LLM: validation framework

"The consequences of AI going wrong are serious, so we need to be proactive rather than reactive".
Elon Musk[94]

Framework

Large Language Models (LLMs) have great potential to transform various industries and applications, but they also pose significant risks that must be addressed. These risks include the generation of misinformation or hallucinations, perpetuation of biases, difficulty in forgetting learned information, ethical and fairness concerns, privacy issues due to misuse, difficulty in interpreting results, and the potential creation of malicious content, among others.

Given the potential impact of these risks, LLMs must be thoroughly validated before deployment in production environments. Validation of LLMs is not only a best practice, but also a regulatory requirement in many jurisdictions. In Europe, the proposed AI Act requires risk assessment and mitigation of AI systems[95]. At the same time, in the United States, the NIST AI Risk Management Framework[96] and the AI Bill of Rights highlight the importance of understanding and addressing the risks inherent in these systems.

Validation of LLMs can be based on the principles established in the discipline of model risk, which focuses[97] on assessing and mitigating the risks arising from errors, poor implementation or misuse of models. However, in the case of AI, and particularly LLMs, a broader perspective needs to be taken that encompasses the other risks involved. A comprehensive approach to validation is essential to ensure the safe and responsible use of LLMs.

This holistic approach is embodied in a multidimensional validation framework for LLMs that covers key aspects (Figure 9) such as model risk, data and privacy management, cybersecurity, legal and compliance risks, operational and technology risks, ethics and reputation, and vendor risk, among others. By systematically addressing all of these issues, organizations can proactively identify and mitigate the risks associated with LLMs and lay the foundation for unlocking their potential in a safe and responsible manner.

In LLMs, this risk assessment can be anchored in the following dimensions used in the model risk discipline, adapting the tests according to the nature and use of the LLM:

- **Input data:** text comprehension[98], data quality[99].
- **Conceptual soundness and model design:** selection of the model and its components (e.g., fine-tuning methodologies, database connections, RAG[100]), and comparison with other models[101].

[94] Elon Musk (n. 1971), CEO of X, SpaceX, Tesla. South African-American entrepreneur, known for founding or co-founding companies such as Tesla, SpaceX and PayPal, owner of X (formerly Twitter), a social network that has its own LLM, called Grok.

[95] European Parliament (2024) AI Act Art. 9: "A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems. The risk management system [...] shall [...] comprise [...] the estimation and evaluation of risks that may arise when the high-risk AI system is used in accordance with its intended purpose, and under reasonably foreseeable conditions of misuse".

[96] NIST (2023): "The decision to commission or deploy an AI system should be based on a contextual assessment of reliability characteristics and relative risks, impacts, costs, and benefits, and should be informed by a broad set of stakeholders".

[97] Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects.

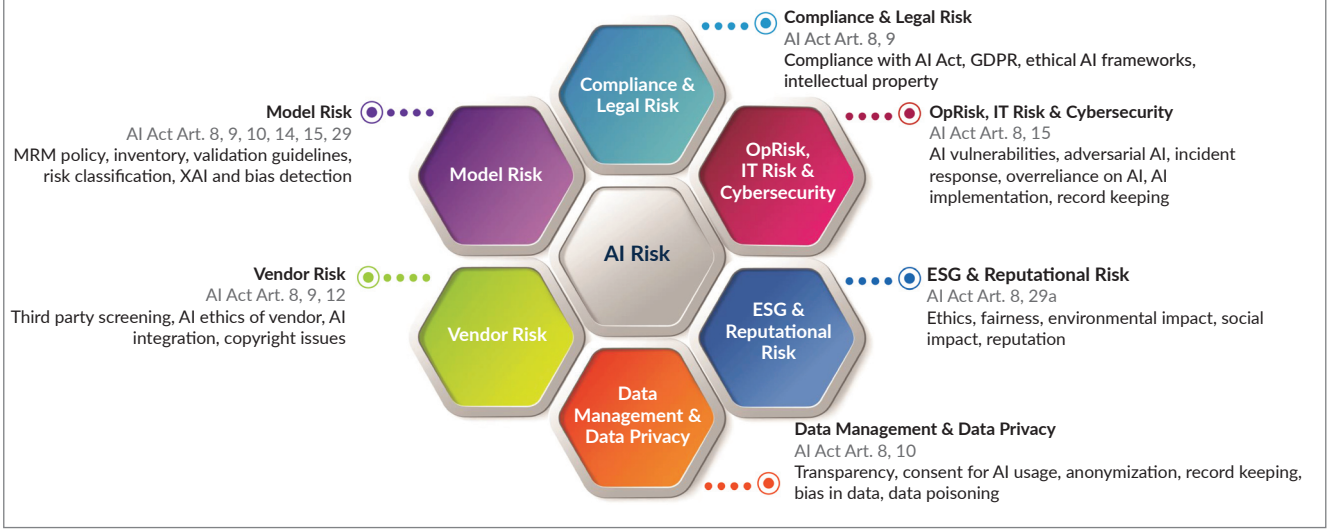
[98] Imperial et al. (2023).

[99] Wettig et al (2024).

[100] RAG (Retrieval-Augmented Generation) is an advanced technique in which a language model searches for relevant information from an external source before generating text. This enriches answers with accurate and current knowledge by intelligently combining information search and text generation. By integrating data from external sources, RAG models, such as the RAG-Token and RAG-Sequence models proposed by Lewis et al. (2020), provide more informed and consistent responses, minimizing the risk of generating inaccurate content or 'hallucinations'. This advance represents a significant step towards more reliable and evidence-based artificial intelligence models.

[101] Khang (2024).

Figure 9. AI Risks and Regulatory References in the AI Act.



- **Model evaluation and analysis of results:** privacy and security of the results[102], model accuracy[103], consistency[104], robustness[105], adaptability[106], interpretability (XAI)[107], ethics, bias and fairness[108], toxicity[109], comparison against challenger models.
- **Implementation and use:** human review in use (including monitoring for misuse), error resolution, scalability and efficiency, user acceptance.
- **Governance**[110] **and ethics**[111]: governance framework for generative AI, including LLMs.
- **Documentation**[112]: completeness of the model documentation.
- **Regulatory compliance**[113]: assessment of regulatory requirements (e.g., AI Act).

To ensure the effective and safe use of language models, it is essential to perform a risk assessment that considers both the model itself and its specific use. This will ensure that the model, regardless of its origin (in-house or from a vendor) or customization (fine-tuning), will function properly in its context of use and meet the necessary security, ethical, and regulatory standards.

Validation techniques

When an organization is considering implementing an LLM for a specific use case, it may be beneficial to take a holistic approach that encompasses the key dimensions of the model's lifecycle: data, design, assessment, implementation and use. It is also necessary to assess compliance with applicable regulations, such as the AI Act in the European Union, in a cross-cutting manner.

In each of these dimensions, two sets of complementary techniques allow for a more complete validation (Figure 10):

- Quantitative evaluation metrics (tests): These standardized quantitative tests measure the model's performance on specific tasks. They are predefined benchmarks and metrics for evaluating various LLM performance aspects after pre-training or during the fine-tuning or instruction tuning (i.e., reinforcement learning techniques), optimization, prompt engineering, or information retrieval and generation phases. Examples include summarization accuracy, robustness to adversarial attacks, or consistency of responses to similar prompts.
- Human evaluation: involves qualitative judgment by experts and end users, such as a human review of a specific sample of LLM prompts and responses to identify errors.

The validation of a specific use of an LLM is therefore carried out by a combination of quantitative (tests) and qualitative (human evaluation) techniques. For each specific use case, it is necessary to design a tailor-made validation approach consisting of a selection of some of these techniques.

[102] Nasr (2023).
[103] Liang (2023).
[104] Elazar (2021).
[105] Liu (2023).
[106] Dun (2024).
[107] Singh (2024).d
[108] NIST (2023), Oneto (2020), Zhou (2021).
[109] Shaikh (2023).
[110] Management Solutions (2014). Model Risk Management.
[111] Oneto (2020).
[112] NIST (2023).
[113] European Parliament (2024). AI Act.

Figure 10. LLM evaluation tests.

Dimensions	Validated aspects	Description	Validation metrics (examples)	Human evaluation (examples)
1. Input data	**1.1 Data quality**	Degree of quality of modeling or application data.	• Flesch-Kinkaid Grade	• Case-by-case review
2. Model design	**2.1 Model design**	Choice of appropriate models and methodology	• Review of LLM elements: RAG, input or output filters, prompts definition, finetuning, optimization... • Comparison with other LLMs	• A/B Testing
3. Model evaluation	**3.1 Privacy and security**	Respect confidentiality and do not regurgitate personal information.	• Data leakage • PII tests, K-anonymity	• Registrations • Ethical hacking
	3.2 Accuracy	Correctness and relevance of model responses	• Q&A: SummaQA, Word error rate • Information retrieval: SSA, nDCG • Summary: ROUGE • Translation: BLEU, Ruby, ROUGE-L • Others: QA systems, level of overrides, level of hallucinations... • Benchmarks: XSUM, LogiQA, WikiData...	• Backtesting of overrides • Case-by-case review
	3.3 Consistency	Correctness and relevance of model responses	• Cosine similarity • Jaccard similarity index	• Case-by-case review • A/B Testing
	3.4 Robustness	Resilience to adverse or misleading informationa	• Adversarial text generation (TextFooler), Regex patterns • Benchmarks of adversarial attacks (PromptBench), number of refusals	• Ethical hacking • Incident drills
	3.5.Adaptability	Ability to learn or adapt to new contexts	• LLM performance on new data by Zero/One/Few-shot learning	• A/B Testing • Case-by-case review
	3.6 Explainability	Understanding the decision making process	• SHAP • Explainability scores	• UX tracking • Focus groups
	3.7 Biases and fairness	Responses without demographic bias	• AI Fairness 360 toolkit • WEAT score, demographic parity, word associations... • Benchmarks of biases (BBQ...)	• Ethical hacking • Focus groups
	3.8 Toxicity	Propensity to generate harmful content.	• Perspective API, Hatebase API • Toxicity benchmarks (RealToxicityPrompts, BOLD, etc.)	• Ethical hacking • Focus groups
4.Implementation and use	**4.1 Human review and safety of use**	Avoid harmful or illegal suggestions and include a 'human-in-the-loop' review.	• Risk protocols, safety assessments • Human control	• Ethical hacking • Focus groups
	4.2 Recovery and error handling	Ability to recover from errors and handle unexpected inputs	• System recovery tests • Error processing metrics	• Incident drills
	4.3 Scalability	Maintain performance with more data or users	• Stress testing of the system, Apache Jmeter... • Scalability benchmarks	• Incident drills • A/B Testing
	4.4 Efficiency	Resource utilization and speed of response	• Time-to-first-byte (TTFB), GPU/CPU utilization, broadcast inference, memory, latency	• Incident drills
	4.5 User acceptance	User acceptance testing.	• User requirements checklist, user opt-out • User Satisfaction (Net Promoter Score, CSAT)	• UX tracking • A/B Testing

The exact selection of techniques will depend on the particular characteristics of the use case; and, in particular, several important factors to consider when deciding on the most appropriate techniques are:

- The level of risk and criticality of the tasks to be entrusted to the LLM.
- Whether the LLM is open to the public (n which case ethical hacking becomes particularly relevant) or its use is limited to the internal scope of the organization.
- Whether the LLM processes personal data.
- The line of business or service the LLM will be used for.

Careful analysis of these factors will allow the construction of a robust validation framework tailored to the needs of each LLM application.

Quantitative evaluation metrics

Although this is an emerging field of study, there is a wide range of quantitative metrics that can be used to evaluate LLM performance. Some of these metrics are adaptations of those used in traditional machine learning models, such as accuracy, recall, F1 score, or area under the ROC curve (AUC-ROC). Other metrics are specifically designed to evaluate unique aspects of LLMs, such as the coherence of the generated text, factual fidelity, or language diversity.

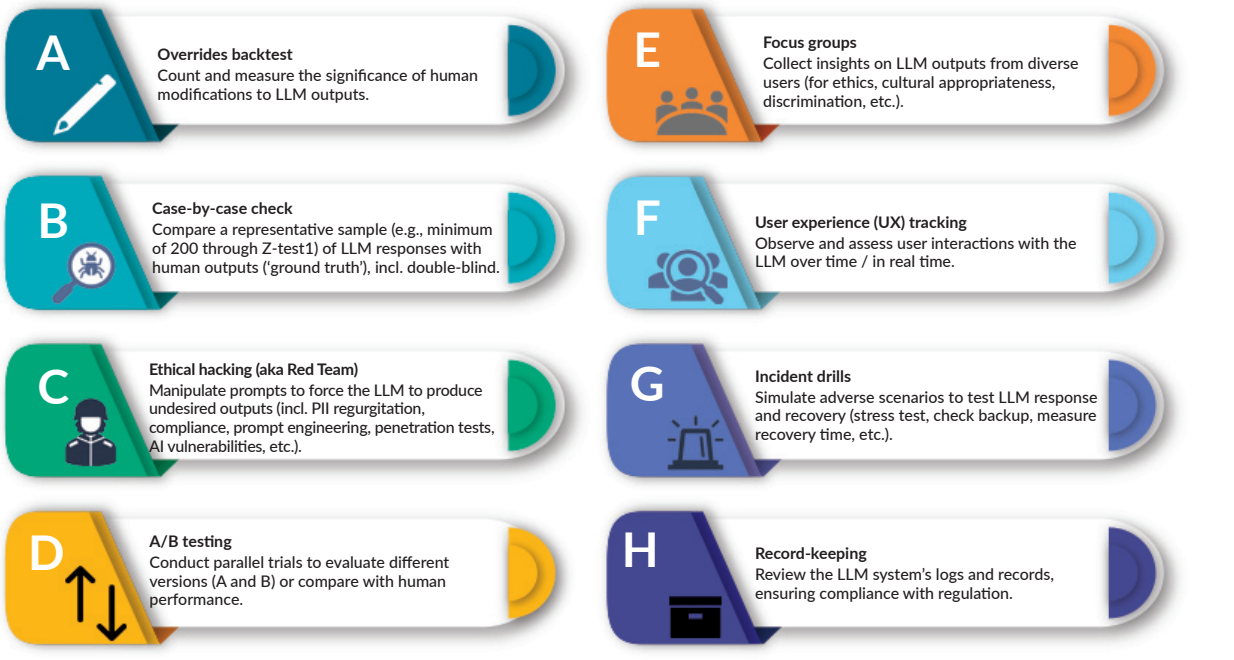
In this context, holistic quantitative LLM testing frameworks already exist in Python programming environments, which facilitate the implementation of many of the quantitative validation metrics, such as:

- **LLM Comparator**[114]: a tool developed by Google researchers for automatically evaluating and comparing LLMs, which checks the quality of LLM answers.
- **HELM**[115]: Holistic Evaluation of Language Models, which compiles evaluation metrics along seven dimensions (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) for a set of predefined scenarios.
- **ReLM**[116]: LLM validation and query system using language usage, including evaluation of linguistic models, memorization, bias, toxicity and language comprehension.

At present, certain validation techniques, such as SHAP-based explainability methods (XAI), some metrics such as ROUGE[117] or fairness analyses using demographic parity, do not yet have widely accepted predefined thresholds. In these cases, it is the task of the scientific community and the industry to continue research to establish clear criteria for robust and standardized validation.

[114] Kahng (2024).
[115] Liang (2023).
[116] Kuchnik (2023).
[117] Duan (2023).

Figure 11. Some LLM human evaluation techniques.



Human evaluation techniques

While quantitative assessment metrics are more directly implementable due to the multitude of online resources and publications in recent years, human assessment techniques[118] are varied and must be constructed based on the specific task[119] being performed by the LLM, and include (Figure 11):

- **User override backtesting:** counting and measuring the importance of human modifications to LLM results (e.g., how many times a sales manager must manually modify customer call summaries generated by an LLM).
- **Case-by-case review:** comparing a representative sample of LLM responses to user expectations ("ground truth").
- **Ethical hacking (Red Team):** manipulating prompts to force the LLM to produce undesired results (e.g., regurgitation of personal information, illegal content, penetration testing, vulnerability exploitation).
- **A/B testing:** comparison to evaluate two versions of the LLM (A and B), or an LLM against a human being.
- **Focus groups:** gathering opinions from various users on LLM behavior, e.g., ethics, cultural appropriateness, discrimination, etc.
- **User experience (UX tracking):** observing and evaluating user interactions with the LLM over time or in real time.
- **Incident drills:** simulating adverse scenarios to test LLM response (e.g., stress test, backup check, recovery time measurement, etc.).
- **Record keeping:** reviewing LLM system logs and records to ensure compliance with regulations and the audit trail.

[118]Datta, Dickerson (2023).
[119]Guzmán (2015).

Benchmarks for LLM Evaluation

Most generative artificial intelligence models, including LLMs, are tested against public benchmarks to evaluate their performance on a variety of tasks related to natural language understanding and usage. These tests are used to measure how well the LLM handles specific tasks and mirrors human understanding. Some of these benchmarks include:

- GLUE/SuperGLUE: assesses language comprehension through tasks that measure a model's ability to understand text.
- Eleuther AI Language Model Evaluation Harness: performs "few-shot" model evaluation, that is, evaluates model accuracy with very few training examples.
- ARC (AI2 Reasoning Challenge): tests the model's ability to answer scientific questions that require reasoning.
- HellaSwag: evaluates the model's common sense through tasks that require predicting a coherent story ending.
- MMLU (Massive Multitask Language Understanding): tests the model's accuracy on a variety of tasks to assess its understanding of multitasking.
- TruthfulQA: challenges the model to distinguish between true and false information, assessing its ability to handle truthful data.
- Winogrande: another tool to assess common sense, similar to HEllaSwag, but with different methods and emphasis.
- GSM8K: uses mathematical problems designed for students to assess the model's logical-mathematical capability.

New trends

The field of LLM validation is constantly evolving, driven by rapid advances developing these models and a growing awareness of the importance of ensuring their reliability, fairness and alignment with ethics and regulation.

Below are some of the key emerging trends in this area:

- **Explainability of LLMs:** As LLMs become more complex and opaque, there is a growing need for mechanisms to understand and explain their inner workings. XAI (eXplainable AI) techniques such as SHAP, LIME, or assigning importance to input tokens are gaining importance in LLM validation. Although a variety of post-hoc techniques for understanding the operation of models at the local and global level are available for traditional models[120] (e.g., Anchors, PDP, ICE), and the definition and implementation of inherently interpretable models by construction has proliferated, the implementation of these principles for LLMs is still unresolved.

- **Using LLMs to explain LLMs:** An emerging trend is to use one LLM to generate explanations for the behavior or responses of another LLM. In other words, one language model is used to interpret and communicate the underlying reasoning of another model in a more understandable way. To enrich these explanations, tools are being developed[121] that also incorporate post-hoc analysis techniques.

- **Post-hoc interpretability techniques:** These techniques are based on the interpretability of the results at the post-training or fine-tuning stage, and allow to identify which parts of the input have most influenced the model response (feature importance), to find similar examples in the training data set (similarity based on embeddings) or to design specific prompts that guide the model towards more informative explanations (prompting strategies).

- **Attribution scores:** As part of post-hoc interpretability[122], techniques are being developed to identify which parts of the input text have the greatest influence on the response generated by an LLM. They help to understand which words or phrases are most important for the model. There are different methods for calculating these scores:

  - Gradient-based methods: Analyze how the gradients (a measure of sensitivity) change for each word as it moves back through the neural network.

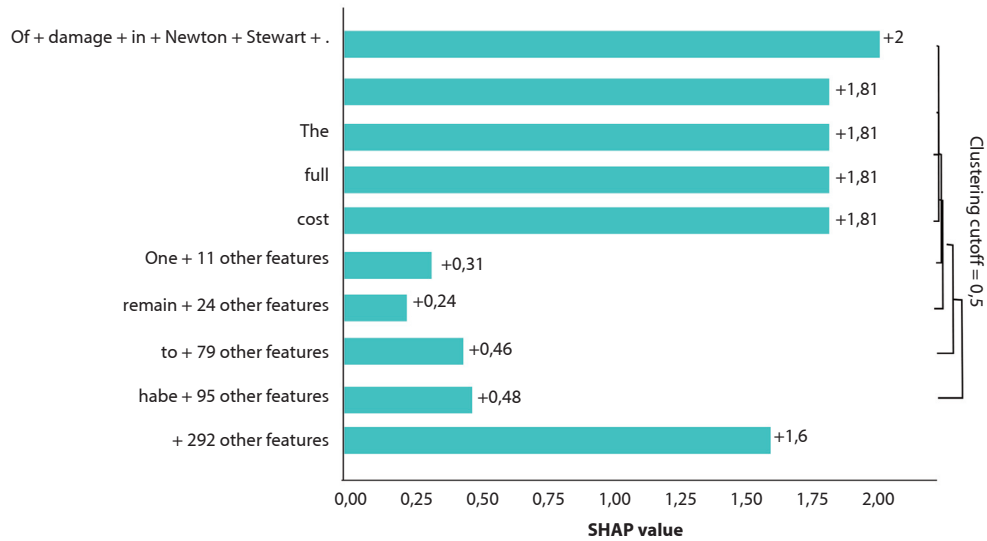
  - Perturbation-based methods: Slightly modify the input text and observe how the model response changes.

  - Interpretation of internal metrics: Use metrics calculated by the model itself, such as attention weights in transformers, to determine the importance of each word.

[120] Management Solutions (2023). Explainable Artificial Intelligence.
[121] Wang (2024).
[122] Sarti (2023).

Figure 12¡. Implementation of SHAP values for text summarization.

Output summary: "The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed . First Minister Nicola Sturgeon visited the area to inspect the damage. Labour Party 's deputy Scottish leader Alex Row ley was in Haw ick on Monday to see the situation first hand. He said it was important to get the flood protection plan right"



An example of attribution scoring is the use of the SHAP technique to provide a quantitative measure of the importance of each word to the LLM output, which facilitates its interpretation and understanding (Figure 12).

- **Continuous validation and monitoring in production:** In addition to pre-deployment evaluation, the practice of continuously monitoring the behavior of LLMs in production, as is done with traditional models, is growing. This makes it possible to detect possible deviations or degradations in their performance over time, and identify biases or risks that were not initially anticipated.

- **Collaborative and participatory validation:** Greater involvement of different stakeholders in the validation process is encouraged, including not only technical experts but also end users, regulators, external auditors and representatives of civil society. This plural participation allows for the inclusion of different perspectives and promotes transparency and accountability.

- **Ethical and regulatory-aligned validation:** In addition to performance metrics, it is becoming increasingly important to assess whether LLM behavior is ethical and in line with human values and regulations. This involves analyzing issues such as fairness, privacy, security, transparency, or the social impact of these systems.

- **Machine unlearning:** This is an emerging technique[123] that allows unlearning "known information from a LLM without retraining it from scratch. This is achieved, for example, by adapting the hyperparameters of the model to the data to be unlearned. The same principle can be used to remove identified biases. The result is a model that retains its general knowledge but has problematic biases removed, improving its fairness and ethical orientation in an efficient and selective way. Several machine unlearning methods are currently being explored, such as gradient ascent[124], the use of fine-tuning[125] or selective modification of certain weights, layers or neurons of the model[126].

[123]Liu (2024).
[124]Jang (2022).
[125]Yu (2023).
[126]Wu (2023)

SHAP (SHapley Additive exPlanations) applied to an LLM

SHAP is a post-hoc explainability method based on cooperative game theory. It assigns each feature (token) an importance value (Shapley value) that represents its contribution to the model prediction.

Formally, let $x = (x_1,\ldots,x_n)$ be a sequence of input tokens. The prediction of the model is denoted by f(x). The Shapley value φ value for the token x_i is defined as:

$$\phi_i = \sum_{\{S \subseteq N_i\}} \frac{\{|S|!\,(n - |S| - 1)!\}}{\{n!\}} [f(S \cup \{i\}) - f(S)]$$

where N is the set of all tokens, S is a subset of tokens, and f(S) is the model prediction for subset S.

Intuitively, the Shapley value φi captures the average impact of token xi on the model prediction, considering all possible subsets of tokens.

Example: Consider an LLM trained to classify corporate emails as "important" or "unimportant". Given a vector of input tokens:

x = [The, Q2, financial, report, shows, significant, increase, in, revenue, and, profitability].

The model classifies the mail as "important" with = 0.85.

Using SHAP, the following Shapley values are obtained:

φ_1 = 0.01 (The)

φ_2 = 0.2 (report)

φ_3 = 0.15 (financial)

φ_4 = 0.02 (from)

φ_5 = 0.1 (Q2)

φ_6 = 0.05 (show)

φ_7 = 0.01 (a)

φ_8 = 0.15 (increase)

φ_9 = 0.1 (significant)

φ_{10} = 0.01 (in)

φ_{11} = 0.02 (th)

φ_{12} = 0.12 (income)

φ_{13} = 0.01 (and)

φ_{14} = 0.02 (the)

φ_{15} = 0.08 (profitability)

Interpretation: The tokens "report" (0.2), "financial" (0.15), "increase" (0.15) and "revenue" (0.12) have the highest contribution to the classification of the mail as "important". This suggests that the LLM has learned to associate these terms with the importance of the message in a business context.