data is scarce or difficult to generate. These capabilities include applying deep domain knowledge for reasoning and performing mathematical and logical problem-solving using theorem-proving techniques.

The final step of the fine-tuning plan also requires multimodal learning, as parts of the requirements may be presented in data tables, graphs, or schemas, for example in PDF documents.

Base model selection

Model selection will be based on previous multilingual support, natural language comprehension, instruction-following, domain-specific nuclear energy knowledge, reasoning capabilities, openness, and compliance with industry standards. Candidate base models include OpenEuroLLM, EuroLLM, Ministral-Reasoning [53], and GPT-OSS [54]. The requirement extraction capability may apply a visual LLM for complex PDF parsing (e.g., LighOnOCR, Molmo [55], Qwen-VL [56], Granite-Docling [57]). Additionally, unsupervised fine-tuning of embedding models such as xlmRoberta [58], mDeberta [59], ModernBert [60], EuroBert [61], mmBert, NukeBert [62], and RE-BERT [63] will support similarity search and information retrieval in multiple use cases.

Data

Adaptation to the Energy domain involves extracting data from multiple sources, including the FinePdfs³³ dataset on nuclear energy and plant operations (filtered and deduplicated using Datatrove³⁴, MinHash [64] and SemDeDup [65]), public websites and scientific document APIs in multiple languages, the Wikipedia Nuclear portal, and confidential corporate documents. The resulting raw dataset is augmented via back-translation and generation of QA pairs to expand it from a few million tokens to several hundred million, forming the *energy domain corpus*. This

³³ <https://huggingface.co/datasets/HuggingFaceFW/finepdfs>

³⁴ <https://github.com/huggingface/datatrove>

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

corpus supports long-term memory of nuclear domain knowledge in the fine-tuned model. For the NER use case, the model will be fine-tuned on the Nuclear Requirement Engineering corpus, with around 100M tokens. Confidential nuclear data from CEA or EDF documents will be considered where possible, although such data cannot in principle be shared within the consortium.

Computation demand, training techniques and deployment strategies

Since no further pre-training phase is planned in the Energy application domain, the overall training requirements are significantly lower compared to those estimated for the Science and Telecommunications domains. The current plan estimates approximately 24 GPU days for domain adaptation to nuclear energy, followed by around six GPU days for each NER- use case downstream task.

For deployment and operationalization, the models will require dedicated inference infrastructure, including servers equipped with NVIDIA GPUs (L40s or higher), a containerized orchestration environment based on K8s, S3-compatible object storage for datasets and model artifacts, and optimized inference and serving frameworks such as LiteLLM and vLLM to ensure scalable, efficient, and low-latency model hosting.

Table 7. Overview of the use case capabilities in the Energy domain

Task/capability	Fine-tuning types	Base model	Datasets	GPU scheduling	Training techniques	Deployment
Extraction	Domain-specific supervised instruction tuning; Domain alignment and enrichment; Downstream task fine-tuning; Other forms of learning	EuroLLM, OpenEuroLLM, Mistral, Gemma, Salamandra, GPT-OSS, Embedding model, visual LLM	Custom energy domain corpus (FinePdf, Wikipedia Nuclear, etc.), Nuclear Requirement Engineering, proprietary CEA or EDF data	24 days (4 x A100) for domain energy model + 6 days (4 x A100) for each use case	PEFT (LoRA)	Servers with NVIDIA GPUs (L40s or better), container management environment (k8s), S3 compatible Object storage, LiteLLM, vLLM
Elicitation						
Formalisation						
Comparison						

5. Use case statistics

This section of the document presents key statistics derived from all the use cases fully defined within the project. The analysis focuses on the most relevant aspects for LLMs4EU, including language coverage, which represents a core pillar of the initiative, as well as an overview of the datasets, the different fine-tuning strategies to be applied, and the thematic areas addressed by the use cases. These thematic areas, identified in T4.2, constitute the theoretical backbone of the project and support the development and optimization of the model fine-tunings.

5.1. Language coverage

Since the project is strongly committed to providing tools for European languages, particularly those currently underrepresented in GenAI solutions, this section provides an overview of language support for the use cases that have been fully defined so far, considering their specific linguistic requirement.

Table 8. Project-relevant languages categorized by group

EU24	EU+	extraEU
• Bulgarian (bg) • Croatian (hr) • Czech (cs) • Danish (da) • Dutch (nl) • English (en) • Estonian (et) • Finnish (fi) • French (fr) • German (de) • Greek (el) • Hungarian (hu) • Irish (ga) • Italian (it) • Latvian (lv) • Lithuanian (lt) • Maltese (mt) • Polish (pl) • Portuguese (pt) • Romanian (ro) • Slovak (sk) • Slovenian (sl) • Spanish (es) • Swedish (sv)	• Albanian (sq) • Basque (eu) • Catalan (ca) • Galician (gl) • Icelandic (is) • Luxembourgish (lb) • Norwegian (no) • Scottish Gaelic (gd) • Welsh (cy) • Serbian (sr) • Turkish (tr) • Ukrainian (uk)	• Arabic (ar) • Armenian (hy) • Bengali (bn) • Hebrew (he) • Hindi (hi) • Kurdish (ku) • Mandarin Chinese (zh) • Nepali (ne) • Persian (fa) • Russian (ru) • Turkish (tr) • Urdu (ur)

For the purposes of this analysis, languages are classified into three main groups, as shown in Table 8: (i) the 24 official languages of the European Union (EU24); (ii) 12 languages, including other European co-official languages and languages of EU candidate countries (EU+); (iii) 6 languages spoken by EU immigrant communities representing significant resident groups

D4.1 Definition of the use cases for the 5 application domains: data and requirements, LLMs4EU

(extraEU). Table 3 also includes the two-letter ISO 639:2023³⁵ language codes in parentheses, which are occasionally used in this section for brevity.

Considering the three groups, Table 9 presents an overview of the distribution of languages across the project's use cases. Since not all groups are fully represented in every use case, the specific languages involved are indicated within the table, as shown for groups A and B.

Table 9. Overview of language support across use cases

Domain	Use case	EU24																EU+	extraEU
		cs	de	el	en	es	fi	fr	ga	it	lv	nl	pl	pt	ro	sl	sv		
Science	SciDev		X	X	X	X	X	X		X			X	X	X				
Science	ClaimAnalysis		X		X	X		X		X									
Science	Research_SSH				X			X		X									
Public Services	PADA									X									
Public Services	LuxHubAI		X		X			X						X				X	
Public Services	LuLoBots		X		X			X										X	
Public Services	AI4LC										X								
Public Services	ELISA														X				
Public Services	IRLT_LLM				X				X										
Public Services	ARTE-PT		X		X	X		X		X				X				X	X
Public Services	GenAI4Med		X		X	X		X		X									
Public Services	Diavgeia			X															
Tourism	HeritAI			X	X					X		X	X	X		X			
Telecom	DomainAdapt				X			X		X		X	X			X			
Telecom	ConvAnalytics							X		X		X	X			X			
Telecom	PIID							X		X		X	X			X			
Telecom	KGGen							X				X							
Energy	NER	X			X		X	X									X		

As shown in the table, 17 of the 24 EU languages (Group EU24) are represented at least once across the use cases, with French having the highest presence, appearing in 15 use cases. The average number of languages per use case is 4.7, highlighting the broad variety of multilingual use cases and reflecting the project's effort and commitment to developing multilingual models covering the widest possible range of European languages, particularly those currently underrepresented in GenAI tools.

Particular attention has been given to language support in the ARTE-PT use case, which focuses on tools that help citizens, including foreign residents and newcomers, navigate administrative

³⁵ <https://www.iso.org/standard/74575.html>

information. In addition to numerous European languages, the solution will also provide support for a language from Group EU+ (*uk*), as well as several languages included in Group extraEU (*bn*, *hi*, *ne*, *ru*, *ur*, *zh*) although full evaluation support for these languages may not be guaranteed due to resource constraints.

It is important to note that this coverage is based on the use case fine-tuning plans and may be updated in later phases of the project through the addition of new use cases or the modification of existing ones, potentially decreasing or increasing the number of languages considered through the acquisition or creation of additional data.

5.2. Data overview

This section presents an overview of the data that will be used in the project, providing insights into its nature, including both the data already available and the data that will be generated as part of the project, in accordance with the WP2 data creation plans.

Table 10 presents the project data divided into two main categories: (i) available data and (ii) data to be additionally created. The first category includes datasets identified, selected, or proposed by use-case developers, and subsequently divided into public data and data provided by partners involved in the use-case development. The second category refers to data identified as necessary for model training but not currently available and therefore requiring generation within the context of the project. As with both the already available datasets and the data to be created, these are schematically divided into unlabelled data, annotated data, and synthetic data.

Unlabelled data refer to datasets that contain raw information without associated labels or predefined outputs, meaning the data have not been manually categorized or tagged. These data are commonly used in unsupervised or self-supervised learning to discover underlying patterns, structures, or representations in the dataset. In this project, unlabelled data are used for further pretraining to acquire domain knowledge. Annotated data, in contrast, consist of datasets that have been enriched with labels, metadata, or ground truth information, typically provided by human annotators or automated labelling systems, to enable supervised learning tasks such as instruction tuning, domain and user alignment, as well as specific downstream tasks such as classification or prediction. Synthetic data refer to artificially generated datasets, generally created through generative models, designed to replicate the statistical properties of real-world data or to support supervised fine-tuning, while reducing collection costs, preserving privacy, or addressing data scarcity.

Table 10 highlights patterns in data usage and generation across application domains. Science use cases rely primarily on various types of publicly available data, with synthetic data mainly used to supplement scarce datasets or those specialized for specific instructions and tasks. Public Services and Tourism, in addition to public data, make extensive use of user-generated data, reflecting a strong emphasis on collaboration with specific partners to address unique

citizen needs across different languages and national contexts. Telecom and Energy exhibit a particularly strong reliance on synthetic data: due to stringent privacy requirements and the sensitivity of their data, these sectors depend heavily on generating synthetic datasets to ensure coverage and enable domain-specific augmentation.

Table 10. Overview of the nature of project data across use cases

Domain	Use case	Available data						Creation of additional data		
		Public			User					
		Unlabelled	Annotated	Synthetic	Unlabelled	Annotated	Synthetic	Unlabelled	Annotated	Synthetic
Science	SciDev	X	X	X	X					X
Science	ClaimAnalysis		X	X						X
Science	Research_SSH	X								X
Public Services	PADA				X					X
Public Services	LuxHubAI				X					X
Public Services	LuLoBots				X		X	X		X
Public Services	AI4LC	X	X							X
Public Services	ELISA	X				X		X		
Public Services	IRLT_LLM	X	X		X	X				
Public Services	ARTE-PT				X			X		
Public Services	GenAI4Med	X				X				
Public Services	Diavgeia				X	X				
Tourism	HeritAI		X			X		X		X
Telecom	DomainAdapt				X	X				X
Telecom	ConvAnalytics									X
Telecom	PIID								X	X
Telecom	KGGen					X				X
Energy	NER	X								X

5.3. Fine-tuning types

This section provides an overview of the specific fine-tuning techniques, as detailed in Section 4.1 and summarized in Table 2, considered for each defined use case.

As expected for a project focused on multilingual LLM fine-tuning for specific application domains, Table 11 shows that the most commonly required techniques aim to specialize the capabilities of pre-trained general models, including: (i) *domain-specific supervised instruction tuning*, adapting the model's instructions to the target domain; (ii) *domain alignment and*

enrichment, aligning the model with domain knowledge and enhancing it with relevant information; (iii) *downstream task fine-tuning*, optimizing the model for specific tasks within the domain.

The second most required category of techniques involves *other forms of learning*. In this context, Research_SSH, LuxHubAI, GenAI4Med and all RE- use cases, require approaches such as contrastive learning [66] to support retrieval or multimodal tasks. LuLoBots and IRLT_LLM rely on transfer learning [67], leveraging models pre-trained on tasks related to the use case learning objective. Meanwhile, Elisa requires a combination of contrastive learning, knowledge distillation [68], and continual learning techniques [69].

User preference alignment is mainly employed in application domains that specifically emphasize factuality and fairness, such as Science, Tourism, Telecom, and Public Services. Finally, *further training on domain data* is considered for solutions that require extended domain knowledge.

Table 11. Overview of the fine-tuning techniques required across use cases

Domain	Use case	Further pre-training on domain data	Domain-specific supervised instruction tuning	Domain alignment and enrichment	User preference alignment	Downstream task fine-tuning	Other forms of learning
Science	SciDev	X	X	X	X		
Science	ClaimAnalysis					X	
Science	Research_SSH					X	X
Public Services	PADA		X			X	
Public Services	LuxHubAI			X			X
Public Services	LuLoBots			X			X
Public Services	AI4LC	X	X	X	X	X	
Public Services	ELISA		X	X	X	X	X
Public Services	IRLT_LLM	X	X	X	X		X
Public Services	ARTE-PT		X	X			
Public Services	GenAI4Med		X			X	X
Public Services	Diavgeia	X	X	X		X	
Tourism	HeritAI		X	X	X	X	
Telecom	DomainAdapt	X	X	X	X		
Telecom	ConvAnalytics		X	X		X	
Telecom	PIID	X	X	X	X	X	
Telecom	KGGen					X	
Energy	NER		X	X		X	X

5.4. Thematic areas coverage

T4.2 focuses on the development and optimization of fine-tuning adaptation strategies. The task explore state-of-the-art fine-tuning solutions and define methodologies tailored to the different domains, modalities, and languages involved. The initial work concentrated on identifying the most important thematic areas to investigate, providing a theoretical foundation to support use case development. This guidance helps use case developers adopt the most appropriate state-of-the-art techniques, specifically optimized for each case. A comprehensive description of this process will be provided in D4.2 “Fine-tuning Methodologies Applied to Pre-trained Large Foundation Models” (M24).

This section provides a mapping between the specified use cases and the identified thematic areas. The mapping was carried out by considering the specific fine-tuning requirements expressed by each use case, highlighting the bidirectional work performed in WP4: defining training requirements (T4.1) and converging emerging needs into technical thematic areas to guide development and support the use case implementation (T4.2).

Table 12 provides an overview of the mapping between use cases and thematic areas. Overall, *further pre-training/domain alignment* [7], *instruction tuning* [70], and *downstream task optimization* [71] emerge as the most consistently required techniques across domains, while emphasis on *biases/human alignment* [72, 73], *efficiency and scalability* [74], and *multimodality* [75] varies depending on sector-specific priorities.

Further pre-training and *domain alignment* are widely considered, indicating that many solutions plan to leverage methodologies to adapt pre-trained models to specialized knowledge. *Instruction tuning* is also prevalent and is commonly used to refine model outputs for specific tasks, with broad adoption in Public Services, Telecom, Tourism, and Energy to guide responses effectively. Optimization of *downstream tasks* is central, highlighting the need to tailor pre-trained models for precise functions or analyses. *Multimodality* is present in SciDev, Research_SSH, IRLT_LLM, HeritAI, and all Energy capabilities, reflecting the importance of handling multiple data types in certain solutions. Techniques targeting *biases* and *human alignment* are applied selectively, mainly in scenarios involving sensitive data and user interactions, ensuring that outputs remain safe and aligned. Similarly, techniques aimed at *efficiency and scalability* are employed in contexts requiring intensive further pre-training or high computational demands, ensuring that solutions remain performant and resource-efficient.

Table 12. Mapping between use cases and thematic areas

Domain	Use case	Efficiency and Scalability	Biases/Human Alignment	Multimodality	Further Pretra./Domain align.	Instruction tuning	Downstream tasks
Science	SciDev	X	X	X	X	X	
Science	ClaimAnalysis						X
Science	Research_SSH			X			X
Public Services	PADA					X	X
Public Services	LuxHubAI				X		
Public Services	LuLoBots				X		
Public Services	AI4LC	X	X		X	X	X
Public Services	ELISA		X		X	X	X
Public Services	IRLT_LLM	X	X	X	X	X	
Public Services	ARTE-PT				X	X	
Public Services	GenAI4Med					X	X
Public Services	Diavgeia	X			X	X	X
Tourism	HeritAI		X	X	X	X	X
Telecom	DomainAdapt	X	X		X	X	
Telecom	ConvAnalytics				X	X	X
Telecom	PIID				X	X	X
Telecom	KGGen						X
Energy	NER			X	X	X	X

6. Conclusions

This document serves as the technical backbone of the LLMs4EU project, detailing the selection, definition, and analysis of its use cases. D4.1 aims to establish the foundation for WP4, which focuses on methodologies, algorithms, and infrastructure for model fine-tuning. Its goal is to provide EU citizens and SMEs with specialized GenAI solutions, addressing existing gaps in the European industrial landscape.

The document begins by specifying the objectives and defining the criteria for use case selection. This process has produced a set of fully elaborated use cases, along with additional candidates under evaluation, reflecting the partners' substantial effort to date. It establishes a systematic methodology for continuous development, designed to scale with additional use cases. Each use case is described with respect to its technological and innovative capabilities, functional relevance to the project, and compliance with European values, including multilingual support, use of regulated data sources, and adherence to applicable legal frameworks.

T4.1 carried out a systematic process of data collection and requirements analysis, capturing detailed specifications for fine-tuning techniques, base model selection, dataset requirements, GPU resources, training methodologies, and deployment considerations. Analysis of the collected data revealed several key insights, providing detailed information on language coverage, data sources, fine-tuning needs, and connections to thematic areas defined collaboratively with T4.2. These insights are essential for guiding both the technical and methodological support, leveraging state-of-the-art approaches.

The document establishes a comprehensive framework in which concrete use cases are defined, prioritized, and systematically organized. This methodology produces actionable data to guide the development, prioritization, and deployment of LLM-based solutions, ensuring they remain effective, safe, and aligned with both user needs and broader societal requirements.

T4.1, covered by this document, specifies the requirements necessary to proceed with the upcoming WP4 actions, guiding the development of a multi-model API (T4.3), the fine-tuning and adaptation of pre-trained models across EU languages and industry domains (T4.4), and their adoption and assessment by use-case owners (T4.5).

References

- [1] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities," Oct. 2024.
- [2] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," Sep. 2023, doi: 10.1007/s12599-023-00834-7.
- [3] W. X. Zhao *et al.*, "A Survey of Large Language Models," Feb. 2026.
- [4] C. Zhou *et al.*, "A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT," May 2023.
- [5] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 6282–6293. doi: 10.18653/v1/2020.acl-main.560.
- [6] M. Hu *et al.*, "A Survey of Scientific Large Language Models: From Data Foundations to Agent Frontiers," Oct. 2025.
- [7] K. Gupta *et al.*, "Continual Pre-Training of Large Language Models: How to (re)warm your model?," Sep. 2023.
- [8] A. Dmonte, R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein, "Claim Verification in the Age of Large Language Models: A Survey," Feb. 2025.
- [9] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for Fact Extraction and VERification," Dec. 2018.
- [10] D. Wadden *et al.*, "Fact or Fiction: Verifying Scientific Claims," Oct. 2020.
- [11] R. Shor, E. A. Greene, L. Sumberg, and A. B. Weingrad, "AI Tools in Academia: Evaluating NotebookLM as a Tool for Conducting Literature Reviews," *Psychiatry*, pp. 1–10, Aug. 2025, doi: 10.1080/00332747.2025.2541531.
- [12] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 2024.
- [13] R. Sharma, S. Gulati, A. Kaur, A. Sinhababu, and R. Chakravarty, "Research discovery and visualization using ResearchRabbit: A use case of AI in libraries," *COLLNET Journal of Scientometrics and Information Management*, vol. 16, no. 2, pp. 215–237, Jul. 2022, doi: 10.1080/09737766.2022.2106167.

- [14] G. Papageorgiou, V. Sarlis, M. Maragoudakis, and C. Tjortjis, "Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models," *Applied Sciences*, vol. 14, no. 18, p. 8259, Sep. 2024, doi: 10.3390/app14188259.
- [15] Y. Zhu *et al.*, "Large Language Models for Information Retrieval: A Survey," Sep. 2025, doi: 10.1145/3748304.
- [16] N. Hosseini-Kivanani, C. Schommer, and P. Gilles, "Voices of Luxembourg: Tackling Dialect Diversity in a Low-Resource Setting," [Online]. Available: <https://api.semanticscholar.org/CorpusID:282198825>
- [17] C. Lothritz, J. Cabot, and L. Bernardy, "Testing Low-Resource Language Support in LLMs Using Language Proficiency Exams: the Case of Luxembourgish," Jan. 2026.
- [18] F. Dehghani, R. Dehghani, Y. Naderzadeh Ardebili, and S. Rahnamayan, "Large Language Models in Legal Systems: A Survey," *Humanit. Soc. Sci. Commun.*, vol. 12, no. 1, p. 1977, Dec. 2025, doi: 10.1057/s41599-025-05924-3.
- [19] Y. Su, J. Bai, Q. Xu, K. Xu, and Y. Dou, "Audio-Language Models for Audio-Centric Tasks: A survey," Jan. 2025.
- [20] T. Nguyen *et al.*, "Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives," Jul. 2025.
- [21] H. Zhang, P. S. Yu, and J. Zhang, "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models," Jun. 2024.
- [22] S. Gu, "A Survey of Large Language Models in Tourism (Tourism LLMs)," *Qeios*, Feb. 2024, doi: 10.32388/8R27CJ.
- [23] Q. Wei, M. Yang, J. Wang, W. Mao, J. Xu, and H. Ning, "CTourLLM: Enhancing LLMs with Chinese Tourism Knowledge," Sep. 2025.
- [24] H. Zhou *et al.*, "Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities," Sep. 2024.
- [25] T. Deußer, L. Sparrenberg, A. Berger, M. Hahnbück, C. Bauckhage, and R. Sifa, "A Survey on Current Trends and Recent Advances in Text Anonymization," Aug. 2025, doi: 10.1109/DSAA65442.2025.11247969.
- [26] H. Bian, "LLM-empowered knowledge graph construction: A survey," Oct. 2025.
- [27] I. Keraghel, S. Morbieu, and M. Nadif, "Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study," Dec. 2024.

- [28] X. Zhao *et al.*, "A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers," Jun. 2024.
- [29] S. Minaee *et al.*, "Large Language Models: A Survey," Mar. 2025.
- [30] F. Godest, M. El Alaoui, V. Richet, and M. Olivier, "Enhanced LLM for Smart Knowledge Management in Nuclear Industry," in *International Conference Knowledge Engineering and Knowledge Management*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:279266234>
- [31] M. Anwar, D. Lau, M. de Costa, and I. Hammad, "Unlocking the Potential of Large Language Models in the Nuclear Industry with Synthetic Data," Jun. 2025.
- [32] A. Gonzalez-Agirre *et al.*, "Salamandra Technical Report," Feb. 2025.
- [33] P. H. Martins *et al.*, "EuroLLM: Multilingual Language Models for Europe," Sep. 2024.
- [34] A. Pareja *et al.*, "Unveiling the Secret Recipe: A Guide For Supervised Fine-Tuning Small LLMs," Dec. 2024.
- [35] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," May 1996.
- [36] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey," Sep. 2024.
- [37] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021.
- [38] V. Prabhakar *et al.*, "OmniScience: A Domain-Specialized LLM for Scientific Reasoning and Discovery," Apr. 2025.
- [39] A. Q. Jiang *et al.*, "Mistral 7B," Oct. 2023.
- [40] R. Rozis and R. Skadins, "Tilde MODEL - Multilingual Open Data for EU Languages," in *Nordic Conference of Computational Linguistics*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:45870801>
- [41] R. Santos, J. R. Silva, L. Gomes, J. Rodrigues, and A. Branco, "Advancing Generative AI for Portuguese with Open Decoder Gervásio PT*," in *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, M. Melero, S. Sakti, and C. Soria, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 16–26. [Online]. Available: <https://aclanthology.org/2024.sigul-1.3/>
- [42] L. Voukoutis *et al.*, "Meltemi: The first open Large Language Model for Greek," Jul. 2024.
- [43] C.-J. Li *et al.*, "POWSM: A Phonetic Open Whisper-Style Speech Foundation Model," Jan. 2026.

- [44] Y. Peng *et al.*, "OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning," May 2025.
- [45] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," Jul. 2024.
- [46] Z. Nie *et al.*, "When Text Embedding Meets Large Language Model: A Comprehensive Survey," Oct. 2025.
- [47] S. Latif *et al.*, "Transformers in Speech Processing: A Survey," Jun. 2025.
- [48] A. Ghosh, A. Acharya, S. Saha, V. Jain, and A. Chadha, "Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions," Oct. 2025.
- [49] C. Cui *et al.*, "PaddleOCR 3.0 Technical Report," Jul. 2025.
- [50] S. Taghadouini, A. Cavaillès, and B. Aubertin, "LightOnOCR: A 1B End-to-End Multilingual Vision-Language Model for State-of-the-Art OCR," Jan. 2026.
- [51] J. Hu, H. Jia, M. Hassan, L. Yao, B. Kusy, and W. Hu, "LightLLM: A Versatile Large Language Model for Predictive Light Sensing," Nov. 2024.
- [52] M. Marone, O. Weller, W. Fleshman, E. Yang, D. Lawrie, and B. Van Durme, "mmBERT: A Modern Multilingual Encoder with Annealed Language Learning," Sep. 2025.
- [53] A. H. Liu *et al.*, "Ministral 3," Jan. 2026.
- [54] Z. Bi *et al.*, "Is GPT-OSS Good? A Comprehensive Evaluation of OpenAI's Latest Open Source Models," Dec. 2025.
- [55] C. Clark *et al.*, "Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding," Feb. 2026.
- [56] J. Bai *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," Oct. 2023.
- [57] Granite Vision Team *et al.*, "Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence," Feb. 2025.
- [58] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," Apr. 2020.
- [59] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," Mar. 2023.
- [60] B. Warner *et al.*, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast,

Memory Efficient, and Long Context Finetuning and Inference,” Dec. 2024.

- [61] N. Boizard *et al.*, “EuroBERT: Scaling Multilingual Encoders for European Languages,” Mar. 2025.
- [62] A. Jain, N. M. Meenachi, and B. Venkatraman, “NukeBERT: A Pre-trained language model for Low Resource Nuclear Domain,” Aug. 2020.
- [63] A. F. Araujo and R. M. Marcacini, “RE-BERT: automatic extraction of software requirements from app reviews using BERT language model,” *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021, [Online]. Available: <https://api.semanticscholar.org/CorpusID:233354151>
- [64] W. Wu, B. Li, L. Chen, J. Gao, and C. Zhang, “A Review for Weighted MinHash Algorithms,” Nov. 2018.
- [65] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos, “SemDeDup: Data-efficient learning at web-scale through semantic deduplication,” Mar. 2023.
- [66] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, “A comprehensive survey on contrastive learning,” *Neurocomputing*, vol. 610, p. 128645, Dec. 2024, doi: 10.1016/j.neucom.2024.128645.
- [67] F. Zhuang *et al.*, “A Comprehensive Survey on Transfer Learning,” Jun. 2020.
- [68] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” May 2021, doi: 10.1007/s11263-021-01453-z.
- [69] D.-W. Zhou, H.-L. Sun, J. Ning, H.-J. Ye, and D.-C. Zhan, “Continual Learning with Pre-Trained Models: A Survey,” Apr. 2024.
- [70] S. Zhang *et al.*, “Instruction Tuning for Large Language Models: A Survey,” Oct. 2025.
- [71] B. Guo *et al.*, “Sample Design Engineering: An Empirical Study of What Makes Good Downstream Fine-Tuning Samples for LLMs,” Apr. 2024.
- [72] I. O. Gallegos *et al.*, “Bias and Fairness in Large Language Models: A Survey,” Jul. 2024.
- [73] Y. Wang *et al.*, “Aligning Large Language Models with Human: A Survey,” Jul. 2023.
- [74] Z. Wan *et al.*, “Efficient Large Language Models: A Survey,” May 2024.
- [75] S. Yin *et al.*, “A Survey on Multimodal Large Language Models,” Nov. 2024, doi: 10.1093/nsr/nwae403.