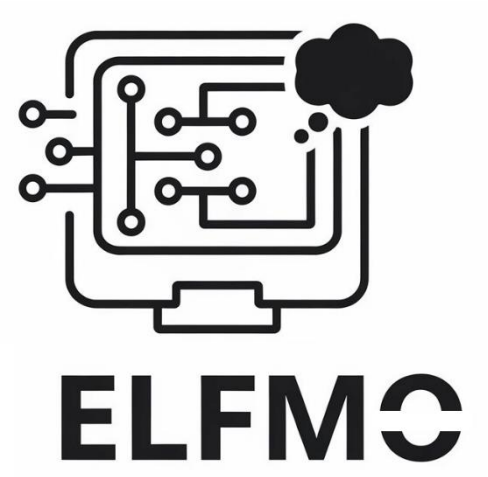

Engineering Large Foundational Models for Enterprise Integration

Deliverable D2.1
Research baseline for model training and benchmarking

Project title	Engineering Large Foundational Models for Enterprise Integration
Project acronym	ELFMo
Project number	23004
Work package	WP2
Deliverable	D2.1
Dissemination level	PU Public
License	CC-BY 4.0
Version	1.0
Date	2025-07-21

Contributors

Editor(s)	Fabio Román (Dextromedica)
Reviewer(s)	Robin Bornoff (Siemens)
Contributor(s)	Tomi Sarni (Nosto) Robin Bornoff (Siemens) Anna Kantosalo (Siili) Manu Setälä (Solita)

	Jukka K. Nurminen (University of Helsinki) Marcos Cobos (CIC)

Abstract

This deliverable D2.1 *Research baseline for model training and benchmarking* provides an overview of the technical foundations, tools, and strategies for the trustworthy adaptation and integration of Large Foundation Models (LFMs) within enterprise contexts, guiding the activities executed in WP2. It outlines detailed methods for model adaptation, including efficient fine-tuning techniques, continuous learning strategies, and automation of data preparation processes. The deliverable also describes the design and validation procedures applied to ensure model performance, safety, and compliance with industrial and regulatory standards. Furthermore, it presents a comparative evaluation of adaptation strategies based on technical metrics and business KPIs, along with initial use cases in healthcare and telecommunications that illustrate practical applications and expected impacts. The conclusions section summarizes the core findings and outlines future directions for the continued development and validation of LFM-based enterprise solutions under the ELFMo project.

Table of contents

1 Introduction

1.1 Context and relevance of reliable adaptation of LFMs

The rapid evolution of Large Foundation Models (LFMs) has opened up new possibilities for automating complex tasks across a wide range of industrial domains. However, their successful adoption requires tailored adaptation to specific use cases, ensuring that performance, security, and reliability meet industry standards and industrial requirements and constraints.

Within the ELFMo project, reliable LFM adaptation addresses the critical need to deploy these models effectively in real-world scenarios. This involves not only technical refinement of the models but also the development of robust tools, infrastructures and dashboards that support scalable, maintainable, and domain-specific integration.

The relevance of this task lies in its direct alignment with the broader goals of the project: improving operational efficiency, supporting high-impact decision-making, and minimizing the computational and economic costs of deploying LFM-based solutions at scale.

1.2 Objectives of the document

This document aims to detail the design, implementation, and expected outcomes of **Task 2.1 – Trustworthy Adaptation and Integration of LFMs**. The main objectives are:

- To **support the efficient and trustworthy adaptation of LFMs** for a variety of industrial use cases.
- To place particular emphasis on **high-yield and cost-effective continuous learning strategies**.
- To document the **methods and tools for automated data preparation**, including data cleaning and normalization, which are essential for accelerating model adaptation.
- To assess the **expected impact** of these activities in improving accuracy, reducing time to deployment, and enabling cost-efficient integration of LFMs into production environments.

The document also outlines the contributions of different partners involved in this task and situates Task 2.1 within the broader innovation framework of the ELFMo project

1.3 Scope and structure of the content

Based on source materials, Task 2.1 titled **"Trustworthy Adaptation and Integration of LFMs"** focuses on the development of tools, methods, and infrastructures to **enable efficient and reliable adaptation of Large Foundation Models (LFMs)** for various use cases.

The **core goals** of this task are:

- To support the reliable and efficient adaptation of LFMs for multiple use cases.
- To emphasize **high-performance, cost-effective continuous learning**.

Key **activities** under this task include:

- The development of **data preparation methods**, including automated data cleaning and normalization, to streamline the adaptation pipeline.

The **expected impact** is to **significantly increase the efficiency** of adapting LFMs to specific use cases, thereby enabling **faster decision-making**, improved **accuracy**, substantial **cost savings** and a **reduced time to market**.

This task directly supports **Innovation Pillar 2** of the ELFMo project: *"Tools, methods, and infrastructures for trustworthy LFM adaptation and integration."*

In summary, Task 2.1 is a foundational component of the ELFMo project, as it addresses the **technical challenges of adapting LFMs to meet industry-specific needs**, with a focus on efficiency, reliability, and affordability through advanced tools and methods—especially in the domains of data preparation and fine-tuning.

1.4 Related Documents

The following abbreviations are used to describe other documents related to the project to this deliverable.

- **[FPP] ELFMo – Full Project Proposal:** describes the full project proposal

2 Development of tools and methods for reliable adaptation of LFMs

2.1. Infrastructures and methodologies for efficient integration of LFMs

This research is fundamental to build the necessary tools, methodologies and infrastructures to adapt these models efficiently, reliably and securely to the specific use cases of several companies.

Infrastructures and Methodologies for Efficient LFM Integration

- Specific Aim: To define the base architecture and operational practices for deploying and managing tailored LFMs.
- Research Process: State of the Art Review: analysis of reference architectures (AWS, Azure, GCP), patterns (Serverless, EDA, Microservices), and MLOps frameworks for LFM lifecycle management.
- Technology Assessment: comparison of specific cloud services (AWS Bedrock, SageMaker, Lambda, Step Functions, DynamoDB, etc.) and relevant open-source tools.
- Architecture Design: Elaboration of a detailed architectural blueprint for the ELFMo platform, considering scalability, resilience and costs. Include diagrams of components and data flows.
- Maintenance Strategies: Investigation of continuous monitoring models, model versioning, retraining/upgrade strategies (continuous learning) and operational cost management.
- Proof of Concept (PoC): Implementation of small prototypes to validate key architectural decisions (e.g. inference latency, scalability of Lambda functions).
- Methodologies: Literature review, benchmarking, architecture design, rapid prototyping, benchmarking.
- Expected Results: Architecture design document, technology assessment report, MLOps best practices guide for LFMs, PoCs results.

Advanced Model Customization Techniques (LFM/LLM and SLM)

- Specific Aim: Master and validate techniques to adapt language models specific knowledge and tasks.
- Research Process:
  - Prompt Engineering:
    - Research techniques (Zero-shot, Few-shot, Chain-of-Thought, Structured Prompts).
    - Development of an internal library/guide of effective prompts for common CX tasks (summarization, classification, response, sentiment analysis).
    - Experimentation with automatic prompt optimization.
    - Evaluation of its effectiveness in real Konecta use cases.

    - Retrieval-Augmented Generation (RAG):
    - Evaluation of vector databases (e.g., Pinecone, ChromaDB, OpenSearch, Neo4J).
    - Research of chunking and embedding strategies for Konecta documents (FAQs, scripts, policies).
    - Development and testing of RAG pipelines integrated with selected LFMs.
    - Benchmarking of the quality (relevance, accuracy) and latency of responses generated with RAG vs. without RAG.
- Fine-tuning:
    - Investigation of fine-tuning techniques (Full fine-tuning vs. Parameter-Efficient Fine-Tuning - PEFT, such as LoRA, QLoRA).
    - Preparation of Konecta specific datasets for fine-tuning (requires collaboration with Area 3.4).
    - Performance of fine-tuning experiments on base LFMs (open-source or commercial via API if applicable).
    - Comparative analysis of performance, computational cost and training time between different techniques.
- SLM exploration:
    - Identification of high volume and specificity tasks where an SLM could be advantageous (e.g. very specific intent classification, concrete entity extraction).
    - Investigation of relevant pre-trained SLMs or knowledge distillation techniques from LFMs.
    - Design of experiments to train/refine SLMs using Konecta-specific data, taxonomies and ontologies.
    - Comparison of performance (accuracy, latency) and cost (training, inference) between SLMs and LFMs adapted for selected tasks.

- Methodologies: Experimental design, benchmarking (metrics such as BLEU, ROUGE, F1-score, accuracy, latency), A/B testing, cost-benefit analysis.
- Expected results: Best practice guidelines for Prompting, RAG and Fine-tuning; Comparative benchmarking reports; Curated datasets for fine-tuning; Feasibility study and prototypes (if applicable) of SLM-based solutions; Adapted models (prototypes).

Data Preparation and Specific Knowledge Management

- Specific Aim: Create an efficient and sustainable workflow to prepare data and structure the knowledge needed for model customization.
- Research Process: Inventory and Analysis of Data Sources: identification and characterization of all relevant data sources (CC, CRM, internal DBs, chat/voice logs, knowledge bases, process documentation)

Quality and format analysis.

- ETL/ELT Pipeline Development: Design and implementation of automated flows for extraction, transformation (cleansing, PII anonymization, normalization) and loading of data into a centralized repository or data lake.
- Feature Engineering: Creation of relevant features from raw data to improve model performance.
- Knowledge Structuring (Taxonomies/Ontologies): Research on the feasibility and value of defining formal taxonomies and ontologies for specific business verticals or countries.
- Development (if deemed feasible) of prototypes of these structures to organize knowledge and improve the accuracy of RAG or SLM training.
- Data Governance: Definition of policies and processes for quality management, updating, versioning and access to data used by the IA.

- Methodologies: Data analysis, data pipeline design, software engineering, knowledge modeling (if applicable ontologies), definition of governance policies.
- Expected Results: Catalog of evaluated data sources, automated data preparation pipelines (code and documentation), prototype taxonomy/ontology (if applicable), proposed data governance framework for AI.

2.2. Safety and reliability considerations in model adaptation

Security, Reliability and Bias Mitigation

- Specific Aim: To establish a framework for the responsible and reliable development and deployment of generative AI at Konecta.
- Research Process:
  - Risk and Threat Analysis: Identification of specific vulnerabilities of LFMs in CX environments (Prompt Injection, Jailbreaking, inappropriate content generation, data leaks) 2. Threat modeling.
  - Security Measures Evaluation: Research and implementation of input filtering/sanitization techniques, output guarding/filtering, and monitoring of anomalous usage patterns.
  - Bias Detection and Mitigation:
  - Audit of biases in training/adaptation datasets.
  - Investigation of metrics and tools to detect biases (linguistics, demographic, etc.) in model outputs.
  - Experimentation with mitigation techniques (data rebalancing, prompts adjustment, debiasing algorithms).
  - Verification and Validation: Definition of robust testing protocols, including functional performance, security (simulated network teaming) and response quality/coherence assessment tests. Establishment of reliability metrics.

    - Traceability and Compliance: Design of mechanisms to record interactions, model decisions and data/model lineage for auditing and compliance (GDPR, AI Act).
- Methodologies: Threat modeling, simulated penetration testing (network teaming), statistical bias analysis, fairness audits, test protocol design, regulatory requirements analysis.
- Expected Results: IA risk assessment framework, catalog of security and bias mitigation measures, model validation and verification protocol, traceability and audit system design.

Design and Prototyping of Platform Components

- Specific Aim: Validate the technical and functional feasibility of key components of the ELFMo platform.
- Research Process:
    - Process Orchestrator: detailed design of the orchestration engine (based on AWS Step Functions), including the definition of states, transitions and context management 2. Prototyping of the core of the orchestrator.
    - Connectors: Design of a standard connector pattern. Development of connector prototypes for 1-2 key systems (e.g. a CRM, a messaging platform).
    - Adaptive Interface: Investigation of UI/UX design principles for AI-powered interfaces. Design of mockups and interactive prototypes (low-fidelity and high-fidelity) for agent co-pilot and potentially customer interfaces.
    - AI Reporting Module: Module architecture design. Prototyping of the natural language-based reporting capability (NLP query -> Report).
    - Contact Center Copilot: Detailed functional design. Prototyping of key functionalities (e.g. response suggestion, interaction summary) integrating adapted models (from Area 3.2).
- Methodologies: Software design, microservices/serverless architecture, API design (REST), rapid prototyping, user-centered design (UCD), usability testing.

2.3. Challenges and solutions in adapting models to different use cases

Research Risk Management

- Technology Uncertainty: Mitigation through continuous review of the state of the art, rapid prototyping and modular design to allow interchangeability of components/models.
- Data Availability/Data Quality: Mitigation through early analysis and development of robust readiness pipelines. Investigation of techniques robust to sparse/noisy data.
- Model Performance/Cost: Mitigation through comprehensive benchmarking, SLM exploration and infrastructure optimization.

- Adoption and Usability: Mitigation through user-centered design and early prototype testing.

The research will require a multidisciplinary team with experience in:

- Product Owners specialized in product definition, understanding this as the platform that we will make available to automate and industrialize customer service processes through GenAI.
- Artificial Intelligence and Machine Learning (specialization in NLP and LLMs).
- Software Engineering (Backend, Frontend, Cloud - especially AWS).
- Data Engineering and Architecture.
- Information Security.
- User Experience Design (UX/UI).
- Domain experts (Konecta CX/BPO). Computing resources (Cloud), access to model APIs (if applicable), software licenses and specific tools will be required.

Additionally, there will be interaction and collaboration with Konecta BTO business support areas including legal, HR, finance, compliance, cybersecurity, technology, etc....

Ethical and Sustainability Considerations

The implementation of an AI-based customer service automation platform such as the one proposed in the ELFMo project carries profound implications that go beyond operational efficiency. Konecta BTO is committed to approaching this project with the highest ethical and sustainability standards, recognizing the potential impacts on consumers, business customers, employees and society at large. This commitment will be materialized through the following considerations and actions

An internal ethics review committee or process will be established to evaluate key phases of the project, from design through deployment and ongoing operation. Ethical and sustainability impacts will be actively monitored, and adjustments to the project will be made as necessary, while maintaining compliance with current legislation (such as the EU AI Act) and international best practice.

Environmental Sustainability:

- The use of energy-efficient AI models will be prioritized (investigating SLMs for specific tasks).
- Optimize algorithms and cloud infrastructure (e.g., use of appropriate instances, serverless architectures to minimize computational resource consumption).
- Consideration will be given to choosing data center regions that use renewable energy, in line with the sustainability policies of Konecta and its cloud providers (e.g. AWS).

Social Sustainability:

- Investment in professional development of employees to adapt to new roles generated by AI.

- Active promotion of diversity and inclusion in development teams and AI deliverables.
- Ensuring that automation does not degrade service quality or exclude vulnerable groups.

Economic Sustainability:

- Development of a technologically robust solution that is economically viable in the long term.
- Creation of added value for business customers and improvement of the experience for end consumers, ensuring Konecta's sustainable competitiveness.

3 Efficient and cost-effective continuous learning

Large Foundation Models (LFMs) undergo a multi-stage training process consisting of pre-training, instruction tuning, and alignment phases. During pre-training, LFMs learn through a self-supervised manner on vast text corpora, predicting masked tokens to build foundational language understanding. The instruction tuning stage employs supervised fine-tuning on curated instruction-output pairs, teaching models to generate appropriate responses to task-specific instructions. Finally, the alignment stage refines LFMs using human feedback and human preference datasets to ensure outputs conform to ethical standards and societal norms (Wu et al 2024).

In the rapidly evolving world of Generative AI, these learning phases must be considered continuously to ensure models maintain factual accuracy, adapt to new domains, and master new tasks without catastrophic forgetting, i.e., losing their prior knowledge and skills. Here, we focus on three main areas, continuous knowledge base updates that keeps the model's information current, incremental task-specific training for learning new tasks without forgetting old ones, and continuous model alignment for maintaining human values and preferences.

Section 3.1 focuses on methods how to learn efficiently, and how to keep models current in production, while section 3.2 discusses how to evaluate the performance of continuous learning strategies.

3.1 Continuous learning strategies to minimize costs

This section focuses on providing methods for keeping a model up to date without extensive retraining, and answering the following: What should we train, how often, and how to perform this in a cost-efficient way.

The ability of LFMs to integrate and adapt to recent information is crucial for maintaining their relevance and continued efficacy. Effective strategies involve constructing dynamic datasets from diverse sources such as news feeds and scientific articles, or even social media then updating existing models with new information. Several approaches have been used for frequent model updates.

Parameter-Efficient methods

Low-rank adaptation (Hu et al. 2021) adds trainable low-rank matrices to a subset of layers in the pre-trained model while keeping the original model weights frozen. Only the added parameters are trained, which drastically reduces the number of parameters that must be updated. This method is architecture agnostic and works with encoder-only, decoder-only, encoder-decoder architectures.

K-Adapter is a parameter-expansion technique that keeps the original language model parameters frozen while adding k new adapter layers that get updated during continued pretraining. Originally developed by Wang et al. (2021), this method was successfully tested on encoder-only models like BERT and RoBERTa to inject factual and linguistic knowledge. The method can be extended to encoder-decoder and decoder-only models.

Regularization approaches

These methods incorporate additional term in the loss function to penalize changes in crucial weights of the network. Elastic Weight Consolidation (EWC) (Kirkpatrick et al. 2017) constrains important parameters to remain close to previous values, while Orthogonal Gradient Descent (OGD) (Farajtabar et al. 2019) restricts parameter movement within orthogonal spaces defined by previous task gradients. Gradient Episodic Memory (GEM) (Lopez-Paz et al. 2017) leverages episodic memories to prevent catastrophic forgetting. Regularization methods are suited for open-source models but may be difficult to incorporate with commercial LFMs.

Rehearsal-based methods

Mix-Review (He et al., 2021) is a rehearsal-based method that requires access to the original pretraining data. During continued pretraining, it blends random samples from the initial corpus with new data according to a mix-ratio that decays over time. As training advances, this ratio approaches zero, progressively reducing the amount of original data included in each training step. As the initial corpus of the LFMs can be difficult to obtain, mix-review may be challenging to use in LFM tuning.

Task-incremental training

Beyond knowledge updates, LFMs require adaptation to new tasks while preserving performance on previously learned skills. Two prompt-based approaches may address this challenge:

Prompt Tuning: Traditional prompt tuning (Lester et al. 2021) trains a single shared soft prompt across all tasks sequentially while freezing underlying model parameters. This approach provides parameter efficiency but may suffer from task interference.

Progressive Prompts: This strategy learns dedicated soft prompts for each new task, concatenating them with previously learned prompts while maintaining a frozen base model (Razdaibiedina et al. 2023). Empirical evaluation demonstrates significant improvements, achieving over 20% better average test accuracy compared to traditional approaches when evaluated on T5 models.

Continuous alignment

It is also important to align LFMs with human preferences and values. One approach for alignment is reinforcement learning from human feedback (RLHF). In the first phase of RLHF, the base model is fine-tuned on high-quality human examples of desired behaviour. Human evaluators then compare and rank multiple model outputs for the same prompt. This comparison data trains a reward model that learns to predict human preferences and assign scores to different responses. Finally, the model is trained to maximize the reward model's scores using algorithms like Proximal Policy Optimization (PPO) (Schulman et al. 2017).

3.2. Evaluating the impact of continuous learning strategies

While training models and executing continuous learning strategies, it's important to evaluate how the taken actions affected the model performance. In this section, we focus on providing tools to evaluate if the model learning had any help.

Benchmarking knowledge base updates

Evaluating the effect of continuous knowledge base updates requires a set of key metrics applied to time-dependent datasets. Several benchmark datasets have been tailored for this purpose – some of them continuously updating while some of them static:

- **TemporalWiki** provides automatically updating datasets using sequential Wikipedia and Wikidata snapshots (Jang et al. 2022a)
- **Firehose** offers large-scale social media data comprising six years with 100 million tweets (Hu et al. 2023)
- **CKL** targets web and news domains, evaluating preservation of time-invariant knowledge while acquiring new information (Jang et al. 2022b)

TRACE provides comprehensive evaluation across eight datasets covering specialized domains, multilingual capabilities, code generation, and mathematical reasoning (Wang et al. 2023b)

However, the rapid evolution of data presents a fundamental challenge: time-sensitive datasets quickly become obsolete, requiring frequent benchmark updates to maintain relevance of the datasets. From the datasets presented above only **TemporalWiki** updates in a continuous manner. In addition to benchmark data, a set of key metrics are used to evaluate model performance on updates:

Perplexity quantifies a model's uncertainty or confusion when making predictions. Perplexity measures how confident a language model is when predicting the next word in a sequence given the preceding context. A lower perplexity score indicates greater model confidence and more accurate predictions, while higher perplexity reveals increased uncertainty and reduced predictive accuracy.

In the specific context of knowledge-based updates, perplexity is calculated separately for two distinct categories of information: changed data i.e. newly updated or added knowledge

and unchanged data i.e. previously existing knowledge that should be retained. This dual measurement allows quantifying model's ability to acquire new information (plasticity) and maintain existing knowledge (stability). The final evaluation metric combines these measurements by averaging the individual perplexity scores across all categories, providing a comprehensive view of the model's overall performance in the continual learning setting (Jang et al., 2022a).

Exact match and precision are metrics used to quantify the effect of information updates. In **exact matches**, one forms factual datasets for instance with fact-based sentences and statements with historical (time-invariant) data, updated data and new data. Then models' ability to predict missing facts in the sentences is quantified. Exact match is hence ratio of exactly matching sentence predictions and number of statements. **Precision@k** measures whether the correct answer appears within the top-k predictions generated by the model, rather than requiring it to be the top prediction as in exact match.

Benchmarking task-incremental learning

LFMs can be trained incrementally to master new tasks. This process requires evaluation of performance across both previously learned and newly acquired skills with task-centric benchmark datasets. Examples of such task-oriented datasets are:

- **NATURAL INSTRUCTIONS** including 61 tasks across six categories, such as question generation and classification (Mishra et al. 2021)
- **SUPER-NATURAL-INSTRUCTIONS** consisting of approximately 1600 natural language processing tasks with expert-written instructions. The dataset spans across 76 types of tasks such as text generation, extraction and classification (Wang et al. 2022).

Three established metrics from (Lopez-Paz et al. 2017) provide a comprehensive framework for evaluating incremental task-specific learning performance. **Forward transfer rate** (FWT) measures how effectively prior knowledge accelerates learning on new tasks, capturing performance gains before any dedicated training on those tasks begins**. Backward transfer rate** (BWT) quantifies catastrophic forgetting by tracking how much performance on old tasks deteriorates after new learning occurs. Finally, **average accuracy** gives overall performance across all tasks. Together, these metrics reveal whether a model successfully assumes new skills or suffers from interference between old and new learning.

Benchmarking model alignment

For benchmarking model alignment, one can use the following datasets:

- **Stanford human preferences (SHP)** is a dataset comprises 385 000 collective human preference judgments across responses to questions and instructions in 18 diverse subject areas from cooking to legal advice. These preferences capture which responses users find more helpful when comparing different model outputs. SHP

focuses on helpfulness preferences rather than safety or harmlessness, so it is best suited for benchmarking helpfulness alignment rather than safety alignment (Ethayarajh et al. 2022).

- **Helpful and Harmless (HH)** data is gathered through a two-step approach. For the helpfulness component, crowd-workers engage in open-ended conversations with models, seeking assistance, advice, or task completion, and then select the more helpful response from paired model outputs. For the harmlessness dataset crowd-workers attempt to prompt harmful responses from models and then identify the more problematic response from the paired responses (Bai et al. 2022).

For alignment benchmarks, one can measure accuracy as the primary metric, comparing predicted preferences against human judgments. While interest in alignment research continues to expand, the field currently lacks comprehensive benchmarking datasets and methods needed for robust alignment evaluation (Wu et al. 2024).

4 Development of data preparation methods.

The performance of general LFMs is significantly influenced by the composition of its pre-training data. Most of the training corpus for LFMs constitute web content crawled from different websites on the Internet, including textbooks. This diversely large training corpus typically covers a broad range of topics and domains, which is important for developing task-agnostic LFMs with broad capabilities. Adaptation of general purpose LFMs can be done in several ways, including fine tuning the models using highly curated, domain/task-specific large dataset (e.g., Gu et al., 2023) or instruction-response datasets (Wang et al., 2022) to ensure effectual handling of specific domain terminologies and complexities or desired final task, such as question answering, reading comprehension and summarization.

Regardless of whether one is developing a new, or adapting existing, LFMs non-trivial data preparation steps will be performed. Research has shown that the quality of LFMs improves as the scale of carefully curated data increases (Gao et al., 2021; Zhang et al., 2022). Publicly available datasets, such as C4 (Raffel et al., 2020) and The pile (Gao et al., 2021) used to train general purpose LFMs typically undergo rigorous filtering, cleaning and transformation to compose training data. This section provides an overview of practices and tools used to prepare data for training and adapting LFMs.

4.1. Automation of data preparation tasks

- Tools for automated data collection, cleaning, and preprocessing.
- Implementation of data pipelines

Depicts a typical data pipeline for LFMs. In addition to existing data sources and collection mechanisms, web crawlers are implemented to gather web content from different sites on the Internet. Alternatively, web data can be acquired from public datasets that crawled data used to train some of the existing open-source LFMs.

One example of an open-source tool designed to support the automation of data preprocessing is the **Data Prep Kit** (https://github.com/data-prep-kit/data-prep-kit). This toolkit facilitates the **cleaning**, **transformation**, and **semantic enrichment** of unstructured, use case-specific data, enabling its direct use in **pre-training**, **fine-tuning**, or **Retrieval-Augmented Generation (RAG)** pipelines. It streamlines the preparation of heterogeneous textual sources by automating common preprocessing steps such as format normalization, duplicate removal, metadata tagging, and content filtering—thus reducing manual workload and ensuring consistent data quality across the LFM adaptation lifecycle.

4.2. Data cleaning processes and adaptation to use cases

Typically, methods such as filtering and duplication are employed to clean and pre-process data that is used to train LFMs. Despite their overlaps, public datasets e.g., C4 and The Pile, are cleaned and processed differently. For example, duplicate content was deliberately chosen to be included in The Pile dataset.

Data Filtering focuses on removing unwanted content that can negatively affect model behavior. Language text filters are used to remove sentences and documents that are falling below a certain threshold for a given language. For example, Langdetect and FastText operate at document level are often used to remove non-Engilsh documents. Content text filters remove any toxic (e.g., rude and disrespectful) and distract (e.g., HTML tags) content. Image-level filtering removes low resolution images that convey little information as well as those with inappropriate aspect ratios. Video filtering leverages techniques in image and text filtering.

Data Deduplication focuses on removing repeated data which can help to fasten model training as well as make the model less likely to exhibit memorization and have the model achieve similar or better perplexity compared to models trained with repeated data. Deduplication methods consider exact duplicates, approximate duplicates and semantic duplicates mostly at sentence-level and document-level. At document-level, approximate deduplication methods are mostly hashing-based, such as Locality Senstive Hashing based MinHash or SimHash. Recently, there is more interest in using pre-trained foundational models for semantic deduplication as semantic embedding metric for document-level embedding. The semantic deduplication process comes after the exact deduplication and approximate deduplication processes, to remove the semantic duplicates by clustering the embedding points and keeping representative data in each cluster.

Data Enhancement focuses on enhancing text data or the modality of data typically using traditional data augmentation methods. For example, for vision-language models, it includes enhancing the quality of image-caption datasets by rewriting the captions using BLIP2 or LLMs with carefully designed prompts to generate diverse captions. Additionally, domain-specific data can be augmented with public data used to train LFMs resulting in a combination of domain-specific text and general-purpose text (Wu et al, 2023).

Data for the adaptation of LFMs. When adapting LFMs to specific domains and use cases, carefully curated datasets that align with the target domain and use case are used to help the models acquire knowledge and abilities necessary to excel in the target application. However, very often instruction-response pairs datasets are constructed and used in adapting LFMs (refs) and enhance their ability to perform well in downstream tasks, such as question-answer, reasoning, captioning, classification etc. As such, primary knowledge of LFMs is acquired during the pre-training stage, and the purpose of Instruction tuning is to enable LFMs to learn

how to perform well on specific tasks and, for example, interact with humans. Research shows that only a small set of carefully crafted high-quality instructions is sufficient to endow LFMs with powerful instruction-following capabilities (Singhal et al, 2022). There exist small LFMs that are trained exclusively on domain-specific data (refs) or combination of both domain-specific data and general data sources (Wu et al. 2023) to ensure the resulting models not only do well in domain-specific tasks but are also able to maintain strong performance on general-purpose benchmarks.

- ***Instruction-response pairs of datasets***. Generally, the steps to curate instruction-response pairs datasets include collecting X-text pairs from various sources; cleaning the collected data through filtering, deduplication etc.; and constructing the cleaned X-text pairs into instruction-response form (see example Z below).
- **Domain-specific datasets**. Institutions can leverage their existing data sources and collection mechanisms to gather and construct large domain-specific datasets, e.g. financial data (Wu et al, 2023). Existing public datasets can be added to enhance and create a large training corpus (Wu et al, 2023).

5 Implementation and validation of LFM adaptation

The implementation and validation of **Large Foundational Models (LFMs)** in enterprise contexts necessitate a comprehensive approach that balances model quality, operational performance, and business relevance. Within the **ELFMo project**, this chapter explores how adapted LFMs are evaluated, not only through conventional technical metrics (such as accuracy and latency) but also by considering architectural dimensions (e.g., agent-based system design), cost-efficiency, and measurable impact on real-world business outcomes.

5.1. Testing and validation of adapted models

Testing within the **ELFMo** project encompasses technical accuracy, operational robustness, and enterprise relevance. The evaluation framework integrates multiple layers of validation to ensure that deployed models meet both performance and compliance requirements. Key components include:

- **Data Evaluation**: The use of both synthetic and real-world test sets, designed to reflect relevant enterprise use cases and operational scenarios.
- **Unit and Integration Testing**: Emphasising agentic and service-oriented workflows, particularly those implemented using orchestration frameworks such as **LangChain**.
- **Trust and Safety**: Incorporating bias audits, hallucination detection, and adversarial input testing to ensure safe and reliable interactions.

Security validation, aligned with guidance from the OWASP Foundation (2024), addresses several critical dimensions:

- Ensuring **data isolation** across domains (e.g. merchant-specific or healthcare-specific environments).
- Establishing **resilience against prompt injection** and adversarial query techniques.
- Enforcing **governance over RAG content**, ensuring that retrievers and vector stores are constrained to appropriately scoped datasets.
- Preventing **prompt memory leakage** and **cross-domain contamination** in agent-based systems.

In addition, **legal and regulatory compliance**, such as alignment with the **EU AI Act**, necessitates comprehensive auditing across these areas of concern.

Finally, **Observability** has become increasingly vital, particularly in multi-agent systems where a single user query may trigger orchestrated execution across LLMs and specialised service components. Protocols such as **MCP**, **A2A**, and **ACP** define modular, traceable execution paths. Tools like **LangChain** (for agent orchestration) and **Langfuse** (for end-to-end tracing and

debugging) are instrumental in mitigating common challenges related to latency, transparency, and traceability within these distributed pipelines.

5.2. Performance metrics and evaluation of adaptation

Evaluation within the **ELFMo** project incorporates a comprehensive set of metrics, encompassing both technical performance and deployment efficiency to ensure models are suitable for real-world enterprise integration.

Technical KPIs and SLOs

Predictive Accuracy: Standard evaluation metrics such as accuracy, F1-score, and task-specific measures (e.g. BLEU and ROUGE for generative tasks; MAPE for time series forecasting) are employed to quantify model effectiveness.

Latency and Throughput: These are increasingly critical for interactive or real-time use cases, particularly in domains such as e-commerce personalisation or security monitoring.

Resource Utilisation: Metrics include GPU hours, memory footprint, and cost-per-inference. These are evaluated across various adaptation strategies, including Parameter-Efficient Fine-Tuning (PEFT), full fine-tuning, Low-Rank Adaptation (LoRA), and Quantised LoRA (QLoRA).

Cost-Efficiency and Trade-offs

A central consideration in model deployment is managing the trade-off between cost, predictive quality, and latency. While full fine-tuning may enhance accuracy, it often incurs significant computational and serving costs. Retrieval-Augmented Generation (RAG) techniques can reduce hallucinations and improve factuality but introduce additional latency and architectural complexity due to external retrieval operations. These trade-offs become particularly pronounced in agentic deployments, where latency accumulates across chained agent calls and decision pathways (Zhao et al., 2023).

Accordingly, each architectural decision—be it lightweight prompt-based strategies, RAG-enhanced pipelines, or smaller task-specific LLMs—must be validated not only by technical metrics but also by its economic feasibility within the target production environment.

5.3. Comparison with base models and industrial benchmarks

To assess the benefits of model adaptation, evaluation should be conducted against a range of relevant benchmarks.

- **Baseline LFM Performance**: Comparative assessments should include traditional zero-shot and few-shot prompting, as well as **Retrieval-Augmented Generation (RAG)** approaches (Brinkmann & Bizer, 2024; Zhang, Nakatani, & Walter, 2024).
- **Commercial vs Open-Source Models**: Evaluation should consider the relative performance of proprietary models (e.g. GPT-4, Claude) against leading open-source baselines such as LLaMA and Falcon (Roumeliotis, Tselikas, & Nasiopoulos, 2024).

While standard benchmarks remain useful for technical validation, **business relevance** is ultimately paramount (Fang et al., 2024). In the context of e-commerce, for example, the performance of personalisation models must be validated, not only through predictive accuracy but also through their measurable impact on key business outcomes. Relevant metrics include:

- **Average Order Value (AOV)**
- **Conversion Rate (CR)**
- **Cart Abandonment Rate (CAR)**
- **Click-Through Rate (CTR)**

Although these metrics are domain-specific, the broader **LLMOps** practice tracking LFM versions against longitudinal business performance applies across industry verticals (Google Cloud, n.d.). This approach reflects a strategic shift from static A/B testing to **continuous, KPI-aligned evaluation** of large language model performance, enabling organisations to monitor effectiveness in real-time production settings.

6 Expected impact of reliable LFM adaptation

Reliably adapting Large Foundation Models (LFMs) offers substantial potential advantages in enhancing enterprise operations by improving efficiency, decision-making quality, and reducing operational costs. This section explores the expected benefits across these key dimensions and highlights sector-specific applications.

6.1. Improved model adaptation efficiency.

Traditional approaches to model adaptation often involve exhaustive data collection and complete retraining processes, resulting in high computational expenses and lengthy development cycles. Modern adaptation methods leverage pre-trained LFMs to significantly streamline the fine-tuning process.

Technical considerations include:

- **Parameter-Efficient Fine-Tuning (PEFT)**: Techniques like Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA), previously detailed in Section 5.2, fine-tune fewer than 1% of model parameters. Studies (Hu et al., 2021; Dettmers et al., 2023) indicate these methods reduce computational requirements by approximately 90–95% compared to full model retraining, yielding substantial savings in GPU usage and memory consumption, thereby enabling quicker iteration cycles with large-scale models (e.g., LLaMA-4, Qwen-3).
- **Reduced Data Requirements**: Recent studies demonstrate that instruction-tuned LLMs can achieve strong performance with limited supervision, sometimes using as few as 500–1,000 high-quality examples when carefully curated (Zhou et al., 2023). This significantly shortens project timelines, accelerating deployments significantly.
- **Accelerated Deployment Cycles**: Reduced training and annotation overhead allow organizations to transition rapidly from prototyping to production, often shortening development from months to weeks. Incremental fine-tuning further enhances agility, enabling quick adaptations to changing business conditions without extensive retraining.
- **Automation of Data Preparation Pipelines**: Tools like the Data Prep Kit automate data ingestion, cleaning, schema alignment, and validation processes, cutting manual workload, improving data quality, consistency, and accelerating the overall adaptation timeline (Section 4.1).

6.2. Optimizing decision making through adapted models

Adapted LFMs significantly enhance enterprise decision-making by delivering more accurate, context-aware, and consistent outputs, reducing uncertainty and aligning closely with domain-specific objectives.

Key technical validations include:

- **Domain Specialization**: Training LFMs on curated, domain-specific datasets enhances accuracy and reduces hallucinations. For instance, in financial contexts, such tailored models decrease false positives in fraud detection, while healthcare applications improve clinical decision-making safety. Validation against established benchmarks, such as FinQA (financial reasoning), BioASQ (biomedical QA) or MIMIC-III (clinical data), quantitatively confirms these improvements.
- **Improved Task Performance**: Adapted LFMs consistently outperform general-purpose counterparts in specific applications (Chung et al., 2022; Fang et al., 2024). For instance, specialized fine-tuning improves precision, recall, and F1-scores in anomaly detection or predictive maintenance tasks. In manufacturing, integrated vision-LFM systems have notably increased defect detection accuracy, achieving improvements upwards of 20% (Brinkmann & Bizer, 2024).
- **Measurable Business Outcomes**: These technical improvements directly translate into measurable business KPIs. E-commerce applications, for example, report higher Average Order Value (AOV), improved Conversion Rate (CR), decreased Cart Abandonment Rate (CAR), and increased Click-Through Rate (CTR). In logistics and manufacturing, metrics like reduced defect rates, downtime, and improved SLA compliance highlight the tangible impact of adapted LFMs on operational effectiveness.

By systematically embedding domain knowledge and rigorously benchmarking performance against realistic scenarios, adapted LFMs significantly enhance enterprise workflows and decision quality.

6.3. Reducing operational and computational costs

Reliable model adaptation inherently optimizes resource use, significantly lowering computational demands for development training and deployment inference phases. Parameter-efficient fine-tuning methods substantially decrease infrastructure expenses, especially relevant for cloud-based deployments.

Moreover, increased efficiency in the adaptation process itself directly reduces operational overhead. Maintenance and incremental updates, facilitated by continuous learning strategies, eliminate the frequent need for full retraining, further lowering long-term costs.

Additionally, optimization techniques such as pruning and quantization further minimize computational load, enabling deployment even in resource-constrained environments like edge devices. The resultant cost savings broaden the feasibility of LFMs, allowing enterprises of varying scales to leverage advanced capabilities without prohibitive investment (Algomox, 2024).

6.4. Use cases and applications in different sectors

Reliable adaptation of LFMs unlocks targeted solutions across multiple industries, addressing specific challenges and seizing unique opportunities:

- Telecommunications: Customized LFMs can enhance customer support, automate incident management, and generate actionable insights from complex network data. In this sector, partners like CIC Consulting Informatico explore applications such as intelligent assistants for SOC/NOC operations, log analysis, and network diagnostics, demonstrating the potential of domain-adapted models in telecom environments.
- Healthcare: LFMs can support clinical documentation, decision-making, and patient interaction through reliable summarization and controlled generation. Dextro, as a partner operating in this domain, exemplifies the integration of LFMs in medical workflows, where explainability, compliance, and bias mitigation are critical factors.
- Manufacturing: Domain-adapted LFMs can assist in translating natural language requirements into simulation or design scripts, automate documentation, or support predictive maintenance tasks. Siemens, an industry partner in ELFMo, investigates such possibilities within engineering workflows to improve productivity and reliability.
- Engineering: In this sector, LFMs can assist in extracting and organizing technical information from documents, supporting maintenance planning, and improving asset data quality within enterprise resource planning (ERP) systems. FTP LDA, representing this domain, explores how adapted language models can streamline documentation processes and enhance the integration of structured equipment data into operational platforms.
- E-Commerce: LFMs can be leveraged to automate product categorization, personalize recommendations, and assist customers via multimodal shopping agents. Nosto Solutions, operating in this domain, illustrates the potential of integrating LLMs into digital commerce platforms to optimize both merchant operations and user experiences.
- Cybersecurity: In cybersecurity, LFMs can enable proactive threat detection, scam prevention, and context-aware alerting while respecting data privacy constraints. F-Secure, partner in this sector, exemplifies how lightweight, privacy-preserving LLMs can serve millions of end-users with minimal infrastructure overhead.

Documenting these practical implementations and outcomes across sectors will be vital to demonstrating and validating the transformative potential and measurable benefits of reliably adapted LFMs in enterprise contexts.

7 Conclusions

This deliverable has established a solid foundation for the reliable, efficient, and secure adaptation of Large Foundation Models (LFMs) in industrial environments. Through a comprehensive analysis of methodologies, tools, and representative use cases, it presents an integrated view of the state of the art and the technical innovations needed to integrate LFMs into complex enterprise systems.

The main contributions include:

- **Reliable and efficient model adaptation**: The document outlines parameter-efficient fine-tuning (PEFT) techniques, continuous learning strategies, and knowledge injection methods via Retrieval-Augmented Generation (RAG). These approaches aim to reduce computational costs while ensuring domain-specific relevance and adaptability.
- **Automation of data preparation**: It details pipelines for automated data collection, cleaning, and normalization—including filtering, semantic duplication, and instruction dataset construction—essential for accelerating model adaptation and improving data quality.
- **Continuous and incremental learning**: The report introduces advanced methods such as LoRA/QLoRA, hot-patching, and federated continual fine-tuning, enabling models to stay up to date without requiring full retraining cycles.
- **Comprehensive evaluation framework**: Beyond technical accuracy, the proposed framework emphasizes evaluation along multiple dimensions, including cost, latency, robustness to distributional shifts, and alignment with business-specific KPIs.
- **Cross-sector applicability**: The use cases from DEXTROMEDICA and CIC illustrate the potential of adapted LFMs to transform critical sectors such as healthcare and telecommunications, addressing challenges around traceability, explainability, and regulatory compliance.
- **Commitment to safety, ethics, and sustainability**: Mechanisms for validation and verification, bias mitigation, traceability, and legal compliance are designed in accordance with European regulations (e.g., the AI Act), promoting responsible adoption of generative AI technologies.

The technologies described herein span various levels of maturity, from experimentally validated components to functional prototypes currently under integration. Full-scale deployment will require ongoing validation with real-world data, scalability assessments, and alignment with the specific operational needs of each industrial use case.

Looking ahead, future work will focus on cross-partner validation of the proposed methods, integration of adapted models into operational platforms, and longitudinal assessment of their business impact. The project also aims to release reusable tools, curated datasets, and technical documentation to support wider adoption and long-term sustainability across the research and industrial communities.

8 References

Bai, B. 2022. Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. CoRR, 2022.

Brinkmann, B. (2024). Self-refinement strategies for LLM-based product attribute value extraction. arXiv. https://arxiv.org/abs/2501.01237. 2024.

Ethayarajh, E. 2022. Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with V-usable information. In ICML, volume 162, 2022.

Farajtabar, F. 2019. Mehrdad Farajtabar, Navid Azizan, Alex Mott, Ang L . Orthogonal Gradient Descent for Continual Learning. arXiv. https://arxiv.org/abs/1910.07104. 2019.

Hu, H. 2021. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

Hu, H. 2023. Hexiang Hu, Ozan Sener, Fei Sha, and Vladlen Koltun. Drinking from a firehose: Continual learning with web-scale natural language. IEEE Trans. Pattern Anal. Mach. Intell., 45(5), 2023.

Jang, J. 2022a. Joel Jang, Seonghyeon Ye, Sohee Yang, et al. Towards continual knowledge learning of language models. In ICLR, 2022.

Kirkpatrick, K. 2017. James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 2017.

Lester, L. 2021. Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691v2. 2021.

Lopez-Paz, L. 2017. David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In NeurIPS, 2017.

Mishra, M. 2021. Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi: Cross-Task Generalization via Natural Language Crowdsourcing Instructions, arXiv:2104.08773, 2021.

Mishra, M. 2021. Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi: Cross-Task Generalization via Natural Language Crowdsourcing Instructions, arXiv:2104.08773, 2021.

Razdaibiedina, R. 2023. Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. arXiv:2301.12314, 2023.

Razdaibiedina, R. 2023. Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. arXiv:2301.12314, 2023.

Schulman, S. 2017. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. CoRR, 2017.

Schulman, S. 2017. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. CoRR, 2017.

Wang, W. 2021. Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. In Findings of ACL, 2021.

Wang, W. 2022. Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar et al., Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, arXiv:2204.07705, 2022.

Wang, W. 2021. Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. In Findings of ACL, 2021.

Wang, W. 2022. Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar et al., Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, arXiv:2204.07705, 2022.

Wang, W. 2023) Xiao Wang, Yuansen Zhang, Tianze Chen, et al. Trace: A comprehensive benchmark for continual learning in large language models. CoRR, 2023.

Wenxiao, W. Model Compression and Efficient Inference for Large Language Models: A Survey. https://arxiv.org/abs/2402.09748. 2024.

Wu, W. 2024. Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari Continual Learning for Large Language Models: A Survey. arXiv:2402.01364v2, 2024.

9 Appendixes and case studies

A separate document has been generated so that all partners can generate content that applies only to their use case. This document is based on the use cases described in the Work Package 1 use case document.

Use case 1: DEXTROMEDICA Reliable Adaptation of LLMs for the Healthcare

To develop a framework for the safe and efficient application of Large-Scale Language Models (LLM) in the healthcare sector. The primary goal is to ensure the verification, validation (V&V), safety, and effectiveness of these models for their integration into critical industrial processes.

DEXTROMEDICA focuses its contribution to Task 2.1 of the ELFMo project on addressing the critical need for **reliable and domain-specific adaptation of Large Language Models (LLMs)** within the highly regulated healthcare environment. The use case revolves around the design and implementation of a **framework for the verification, validation, and trustworthy integration of LLMs** tailored to medical applications, ensuring compliance with ethical and legal standards such as the **EU AI Act** and GDPR.

Objectives

- Develop tools and methods for **systematic risk evaluation** and classification across different healthcare use cases involving LLMs.
- Build a modular, risk-aware **V&V framework** that enables the integration of specific models, datasets, and fine-tuning techniques, adapted to the **sensitivity and complexity of clinical environments**.
- Implement components that support **real-time monitoring, explainability, bias detection, and adversarial robustness**, ensuring trustworthy deployment in production healthcare systems.

Key Activities and Components

DEXTROMEDICA's implementation includes:

- **Risk Manager**: associates healthcare-specific use cases with potential risks during both training and inference stages.
- **Model Manager**: enables secure loading, fine-tuning, evaluation, and versioning of LLMs adapted for diagnosis support, patient communication, and medical documentation.

- **Test Manager**: automates validation via diverse testing strategies, including **metamorphic testing** and stress testing.
- **Bias and Fairness Modules**: based on recent academic work, these components integrate fairness metrics and mitigation strategies specifically targeting healthcare datasets.
- **Real-Time Monitoring Module**: includes alerting, data collection, and anomaly detection systems compatible with LLMOps practices.
- **XAI Module**: employs RAG (Retrieval-Augmented Generation) and vector databases to enhance interpretability and traceability of model outputs for clinical professionals.

Technological Stack

The use case leverages modern technologies and tools including:

- Python (FastAPI, Flask, SciPy, PyTorch, Hugging Face Transformers)
- Databases: PostgreSQL, MongoDB, LanceDB
- Container orchestration: Docker, Kubernetes
- Monitoring and visualization: Prometheus, Grafana, D3.js, Chart.js
- Fairness libraries: Fairlearn, AI Fairness 360

Strategic Contribution to WP2

This use case exemplifies **Innovation 2** of ELFMo by contributing a reproducible, domain-adapted toolkit for LLM adaptation in high-risk environments. It also aligns with **Innovation 3**, ensuring **evidence-based quality assurance** through robust V&V methods. DEXTROMEDICA also supports **open-source practices** and commits to feeding back validated modules and tools to the community where feasible.

Expected Impact

- **Faster and safer adoption of LLMs** in medical workflows.
- **Reduction of clinical risk and operational costs** through high-fidelity model validation.
- **Enhanced transparency and explainability**, fostering trust from healthcare providers and patients.
- **Contribution to European sovereignty in AI** through healthcare-specific technological innovation aligned with EU legislation.

Use case 2: CIC Consulting Informático Reliable Adaptation of LLMs for the Telecommunication

The main objective of this use case is to enable the secure, efficient and scalable integration of Large Language Models (LLMs) in enterprise environments of the telecommunications sector. In a context characterized by high operational criticality, massive data volume and increasing regulatory pressure, this use case seeks to create the methodological and technological foundations necessary to exploit the potential of LLMs while ensuring reliability, traceability, and regulatory compliance.

CIC focuses on three strategic lines: the customization of expert virtual assistants for operational and cybersecurity tasks, the synthetic generation of data to improve the training of telco domain-specific models, and the establishment of a continuous validation and monitoring framework to maintain the quality and security of the models during their entire lifecycle, ensuring alignment with ethical and legal requirements such as the EU AI Act and GDPR.

Objectives

- Domain adaptation of LLMs for telco operations: Leverage state-of-the-art models such as Mixtral, Llama-3 or Gemma and fine-tune them on telecom-specific data to enable more relevant and precise technical responses.
- Development of intelligent assistant systems: Build a modular architecture to deploy expert-driven conversational agents for key roles (e.g., cybersecurity analysts, network operators, and SGRwin users), enabling natural, effective user interactions.
- Synthetic data generation for training enhancement: Address the shortage of labeled telecom data by implementing secure, realistic data augmentation strategies that improve model generalization and fairness.
- Continuous validation and monitoring of deployed models: Implement tools and processes to track model behavior, detect anomalies, ensure performance stability, and verify compliance with ethical, legal, and operational standards.

Key Activities and Components

- Fine-Tuning Module: Responsible for adapting pretrained LLMs using domain-specific datasets and performance metrics. Enables specialization in areas like network operations, incident analysis, and platform-specific tasks.
- Assistant Creation Module: Enables the creation and routing of queries to expert-level assistants tailored to specific roles. Supports hybrid input processing (technical

questions vs. operational commands), breaking down complex instructions into execution graphs when needed.

- Data Generation Module: Implements telecom-specific synthetic data pipelines. Includes a Data Augmentation System and Dataset Management Tool, using tools like PyTorch, NLTK, and SMOTE to ensure diversity, privacy, and traceability.
- Monitoring and Validation System: Integrates MLflow, Grafana, and automated testing pipelines (e.g., robustness and stress testing). Enables real-time monitoring, lifecycle tracking, and alerting, aligned with LLMOps best practices.
- LLM Integration & Service Exposure: Orchestrates LLMs using containerized environments (Docker/Kubernetes) and exposes assistant functionalities via secure RESTful APIs, ensuring interoperability and seamless deployment within telco infrastructure.

Technological Stack

- Modeling & Training: PyTorch, Hugging Face Transformers, Fairlearn
- Data: NLTK, DVC, PostgreSQL
- Orchestration: Docker, Kubernetes
- Monitoring: Timescale, Prometheus, Grafana
- Service Exposure: FastAPI, Flask, REST

Strategic Contribution to WP2

This use provides a concrete contribution to the responsible deployment of LLMs in complex, high-stakes industrial environments. By focusing on domain-specific adaptation, continuous validation, and operational integration, it establishes a practical and scalable framework for applying generative AI in telecommunications. Through modular components, traceable processes, and alignment with open and auditable technologies, the use case promotes both technical robustness and regulatory trust. CIC also reinforces its commitment to collaboration by designing solutions that can be reused, extended, and shared across sectors and communities.

Expected Impact

- Enhanced user experience through intelligent assistants capable of natural conversations and precise technical guidance in telco operations.
- Smart automation of tasks such as network troubleshooting, fault detection, and command execution within platforms like SGRwin, reducing operational overhead.

- Improved resilience and adaptability through synthetic data-driven training and continuous monitoring, ensuring high model accuracy and regulatory compliance.
- Alignment with European standards and regulations such as GDPR and the AI Act, ensuring lawful, transparent, and auditable AI integration.
- Support for European digital sovereignty, by fostering secure, open, and responsible GenAI adoption in critical industries through reusable frameworks and open collaboration.