

INTEGRATING GENERATIVE AI INTO HSE LEGAL COMPLIANCE: CONTRIBUTIONS, LIMITS AND ORGANISATIONAL IMPACTS

Adelin Taddei

Center for Research on Risk and Crisis, Mines Paris/PSL, France. Email : adelin.taddei@minesparis.psl.eu

This article analyzes the contribution of generative artificial intelligence (GAI), and more specifically large language models (LLMs), to legal work in the field of health, safety, and environmental (HSE) compliance within a medium-sized LegalTech company. The study focuses on a French SME specializing in regulatory monitoring, compliance management, and HSE risk assessment.

The research is based on an action research approach, combining field observations, interviews with HSE legal experts, and experiments using a proprietary database of legal questions. Three complementary evaluations were conducted: a qualitative evaluation of the generated responses, an evaluation of the impact of prompt engineering, and a quantitative performance analysis.

The results demonstrate that GAI significantly improves the responsiveness and efficiency of regulatory monitoring and information retrieval, with performance varying depending on the legal area. The structured design of prompts considerably improves the consistency, accuracy, and usefulness of the generated responses. However, limitations remain regarding reliability, risk of error, interpretability, and the need for expert validation. Integrating GAI also raises strategic, organizational, and ethical challenges within legal teams.

The study concludes that a gradual and structured integration of GAI, supported by a new organization of legal work, can improve the performance of an HSE legal department while creating sustainable organizational value, provided that human expertise remains central.

Keywords: AI, LLM, GAI, Compliance.

1. Introduction

This study is conducted within a French LegalTech SME specializing in legal compliance support in matters of health, safety and environment (HSE).

Legal work in the HSE field is characterized by a strong reliance on regulatory monitoring, legal interpretation, and the ability to respond quickly to complex compliance issues. In this context, LegalTech solutions have progressively transformed legal work by improving access to legal information and facilitating regulatory monitoring, document retrieval, and compliance tracking. In fields such as HSE, these tools play a crucial role in ensuring the consistency and reliability of the legal information provided to organizations.

LegalTech systems rely on structured legal databases (Bourreau et al., 2012) and expert systems designed to formalize legal standards and reasoning (Audiffren et al., 2012). While these approaches offer strong guarantees of legal reliability, they require significant modeling efforts and continuous updates to remain compliant with evolving regulatory frameworks.

The emergence of generative artificial intelligence (GAI), and in particular large language models (LLMs), represents a major technological evolution in legal practice. These models open new perspectives for assisting legal work. They are capable of synthesizing regulatory information, reformulating legal content, and providing concise answers to legal questions.

These models function as probabilistic models. This leads to limitations such as hallucinations,

algorithmic biases, and inaccurate legal references. These are particularly critical in HSE compliance, where regulatory errors can have significant financial, legal, safety, and organizational consequences. Therefore, the use of AI in the legal field requires rigorous oversight and systematic validation by legal experts.

In the field of HSE compliance, legal departments must deal with large volumes of regulatory texts, technical standards, and institutional guidelines, while ensuring that the information provided to client companies remains accurate, up-to-date, and legally sound. This situation makes HSE legal departments a particularly relevant context for analyzing the contribution and limitations of GAI as a legal support tool. The central question is whether, and under what conditions, GAI can assist legal experts in tasks such as regulatory monitoring, legal research, and legal analysis.

This article examines the integration of GAI into the legal work of a French LegalTech SME specializing in regulatory monitoring and HSE compliance support. It focuses on the use of ChatGPT as a support tool, evaluated through experiments conducted on a database of legal questions extracted from the partner company. ChatGPT was selected as a representative conversational large language model (LLM) in order to examine the integration of GAI into HSE legal workflows. The objective of this study is not to compare different LLM architectures, but to analyse the practical implications and limits associated with the use of an open conversational model in a professional compliance context.

French HSE regulations are structured in conjunction with the European framework. The regulatory requirements set out in European regulations are directly applicable, while European directives are transposed into national law through the adoption of laws, decrees, or orders before being integrated and structured within French legal codes. In the area of HSE,

these provisions are primarily incorporated into the Labor Code and the Environmental Code. This structure influences the compliance practices analyzed in this study.

The research methodology is first presented (2), followed by the evaluations carried out and the results obtained (3). The contributions and limitations of integrating GAI into legal work in the field of HSE compliance are then analyzed (4), before an opening to future perspectives (5).

2. Research Methodology

The applied research methodology is developed in two parts: the research-intervention methodology with the partner company (2.1) and the methodology of the evaluations carried out (2.2)

2.1 Research-intervention methodology

The study adopts an action research methodology, combining observation of professional legal practices with experimentation conducted in real working conditions.

This foundation was also based on interviews with HSE legal experts and an analysis of the workflow of the company's legal department in order to contextualize legal practices, evaluation criteria and organizational constraints.

Based on this empirical data, experiments were conducted to evaluate the contribution of GAI as an assistive tool. The experiments were carried out using ChatGPT. No retraining or fine-tuning of the model was performed, and the partner company's database was not integrated into the model. The evaluation reflects the performance of the LLM in its default configuration.

2.2 Evaluation Methodology

Three evaluations were carried out. A qualitative evaluation of the use of ChatGPT, an evaluation

of the use of prompt engineering and a quantitative evaluation of the use of GAI.

The data used in this study comes from a database of 110 legal questions from the internal legal support of the partner company. The questions used were anonymized and did not contain personal data or confidential client information.

Several levels of complexity exist. For these assessments, two levels of legal complexity were explored. "Level 1" questions correspond to requests for legal information, such as identifying applicable regulatory obligations or clarifying well-defined compliance requirements. "Level 2" questions present a higher level of complexity and require in-depth legal reasoning, including contextual interpretation, consultation of multiple legal sources, and consideration of specific operational constraints.

For each question, the answers generated by ChatGPT were systematically compared to the reference answer provided by the partner company. This ensured their compliance with the company's quality standards and professional practices, and they therefore served as the benchmark for evaluations and were rated 5/5.

A rating scale from 0 to 5 was used to reflect the remaining effort required from the legal expert to validate, correct, or supplement the generated response. A score of 5/5 corresponds to a response that can be used directly as a basis for validation by an expert, while lower scores indicate increasing levels of intervention required by lawyers, down to results unusable in professional practice.

In addition to this rating, a thematic analysis was conducted to identify recurring trends in ChatGPT responses. This analysis is based on several criteria, including legal accuracy, response structure, explanation of concepts, correct legal contextualization of the response, and completeness of the information provided.

The structure of the response is examined in terms of the reformulation of the question, the organization of the argument, and its clarity. Legal accuracy is assessed based on the relevance of the legal references and the correctness of the legal citations.

The correct legal contextualization of the response is analyzed by its ability to fit within the current legislative and regulatory framework, specifying and interpreting the applicable standards. The clarity of the concepts is assessed based on ChatGPT's ability to provide a sufficient level of understanding for non-lawyers.

Finally, the completeness of the response is assessed based on the legal information provided and the possible presence of useful documentation reinforcing the argument.

3. Assessments and results

The results are based on a qualitative assessment of the GAI responses (3.1), an analysis of the contribution of prompt engineering (3.2), and a quantitative assessment of the GAI performance (3.3).

3.1 Qualitative assessment of GAI

An example of a qualitative evaluation was conducted using a question from the partner company's (PC) database, one of the 110 legal questions analyzed in this study. The objective was to assess the professional usability of the responses generated by two ChatGPT LLMs (GPT-3.5 and GPT-4) by comparing them to the reference response produced by the PC's expert lawyers.

The question under consideration is: *"Is the explosion protection document (DRPCE) mandatory when an area with potentially explosive atmospheres (ATEX) is present?"*

The response provided by the PC is based on a precise reformulation of the question, identification of the scope of the ATEX regulations, the application of relevant articles of the French Labor Code (specifically R. 4227-42,

R. 4227-46, and R. 4227-52), as well as institutional references such as the Research and Safety Institute for the Prevention of Occupational Accidents and Diseases (INRS) methodological guides. This response is validated by a double internal review by expert legal professionals and constitutes the reference rated 5/5.

The same question was then presented to LLM ChatGPT-3.5 and ChatGPT-4, with the following prompt: "cite applicable legal references." The generated responses show a progressive improvement in terms of structure and informational depth between versions of the model. However, several significant limitations are observed.

ChatGPT-3.5's responses contain errors in its references to the Labor Code, bordering on legal hallucinations, rendering the answer unusable in its current state. ChatGPT-4 provides a generally correct answer by citing the appropriate regulatory sections and relevant European directives.

Adding a more structured prompt: "- Specify the applicable legal references. - Specify relevant documents from institutions such as the National Research and Safety Institute for the Prevention of Occupational Accidents and Diseases (INRS), the National Institute for Industrial Environment and Risks (INERIS), or the Institute for Radiological Protection and Nuclear Safety (IRSN)." improves the overall quality of the response, particularly through the integration of institutional resources (INRS) and better alignment with the single document for assessing occupational risks (DUERP). The response requires analysis and validation by a legal expert to precisely identify applicable articles, exclude irrelevant references (private sources, standards, foreign sources), and correct inaccurate legal interpretations.

This example highlights that, while the GAI can provide a coherent overview of the ATEX regulatory framework, it does not guarantee the expected legal accuracy or completeness. The value of the GAI lies in structuring the information and reducing the search effort, while expert validation remains essential to ensure safe

use and prevent the risk of legal errors. These observations indicate that hallucinations in this context primarily concern the precision and attribution of legal references rather than the overall thematic identification of the regulatory framework. While the model generally identifies the relevant legal domain, inaccuracies tend to emerge at the level of specific article citation.

3.2 Evaluation of Prompt Engineering

The evaluation of prompt engineering focuses on the impact of prompt structuring on the quality of the legal responses generated.

The prompt structuring was applied iteratively throughout the experiment.

The various iterations of prompts were a crucial part of the evaluations. Improvements were guided by the differences in completeness observed between the responses produced by the GAI and those provided by the PC. These different formulations correspond to more structured levels of knowledge-based systems (Bourreau et al., 2012) and legal ontologies (Vigneron et al., 2012).

Several iterative tests were carried out to determine the limitations of ChatGPT in exploring questions in depth through structure.

From this we were able to see which prompts were effective (3.2.1), which prompts could be reformulated (3.2.2) and which prompts did not work (3.2.3).

3.2.1 Prompts for effective supervision

The following prompts proved effective in all database tests. Of the 110 PC database questions posed to ChatGPT, the responses produced by ChatGPT-4 all complied with the instructions of the following prompts: "- Exclude unofficial sources (private organizations) - Exclude foreign sources - Exclude standards"

The results of these tests show that these prompts effectively frame the search scope for the answer and restrict the GAI to relevant legal sources.

This finding validates the hypothesis that prompt engineering enhances the legal accuracy of the proposed answer by introducing logical filtering instructions that closely resemble the legal reasoning performed by a legal expert.

The question was initially posed with the following prompts determined jointly with the expert lawyers of the partner company:

"Is a document relating to explosion protection mandatory as soon as an ATEX zone is established? - Cite the applicable legal references - Exclude unofficial sources (private organizations) - Exclude foreign sources - Exclude standards"



V.4: The LLM correctly identifies the obligation to implement a document relating to explosion protection and refers to the section of the French Labor Code concerning explosion prevention (articles R. 4227-42 to R. 4227-54). However, it does not mention the European directives that led to the transposition of these provisions into national law, unlike the partner company's response.

Rating: 3/5

This case illustrates that using a generic prompt allows for the identification of the applicable regulatory framework but does not guarantee complete normative coverage as in the reference response.

3.2.2 Prompt reformulation

A test was conducted asking the supervisor to cite the legal references in force under French and European law as follows:

"Is a document relating to explosion protection mandatory as soon as an ATEX zone is established? - Cite the legal references in force under French and European law - Exclude unofficial sources (private organizations) - Exclude foreign sources - Exclude standards"



V.4: The LLM draws on both European Directive 1999/92/EC and the section of the French Labor Code relating to explosion prevention (Articles R. 4227-42 to R. 4227-54), thus offering a more comprehensive regulatory framework than in the previous case. However, referencing an entire section of the Labor Code still necessitates further analysis to precisely identify the relevant articles.

Rating: 3/5

This case shows that reformulating the prompt improves the comprehensiveness of the legal sources used, without eliminating the validation effort required for operational use.

3.2.2 The prompts that have shown limitations

The tested prompts that revealed the limitations of the LLM used were those related to Copyright (3.2.2.1), the focus on a single legal field (3.2.2.2) and the search for increased precision (3.2.2.3).

3.2.2.1 Copyright limitations

The prompt "Cite the content of articles related to the question asked" cannot be used because ChatGPT is subject to the copyright of other sites. It is only able to rephrase the text of articles found in its search using the prompt "Cite current legal references".

3.2.2.2 Framework for a single legal field

It was tested to limit the response to a single targeted field. Here is an example where the request was to cite only references related to the field of occupational health and safety:

"Is a document relating to explosion protection mandatory as soon as an ATEX zone is established? – Cite only articles relevant to occupational health and safety – Exclude references to other legal fields – Exclude unofficial sources (private organizations) – Exclude foreign sources – Exclude standards"



V.4: the model correctly identifies the provisions of the Labor Code relating to the prevention of explosions, but also includes Directive 2014/34/EU, which does not fall under the field of occupational health and safety but under environmental law.

Rating: 3/5

The answer remains usable by a legal expert, but this case illustrates the model's inability to reliably refer to a single legal field, despite explicit constraints formulated in the prompts.

3.2.2.3 The search for increased precision

It was tested to obtain a more precise answer to the question posed previously by requesting him, after his last answer provided, to specify the items precisely as follows:

"Can you provide me with the following details regarding articles R. 4227-42 to R. 4227-54 of the labor code: - Specify which one relates to the scope of the section - Specify which one provides for risk assessment - Specify which one provides for the employer's obligation to create a DRPCE."



V.4: the model correctly identifies the scope of application in article R. 4227-42 and the risk assessment in article R. 4227-46, but erroneously attributes the obligation to create the DRPCE to article R. 4227-54, instead of article R. 4227-52 of the labor code.

Rating: 0/5

This case illustrates that reducing legal reasoning to the article level can decrease the usability of responses and introduce critical errors, even when the overall regulatory framework is correctly identified.

This situation corresponds to a hallucination where the model produces a confident but incorrect legal attribution despite identifying the appropriate regulatory context.

Structuring prompts improves the framing, relevance, and filtering of the legal responses produced. However, the examples also highlight obvious limitations: structuring prompts cannot guarantee the accuracy of the response and can introduce errors when excessive specificity is required.

Its contribution lies in reducing research efforts and better aligning with legal experts' reasoning, while preserving the indispensable role of human validation.

3.3 Quantitative evaluation of the GAI

A quantitative performance analysis was conducted to complement the qualitative assessments and prompt engineering.

The analysis is based on 100 level 1 questions divided into 24 categories of themes in the field of HSE compliance and 10 level 2 questions divided into 8 themes in the same field.

For Level 1 questions, performance varies depending on the HSE topic, such as employer liability and chemical risk. In these categories, the average score is between 3 and 4 out of 5, indicating legally usable answers, but requiring further refinement by the legal expert. Conversely, a topic like the health and safety of workers involved in the transport of hazardous materials has an average score below 2 out of 5, rendering it unusable for the legal expert. For Level 2 questions, the average score for HSE topics such as fire/explosion risk, employer liability, and regulations for Classified Installations for Environmental Protection (ICPE) is around 3 out of 5. These results, while usable, require some additional work from the legal expert to obtain an answer that meets the expected quality.

Quantitative analysis reveals divergent contributions depending on the legal topics and levels of complexity. However, the results confirm that GAI improves efficiency and responsiveness in legal work within the HSE field, while remaining dependent on the supervision of legal experts.

4. Contributions and limitations

The results of this study show that GAI can contribute significantly to legal work in health, safety and environment matters, while presenting structural limitations that prevent its use as a standalone legal solution.

The main contribution of GAI lies in its ability to improve the responsiveness and efficiency of legal tasks requiring intensive information processing. Regulatory monitoring, legal research, text summarization, and initial drafting directly benefit from the rapid processing capabilities of large language models. In these contexts, GAI reduces the time required to access and structure legal information, thus allowing legal experts to focus on higher-value activities such as interpretation, validation, and risk analysis.

However, the results indicate that the performance of the GAI is not uniform. It varies considerably depending on the legal topic and the level of contextual complexity.

One limitation concerns reliability and the risk of error. Despite improvements brought about by the structured design of prompts, the generated responses may contain inaccuracies, incomplete legal references, or hallucinations. Therefore, systematic validation by experts remains an essential requirement for any professional legal use.

From an organizational perspective, these results reinforce the central role of the legal expert. The GAI restructures legal work by redistributing tasks: the GAI takes charge of the preliminary processing of the issue, while legal experts retain responsibility for validation. This redistribution requires an explicit redesign of workflows.

The integration of the GAI has broader organizational and managerial repercussions. Skills development, ethical and professional safeguards, as well as change management within legal teams emerge as key factors. It requires the establishment of a governance charter, shared practices, and managerial oversight, as well as a training framework covering technical, ethical, and supervisory dimensions in order to ensure that

efficiency gains do not come at the expense of legal certainty. (Bensoussan, 2024)

5. Conclusion and perspectives

Integrating GAI into legal work within the HSE field has operational, organizational, and managerial impacts. The results demonstrate that GAI can improve the productivity of regulatory monitoring and legal information retrieval, while exhibiting varying performance levels depending on the specific legal field.

One of the main contributions of this work lies in clarifying the conditions under which GAI creates value in legal services within the occupational health and safety sector. The study shows that this value emerges when GAI is integrated into supervised workflows, supported by prompt engineering and validation by legal experts.

From an organizational perspective, the results highlight the need for a phased deployment strategy. The study underscores the central importance of training and governance as essential conditions for the integration of GAI. Training should not be limited to the use of technical tools but must encompass an understanding of the limitations of GAI, as well as ethical and professional obligations. In parallel, a governance charter outlining the distribution of responsibilities and validation processes is necessary to ensure accountability and maintain trust in legal decision-making.

The results of the various assessments confirm that a progressive and structured integration into HSE legal work can enhance the productivity and responsiveness of the legal department of an SME while creating sustainable organizational value, provided that it is embedded in supervised workflows, supported by a new managerial organization and the implementation of continuous training covering technical, ethical, and deontological dimensions, and that human legal expertise remains central.

Acknowledgment

I would like to thank PREVENTEO for its support of this research and development work.

References

Bourreau, Léa, Thomas Audiffren, Jean-Marc Rallo, Franck Guarnieri (2012). The contribution of knowledge bases to compliance assessment: a case study of industrial maintenance in the gas sector. ESREL Conference 2012.

Audiffren, Thomas, Jean-Marc Rallo, Franck Guarnieri (2012). The contribution of case law to compliance management in Occupational Health and Safety (OHS) in France. ESREL Conference 2012.

Vigneron, Jonathan, Franck Guarnieri, Jean-Marc Rallo (2013). The Contribution of Ontologies to the Creation of Knowledge Bases for the Management of Legal Compliance in Occupational Health and Safety. ESREL 2013.

Bensoussan, Alain (2024). *ChatGPT dans le monde du droit : la cobotique juridique*. Bruylant.