Engineering Large Foundational Models for Enterprise Integration

Deliverable D3.2
Risk, quality and conformity assessment methods, risk indicators and quality metrics

Project title	Engineering Large Foundational Models for Enterprise Integration
Project acronym	ELFMo
Project number	23004
Work package	WP3
Deliverable	D3.2
Dissemination level	PU (public)
License	CC-BY 4.0
Version	1.0
Date	2025-10-01

Contributors

Editor(s)	Juhani Kivimäki (University of Helsinki)
Reviewer(s)	Marcos Cobo Carrillo (CIC), Diogo Emanuel Martinho (ISEP), Patrícia Alves (ISEP)
Contributor(s)	Afnan Baig (University of Helsinki), Robin Bornoff (Siemens), Igor Casado Moreno (Konecta), Isabel Ribeiro (FTP), Tomi Sarni (Nosto), Davor Stjelja (Granlund), Andrea Vianello (Siili)

Abstract

This document describes the initial version of the assessment methods, risk indicators and quality metrics. The methods, indicators and metrics covered in this document are Human-in-the-loop, Business Key Performance Indicators (KPIs) in telemarketing, Autonomous LLM Evaluation and Remediation Framework, Measuring internal business KPIs, Measuring Service Level Objectives, Evaluation Framework Comparison, Trustworthiness, and Large Foundation Models (LFM) for Enterprise Resource Planning (ERP). The methods, indicators and metrics will be tested in the use cases of the ELFMo project. The implementation of these methods, indicators and metrics will be detailed later in another deliverable, and refined and extended, possibly adding new methods, indicators and metrics in another deliverable.

Table of contents

1 Introduction

1.1 Role of this Document

The purpose of this document is to describe the initial versions of the risk, quality and conformity assessment methods, risk indicators and quality metrics developed within the ELFMO project. The document focuses on risk, quality and conformity assessment in parallel with the model training and benchmarking -focused deliverable "D2.2 Initial release of benchmarking techniques". The methods and techniques will be tested in the use cases of the ELFMo project. The implementation of the assessment methods, risk indicators and quality metrics will be detailed later in deliverable "D3.3 The ELFMO risk management, monitoring and conformity assessment tools and dashboards V1", and the assessment methods, risk indicators and quality metrics will be refined and extended, possibly adding new methods, indicators and metrics, in "D3.4 The ELFMO risk management, monitoring and conformity assessment tools and dashboards V2".

1.2 Intended Audience

The intended audience of the present document is composed primarily of the ELFMo consortium for the purpose of understanding methods and techniques and advancing LFM risk management, monitoring and conformity assessment practices. However, this document is public and can provide an overview of the advances in the ELFMo project. This document describes methods and technologies for the technically oriented audience rather than the general public.

1.3 Definitions and Interpretations

The terms used in this document have the same meaning as in the contractual documents referred to in [FPP] with Annexes and [PCA] unless explicitly stated otherwise.

1.4 Related Documents

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

- [FPP] *ELFMo – Full Project Proposal 23004* describes the full project proposal
- [PCA] *ELFMo Project Consortium Agreement* outlines the common agreement between project participants
- [D3.1] *Research baseline for risk, quality and conformity assessment tools and procedures* describes the baseline to be advanced.

2 Human-in-the-loop

Title	Human-in-the-Loop -Driven, Knowledge-Enhanced LFM Workflow
Description	We are developing a Human-in-the-Loop (HITL) AI workflow for building design and consultancy, supported by a comprehensive Proprietary Knowledge Layer that integrates drawings, schematics, design documents, and time-series data. This approach ensures AI output is transparent, trustworthy, and domain-aware while preserving expert control. Security, oversight, and UX are core pillars, alongside training human experts to effectively guide and review AI assistance.
Corresponding contact	Davor Stjelja
Contributors	Davor Stjelja
Life-cycle stage	Exploratory PoC / TRL 4-5
EFLMo innovations	2 innovations, Human-in-the-Loop governance layer and Proprietary Knowledge Layer for multimodal building data
Technological environment	Hybrid edge + cloud; Azure OpenAI & on-prem GPUs; OSS frameworks (LangChain...) with proprietary integrations
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☒ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☒ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☒ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Mixed: methodology & templates CC-BY; code and data layer components proprietary

2.1 What is it?

We are currently on the exploratory stage, setting foundations for integrating Human-in-the-Loop (HITL) methodologies into Granlund's workflows. Initial proof-of-concepts have already demonstrated the essential role of HITL for effectively deploying LLMs in Granlund's domain-specific tasks. These early explorations indicate clear benefits when embedding human expertise centrally in the AI-assisted processes. Additionally, significant preliminary work has

started on developing a comprehensive Proprietary Knowledge Layer to ensure AI applications fully leverage Granlund's extensive operational datasets and domain expertise.

2.2 Why have we done this?

Our motivation comes from the observed limitations of LLMs regarding accuracy and reliability in complex, domain-specific tasks. Typically, LLM performance remains below that of human experts, especially in specialized areas critical to Granlund's operations. The primary goal is to leverage AI to handle routine, repetitive tasks, thereby enhancing human productivity and job satisfaction. Crucially, the human expert retains full oversight and transparency over AI-generated outputs, including decision rationales and information provenance. The Proprietary Knowledge Layer further supports this goal by enhancing the accuracy, context-awareness, and reliability of AI outputs, ensuring alignment with Granlund's operational standards and requirements.

Security considerations are also important for HITL systems, as complete reliance on human responses introduces potential vulnerabilities. Incorrect or maliciously injected human inputs could propagate through the system, negatively impacting future AI-generated responses. Thus, additional safeguards and verification mecha-nisms are necessary to mitigate these risks.

2.3 How it works?

Currently, we are actively assessing various technological strategies and platforms, ranging from fully open-source solutions to commercial services, including internal developments. Our exploration covers diverse deployment scenarios such as hosting proprietary LLMs on-premises or using cloud services like Azure OpenAI.

A major component of our current focus is developing a comprehensive Proprietary Knowledge Layer. This layer is designed to integrate and process the broad variety of data Granlund works with, including technical drawings, schematics, detailed design documents, and extensive time-series datasets. This knowledge layer significantly enriches the context-awareness, precision, and relevance of AI recommendations, fully leveraging Granlund's unique domain knowledge and extensive operational datasets.

Development frameworks such as LangChain, LangSmith, and Langfuse are being evaluated to accelerate solution building and monitoring.

Education of human experts is also a critical component. We are planning targeted training programs and interactive workshops to ensure human experts can effectively collaborate

with AI systems, interpret AI-generated outputs accurately, and apply best practices for secure and productive human-AI interactions.

Significant attention is also placed on UX/UI design, as the interface between human experts and AI systems is critical for seamless collaboration, transparency, and usability. Our final implementation aims for an intuitive, efficient interface that simplifies human oversight while maximizing the AI's productivity benefits.

2.4 Further Reading

For deeper dive into HITL governance, knowledge-layer architectures and user-centered AI design, the resources below provide practical guidance and real-world examples:

- **IAPP / Marsh – "Human-in-the-loop in AI risk management — not a cure-all approach."** Discusses why Article 14 of the EU AI Act mandates human oversight and explains that simply adding a human checkpoint is insufficient without clear processes and the right expertise. https://www.marsh.com/en/services/cyber-risk/insights/human-in-the-loop-in-ai-risk-management-not-a-cure-all-approach.html
- **Nature Scientific Reports (2025) – "Human-in-the-loop approach to improve safety and robustness of large foundation models."** Presents empirical evidence that combining human review with automated monitoring significantly reduces harmful outputs.
  https://www.nature.com/articles/s41598-025-92889-7
  IBM Think Blog – "A practical framework for AI risk management." Offers step-by-step guidance for identifying, assessing and mitigating technical and organizational risks throughout the AI lifecycle.https://www.ibm.com/think/insights/ai-risk-management
- **Google Cloud – "Human-in-the-Loop."** Explains Google's managed HITL service and design patterns for routing low-confidence predictions to human reviewers. https://cloud.google.com/discover/human-in-the-loop
- **ITResearches – "Designing seamless human-AI interfaces: principles for success."** Summarizes UX heuristics that keep experts in control while benefitting from automation.
  https://itresearches.com/designing-seamless-human-ai-interfaces-principles-for-success/
- **Microsoft Copilot UX Guidance for ISVs.** Provides concrete UI patterns, content moderation tips and accessibility considerations for generative-AI applications. https://learn.microsoft.com/en-us/microsoft-cloud/dev/copilot/isv/ux-guidance

- **Outwitly – "AI for UX Design."** A designer-oriented view on integrating AI tools into product research and wire-framing workflows. https://outwitly.com/blog/ai-for-ux-design/
- **Permit.io – "Human-in-the-loop for AI agents: best practices, frameworks & demo."** Describes enforcement hooks and policy engines that ensure human approval before agents commit high-stakes actions. https://www.permit.io/blog/human-in-the-loop-for-ai-agents-best-practices-frameworks-use-cases-and-demo

3 Business KPIs in telemarketing

Title	Business KPIs in telemarketing
Description	The objective is to identify the basic KPIs by type of service/conversation/sector. The set of basic KPIs that make up the operation is analyzed, taking into account the sector and establishing those that are most significant.
Corresponding contact	Igor Casado
Contributors	Konecta BTO
Life-cycle stage	Development
EFLMo innovations	AI Governance and Regulatory Compliance and System Adaptation and Scalability Infrastructure for LFM
Technological environment	Cloud
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☒ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☐ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☒ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Proprietary

3.1 What is it?

Key performance indicators (KPIs) have been defined based on the specifications of the use cases associated with telemarketing. These requirements and KPIs will guide the evaluation and validation of the LLM-augmented system.

3.2 Why have we done this?

In BPO/contact center environments, KPIs make it possible to establish performance and determine strategies for improvement or evolution. These indicators are composed of variables that come from daily operations and activities, but do not include any vocal variables that categorize the service provided.

The objective is to identify the basic KPIs by type of service/conversation/sector. The set of basic KPIs that make up the operation is analyzed, taking into account the sector and establishing those that are most significant.

3.3 How it works?

Indicators have been defined based on use case specifications. The impact of the most representative indicators within the call-center sales process has been analyzed. The following metrics have been defined for telecontact campaigns:

KPI	Description
FCR	First Contact Resolution, First Call Resolution, resolution on the first call or resolution on the first contact: FCR measures the percentage of cases or incidents that are resolved on the first call. It is undoubtedly one of the two most relevant inbound indicators, which are measured in every call center, because it tells us about two key aspects: the quality of customer service and the cost of the operation.
TMO	Average Handling Time (AHT) measures the average time spent on custo-mer interactions and, together with FCR, is probably the most important inbound indicator for a contact center.
NPS	Net Promoter Score measures customer loyalty and predicts the likelihood of them recommending you. To do this, the NPS requires asking the customer directly about their intention: How likely are you to recommend [product, service, brand] to a friend?
CES	Customer Effort Score measures the effort made by a user to resolve their query or incident during a call with an agent: How easy do you make it for your customers. It is measured through satis-faction surveys, including questions at the end of the communication such as: “To what extent do you agree or disagree with the following statement: It was easy to resolve my problem.”
CRR	Customer Retention Rate measures the percentage of customers that the company has retained; that is, how many customers have remained loyal over a given period of time. To measure the retention rate, you need to take three factors into account: 1. Number of customers at the beginning of the period (x) 2. Number of customers at the end of the period (y) 3. Number of new customers acquired during that period (z)
	Conversion Rate measures the percentage of calls in which the objective has been achieved.

	Number of records (leads) consumed per sale: This is the main factor determining cost differentials per sale between different agents.
Churn	Customer Churn Rate, CRR, churn rate, cancellation rate, or termination rate: Customer churn measures the percentage of users who cancel a contract or stop using the company's services during a given period. It could be said that it is the opposite metric to the retention rate.
SL	Service level measures the percentage of calls answered within a given time period (typically 20 seconds) and is used to determine whether agents are moving quickly enough from one call to the next. Service level is one of those KPIs that is more than essential for productivity because, among other things, it helps to correctly size the call center.
CAR	Call Abandon Rate, Abandon Rate (AR): Abandonment rate measures the percentage of abandoned calls, i.e., when the customer or consumer hangs up before speaking to an agent.
CSAT	Customer satisfaction level, through a survey asking them how satisfied they are with the interaction they have just had with the agent: The possible answers are organized on a scale of 1 to 5, where 1 is “not at all satisfied” and 5 is “very satisfied.”
AT	Absenteeism rate: This has a direct impact on costs and an indirect impact on other KPIs related to productivity and customer experience, as well as negatively affecting the working environment and the well-being of other agents.

The main KPIs are used to measure the success of the campaign's operational improvement. The current challenge we face once this improvement has been achieved is to maintain it over time, which is why it was essential to create a monitoring dashboard that allows the operation to detect any deterioration in optimization and improvements to be made.

Our goal must be to create a management model based on a set of indicators that, while relevant to the improvement of the operation, are easily understandable and usable (operationalizable) by the Contact Center operation.

Here we quickly encounter the reality of the operation, for which it is more useful to have different indicators for “what to say” and “how to say it.” In other words, the KPIs for “how to speak” and “what to speak,” or form of speech and content of speech, must be differentiated and shown as such in order to be able to recommend effective actions to agents.

It is more than reasonable to think that there is a direct interrelationship between the vocal variables defined throughout the project and the success of the operation, so statistical prediction work will be necessary to generate a new way of interpreting the ratios.

4 Autonomous LLM Evaluation and Remediation Framework

Title	Autonomous LLM Evaluation and Remediation Framework
Description	A modular framework that continuously monitors large language model (LLM), RAG Specific pipelines, detects failures (low relevancy, hallucinations, semantic drift, poor recall), and triggers automated corrective actions. It integrates retriever/LLM metrics with a decision engine for severity-aware, cost-efficient remediation. The system combines open-source tools (LangChain, ChromaDB, NLI models) with dashboards (Streamlit, Grafana) for real-time observability, aiming to improve reliability, faithfulness, and trust in AI outputs.
Corresponding contact	Afnan Baig, Mikko Raatikainen, Jukka Nurminen
Contributors	Afnan Baig
Life-cycle stage	Prototype Designing and Early Implementation
EFLMo innovations	3
Technological environment	Local hardware, Flexible setup allows edge/cloud deployments.
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☒ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☒ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☒ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Proprietary

4.1 What is it?

We have designed and implemented a modular monitoring and remediation framework for Large Language Models (LLMs), with a special focus on retrieval-augmented gene-ration (RAG) and prompt engineering pipelines. The framework continuously monitors model behavior during inference, computes a comprehensive set of diagnostic metrics (such as relevancy, faithfulness, hallucination, and semantic drift), and maps these failures to automated or semi-automated corrective actions. A decision engine evaluates severity, cost, and benefit to determine the best intervention. The prototype integrates LangChain,

ChromaDB, NLI/fact-checking models, and Streamlit demon-strating real-time observability and governance for LLM pipelines.

4.2 Why have we done this?

Current evaluation frameworks like HELM and LM Harness focus on offline bench-marking, not real-time monitoring. In production, RAG and prompting introduce hidden dependencies, hallucinations, and semantic misalignment. Failures are often unde-tected, and even when detected, most systems only report issues and do not provide autonomous remediation. This creates trust, safety, and operational risks for enterprises.

Production LLM systems should have continuous observability and not only detecting issues but also diagnosing why they occur and taking closed-loop corrective actions (e.g., re-ranking, context compression, OOD flagging). This is critical for domains like legal, medical, or finance, where errors carry high risks.

By bridging the gap between monitoring and remediation, this framework provides a path toward trustworthy, resilient, and cost-efficient deployment of LLMs. It enables enterprises to reduce hallucinations, manage brittleness in prompts and retrieval, improve grounding, and maintain user trust through transparent and auditable inter-ventions.

4.3 How it works?

The framework is structured around four main layers. The **Metric Monitoring Layer** continuously evaluates both prompt-level and RAG-level metrics to detect failures during inference. At the prompt level, the system measures out-of-distribution (OOD) detection to identify whether a user query falls outside the model's domain, prompt alignment to ensure that inputs are clear and relevant, and prompt-response consistency to check that generated outputs logically follow from the given prompt. At the RAG level, the system monitors answer relevancy to assess how directly responses address user questions, context relevancy to verify whether retrieved documents are appropriate to the query, and faithfulness to ensure that outputs remain grounded in retrieved context. Additional metrics include hallucination detection to flag unsupported claims, retriever precision and recall quantifying retrieval quality, and semantic drift to detect subtle meaning shifts between retrieved knowledge and generated answers. Together, these metrics provide a comprehensive, multi-layered perspective on both retrieval and prompt performance.

The framework also integrates the **DeepEval evaluation library**, which provides a modular suite of metrics specifically designed for testing and benchmarking LLM pipelines. DeepEval supports both retrieval-augmented generation (RAG) and prompt-response evaluation by offering ready-to-use metrics. By incorporating DeepEval into the monitoring layer, the

system gains a standardized and reproducible way to compute diagnostic signals during inference, enabling fair comparisons across different retrievers, prompts, and models. This not only reduces implementation overhead but also ensures that the evaluation process is aligned with industry practices for reproducibility, transparency, and continuous improvement.

The **Decision Engine** transforms metric observations into corrective actions. When a monitored metric falls below a threshold (for example, low answer relevancy or high semantic drift) the engine evaluates a set of candidate actions. These actions are prioritized based on expected benefit, computational cost, and severity of the detected failure. Light interventions, such as re-ranking retrieved documents, are applied for mild deviations, while heavier interventions, such as corrective retrieval or abstention, are reserved for severe cases. Light interventions are immediately actionable with existing tooling and can be deployed in pipelines while heavier interventions are technically feasible but may introduce latency and additional costs that projects or organizations need to consider. More advanced learning-based policies like reinforcement learning and multi-armed bandit strategies, remain largely experimental and would require significant investment in offline evaluation, feedback collection, and retraining infrastructure before adoption at scale. From an LLM-agent perspective, the engine functions as a governance layer that autonomously selects corrective measures, compares outcomes against evaluation criteria, and updates policies through continuous logging, A/B testing, and threshold refinement.

The **Remediation Executor** is responsible for carrying out corrective actions once the decision engine identifies a failure. Depending on the nature of the issue, it can apply interventions such as context compression to reduce noise in retrieved documents, semantic splitting to refine document granularity, re-ranking to prioritize more relevant context, or generation-level strategies such as contrastive decoding and controlled generation to enforce factual grounding. In cases where outputs involve high-risk or ambiguous content (for example, in legal, medical, or financial domains) the executor does not act autonomously but instead escalates the case for human review. This ensures that sensitive scenarios are handled with an additional layer of oversight, balancing automation with safety.

The **Monitoring Dashboard** provides real-time observability of the pipeline. During development, lightweight interfaces are implemented using Streamlit, while in industrial production environments, more robust solutions such as Grafana and Prometheus could enable continuous monitoring and alerting. Whenever anomalies are detected, alerts are triggered, and system behavior is logged in a structured format. Logs can be captured as JSON files during rapid prototyping and for enterprise level analysis scalable databases such as Postgres and modern cloud-based warehouses could be used. The monitoring system

integrates seamlessly with RAG pipelines built on LangChain and ChromaDB, allowing it to be adopted with minimal engineering effort and ensuring compatibility across research and production settings.

Deploying a fully closed-loop engine in enterprise settings would require scalable retriever indexing, dependable fact-checking services, and production-grade observability tools like Grafana or Prometheus. These technologies are already available, but they introduce additional costs and integration challenges. By recognizing these trade-offs, the framework highlights what can be adopted today while also pointing toward longer-term innovations that remain under development.

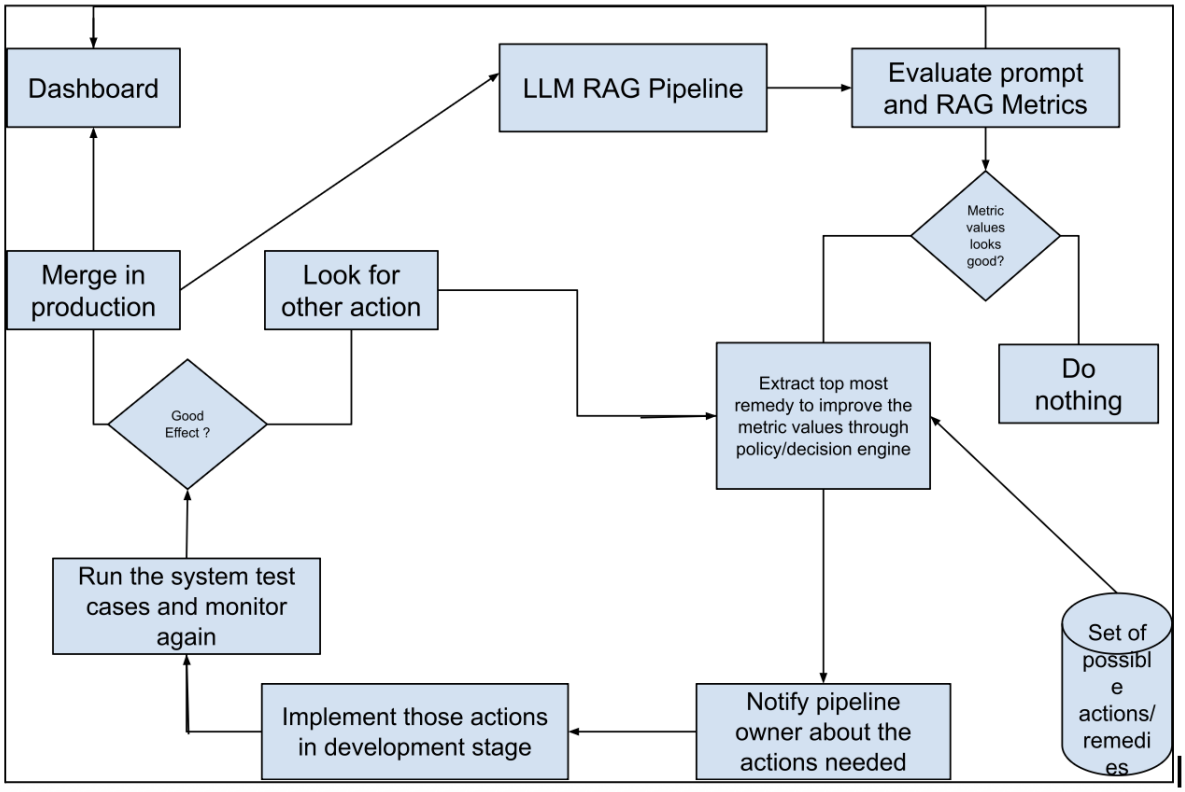

Figure 1. Overview of the modular LLM monitoring framework with metric monitoring, decision engine, and remediation layers.

4.4 Further Reading

- DeepEval framework: https://github.com/confident-ai/deepeval
- RAGAS framework: https://github.com/explodinggradients/ragas
- Chase, H. (2023). LangChain: Building applications with LLMs through composability. GitHub. https://github.com/langchain-ai/langchain

- Chroma (2023). The AI-native open-source embedding database. GitHub. https://github.com/chroma-core/chroma
- Streamlit Inc. (2023). Streamlit: The fastest way to build data apps in Python. https://streamlit.io/
- OpenAI (2024). GPT-4o Technical Report. https://openai.com/research
- PostgreSQL Global Development Group. (2023). PostgreSQL: The world's most advanced open-source relational database. https://www.postgresql.org/
- Grafana Labs (2023). Grafana: The open-source observability platform. https://grafana.com/
- Prometheus Authors (2023). Prometheus: Monitoring system and time series database. https://prometheus.io/
- Blog: RAG Failure Points and Optimization Strategies – A Deep Dive (Medium) https://medium.com/@ajayverma23/rag-failure-points-and-optimization-strategies-a-deep-dive-b39ceb7d11c5

5 Measuring internal business KPIs

Title	Feature Adoption
Description	By measuring weekly or monthly active users engaging with the LFM features such as specialized agents through conversational interface is what ultimately measures the value of what a feature brings to the customer. This is more of an internal business KPI and not a survival level objective measuring the quality of service towards customers. ITARS framework (Target, Adopted, Retained, Satisfied) as the measurement where we would measure active adopted users / active target users > 50%.
Corresponding contact	Tomi Sarni
Contributors	Anatoly Soldatov
Life-cycle stage	Proof of concept
EFLMo innovations	3 (evidence-based quality & compliance assessment), 4 frameworks and methodology
Technological environment	This can be deployed to production environments utilizing open source or proprietary tools to log user interactions within an environment where the LLM is present to measure how many conversational interactions with an agent there were.
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☐ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☐ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☐ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Proprietary

5.1 What is it?

We have identified that a qualitative measure of measuring the business impact of a LFM would be feature adaptation. We have a basic plan on how to measure and assess the results.

5.2 Why have we done this?

We are already monitoring this for other features; therefore, a comparison could help us contextualize the success metrics of LFM based features and compare their impact to other features.

5.3 How it works?

We can measure active adopted users / active target users on a daily, weekly and monthly basis to see whether there are measurable trends and can those be connected to the changes made to the underlying LFM based feature. We can also analyze the users deeper in terms of the impact on the business. For instance, we can have cohorts for churned customers vs non-churned customers and their adaptation of the features.

5.4 Further Reading

TARS — How to execute and evaluate a feature strategy.
https://medium.com/@niklas2106_71245/tars-how-to-execute-and-evaluate-a-feature-strategy-f7a965cc1fb9

6 Measuring Service Level Objectives

Title	SLOs for Conversational Agent
Description	Measuring the acceptable or agreed service level for interaction. This is not measuring the quality of the answers, but rather that a response is delivered within a certain time. For instance, SLO could be defined that p99 of user prompts should be responded within XX seconds. And p99 of user interactions should result in agent response.
Corresponding contact	Tomi Sarni
Contributors	Anatoly Soldatov
Life-cycle stage	Proof of concept
EFLMo innovations	3 (evidence-based quality & compliance assessment)
Technological environment	It should be implemented in a real production system and connected to the logging and monitoring stack.
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☐ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☐ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☐ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Proprietary/MIT/CC-BY/What

6.1 What is it?

We have identified the need to monitor performance of Agentic interactions similarly to standard practices for monitoring the performance of customer facing services.

6.2 Why have we done this?

We have identified the need to provide constant feedback to the customer engaging with an agentic interface. We have also noticed that in some cases where a system error has surfaced to a user and we have identified the need to monitor the service level and set service level objectives. SLOs have been used.

6.3 How it works?

Instrumenting the interactions as metrics data to any open source or proprietary logging and monitoring stack.

6.4 Further Reading

Google introduced the concept of the Service Level Objective (SLO) to formalize a quality-oriented culture that emphasizes reliability and acceptable service behavior. In contrast to frameworks like ITIL, which prioritize predefined processes to prevent incidents, the Site Reliability Engineering (SRE) approach embraces SLOs and error budgets to ensure that monitoring is aligned with what truly matters to service consumers. This shift in focus promotes responsiveness to actual service experience rather than rigid process adherence (Beyer et al., 2016).

In the context of large foundational models, SLOs can serve as meaningful indicators of the usability and responsiveness of deployed systems. While a continuous integration and deployment (CI/CD) pipeline might track metrics such as accuracy or F1-score on a holdout dataset, an SLO might instead define a latency objective, such as requiring that 99.9% of inference requests be completed within 5 seconds. Under this objective, if a system handles 1,000,000 requests in a 30-day period, a violation will occur if more than 1,000 requests exceed the 5-second threshold. This remaining tolerance—known as the error budget—provides a quantitative margin for controlled degradation and informs operational decision-making (Beyer et al., 2016).

SLOs can be defined based on any measurable indicator that meaningfully reflects the quality-of-service delivery. Importantly, these indicators should align with the end-user's experience and expectations. For example, in the context of machine learning, an SLO could be based on user-perceived model performance rather than traditional validation metrics. In the case of a Large Language Model (LLM)-powered Shopping Assistant, a simple post-interaction feedback mechanism—such as a thumbs-up or thumbs-down rating—could be used to classify events as successful or failed. An SLO could then be specified to require that a certain percentage of interactions (e.g., 95%) receive positive feedback within a defined time window (e.g., 30 days), thereby directly tying system objectives to observed user satisfaction (Wang et al. 2024).

Langfuse. (2024, July). AI agent observability with Langfuse.
https://langfuse.com/blog/2024-07-ai-agent-observability-with-langfuse

7 Evaluation Framework Comparison

Title	Evaluation Framework for Quality-Aware Assessment of LFM-Based Services
Description	We compared multiple evaluation frameworks for assessing the performance and trustworthiness of LFM-based applications. Our study covered traditional NLP metrics, retrieval-based methods, uncertainty estimation, attention analysis, semantic similarity, and LLM-as-a-judge approaches. This comparison identified strengths, limitations, and complementarities, and recommended '**LLM-as-a-Judge**' for our industrial use cases.
Corresponding contact	Robin Bornoff (Siemens)
Contributors	Robin Bornoff, Kefan Sun (Siemens)
Life-cycle stage	Prototype / Early validation
EFLMo innovations	3 (evidence-based quality & compliance assessment)
Technological environment	Hybrid OSS/proprietary evaluation frameworks (RAGAS, Siemens internal pipelines) integrated with LFM-based services.
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☒ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☒ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☐ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Internal prototype (with use of open-source components under MIT/Apache licenses).

7.1 What is it?

Siemens evaluated a broad range of methods for assessing the performance and reliability of LFM-enabled applications. These included traditional NLP metrics (BLEU, ROUGE), retrieval-based measures, semantic similarity scoring, uncertainty estimation, attention-weight analysis, and the use of large language models themselves as evaluators ("LLM-as-a-Judge"). After systematic comparison of these approaches against industrial needs, we selected **LLM-as-a-Judge** as the preferred evaluation method for further development and integration into our workflows.

7.2 Why have we done this?

Enterprises adopting LFMs need evaluation methods that are not only technically sound but also practical, explainable, and aligned with industrial quality requirements. Traditional automatic metrics provide speed but fail to capture semantic nuance; retrieval, attention, and uncertainty methods are limited by data availability or scope, and semantic similarity scoring lacks explainability. In contrast, **LLM-as-a-Judge** is flexible, can assess multiple quality dimensions (correctness, reasoning, grounding), and provides evaluations that align more closely with human judgement. Despite its computational cost, we judged it to be the most suitable choice for ensuring trustworthiness and accountability in enterprise deployments. This aligns with ELFMo's goal of enabling **evidence-based quality assurance and conformity assessment**.

7.3 How it works?

The selected approach uses advanced LLMs as evaluators. In practice, generated outputs (together with their context and reference answers) are passed to a dedicated evaluation LLM. This model scores or categorizes responses according to predefined dimensions such as factual correctness, faithfulness to source documents, etc. The setup supports both automated batch evaluation (for regression testing) and targeted assessment (for high-value use cases).

Approach	Strengths	Weaknesses	Example
Traditional NLP metrics	Fast, standardized	Surface-level only	ROUGE/BLEU
Retrieval-based	Measures grounding	Needs curated gold datasets	MRR
Uncertainty-based	Confidence estimation	Limited scope (faithfulness)	Log-probs
Attention-based	Explainable focus analysis	Hard to interpret automatically	RAGViz
Semantic similarity	Captures semantic nuance	Score not explainable	Cosine sims
LLM-as-a-judge	Flexible, covers many aspects. Explainable score	Costly	RAGAS/DeepEval

Our experimentation included open-source frameworks such as **RAGAS** and **DeepEval**, which provide integration scaffolding, test definitions, and CI/CD compatibility. We confirmed that LLM-as-a-Judge can be embedded into continuous evaluation pipelines, supporting industrial

requirements for monitoring quality and detecting drifts. This approach is now the Siemens-selected baseline for ELFMo evaluation activities.

7.4 Further Reading

The BLEU metric was introduced in Papineni et al. (2002). Retrieval-Based Evaluation using metrics designed for information retrieval was examined in Järvelin & Kekäläinen (2002). Uncertainty based metrics for NLP tasks were considered as early as 1977 in Jelinek et al. (1977). Measuring semantic similarity with embedding-based metrics was introduced in Zhang et al (2019). The LLM-as-a-Judge approach has been considered for example by Jiang et al. (2025). Further resources are listed below:

- RAGAS framework: https://github.com/explodinggradients/ragas
- DeepEval framework: https://github.com/confident-ai/deepeval
- Attention-Based Evaluation / Interpretability: https://en.wikipedia.org/wiki/Attention_Is_All_You_Need

8 Trustworthiness

Title	A framework for designing and developing trustworthy AI and data solutions.
Description	We are defining a framework for designing and developing trustworthy AI and data solutions that deliver value and address real business challenges. The process starts by identifying the trustworthiness factors most critical to each AI opportunity. For these factors, risks are anticipated and translated into guardrails with accountability, KPIs, and review cycles. Guardrails are then synthesized into governance policies, enabling organizations to anticipate blockers, ensure compliance, build trust, and scale AI responsibly across use cases.
Corresponding contact	Andrea Vianello
Contributors	Andrea Vianello, Caroline Liu
Life-cycle stage	Research and definition
EFLMo innovations	• 1 (A risk-based approach to informed decision making for the rapid integration of LFMs into one's business environment) • 3 (Evidence-based procedures for quality and compliance assessment for LFM-based applications and services)
Technological environment	None (the innovation includes practical tools to be used during workshop and co-creation sessions)
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☒ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☐ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☒ Enables fair and inclusive access to LFM technologies (G10: T2) ☒ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☐ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	Proprietary

8.1 What is it?

We are defining a framework to design and develop trustworthy AI and data solutions. It aims at supporting our customers in identifying opportunities where AI can address real business challenges and deliver tangible value, considering trust-building and risk factors since the

beginning. To support this, we have conducted a brief literature review to identify key trustworthiness factors for Large Foundational Models and developed practical tools for workshop and co-creation sessions. In the future, we will tailor these tools to specific business domains to ensure relevance and applicability.

8.2 Why have we done this?

Despite the benefits that AI can provide to business, many projects – especially those involving GenAI – fail to progress beyond the proof-of-concept stage. Common causes include unclear or inadequate business value, poor data quality, or insufficient risk controls (Gartner, 2024; IBM, 2025).

The goal of our framework is to provide our customers with a structured and effective approach to identify where AI can provide real business value, while addressing factors that can build trust, encourage AI adoption, and anticipate possible blockers before major investments.

In sum, this could offer our customers a low-barrier path to adopting trustworthy AI in their business processes while maximizing its potential benefits.

8.3 How it works?

The current draft of the framework includes the following steps and tools:

- For each envisioned AI opportunity, the **Trustworthiness Canvas** helps identify the trust factors most critical to the business case, weighing positive outcomes against negative impacts.
- For each key trustworthiness factor, the **Risk Card** helps anticipate risks, along with their likelihood and impact.
- For each risk, the **Guardrail Card** turns initial mitigation ideas into concrete safeguards, adding accountability, success KPIs, and review cycle.
- The **Governance Policy Card** synthetizes multiple guardrails into organization-wide governance policies, defining success KPIs, ownership, and review cycle.

8.4 Further Reading

- Gartner, 2024. Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025. Retrieved from: https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025

- IBM, 2025. The 5 biggest AI adoption challenges for 2025. Retrieved from: https://www.ibm.com/think/insights/ai-adoption-challenges

9 LFM for ERP

Title	Intelligent and Modular ERP with LFM Integration
Description	Modern ERP systems are migrated from monolithic architectures to modular and microservices-based models, enhanced with Large Foundation Models (LFMs. The contribution provides conversational interfaces, predictive analytics, and automation tools for industrial and e-commerce enterprises. It ensures compliance with GDPR and the AI Act, while delivering scalable, explainable, and secure ERP capabilities.
Corresponding contact	Isabel Ribeiro
Contributors	Isabel Ribeiro, André, Diogo Martinho, Patrícia Alves
Life-cycle stage	Prototype Definition
EFLMo innovations	- Innovation 2: Tools and infrastructures for trustworthy adaptation and integration of LFMs into ERP workflows. - Innovation 3: Evidence-based procedures for transparency, GDPR compliance, and explainability in AI-driven business processes. - Innovation 4: Open-source architectural modelling and reusable API patterns for ERP modernization.
Technological environment	Cloud-native microservices, hybrid use of open-source LFMs and enterprise APIs.
Contributions to sustainability? (In relation to UN SDG Goals & Tasks)	☒ Supports responsible data governance and transparent AI use (G16: T.6 & T.10) ☒ Improves energy and resource efficiency of model training and/or inference (G12: T2 & T6) ☐ Enables fair and inclusive access to LFM technologies (G10: T2) ☒ Strengthens safety and robustness of AI systems (G9: T.1 & T.5) ☒ Enhances workforce upskilling and human-AI collaboration (G4: T4 & G8: T.2)
Access	CC-BY 4.0 (open contributions where possible; proprietary ERP-specific datasets remain restricted)

9.1 What is it?

We have designed a modular Enterprise Resource Planning (ERP) modernization framework that integrates LFMs for intelligent automation, forecasting, and user interaction. The

contribution defines the architectural migration path (monolithic → modular monolith → microservices), API-first integration, and AI-enabled ERP functio-unalities.

9.2 Why have we done this?

Current ERP systems are rigid, monolithic, and poorly aligned with the requirements of digital transformation. Enterprises require scalable, interoperable, and AI-enhanced platforms to remain competitive. By embedding LFMs into ERP systems, we provide added value through natural language interfaces, predictive analytics, and process automation — while ensuring compliance with GDPR and the AI Act.

9.3 How it works?

- **Architecture Migration**: Existing ERP modules are restructured into modular services and exposed via secure APIs.
- **AI Integration**: LFMs are used for conversational interfaces (HR, finance, logistics) and decision-support analytics.
- **Compliance & Transparency Layer**: Includes audit logging, explainable AI modules, and bias mitigation mechanisms.
- **Monitoring & Validation**: Continuous validation ensures robustness, fairness, and regulatory alignment, with real-time performance monitoring.

9.4 Further Reading

- ELFMo project website (https://elfmo.ftpporto.com/)
- Official documentation for cloud-native architectures and microservices: https://cloudnative-pg.io/documentation/1.15/api_reference/
- CEGID-PHC: https://phcsoftware.com/pt/
- Open-source tools and libraries for LFM fine-tuning and adaptation: https://huggingface.co/
- Responsible AI and transparency guidelines: https://www.microsoft.com/en-us/corporate-responsibility/responsible-ai-transparency-report
- European Commission AI Act overview: https://artificialintelligenceact.eu/

10 Conclusions

This document has provided a comprehensive overview of the initial risk, quality, and conformity assessment methods, risk indicators, and quality metrics essential for the effective integration of Large Foundation Models (LFMs) in enterprise settings. By establishing a structured framework for evaluating and monitoring these aspects, the document lays the groundwork for future developments in risk management and quality assurance, as outlined in subsequent deliverables D3.3 and D3.4. The methods listed in this document demonstrate the project's commitment to enhancing the reliability and trustworthiness of AI systems and ensuring compliance with regulatory standards. The collaborative efforts of the ELFMo consortium will continue to drive innovation and excellence in the deployment of AI technologies, paving the way for enhanced business outcomes and user satisfaction.

References

Beyer, B., Jones, C., Petoff, J., & Murphy, N. R., 2016. *Site reliability engineering: how Google runs production systems*. " O'Reilly Media, Inc.".

Jelinek, F., Mercer, R.L., Bahl, L.R. and Baker, J.K., 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1), pp.S63-S63. https://www.semanticscholar.org/paper/Perplexity%E2%80%94a-measure-of-the-difficulty-of-speech-Jelinek-Mercer/8d350f2d767a70d55275a17d0b3dfcc80b2e0fee

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T. and Shu, K., 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge, *arXiv preprint arXiv:2411.16594*. https://arxiv.org/abs/2411.16594

Järvelin, K. and Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* (TOIS), 20(4), pp.422-446. https://faculty.cc.gatech.edu/~zha/CS8803WST/dcg.pdf

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). https://aclanthology.org/P02-1040.pdf

Wang, Z., Li, S., Zhou, Y., Li, X., Gu, R., Cam-Tu, N., Tian, C. & Zhong, S., 2024. Revisiting SLO and Goodput Metrics in LLM Serving. *arXiv preprint arXiv:2410.14257*. https://arxiv.org/pdf/2410.14257

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*. https://arxiv.org/pdf/1904.09675