Leveraging LLMs to Streamline the Review of Public Funding Applications

João D.S. Marques[1,2*], André V. Duarte[1,2,3*],
André Carvalho[2], Gil Rocha[2], Bruno Martins[1,2], Arlindo L. Oliveira[1,2]
[1]Instituto Superior Técnico, [2]INESC-ID, [3]Carnegie Mellon University
{joao.p.d.s.marques, andre.v.duarte, arlindo.oliveira}@tecnico.ulisboa.pt

Abstract

Every year, the European Union and its member states allocate millions of euros to fund various development initiatives. However, the increasing number of applications received for these programs often creates significant bottlenecks in evaluation processes, due to limited human capacity. In this work, we detail the real-world deployment of AI-assisted evaluation within the pipeline of two government initiatives: (i) corporate applications aimed at international business expansion, and (ii) citizen reimbursement claims for investments in energy-efficient home improvements. While these two cases involve distinct evaluation procedures, our findings confirm that AI effectively enhanced processing efficiency and reduced workload across both types of applications. Specifically, in the citizen reimbursement claims initiative, our solution increased reviewer productivity by 20.1%, while keeping a negligible false-positive rate based on our test set observations. These improvements resulted in an overall reduction of more than 2 months in the total evaluation time, illustrating the impact of AI-driven automation in large-scale evaluation workflows.

1 Introduction

In the last few years, Large Language Models (LLMs) have dramatically reshaped the capabilities and expectations around automated text and image processing tasks, having shown remarkable proficiency across a wide range of domains, including translation, code generation, and mathematical reasoning, among others (Zhang et al., 2023; Alves et al., 2024; DeepSeek-AI et al., 2025).

Despite these impressive capabilities, their real-world adoption has been approached with caution due to concerns over reliability and potential misapplications (European Commission, 2021). Without proper safeguards, automated systems may introduce biases, propagate misinformation, or operate with unintended autonomy, leading to unforeseen consequences (Bommasani et al., 2022; Ruan et al., 2024). However, these risks can be mitigated by using LLMs as a supportive tool rather than a standalone decision-maker, ensuring that human oversight plays a key role in the process. This way, AI can be leveraged to enhance efficiency, while keeping critical processes accountable and reliable.

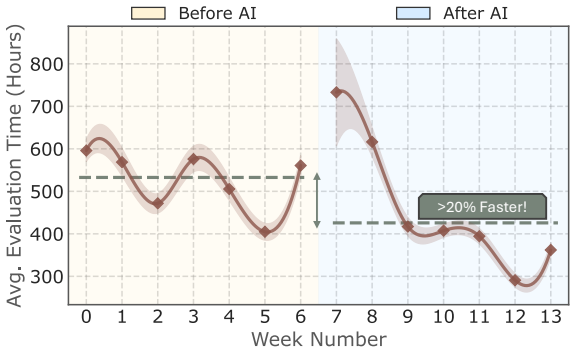

Figure 1: Average application evaluation time within the *ReClaim* initiative, demonstrating a reduction of over 20% following the deployment of our solution.

One particular domain where automated assistance can offer tangible benefits is in public funding allocation programs across the European Union (EU). Annually, the EU and its member states allocate millions of euros through various development initiatives aimed at supporting economic growth, sustainability, and innovation (European Commission, 2019). These initiatives generate substantial interest, prompting thousands of individuals, businesses, and institutions to submit applications. Unfortunately, this high volume of submissions often surpasses the current processing capacities of human evaluation teams, leading to significant bottlenecks and delays in application processing (Silva et al., 2025). Such delays translate into prolonged evaluation times, widespread dissatisfaction, and

* Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2041–2060

ultimately a deterioration of public confidence in the efficiency of these initiatives (Dias, 2024).

To address these challenges and explore the potential for AI-driven automation in public funding allocation, we introduce two distinct LLM-based systems designed to enhance human-in-the-loop evaluation processes. These were deployed in two EU-funded initiatives in Portugal: (i) corporate applications aimed at international business expansion (COMPETE, 2023) and (ii) citizen reimbursement claims for investments in energy-efficient home improvements (Environmental Fund, 2023).

In the first initiative, the main focus is on producing high-quality summaries of applications - the task reviewers identified as the biggest bottleneck. With each submission averaging 30,000 tokens (over 50 pages), manual summarization is both time-consuming and resource-intensive.

In the second initiative, the objective is to ensure consistency between claimed expenses and the supporting documents submitted by applicants. These documents contain unstructured information, requiring a combination of classical Optical Character Recognition (OCR) techniques and LLMs to extract and structure key details. This enables the automatic pre-filling of a verification checklist with over 80 mandatory review items per application. While the task itself is highly deterministic, due to strict submission guidelines, the real challenge lies in its scale: processing around 80,000 applications, each averaging eleven user-uploaded documents (totaling $\approx$ 880,000 documents) that would otherwise require full manual review.

At the time of writing, our systems have been deployed in the field for over three months, demonstrating both quantitative and qualitative improvements. We find the most significant gains to be in the reimbursement claims initiative, where reviewer productivity increased by $\approx$20% (Figure 1), highlighting that structured document verification tasks are particularly well-suited for automation.

Our main contributions are as follows:

- We report on the successful deployment of two AI-assisted document evaluation systems, demonstrating how automation can accelerate application analysis while ensuring human oversight for decision-making integrity.
- We provide a discussion on the key lessons learned from the real-world deployment, offering best practices for integrating AI models into similar settings to ours.

2 Related Work

Automating document review has long been a central goal across both the public and private sectors. Today, the landscape is rapidly changing, largely due to the transformative capabilities of LLMs (OpenAI, 2024; DeepSeek-AI et al., 2025). These models offer unprecedented flexibility, adaptability to diverse document types, and the ability to generalize across a wide range of unstructured data formats (Van Veen et al., 2024; Duarte et al., 2024). However, it is important to recognize that automated document processing predates the arrival of LLMs. Traditionally, such automation relied on rule-driven natural language processing (NLP) and optical character recognition (OCR) techniques (Esposito et al., 1995). While these systems could achieve good performance in controlled environments, they were highly sensitive to variations in document structure and content, which made them difficult to maintain and challenging to extend to broader applications (Chiticariu et al., 2013; Rijcken et al., 2025).

The shift from traditional to LLM-driven approaches is now evident across multiple domains. In the legal sector, LLM-based tools enable rapid review and extraction from a wide range of legal texts, streamlining processes that once required extensive manual effort (Shu et al., 2024). In human resources, resume screening has evolved from basic keyword analysis (Daryani et al., 2020) to sophisticated LLM-powered systems capable of nuanced candidate-role matching (Gan et al., 2024). Public administration (European Commission, 2024) offers a further illustrative example: while the European Commission's 2020 AI Watch report (Misuraca and Noordt, 2020) documented widespread adoption of conventional machine learning solutions within government applications, the Commission's 2024 AI@EC strategic vision report (for Digital Services, 2024) signals a clear commitment to generative AI and the integration of LLMs into these processes.

Despite notable successes, the adoption of LLMs in these sectors has also exposed significant challenges. Failures due to bias, lack of transparency, or over-reliance on unsupervised automation have led to widely publicized setbacks (Dastin, 2018; Angwin et al., 2016). These experiences have shaped the current best practices for LLM deployments, with most organizations now embedding explicit human-in-the-loop review at key decision

points, to ensure oversight and maintain accountability of the systems (Sterz et al., 2024).

Our work follows precisely this trajectory by reporting on two instantiations of LLM-based systems developed for Portuguese government initiatives. In both cases, efficiency gains are balanced with continuous human oversight, supporting trustworthy automation in complex and socially significant document evaluation workflows.

3 AI-Assisted Evaluation: Overview of Initiatives and Proposed Pipelines

Both instantiations aim to streamline the evaluation of large-scale funding programs, although each comes with distinct requirements, prompting us to develop custom solutions for each context. Here, we summarize the main objectives of the two programs, followed by an explanation of how we adapted our systems to meet their particular demands.

3.1 Corporate Applications for International Expansion (IExp)

The *IExp* initiative (COMPETE, 2023) aims to strengthen the international competitiveness of Portuguese Small and Medium Enterprises (SMEs). With a total allocation of 32 million euros, this program invites companies to apply for funding to support projects specifically aimed at expanding their business models and integrating more effectively into global value chains.

To qualify for funding, companies must submit a comprehensive application form detailing their business plans, internationalization strategies, historical and projected market analysis, and associated risks, which results in documents spanning more than 50 pages per submission.

To determine the overall eligibility of the application, reviewers complete a comprehensive evaluation process that involves verifying supporting documentation, cross-checking key information against external government databases, extracting and summarizing relevant details, and assigning scores across multiple criteria. While applications are ultimately reviewed with care and precision, the complexity and thoroughness of the evaluation process leads to extended timelines and a significant burden on reviewers.

Our AI-assisted system was designed to address the most time-consuming and objective elements of the process. Through ongoing discussions with the evaluation team, we prioritized the summarization of the application, assisting in detecting internal inconsistencies across documents and in assigning preliminary scores and justifications for selected sections. In total, our approach automates six specific tasks within the reviewer workflow, detailed in Appendix C.

As illustrated in Figure 2, our pipeline leverages GPT-4o (OpenAI, 2024), which is prompted with the same guidelines the human evaluators follow. For each task, the LLM receives the most relevant sections of the application as the only input. The most relevant sections for each task were identified by the evaluation team, who contributed with their domain expertise by sharing this information. This targeted input not only helps to address the LLM's 'lost in the middle' problem (Liu et al., 2024), but also makes the prompt more cost-efficient.

3.2 Reimbursement Claims for Energy-Efficient Home Investments (ReClaim)

The *ReClaim* initiative (Environmental Fund, 2023) is part of Portugal's *More Sustainable Buildings* program, and has the core objective of improving the energy and environmental performance of Portuguese residential buildings. With a total budget of 30 million euros, the program covers a diverse range of intervention categories, such as window replacement, thermal insulation, HVAC system upgrades, solar installations, and water efficiency measures. Applicants can submit multiple applications across these categories, but each application must focus on a single typology and sub-typology, as detailed in Appendix B.

Unlike in the case of the *IExp* initiative, the *ReClaim* process is fundamentally a large-scale verification task. The program received approximately 80,000 applications, each accompanied by a bundle of supporting documents (eleven documents per application on average). The main challenge was not the need for subjective interpretation, but rather managing the substantial variability in how documents were submitted. Documents were provided in a range of formats, essential documents were sometimes missing or incorrectly categorized, and the information reported by applicants did not always correspond to the details found in invoices or official declarations.

To handle this variability, we first standardized the scope of data processed by our solution. The pipeline supports only widely used file types, such

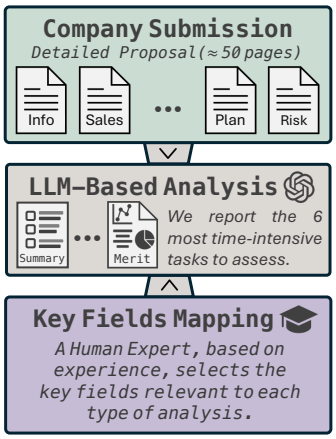

Figure 2: The *IExp* review system leverages GPT-4o to automate the six most time-consuming tasks. Before the analysis, reviewers segment and filter the proposal to avoid overloading the LLM with irrelevant context.

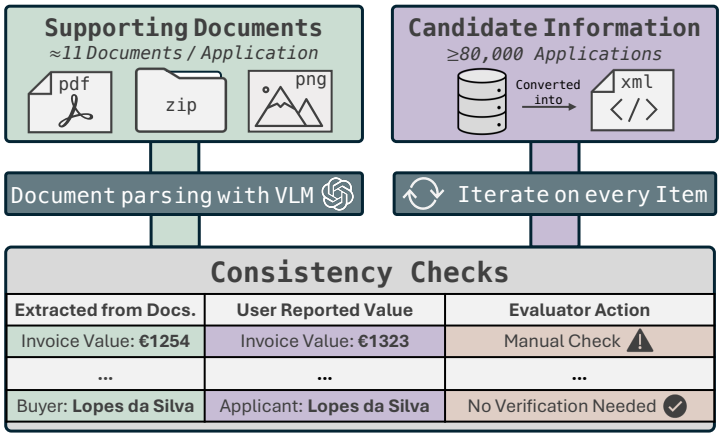

Figure 3: The *ReClaim* evaluation system processes a large volume of supporting documents submitted by citizens in various formats. These documents are automatically parsed using GPT-4o, which extracts critical details and performs automated consistency checks, flagging discrepancies for manual reviewer verification.

as PDF, ZIP, and PNG. Roughly 10% of documents fell outside this criterion and were automatically excluded from automated parsing. For each excluded document, the reviewer received a notification explicitly indicating the presence of unsupported data and the need for manual verification.

The core of the *ReClaim* solution is then a hybrid processing pipeline (Figure 3) that combines classical document parsing and manipulation with VLM-driven information extraction. First, all user-provided form fields are converted into a structured XML format. Then, every supporting document is mapped to its corresponding application. GPT-4o is then used to parse unstructured supporting documents, extracting key details such as invoice values, buyer information, and intervention descriptions. In the final step, the system conducts automated consistency checks by comparing the extracted information against the values reported by the applicant. When a match is found, the item is marked as "No Verification Needed," allowing reviewers to progress efficiently through the checklist. In cases where discrepancies arise (e.g., if the extracted invoice value differs from the user-reported value), the system flags the item for "Manual Check," thereby directing the reviewer's attention to items that require targeted intervention.

4 System Design & Implementation

The deployment of automated systems in real-world settings requires careful consideration of three main aspects: ensuring safety and security, balancing the cost-performance tradeoff, and performing a smooth integration with existing workflows.

4.1 Safety and Security

To comply with the *General Data Protection Regulation* (GDPR) (European Parliament, 2016), our systems are designed to keep the data within the EU borders at all times. Processing occurs exclusively within the region, and storage is managed on encrypted local disks with restricted access, safeguarding confidentiality and integrity.

Beyond data protection, we ensure the reliability of our systems through a human-in-the-loop design. This means that our outputs serve strictly as recommendations, ensuring human reviewers retain oversight and accountability at all times.

4.2 Balancing Cost and Performance

Deploying AI-powered systems at scale requires carefully balancing the available budget with the expected performance, making model selection a critical factor in the process. Given that a significant portion of our tasks consists of Visual Question-Answering (VQA), we initially considered locally hosted open-source models, which offer advantages such as faster inference times and lower operational costs. However, despite promising benchmark results in VQA tasks (Agrawal et al., 2024), we found that VLMs like Qwen2-VL (Wang et al., 2024) or LLaMa-3.2 (Dubey et al., 2024) perform worse in Portuguese, the main language of our data.

Following discussions with the evaluation teams, we prioritized minimizing false negatives, particularly in the reimbursement claims task, where such errors significantly increase the manual verification workload. Given that cost was not a major limiting factor, we opted to use top-performing closed-source models. For that, we conducted a blind test where evaluators assessed anonymized outputs from GPT-4o (OpenAI, 2024) and Gemini-1.5 Pro (Reid et al., 2024). Evaluators consistently preferred GPT-4o, leading to its selection for both initiatives. Further details are given in Appendix A.

4.3 Integration with Existing Workflows

Adapting our solution into the existing workflow of the *IExp* initiative was relatively straightforward, as our method was developed in parallel with the software used by the reviewers. This early involvement ensured a smooth transition, allowing evaluation procedures to be designed with automated assistance in mind.

The *ReClaim* initiative, however, presented more challenges, because the evaluation process was already in progress when we joined. Reviewers were not used to working with AI-generated outputs, so we had to refine the system step by step, working closely with them to ensure compliance with their needs. Another difficulty was that the evaluation platform is not owned by the evaluation team but instead managed by an external provider. This limited how much we could integrate our functionalities directly. Despite these constraints, continuous feedback helped us improve the system over time, increasing reviewers' confidence in the solution and making the process more efficient.

5 Results

We evaluated our systems on both quantitative improvements and reviewer feedback.

5.1 Quantitative Improvements

Corporate Applications (*IExp*): We focus our quantitative analysis on two key metrics: the alignment between AI-generated and reviewer summaries and the classification alignment of company activity types.

Summary alignment was evaluated in two settings: (a) during the proof of concept (POC), using a test set containing 50 applications from past calls reviewed without AI assistance, and (b) in the ongoing evaluation of the most recent call, using a test set containing 11 applications where reviewers are supported by our tool. As shown in Figure 4, the average cosine similarity improved significantly, rising from 0.77 to 0.99 between the two calls. In parallel, all other metrics (ROUGE-L, BLEU, and METEOR) also showed big improvements, increasing from below 0.35 to above 0.9. This demonstrates a substantial gain in alignment quality across all evaluation dimensions, demonstrating that the tool now effectively helps reviewers standardize evaluations, increasing speed and focus on key details. In the case of activity classification (i.e., identifying valid activities and labeling them as either marketing or organizational), we used a test set of 764 applications from past calls that were reviewed without AI assistance. As shown in Figure 5, although LLMs tend to agree more with candidates, agreement with reviewers remains at around 70%. Additional results are provided in Appendix C.

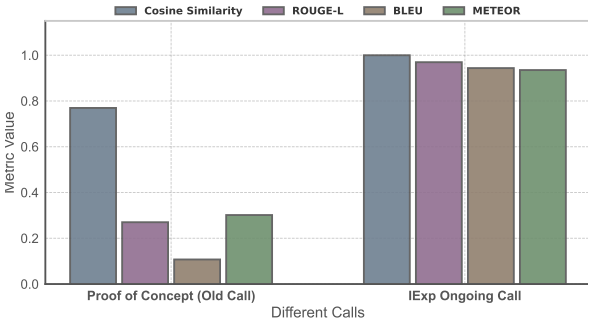

Figure 4: Agreement between application summary during the POC and the current application call.

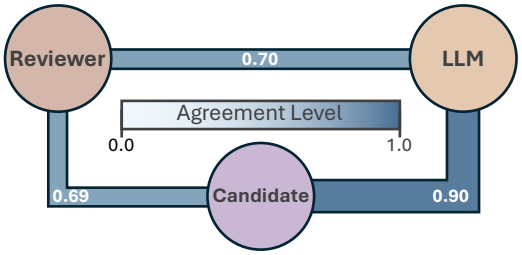

Figure 5: Agreement on classifying activities as either marketing or organizational, between applicants, reviewers, and the LLM.

Reimbursement Claims (*ReClaim*): As shown in Figure 1, reviewer productivity increased by $\approx$ 20%. An evaluation of 200 test samples, distributed evenly across typologies, demonstrated that the LLM-assisted system enabled reviewers to skip manual validation in approximately 76% of field verifications, especially for standardized typologies. Most outputs were accurate (88%), with the remaining errors being either minor or easily

identifiable. Critical issues, such as false positives or reading errors, were rare. Detailed results can be found in Appendix B.

Beyond direct improvements in efficiency, we further analyzed whether the deployment of our system affected evaluator behavior and applicant responses. Specifically, we examined (1) the number of clarification requests that evaluators sent to applicants, and (2) the proportion of appeal requests submitted by applicants following a decision. As shown in Table 1, both measures decreased after the adoption of AI assistance. This suggests that evaluators made fewer human-level mistakes and achieved greater consistency during their assessments, leading to fewer appeals and, consequently, indicating a more positive end-user experience.

Metric	Before AI	After AI
Clarification Requests / Application	2.13	2.05
Applicant Appeal Rate (%)	25.8	20.4

Table 1: Evaluator clarification requests and applicant appeal rates before and after the solution deployment.

5.2 Reviewer Feedback

For both tasks, reviewers were asked for feedback on our solutions. In the *IExp* task, evaluators estimated that AI assistance could accelerate the review process by up to 30%, with the greatest benefit observed in generating application summaries. Secondly, and though not explicitly stated, it can be inferred that the transition toward more standardized, LLM-generated summaries has been positively received by the evaluation team.

In the case of the *ReClaim* task, feedback was more mixed and divided into two main groups. The first group, composed of reviewers who collaborated on the development of the AI tool, expressed strong appreciation for its usefulness. The second group included reviewers who began using the tool after its deployment. Among these, those with extensive evaluation experience often reported it as highly useful, with estimated time savings of up to 40%. However, other reviewers either struggled to understand how to effectively use the tool, or lost confidence after encountering errors. These errors were typically minor or false alerts, but still impacted trust in the system.

Overall, the findings suggest that LLMs are already mature enough to significantly support the application review process. However, their effectiveness is highly dependent on the surrounding ecosystem, including bureaucratic context, reviewer tooling, and the stability of evaluation criteria.

6 Conclusions

This work explored the potential of LLMs to improve evaluation of public funding applications. Our deployments demonstrated that, when properly integrated into human-in-the-loop pipelines, LLMs can substantially accelerate evaluation workflows, reduce manual workload, and promote greater uniformity in the reviewer outputs.

To conclude, we now summarize some of the key lessons learned from this project. We believe that these insights can be valuable for future developments that aim to explore the potential of LLMs in similar scenarios.

Organizational and Regulatory Barriers

Although technical challenges were expected, our experience revealed that bureaucracy is often the main source of delays and reduced solution quality. Critical roadblocks included third-party platform ownership, which restricted our ability to implement system modifications; strict GDPR requirements, which narrowed the pool of viable models; and complex, multi-step authorization workflows that delayed data access. As a result, we believe that successful deployment in this domain requires not only technical flexibility but also careful planning for this constraints, which can fundamentally restrict the available technical options.

Polarized Adoption Patterns

Integrating AI in a human-in-the-loop pipeline is only effective if human reviewers actually decide to use the AI during their evaluations. In our deployment, we observed that reviewers often split into two groups: those who were willing to use the tool and focus on its benefits, and those who became very cautious or critical whenever the system made a mistake. When reviewers lose trust in the system, they may stop using the AI altogether; hence, the potential gains in efficiency are not fully achieved. As a result, we found it is important to give extra attention to reviewers who are less open to using AI. By clearly explaining what the system can and cannot do, it is possible to set realistic expectations and help all reviewers become more tolerant of small system errors, leading to better overall adoption. In practice, effective change management across multiple levels of relevant processes and stakeholders is essential for successful implementation.

High Practical Application Potential

Perhaps the most significant takeaway is the high practical application potential of AI-assisted evaluation systems at scale. While the initial development and adaptation of each solution requires careful planning and close collaboration with the evaluation teams, the large-scale deployment is incomparably faster than manual evaluation. For example, our *ReClaim* system processed the $\approx$ 80,000 applications in less than three weeks. Obviously, this only impacts part of the total evaluation time, since the human-in-the-loop setting still requires significant manual work across other tasks. However, as models continue to improve, it becomes increasingly plausible that fully automated evaluation could become viable, in particular when the underlying processes are reengineered. If and when that point is reached, we can expect even more dramatic speedups in public sector workflows.

Limitations

In the *IExp* initiative, the system was limited by the fact that certain reviewer tasks depended on past applications or external databases, which were not accessible to the LLM. Instead, the LLM is restricted exclusively to relevant sections of the current application (as detailed in Section 3.1). Reviewers were informed of this limitation, which influenced task selection and led us to prioritize those that required little or no external context.

In the *ReClaim* initiative, the system faced challenges due to inconsistent document formats and low-quality file submissions. The permissive nature of the submission platform resulted in cases where the system could not be applied. As a result, we recommended enforcing stricter file format requirements in future calls. Moreover, poor-quality content (e.g., blurry scans) negatively affected performance, emphasizing the importance of more precise applicant guidelines and implementing file verification during submission to enhance system effectiveness.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions. This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and by Fundação para a Ciência e Tecnologia, I.P. (FCT) through the projects with references UID/50021/2025 and UID/PRR/50021/2025.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. In *First Conference on Language Modeling*.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: Risk Assessments in Criminal Sentencing. *ProPublica*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Lucia Zheng, Kaitlyn Zhou, Percy Liang, et al. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. 2013. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA. Association for Computational Linguistics.

COMPETE. 2023. Internationalization of Small and Medium Enterprises (SME). *Portugal 2030*. Published: 2023-11-02.

Chirag Daryani, Gurneet Singh Chhabra, Harsh Patel, Indrajeet Kaur Chhabra, and Ruchi Patel. 2020. An automated resume screening system using natural language processing and similarity. *ETHICS AND INFORMATION TECHNOLOGY [Internet]. VOLKSON PRESS*, pages 99–103.

Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Zhen Zhang, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Mariana Dias. 2024. "Nothing to pay by the end of the year"? Government risks new failure in support of windows and solar panels. *Expresso*. Published: 2024-11-18.

André V. Duarte, João DS Marques, Miguel Graça, Miguel Freire, Lei Li, and Arlindo L. Oliveira. 2024. LumberChunker: Long-Form Narrative Document Segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6473–6486, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Environmental Fund. 2023. Support Program for more Sustainable Buildings. *XXIII Government - Portuguese Republic*. Published: 2023-08-16.

F. Esposito, D. Malerba, and G. Semeraro. 1995. A knowledge-based approach to the layout analysis. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 466–471 vol.1.

European Commission. 2019. Europe 2030. In *Reflection Paper - Towards a Sustainable Europe by 2030*.

European Commission. 2021. Artificial Intelligence Act. In *Regulation (EU) 2024/1689: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on AI and amending certain Union legislative acts*.

European Commission. 2024. Eu study calls for strategic ai adoption to transform public sector services. https://digital-strategy.ec.europa.eu/en/library/eu-study-calls-strategic-ai-adoption-transform-public-sector-services. Accessed: 2025-03-29.

European Parliament. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Directorate-General for Digital Services. 2024. Artificial Intelligence in the European Commission (AI@EC). Technical report, European Comission, Brussels (Belgium).

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of llm agents in recruitment: A novel framework for automated resume screening. *Journal of Information Processing*, 32:881–893.

Nelson F Liu, Kevin Lin, John Hewitt, et al. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Gianluca Misuraca and Colin Van Noordt. 2020. AI Watch - Artificial Intelligence in public services. Scientific analysis or review, Policy assessment KJ-NA-30255-EN-N (online), European Comission, Luxembourg (Luxembourg).

OpenAI. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Emil Rijcken, Kalliopi Zervanou, Pablo Mosteiro, Floortje Scheepers, Marco Spruit, and Uzay Kaymak. 2025. Machine learning vs. rule-based methods for document classification of electronic health records within mental health care—A systematic literature review. *Natural Language Processing Journal*, 10:100129.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. In *The Twelfth International Conference on Learning Representations*.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. LawLLM: Law large language model for the US legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889.

Pedro Silva, Antonieta Duarte, Ana Rita Costa, and Alda Mota. 2025. Support Program for more Sustainable Buildings: Half of the Applications Remain Unpaid. *DECO Proteste*. Published: 2025-01-28.

Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2495–2507, New York, NY, USA. Association for Computing Machinery.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Kexun Zhang, Danqing Wang, Jingtao Xia, William Yang Wang, and Lei Li. 2023. ALGO: Synthesizing Algorithmic Programs with Generated Oracle Verifiers. In *Advances in Neural Information Processing Systems*, volume 36, pages 54769–54784. Curran Associates, Inc.

A Model Selection - Blind Test

To better understand which language model could align more closely with human preferences for our tasks, we conducted a blind evaluation of outputs from two language models: Gemini-1.5 Pro and GPT-4o. For 10 applications from the Corporate Applications (*IExp*) initiative, both models generated the application summary. Each pair of summaries was then blindly evaluated by three reviewers, who were shown the summaries in randomized order and asked to select the one they preferred.

As shown in Table 2, GPT-4o was preferred in 8 out of 10 cases, receiving 21 out of the 30 total votes. Based on this trend, we adopted GPT-4o for our solution, and used it across all tasks to ensure consistency.

Summary ID	**GPT-4o Votes**	**Gemini Votes**	**Majority Winner**
1	3	0	GPT-4o
2	2	1	GPT-4o
3	1	2	Gemini
4	2	1	GPT-4o
5	3	0	GPT-4o
6	2	1	GPT-4o
7	2	1	GPT-4o
8	1	2	Gemini
9	3	0	GPT-4o
10	2	1	GPT-4o
Total Votes	**21**	**9**	**GPT-4o**
Summary Wins	**8**	**2**	**GPT-4o**

Table 2: Human preferences between summaries generated by GPT-4o and Gemini. Each pair of summaries (one from each model) was evaluated by three annotators.

B *ReClaim* Details

ID	Typology	# Sub-Typologies	Applications (%)	Avg Documents
1	Window Replacement	None	27.34%	14
2	Thermal Insulation	4	1.36%	12
3	Heating and Cooling Systems	3	48.06%	10
4	Solar Panels	2	32.72%	11
5	Water Efficiency	None	0.52%	8

Table 3: Typology distribution for more than 80,000 *ReClaim* applications that were analyzed.

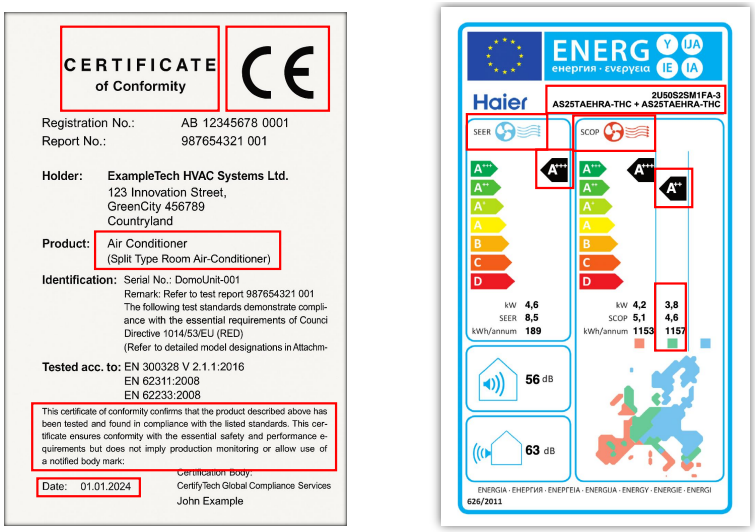

Figure 6: *ReClaim* application files examples.

Report	Cost (€)	Time (s)
Typology 1	0.05	37
Typology 2.1.1	0.06	61
Typology 2.1.2	0.02	34
Typology 2.2.1	0.02	24
Typology 2.2.2	0.09	108
Typology 3.1	0.02	41
Typology 3.2	0.10	87
Typology 3.3	0.09	23
Typology 4	0.04	39
Typology 5.1	0.03	25
Typology 5.2	0.21	173
All Typologies Avg.	0.06	48
Eligibility	0.01	13
Common Core	0.02	29
Total	**0.09**	**90**

Table 4: Cost and time analysis for *ReClaim* applications on average.

B.1 Additional Results

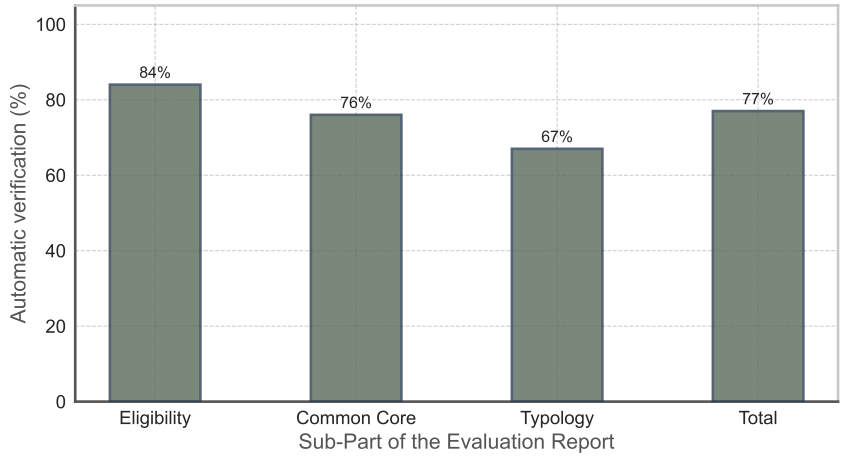

Figure 7: Proportion of field verifications that can suppress manual validation across the main sections of the application. A verification is considered not to require manual review if it is either correct or clearly incorrect in an interpretable way (e.g., a misplaced digit or incorrect capitalization). On average, 76% of verifications did not require human intervention.

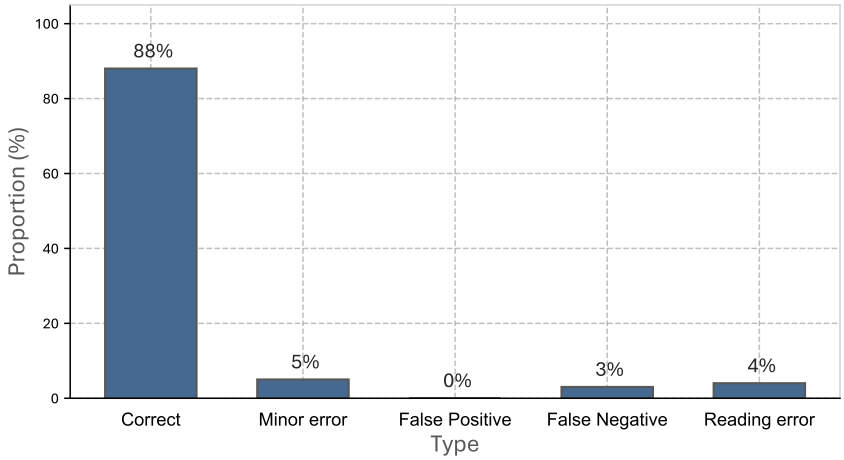

Figure 8: Evaluation of the correctness of $\approx$ 7,000 verification fields obtained from 200 applications. This analysis categorizes the types of errors produced by the LLM-generated reports. Minor errors refer to mistakes that evaluators can immediately recognize and correct, such as misspellings, capitalization issues, or false alerts. False positives occur when the system fails to detect an actual error. False negatives are instances where the system incorrectly flags an error when there is none. Reading errors result from the LLM being unable to read or process a file properly.

B.2 *ReClaim* Prompt Example

System Prompt: You will receive a document as an attachment (MCP). Please help convert part of it into XML. The XML tags to be created are:

<mcp_type> (You must classify the type of document. The main categories are:
1. Submission Receipt – issued by DGEG
2. Screenshot of the MCP platform submission
3. Confirmation email of MCP submission
4. Document recognizing technician or company responsible for private electrical installations
5. Document granting exemption from prior control)

<ID_energy_producer> (ID of the energy producer)
<NIF_NIPC_mcp> (Tax ID of the energy producer)
<address_mcp> (Address of the installation)
<energy_source_mcp> (Source of energy)
<generator_power_mcp> (Nominal power of installed generators. Include units when possible)
<nominal_power_mcp> (Installed capacity/nominal power of the inverter. Include units when possible)
<date_start_mcp> (Date of authorization to begin operation)
<date_submission_mcp> (Date of MCP submission)

If any of the values are missing in the document, return the corresponding tag with 'None'.
All responses should follow the format: <tag_name>value</tag_name>
If a document doesn't match any listed type, use: <mcp_type>None</mcp_type>

Table 5: Example of the system prompt used for the Prior Communication (MCP) document-to-XML task. The user prompt consists of a set of images corresponding to the attached document. Each required tag is extracted using structured JSON query formatting, as shown in the schema. Values are returned in XML-style tags (e.g., <tag_name>value</tag_name>) and are defined through a JSON schema that specifies their type, description, and possible enumerations. Most prompts for similar tasks follow this same structure, varying only in the specific tags and extraction goals. Once extracted, the data is filtered and either passed to subsequent LLM calls or used in downstream logic (e.g., checking if one date is later than another).

B.3 *ReClaim* Reports Examples

B.3.1 Eligibility Report Example

Eligibility Report

Verification - Consistency between Expense and Type:
The invoice mentions a piece of equipment, which appears to be a biomass heat recovery unit. However, the specific model is not mentioned in the application description. There is not enough information to confirm if it is the same equipment due to the difference in power.
Confirm Manually.

Verification - CPU Validity:
Application submission date: 2015-11-04
CPU extraction date: 2015-02-27
Comment - Confirm Manually. CPU extraction date must not be older than 6 months from the application submission date.

Verification - Total Eligible Expense + CE:
Total eligible expense: 3250€
Comment - The total eligible expense is less than 5000€. Energy Certificate was not found.

Verification - Property Address:
Address indicated in the application: Rua da Liberdade, 17, Lisbon
Address from CPU: Rua da Liberdade, 17, Lisbon
Address from CPU: Rua da Liberdade, 18B, Lisbon
Comment - Confirm Manually. The first two addresses seem similar. The third address seems different as it has a different door number.

Verification - Property Use:
Property use from CPU: Housing
Comment - The property use is in accordance with the eligibility criteria.

Verification - Property Type:
Property type from CPU: Building with multiple units. Owner. Independent use.
Comment - Property type is in accordance with the eligibility criteria.

Verification - Invoices or Receipts dated after 01/01/2015:
Invoice date from documents: 1/01/2018
Invoice issue date from documents: 1/01/2018
Comment - The invoice date is in compliance with eligibility criteria.

Figure 9: Eligibility report dummy example.

B.3.2 Common Core Report Example

Common Core Report

Verification - Candidate Name
Candidate name according to the application: Manuel dos Santos
Name extracted from the invoice: Manuel dos Santos
Name of the property owner extracted from the CPU: Manuel dos Santos
Comment - The candidate's name appears to be consistent with all names present in the documents.

Verification - Candidate ID:
Candidate NIF according to the application: 1234567
NIF extracted from the invoice: 1234567
Owner's NIF extracted from the CPU: 1234567
Comment - The candidate's ID appears to be consistent with all IDs present in the documents.

Verification - Property Type:
Property type extracted from CPU: Building under Horizontal Property Regime
Property type indicated in the application: Building under Horizontal Property Regime
Comment - The property type indicated in the application matches the type extracted from the CPU.

Verification - Matricial Article:
Matricial Article of the property as indicated in the application: 1111
Matricial Article of the property extracted from the CPU: 1111
Comment - The matricial article of the property indicated in the application matches the one extracted from the CPU.

Verification - Private Gross Area:
Private gross area of the property indicated in the application: 77.58
Private gross area of the property extracted from the CPU: 77.58
Comment - The private gross area of the property indicated in the application matches the one extracted from the CPU.

Verification - Year of Housing License:
Year of housing license in the application: 1990
Comment - The housing license date appears to comply with the eligibility criteria.

Verification - Invoice Number:
Invoice number indicated in the application: 2015/111
Invoice number extracted from the document: 2015/222
Comment - Confirm Manually. The invoice number in the application does not match the invoice number extracted from the document.

Verification - Receipt Number:
Receipt number in the documents: 2015/333
Comment - The receipt was found in the documents and is numbered: 2015/333

Figure 10: Common core report dummy example.

B.3.3 Typology Report Example

Typology Report

Verification - Energy Source:
Energy source indicated in the application: Solar
Energy source extracted from MCP: Solar
Comment - The energy source indicated in the application matches the energy source extracted from the MCP.

Verification - Installed Power:
Installed Power (Inverters) indicated in the application: 3 kW
Installed Power (Inverters) extracted from the invoice: None kW
Installed Power extracted from MCP: 3 kW
Comment - Installed Power (Inverters) does not match the values presented in the following documents: “Invoice”, “MCP”. **Confirm manually.**

Verification - Generator Power:
Total Power of the Generators (Solar Panels) indicated in the application: 3.1 kW
Total Power of the Generators (Solar Panels) extracted from the invoice: 3.5 kW
Total Power of the Generators extracted from MCP: 3.1 kW
Comment - Confirm manually: Installed Power of the Generators (Solar Panels) does not match the values presented in the following documents: “Invoice”.

Verification - MCP with commissioning date after 01/01/2015:
Commissioning date extracted from MCP: 2020-1-1
Comment - The commissioning date found in the MCP is after Jan 1, 2015, as required by the notice's criteria.

Verification - MCP with issue date prior to application:
Comment - Confirm manually. The MCP date was issued after the application submission date.

Verification - Number of Panels:
Number of panels indicated in the application: 5
Number of panels extracted from the invoice: 5
Comment - The number of panels indicated in the application matches the number extracted from the invoice.

Verification - Brand and Model of Panels:
Brand and Model of Panels extracted from invoice: XPTO 123
Brand and Model of Panels extracted from documentation: LMN 555
Comment - Confirm manually. The brand and model of the solar panels from the invoice do not match the ones extracted from the documentation.

Verification - Brand and Model of Inverters:
Brand and Model of Inverters extracted from invoice: PQR 2S
Brand and Model of Inverters extracted from documentation: MNS 3T
Comment - Confirm manually. The brand and model of the inverters from the invoice do not match the ones extracted from the documentation.

Figure 11: Typology report dummy example.

C *IExp* Details

Reviewer ID	Documents	LLM vs Reviewer (Eligibility)	Reviewer Acceptance Rate	LLM vs Reviewer (Typology)
1	21	0.76	0.95	0.76
2	**216**	**0.85**	**0.96**	**0.68**
3	47	0.78	0.93	0.72
4	23	0.82	0.95	0.73
5	**278**	**0.76**	**0.93**	**0.69**
6	17	0.58	0.64	0.58
7	30	0.83	1.00	0.66
8	**132**	**0.71**	**0.97**	**0.75**

Table 6: Agreement between the LLM and human reviewers in filtering and classifying activities as either marketing or organizational. The *Documents* column indicates the number of applications evaluated by each reviewer. *Eligibility* measures agreement on whether activities qualify for funding. *Acceptance Rate* reflects the proportion of initiatives accepted for funding. *Typology* captures the agreement between the LLM and the reviewer when assigning the same category (marketing or organizational) to a given initiative.

Section	Cost (€)	Time (s)
Full summary	0.06	25
Inovation activities	0.07	12
Qualitative Analysis	0.03	13
Inconsistency report	0.02	10
Global score and rationale	0.02	9
Total	**0.20**	**69**

Table 7: Cost and time analysis for *ReClaim* applications on average.

C.1 *IExp* Prompt Example

System Prompt: A company is applying for funding with the goal of improving its operations. You are an evaluator of applications whose assessments will lead to either the approval or rejection of each funding proposal.

Write a project summary in European Portuguese (pt-PT), based only on the information provided about the company. Do not add external content. The summary should cover the following topics:

1. Business activity, recent developments, and relevant historical milestones;
2. Main products and services and their relative weight in the business;
3. Reference clients, sales structure (concentration/diversification);
4. Export activity before the project;
5. Other relevant attributes (Brands, Certifications, Awards/Insignia, Market positioning...);
6. Main objectives underlying the funding proposal;
7. Types of planned actions (an exhaustive list is not necessary);
8. Export activity after the project (evolution of exports, market diversification, etc.)

The summary must always be structured in exactly five paragraphs, distributing the above topics logically among them.

User Prompt: This company is applying with an internationalization project. The relevant information provided by the company (which may contain minor errors), grouped by topic, is as follows:

General information about the beneficiary: `{benef_info}`
Brief summary of the proposed project: `{cnd_text_info[6]}`
Company's areas of economic activity: `{cnd_text_info[0]}`
Context, history, and evolution of the company: `{cnd_text_info[1]}`
Currently relevant markets: `{cnd_text_info[2]}`
New markets the company intends to reach: `{cnd_text_info[3]}`
How the company intends to operate in these new markets: `{cnd_text_info[4]}`
Extended summary of the project: `{cnd_text_info[5]}`
Strategic diagnosis: `{cnd_text_info[7]}`
Objectives associated with the investment: `{cnd_text_info[8]}`
Technical description of the planned investments: `{cnd_text_info[9]}`
Verified key market information: `{merc_text}`

Table 8: Example of the prompt used for generating a five paragraph application summary. The user prompt provides structured company and project data, using JSON query formatting. Some data, as the market information (stored in merc_text), is pre-processed in order to reduce errors. A sixth paragraph, concerning financial values before and after the project is generated through logical inference based on the company's financial sheets to avoid inaccuracies.

C.2 *IExp* Reports Examples

C.2.1 Summary Report Example

P1

Company A was established on October 1, 2005, in Portugal. Its core activity is the retail of sunglasses, targeting the luxury and sophisticated segments with ties to the healthcare sector.

In terms of market presence, the company owns an internationally registered brand, **Brand X**.

P2

Its main competitors are identified as companies based in Germany, Belgium, France, Ireland, and the United States.

P3

Currently, the company operates in 10 international markets, primarily within Europe, namely Spain, and France. Through the implementation of this project, it aims to strengthen its position in existing markets and expand into new ones such as Ireland, Canada and Brazil.

P4

The strategic goals of the project include achieving 99% of its revenue from international markets, consolidating its presence where it already operates, and entering new markets.

P5

To support its internationalization strategy, the company plans to carry out activities within the following operational areas:

- **Market Knowledge**: Participation in international trade fairs to promote its brand;
- **Web Presence and Digital Economy**: Development of an online management platform with e-commerce capabilities, hiring of international digital marketing consultancy services, and execution of ad campaigns to strengthen its digital presence;
- **International Brand Development and Promotion**: Trademark registration of **Brand X** in various international markets, collection development, and promotional activities for product presentation;
- **Prospection and Presence in International Markets**: Prospecting trips to several target markets to attract new clients;
- **International Marketing**: Preparation of an international marketing plan, production of promotional materials, and placement in magazines;
- **New Organizational Methods in Commercial Practices and External Relations**: Hiring four qualified professionals in sales and marketing to support the internationalization process.

P6

With this internationalization project, the company expects to increase its international revenue from approximately €500,000 pre-project to €1.550 million post-project - an increase of over €1 million.

Figure 12: *IExp* summary report example. The report is structured into six logical sections (referred to as paragraphs), each responsible for presenting a key component of the required information: P1 – Company Introduction, P2 – Market Activity, P3 – Competition, P4 – Strategic Objectives, P5 – Internationalization Activities, and P6 – Revenue before and after funding. Each paragraph is generated using structured JSON query formatting.

C.2.2 Qualitative Analysis Report Example

Based on the analysis of the application submitted by the company X, it was found to be well-structured and includes a solid action plan aligned with the intended internationalization objectives and strategy.

Regarding the company's internal analysis, it presents a consistent SWOT assessment, identifying key strengths such as team specialization, proven experience in the sector, and a leading position in the national market.

Among the weaknesses, the analysis highlights limited financial resources, low production capacity, and challenges in allocating human resources to the internationalization process.

The investment plan outlines actions that are expected to address some of these weaknesses while leveraging the company's strengths. For example, to overcome the challenges of international investment given the company's small size, there is significant investment planned in prospecting and promotional activities across a wide range of international markets.

In terms of cost reasonableness, the total proposed investment (€200,000.00) is below the company's total turnover in the year prior to the project (€400,000.00), representing approximately 50% of it. This is considered a reasonable level of investment given the company's financial context.

With this internationalization project, the company expects to increase its international turnover by approximately €200,000.00. Given that the expected gains exceed the investment, the project is considered to offer clear added value for the company.

Finally, in terms of funding sources, the project will be financed through self-financing and shareholder loans, which suggests a low execution risk based on the company's financial track record.

Figure 13: *IExp* qualitative analysis report example. Each paragraph is generated using structured JSON query formatting.

C.2.3 Innovation Activities Report Example

As part of the project, several marketing innovation activities have been implemented, including:

(i) Organizing *Open Days* with major operators such as X, Y, Z , and others. These events allow the company to showcase its production and innovation capabilities directly to potential clients, helping to build trust and credibility - key factors for securing contracts and developing new solutions.
(ii) Developing a comprehensive digital communication strategy, including the creation of a brand voice and communication guidelines, as well as the analysis and optimization of digital tools. This approach aims to ensure consistent and clear messaging to the target audience, improving the company's digital positioning and marketing campaign effectiveness.
(iii) Participating in international trade fairs, where the company can present new products and technologies to a global audience, promoting its brand and solutions among key players in the railway industry.

In the area of organizational innovation, the project includes:

(i) Engaging a consultancy service to explore the French market, including market research and commercial outreach, with the goal of expanding the company's presence in the French sector;
(ii) Implementing the Salesforce platform to centralize information, enhance the 360º view of customers, and support the sales team - resulting in more integrated and efficient data and process management;
(iii) Creating two highly qualified positions in the commercial and marketing department, requiring engineering or related backgrounds, to strengthen market outreach efforts and improve integration into global value chains.

Figure 14: *IExp* innovation activities report example. The task involves selecting, from a list of proposed activities, those that are eligible, and classifying each as either organizational or marketing. Each paragraph is generated using structured JSON query formatting.