Design, Validation, and Risk Assessment of LLM-Based Generative AI Systems Operating in the Legal Sector

Emanuele Buchicchio
R&D Department
Nomos AI
Napoli, Italy
emanuele.buchicchio@ieee.org

Alessio De Angelis
Department of Engineering
University of Perugia
Perugia, Italy
alessio.deangelis@unipg.it

Antonio Moschitta
Department of Engineering
University of Perugia
Perugia, Italy
antonio.moschitta@unipg.it

Francesco Santoni
Department of Engineering
University of Perugia
Perugia, Italy
francesco.santoni@unipg.it

Lucio San Marco
Editorial Department
Nomos AI
Napoli, Italy
editoriale@nomos-ai.it

Paolo Carbone
Department of Engineering
University of Perugia
Perugia, Italy
paolo.carbone@unipg.it

2024 IEEE International Symposium on Systems Engineering (ISSE) | 979-8-3503-5372-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISSE63315.2024.10741134

***Abstract*—Large Language Models (LLMs) are increasingly capable of supporting user decisions and performing various tasks autonomously in a wide range of professional domains, including the legal sector. In this work, we describe the general lifecycle of LLM-based systems and the architecture of a system designed to support legal experts in the process of generating maxims from judgments. We propose three application-specific performance metrics and a general-purpose assessment method suitable for the comparison of different systems in any content generation task. Finally, we discuss the results achieved by the system prototype that, according to a board of magistrates, outperforms human experts in the task of generating maxims from judgments.**

***Index Terms*—Large language models (LLM), generative AI, legal domain**

I. Introduction

The recent exponential growth in artificial intelligence (AI) applications in industries, businesses, organizations, and households has brought benefits, but also raised concerns regarding potential threats that AI solutions can bring to society. More than 80 bodies around the world have developed ethics guidelines for artificial intelligence (AI) [1]. These guidelines recognize that there is a conflict between the possible advantages of AI systems and their effects on people and society.

Large language models (LLM) and large generative AI models (LGAIMs), such as ChatGPT, GPT-4, or Stable Diffusion, are rapidly transforming the way we communicate, illustrate, and create. Interacting with a contemporary LLM-based conversational agent can create the illusion of being in the presence of a thinking creature. However, in their very nature, such systems are fundamentally not like us [2]. The apparent human-like interaction of some AI systems demands special attention in user trust management, therefore many AI regulations, such as the European AI Act[1], aim to enforce transparency and trustworthiness of these systems.

In this work, we describe the general lifecycle of LLM-based systems and the architecture of a system designed to support legal experts in the process of generating *maxims* from judgments. The *maxim* represents the primary tool for information retrieval in jurisprudence (where jurisprudence refers to the set of court judgments) and corresponds to the *legal principle*, or the *ratio decidendi*, which guides a judge in resolving one or more legal issues. The discovery of jurisprudential information is of utmost importance in the Italian legal context because, on the one hand, precedents increasingly influence judges' decisions; on the other hand, we witness the paradox where the available jurisprudential information, i.e. *maxims*, represents only a small percentage compared to all issued judgements [3].

The design and the operation of a system based on a LGAIM is an inherently interdisciplinary process that involves multiple engineering disciplines. Translating the initial idea into a working product requires a pragmatic systems thinking approach that blends the researcher's result, practice, and domain-specific application domain understanding. Although researchers and open-source communities have proposed numerous application development frameworks or tool components, there is a lack of overall architecture design for LGAIMs systems engineering [4].

The purpose of this research is to assess whether it is possible to develop a system for generating *maxims* of Italian court rulings that achieve a level of quality comparable to that

[1]The EU AI Act was adopted by the Council of the European Union on May 21, 2024. It will be officially published in the EU Official Journal during the second half of July. The final draft of the act is available online. See https://artificialintelligenceact.eu/the-act/

of texts produced by specialized judges. Our goal is to design and develop a functional prototype that demonstrates both the technical feasibility and the economic sustainability of a large-scale process.

The main contributions of this work are the following.

1) We describe with a holistic systems engineering approach [5] the design and the life cycle of an LLM-based system operating in the legal domain
2) We propose performance assessment methods that allow the comparison of different LLMs in any content generation task and three application specific performance metrics
3) The performance of 8 different state-of-the-art LLMs are compared
4) The extensive performance validation process performed by an independent scientific board made up of 59 magistrates through a blind review process produces empirical proof that current LLMs outperform people in the task of generating maxims from judgments
5) We prove the feasibility of an AI-supported maxim generation process at scale.

II. Related Works

Artificial Intelligence and Law began to establish itself as a distinct subfield of research during the 1980s [6] with dedicated conferences, journals, and scientific societies.

Automatic summarization of legal case judgments has traditionally been attempted using *extractive summarization* methods. However, in recent years, the *abstractive summarization* approach has gained popularity in recent works such as [7], [8], as it can generate more natural and coherent summaries.

Despite the great success of LLMs, their performance in high-risk domains, such as legal domain, remains unclear. The ability of recent GPT models, such as GPT-4, to semantically annotate legal texts with respect to previous generation systems is highlighted in [9]. Various use cases, best practices, and concerns about LLMs in legal tasks are discussed in [10].

Very few studies are available on the application of the LLM-based system in the domain of the Italian legal system. An overview of the use of Artificial Intelligence techniques for automated information extraction from Italian case documents is presented in [11]. Several variants and derivatives of the BERT LLM models have been trained and fine-tuning for domain-specific applications such as named entity recognition, sentence classification, semantic similarity , and document classification in the context of Italian legal documents. The results presented in [12], [13], and [14] show that the performances of these models are not comparable to current state-of-the-art LLMs such as those in Tab. I. The application of more advanced LLMs, such as GPT-3.5 and GPT-4 to process Italian legal documents, has been carried out by the PROGIT project [15].

So far, no studies in the existing literature have addressed the feasibility of a large-scale concrete application. With our work, our aim is to fill this gap by employing a systems engineering approach.

III. A LLM-Based Maxim Generation System

Recently, LLM has been widely adopted to refer not only to the generative models themselves, but also to the systems that incorporate them, particularly in the context of conversational agents or AI assistants, such as the popular ChatGPT[2] from OpenAI.

From a systems engineering point of view, it is crucial to maintain the distinction between these two entities, because regulation, metrics apply to the whole system. A fully functional AI assistant system requires many hardware and software components around the LLM model. A complex pipeline is required to handle user inputs and consequential output, in particular in multimodal systems that can parse and produce textual, graphical, audio, or video content.

The bare-bones LLM itself, the core component of an AI assistant, has a highly specific, well-defined function in the whole system context. LLMs are mathematical generative models of the statistical distribution of tokens in the vast public corpus of human-generated text, where the tokens in question include words, parts of words, or individual characters, including punctuation marks. They are generative because we can sample from them, which means that we can ask them questions [16]. But the only question an LLM can answer is "According to your model of the statistics of human language and given this fragment of text, what words are likely to come next?"

The case study system described in this work was designed to support legal experts in the process of generating *maxims* from judgments. The generated content is not limited to the *principle of law* extracted from the judgment but extends to the specific case and the relevant factual elements in the application of the law to the concrete case; and this in-depth analysis also extends to keywords. This particular technique of maxims' extraction is less common in legal practice, but is more valuable ad appreciated by magistrates, as shown by the validation experiment results.

The block diagram of the system prototype is illustrated in Fig.1. The *court judgments* dataset is built from an online archive through a *extraction, transformation, loading (ETL)* process. Depending on the source archive, the process could be fully automated or semi-automated. After some data preparation steps, the judgment documents are transformed in *input documents* to be fed to the LLM-based system. *Input documents* are plain text documents without formatting and structural elements such as HTML and XML tags. The system can use any LLM that exposes a standard API interface.

The LLM elaborates an input document at a time as *user prompt* following the processing instruction given in *system prompt*. The LLM outputs an AI-generated *maxim* document and a computational resource usage report that allow for the calculation of the cost per document of the process.

The system also includes a second pipeline *ETL* that builds a dataset of *maxim* documents from online archives. We use these *maxims*, produced by judges working in a specialized

[2]ChatGPT - https://chat.openai.com/

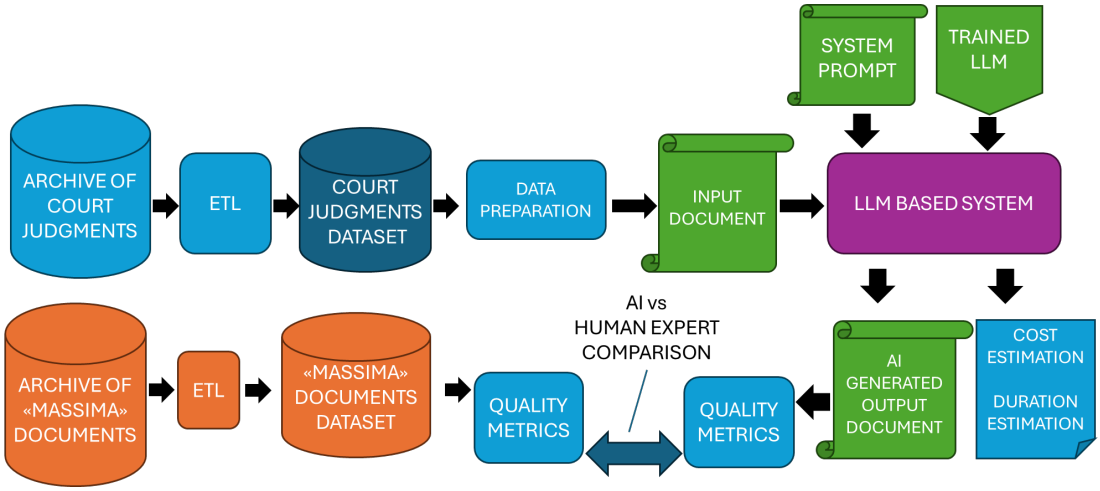


Fig. 1: Block diagram of the LLM-based maxim generation system

branch of the Italian Department of Justice, as a reference *gold standard* to assess the performance of the system.

IV. Lifecycle of LLM-Based Systems

The lifecycle of LLM-based systems, although sharing some similarities with classical software systems and ML systems, requires additional considerations around data handling, model training, ethical implications and continuous learning. Effective management, validation, and risk assessment are crucial for the successful deployment and operation of these advanced systems.

The procedures aimed at streamlining the process of deploying, operating, and maintaining LLM have not yet been fully standardized, and the terminology is not yet harmonized, but there is a set of practices (typically referred to as LLMOps[3] that gained consensus among major IT firms, which can be used as a de facto reference standard.

In this work, we considered five main stages in the LLM-based system lifecycle: development, adaptation, deployment, monitoring, and maintenance. Each stage plays a vital role in the efficient and effective operation of LLMs.

1) **Development**. The LLM foundation model is selected and integrated with other system components. The input and output pipeline are developed to handle the application-specific content format (plain text, XML, JSON, images, video, etc.). Although in theory it is possible to create a foundation model from scratch, this is rarely done because the process demands substantial resources. Closed-source LLM are usually available from their developer through the web API. Open-source LLMs are available through community hubs such as HuggingFace[4] or cloud service providers.
2) **Adaptation**. Adaptation is an iterative process aimed to improve the performance of a machine learning (ML) system. Traditional ML models are usually trained to fit the model parameters to the training dataset. In classic ML systems, the model is trained from scratch or a pre-trained model is fine tuned for the specific application. With LLMs, the adaptation process starts from a *foundation model*. A variety of approaches can be used, and some are specific to LLMs. They include prompt engineering, fine-tuning, embedding, and reinforced learning.
3) **Deployment**. LLM-based systems could be deployed through on-premise, cloud-based, or hybrid solutions. The choice depends on infrastructure considerations such as hardware, software, and networks, as well as the organization's specific needs and security requirements.
4) **Monitoring**. Once the model is available for use, its performance must be continually evaluated for accuracy and biases. This can be accomplished through automated tools, metrics, logs, and alerts that track the LLM while it is in use, ensuring it continues to deliver value to user and meet all compliance requirements.
5) **Maintenance**. Maintenance is an ongoing process that involves fixing bugs and improving performance over time. Given the complexity of LLMs, changes must be tracked. Versioning allows for rollbacks if issues arise and ensures the reproducibility of effective improvements.

Fig. 2 represents a generic LLM process model, including its foundation model (which is shown as a black box). The five basic components are: 1) raw data in the world, 2) inputs, 3) model, 4) inference algorithm, and 5) outputs. Note that in this diagram, processes are shown as ovals, whereas things and collections of things are represented as rectangles.

The most noticeable difference between a classic ML process model and our LLM process model in Fig. 2 is the large black box that hides and leaves out of the systems engineers' control important aspects of the ML process. For example, both the *dataset assembly process* and the *datasets*

[3]Due to space constraints the list of LLMOps referece guideline is available only as additional materials in the companion repository https://electrical-and-electronic-measurement.github.io/LLM-maxim-generation/

[4]https://huggingface.co/models

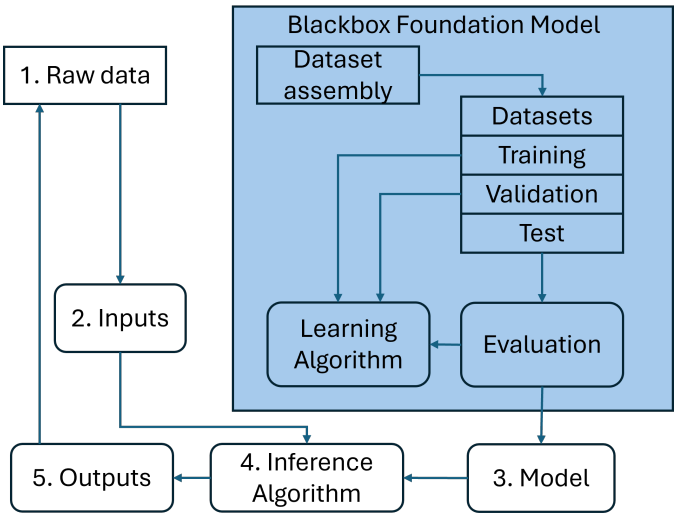


Fig. 2: Components of a generic LLM-based system shown as a process model. Arrows represent information flow.

used in the training of the foundation model are hidden from system engineers, as are the full details of both *learning algorithms* and *evaluation methodologies*. Information about these critical aspects of LLM creation is often incomplete and underspecified, even in some open-source LLMs.

Data play a crucial role in the security and performance of ML systems. In ML systems (and LLMs), when training data are inaccurate, affected by bias, or poisoned with false or inaccurate information, this greatly impacts the performance of the final system.

When a fully trained model is put into production, input data are fed into the trained model. English, which is the main interface language for LLMs, is an ambiguous interface. Natural language can be misleading, making LLMs susceptible to misinformation and manipulation in both training and operations. One of the most important categories of computer security risks is malicious input. The LLM version of malicious input has become known as prompt injection [17]. The stochastic nature of the basic LLM interface exacerbates this control issue.

LLM foundation models already impose restrictions on the prompts, but foundation models are often outside the scope of system engineers. That means that some risk management decisions were made by the LLM foundation model provider and not by the system engineers. Sanity checks and filters integrated into the input and output pipelines, as well as a data cleaning policy, can control this risk.

V. Materials and Methods

A. LLM Selection Criteria

We created a list of candidates whose performance we measured experimentally to select an LLM. The criteria for selecting candidates were as follows:

1) availability of the model in as-a-service (SaaS) mode or on some sort of ”on demand” cloud infrastructure
2) size of *context* compatible with input document size
3) availability of a model version trained on Italian text

These three criteria are related to the amount of technical and economic resources available for system implementation. The SaaS mode allows for easy and cost-effective integration with the system prototype. The availability of a version of the model pre-trained on Italian text does not require to carry on a costly and time-consuming training process of a very big LLM.

An additional criterion is the compatibility with the AI act. This is critical to ensure that the final system can be brought to market by a European-based company. At the time of writing, it is still unclear how big market players such as OpenAI, Microsoft, Meta, and Google will make their foundation LLMs compatible with the requirements of the AI Act such as *data quality* and *transparency*.

For the validation experiment, we selected the last stable release of 8 state-of-the-art LLMs from 3 major providers listed in the Tab. I.

Vendor	Model	Release
OpenAI	GPT4 Omnia	2024-05-13
OpenAI	GPT4 Turbo	2024-04-09
OpenAI	GPT3.5 Turbo	0125
Anthropic	claude3 sonnet	20240229
Anthropic	claude3 haiku	20240307
Anthropic	claude3 opus	20240229
Google	Gemini 1.5 flash	001
Google	Gemini 1.5 Pro	001

TABLE I: Foundational LLM included in the experiment

B. LLM Risk Analysis

Generative artificial intelligence is poised to become a cornerstone of tomorrow’s enterprise architecture, driving innovation and efficiency across industries [18], and organizations that embrace this technology must mitigate key risks to ensure responsible implementation.

As the technological ecosystem rapidly adopts Large Language Models (LLMs), the need for comprehensive threat modeling and risk analysis becomes imperative to safeguard the applications they empower. In [19], a holistic approach to threat modeling is presented, based on the well-known STRIDE and DREAD methodologies. The authors also propose several concrete mitigation strategies tailored for AI systems based on third-party LLM. In [20], a generic machine learning risk analysis framework is tailored to the more specific case of LLMs, identifying an architectural black box with 23 associated risks.

New regulatory institutions related to AI are rapidly expanding. There are specific regulations that deal with activities such as employment, as well as all-encompassing regulations that cover all aspects of AI use [20]. Different jurisdictions, including China, the EU, Japan, the United Kingdom, and the United States, are in varying stages of developing broader regulatory frameworks to address AI governance issues. Tech developers have also begun to explore best practices on their own. For example, Anthropic, Google, Microsoft and OpenAI

have established the Frontier Model Forum[5] to develop standards for AI safety and share best practices across the industry.

LLMs are increasingly capable of supporting user decisions and performing various tasks autonomously in a wide range of professional domains, including legal. However, relying on LLMs for legal application raises concerns due to the significant expertise required and the potential real-world consequences of the advices or decisions. A detailed analysis of the risks associated with the use of LLM-based AI systems in legal applications is available in [21]. The authors, while recognizing the great potential of AI-based systems in legal applications, suggested that involving human-in-the-loop decision-making processes can help mitigate risks and ensure responsible use of this rapidly developing technology.

In the case study system presented in this work, the decision-making process remains a human-guided activity. The aim of the AI-based tool is to support judges in their decision-making process without replacing them in any way.

C. Red Teaming

The purpose of red-teaming exercises is to play the role of the adversary (the red team) and to find hidden vulnerabilities in defenses. AI-based systems are affected by cybervulnerabilities as any other type of software system: they can be hacked by nefarious actors to achieve a variety of objectives including data theft or sabotage. For example, Google[6] uses red teaming to protect its AI models from threats such as prompt attacks, data poisoning, and backdooring. Once such vulnerabilities are identified, they can close the software gaps.

The red team test suite for the proposed system includes a specific test aimed at finding vulnerabilities to LLM-specific attacks such as the example enumerated in [19] and edge cases in the input documents. The edge cases content includes violence, sexual offenses, racism, and discriminatory content. The test suite has been designed to test the response of the system to content that may be censored or limited by the LLM providers but could be present in the text of the court judgement processed by the system. Although offensive content must not be present in the generated document, the output of the system must be clear, complete, and accurate regardless of the presence of offensive content in the input documents.

D. AI Act impact evaluation

The AI Act, proposed by the European Commission, aims to regulate the development and deployment of AI systems within the European Union. The Act categorizes AI systems based on their risk levels and sets specific requirements to ensure safety, transparency, accountability, and non-discrimination. The *Unacceptable risk* category includes harmful uses of AI or uses that violate the values of the European Union and that are consequently prohibited; AI systems that may create an adverse impact on people's safety, health, or fundamental rights under certain circumstances belong to *high risk* category; the *limited risk* classification, apply to some AI systems that are not considered high risk but whose operation shall be informed to the natural persons exposed to them (for example, chatbots or deepfake videos); and *minimal risk*, comprises all other AI systems that can be deployed in the EU without additional legal obligations beyond those already in place [22].

The impact of EU regulation on LLM-based systems has not yet been fully explored. Within the provisions of the recent European Regulation on artificial intelligence, expected to come into effect in July (but applicable from 2026), it can be assumed that the activity carried out in the test is attributable to that of the *deployer*, defined by the Regulation as the entity that uses an AI system under its own authority; conversely, the entity that provides the LLM model is considered, according to the Regulation, a *provider*, as the entity that develops an AI system or a general-purpose AI model and puts such system or model on the market under its own name or brand.

In general, the application of AI in the field of justice represents a significant risk to the fundamental rights of individuals, potentially materially influencing the outcome of a decision-making process and thus being classified under high-risk AI systems: in particular, according to the Regulation, AI systems intended to be used by a judicial authority, or on its behalf, in the research and interpretation of law and in law applying are considered high risk.

The document generation activity (a special case of the *abstractive summarization* task) carried out by the AI case study system described in this work has more the characteristic of a preparatory task for judicial decision-making, which, referring to the wording of the Regulation, is associated to a previous decision-making patterns without replacing human evaluation. In this way, the use of AI tools primarily aims to support the decision-making power of judges without replacing it, leaving the final decision-making a human-guided activity: thus understood, according to the distinction of the Regulation, the activity of generating maxim from court judgements with AI tools would not fall among the high-risk. Therefore, from an initial reading of the Regulation, it can be considered that the system is not to be placed within the scope of a high-risk AI system, for which the Regulation imposes specific obligations, particularly of transparency, on both the *provider* and the *deployer*.

Assuming that the case study system could be included in the class *limited risk*, the EU regulation requirements focus on ensuring *transparency*, *Non-Discrimination*, *Security and Robustness*, *risk management*, and *responsible use*.

1) Transparency: Users must be informed that they are interacting with an AI system and must be informed of the system's capabilities and limitations. This is especially important for situations where it might not be obvious, like chatbots or AI-powered content generation. The proposed system's users are from the editorial staff of Nomos AI, and all users will receive extended training about the intended use and limitations.

[5]https://www.frontiermodelforum.org/

[6]https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/

2) Non-Discrimination: AI systems must not discriminate against users based on sensitive characteristics such as race, ethnicity, gender, sexual orientation, religion or disability. The proposed system focuses on an abstractive summarization of legal principles from judgements. This does not involve specific people or groups. Anyway, the system *deployer* will include a specific test cases in is red team security test suite.

3) Security and Robustness: The AI system should be designed and implemented with security in mind to minimize the risk of malfunctions or damage. Appropriate measures must be taken to prevent these issues, which could potentially harm people or property. The main security measure implemented by the *deployer* is strict human supervision by the editorial staff. All AI generated material will be revised and published under the responsibility of one or more human authors.

4) Risk Management: AI systems need a risk management system in place to identify, assess and mitigate potential risks. A specific risk management strategy, tailored on AI system based on tird party LLM has been defined by the *deployer* and will be in place before the launch of the first commercial application of the system see also Sec.V-B.

E. Prompt Engineering

The construction and refinement of the prompt aimed at maximizing the decisions were achieved through a combination of various Prompt Engineering techniques [23] (*Priming*, *Self-Consistency*, *Least to Most*) and the maximization criteria guidelines published by the *Uffici del massimario*[7].

The model for refining the prompt was ChatGpt 4.0 (updated on May 20, 2024). The Prompt Engineering to achieve the desired result was conducted based on 387 judgements from the Administrative jurisdiction under the supervision of a group of magistrates. Consequently, the choice to conduct the experiment on a dataset of decisions exclusively from the Administrative jurisdiction was dictated by the fact that the decisions of this jurisdiction represented a tested proving ground for prompt engineering; similarly, in the text validation of the judgements, a group of 59 magistrates was involved, who represent - for the reasons described above - the ideal domain experts in evaluating the maxims of the judgements, being the primary users.

F. Validation

The validation experiment involves the generation of about 700 *maxims* (*massime*) and keywords from 90 judgments, specifically from administrative jurisdiction (judgements of *Consiglio di Stato* and *Tribunali Amministrativi Regionali*). The generated *maxims* plus the 90 "official" maxim available from the online archive of the Italian Department of Justice make up the *validation dataset*.

The system validation is carried out by applying the following procedure:

1) AI-generated maxims of sentences are obtained and, for the same sentences, maxims are also produced by specialised magistrates.
2) The quality of the maxims is evaluated in a 'blind' way by experts, through questionnaires. The expert does not know whether the maxim is generated by one of the LLMs in Tab. I or a human magistrate.
3) The results of the questionnaires (5 questions, each with an answer on a scale from 1 to 5, reported in Tab, II) are expressed in an ordinal measurement scale, called the Likert Scale [24]. Such a scale is widely used when defining the results of questionnaires in a variety of fields, including medicine and psychology [25] [26].
4) The scores obtained from the maxims produced by AI are compared with those produced by magistrates. Comparisons are made on both median and dispersion using histograms and box-and-whiskers plots.
5) By summing 3 of the 5 questions, a quality metric *Text Quality score* is obtained, denoted as Q_{TXT}. The remaining questions is of a different nature and must be analyzed separately. Question 4 is about the capability of keyword extraction and is the source of *Keyword Quality Score*, denoted as Q_{KEY}. Question 5 is aimed at assessing the capability of the AI to separate different topics and is the source of the *Multiple Topics Quality Score*, denoted as Q_{MULTI}. The answer to question 5 might not be always present in the questionnaire

Each evaluator, answering the five questions, assigns a set of scores s_{ij} where $j \in [1, \ldots, 5]$ is the question index and $i \in [1, \ldots, N]$ is the evaluation index with $N =$ total evaluation available. The evaluation results were grouped into 9 samples according to the source of the evaluated maxim document. The groups are denoted as "AI system A", "AI system B", "AI system C","AI system D","AI system E", "AI system F", "AI system G", "AI system H", and "Human" (denoted with the code "I" for the blind validation experiment).

Given that a five-value Likert scale is adopted for all questions, the maximum possible score $S_{\max_{ij}} = 5$ for every question and the The metrics *Text Quality Score*, *Keyword Quality Score* and *Multiple Topics Quality Score* are defined as in the equation. 1,2,3 where M is the number of evaluations in which the answer to question 5 has been filled.

$$Q_{\text{TXT}} = \frac{\sum s_{ij}}{\sum s_{\max_{ij}}} \text{ for } j = 1, 2, 3 \text{ and } i = 1, \ldots, N; \quad (1)$$

$$Q_{\text{KEY}} = \frac{\sum s_{ij}}{\sum s_{\max_{ij}}} \text{ for } j = 4 \text{ and } i = 1, \ldots, N; \quad (2)$$

$$Q_{\text{MULTI}} = \frac{\sum s_{ij}}{\sum s_{\max_{ij}}} \text{ for } j = 5 \text{ and } i = 1, \ldots, M; \quad (3)$$

VI. Results

At the time of writing, 630 evaluations were available. Fig. 3 shows the performance of the considered LLMs and

[7]See *Sintesi dei Criteri della Massimazione Civile e Penale* and *Nuovi Standard per il servizio News, Newsletter e massimazione aggiornato al 4 giugno 2024*

Num	Questions (ITA)	Questions (EN)
1	La massima è redatta in modo conciso e chiaro?	Is the maxim written concisely and clearly?
2	La massima individua in modo corretto lo specifico orientamento del giudice?	Does the maxim correctly identify the specific orientation of the judge?
3	La massima descrive correttamente la fattispecie concreta a cui è applicato il principio di diritto?	Does the maxim correctly describe the concrete case to which the principle of law is applied?
4	Se la sentenza ha deciso più questioni di diritto, sono formulate in maniera autonoma tante massime quante sono le questioni decise?	If the judgment decided multiple questions of law, are as many maxims independently formulated as there are questions decided?
5	Le parole chiave sono state estratte in maniera corretta?	Were the keywords extracted correctly?

TABLE II: List of questions in Italian and their English translations

those of the human experts, according to the three quality scores defined in section V-F. Furthermore, Fig.4 shows the distribution of the answer values of Question 1 for the *Human* and *AI system F* groups. The box plots of all the samples for Question 1 are compared in Fig. 5.

The Kruskal-Wallis H test was used to test the null hypothesis that the population medians of all the groups are equal and the result shows that there are statistically significant differences between the medians of the groups. Further comparison of group pairs with the Mann-Whitney U rank test shows that the scores assigned to documents generated by most LLMs are higher than the scores assigned to documents written by human experts. The magnitude of the difference depends on the specific question and LLM considered in the comparison, as shown in Fig. 3. Due to space constraints, some plots it is not possible to include all plots and statistical tests in this paper. All material is available in the companion repository [8].

VII. Discussion

It can be hypothesised that the greater success achieved by some LLMs compared to human-produced maxims, as shown in Fig. 3-5, is due to the use of an engineered and refined prompt, designed to dedicate particular attention to the concrete case. In contrast, in legal practice, attention is not always paid to the concrete case. We think evaluators appreciated the greater depth offered by the LLM models, which was not always present in the maximes created by humans. However, it should be noted that question no. 3 (see Tab. II) specifically aimed to verify whether the maxim correctly describes the concrete case to which the principle of law is applied. Furthermore, the prompt instructs the LLM to identify all legal issues emerging from the judgment, whereas a human expert might deliberately choose to report only one issue in the maxim. Question n.4 specifically addresses the multiple legal issues. In contrast, the better performance achieved by LLM models in questions 1, 2, and 5 should be attributed solely to the higher quality of the maxim produced by LLMs compared to the maxims produced by the human expert.

There are no studies on the time it takes a human expert to write a maxim from a judgment. Anecdotally, we could say that time can move from half an hour to two to three hours for more complex judgments. In the validation experiment, the LLM-based system can complete the whole process from the

[8]Companion repository: https://electrical-and-electronic-measurement.github.io/LLM-maxim-generation/

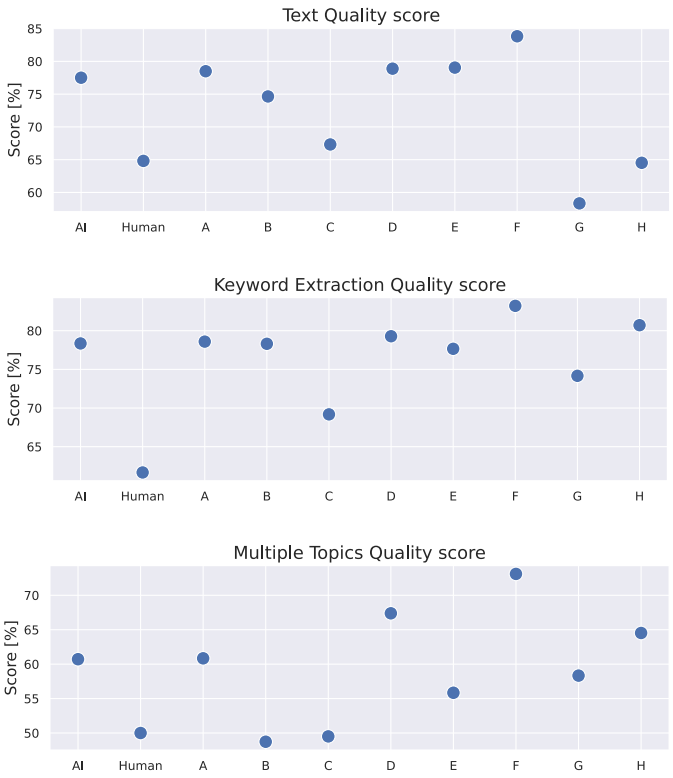
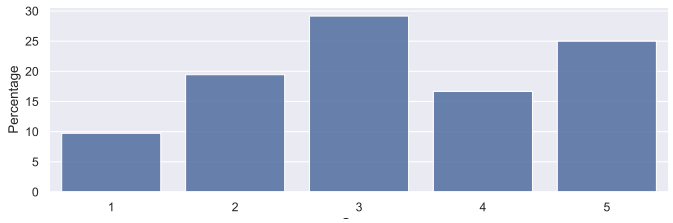
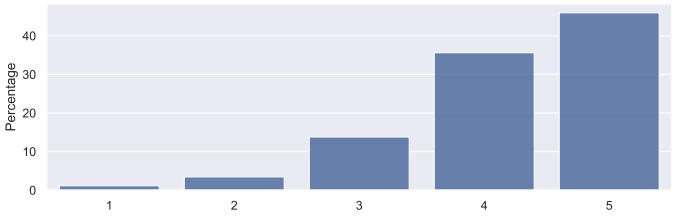


Fig. 3: Comparison of performance of humans and LLMs.



(a) Document written by human expert



(b) Documents generated by LLM system F

Fig. 4: Question 1 evaluation score distribution.

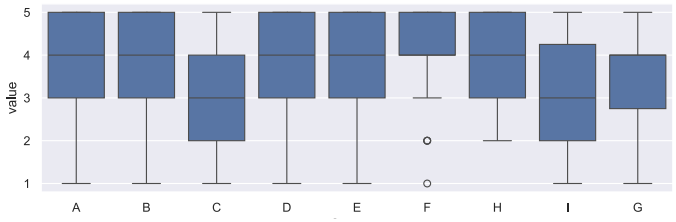


Fig. 5: Boxplot of evaluation scores of question 1. Maxim documents in sample *I* were written by human expert. The document in other samples were generated by one of the 8 LLMs.

judgment to the maxim in less than 1 minute, with limited computation resources.

VIII. Conclusion

The preliminary results indicate that the proposed system is capable of generating maxims that are rated as high quality by industry experts. Managing a generative AI-based system necessitates a bespoke risk management framework, which includes oversight by human domain experts to ensure the system's compliance with regulations. The findings from this case study are promising and suggest that it is feasible to develop and manage such a system in a manner that adheres to EU regulations while keeping the associated risks under control. Further testing will be necessary to select the LLM model that ensures the optimal balance between cost and performance, considering that the quality of the generated document affects the time human experts spend on editorial review before publication.

Acknowledgment

We extend our heartfelt gratitude to Nomos AI and the magistrates of the scientific committee who evaluated the generated maxims. Special thanks go to the committee coordinator, Maurizio Santise, and the other members who agreed to be mentioned in this work. Alessandro Fabbroni, Alessia Annunziata, Antonella Rosato, Antonio Franzese, Carmen Carandente, Claudia Altomare, Elena Sofia Ciccone, Federica Peluso, Federica Sanvenero, Fiorella Todisco, Francesca Bellisario, Giorgia Iurza, Marco Martone, Martina Contieri, Pasquale Pirone, Renata Naso, Rosa Amato, Vittoria Picone, Vittorio Todisco.

References

[1] J. B. Peckham, "An ai harms and governance framework for trustworthy ai," *Computer*, vol. 57, no. 03, pp. 59–68, mar 2024.

[2] M. Shanahan, "Talking about large language models," *Commun. ACM*, vol. 67, no. 2, p. 68–79, jan 2024.

[3] G. Amoroso, "Nomofilachia e massimario," in *Relazione al Convegno L'Ufficio del Massimario e del Ruolo della Corte di Cassazione: il presente che guarda al passato per pensare al futuro, a cura della Struttura decentrata della Corte di Cassazione*, 2017.

[4] W. Chen, L. Yan-yi, G. Tie-zheng, L. Da-peng, H. Tao, L. Zhi, Y. Qing-wen, W. Hui-han, and W. Ying-you, "Systems engineering issues for industry applications of large language model," *Applied Soft Computing*, vol. 151, p. 111165, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623011833

[5] N. Hutchison, "The guide to the systems engineering body of knowledge (sebok), v. 2.10," 2024, accessed July 2024. [Online]. Available: www.sebokwiki.org.

[6] T. Bench-Capon, "Thirty years of artificial intelligence and law: editor's introduction," *Artificial Intelligence and Law*, vol. 30, no. 4, pp. 475–479, 2022.

[7] A. Deroy, K. Ghosh, and S. Ghosh, "How ready are pre-trained abstractive models and LLMs for legal case judgement summarization?" in *3rd Int.l Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace, LegalAIIA*, 2023.

[8] J. Yoon, M. Junaid, S. Ali, and J. Lee, "Abstractive summarization of korean legal cases using pre-trained language models," in *Int.l Conf. on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, 2022.

[9] J. Savelka and K. D. Ashley, "The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[10] D. Zhang, A. Petrova, D. Trautmann, and F. Schilder, "Unleashing the power of large language models for legal applications," in *International Conference on Information and Knowledge Management*, 2023.

[11] P. Bushipaka, D. Licari, G. Marino, G. Comandé, and T. Cucinotta, "Ai-assisted legal holding extraction," in *Proc. of Italia Intelligenza Artificiale - Thematic Workshop*. CINI National Lab, may 2023.

[12] A. Tagarelli and A. Simeri, "Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code," *Artificial Intelligence and Law*, vol. 30, no. 3, pp. 417–473, 2022.

[13] I. Benedetto, L. Cagliero, F. Tarasconi, G. Giacalone, and C. Bernini, "Benchmarking abstractive models for italian legal news summarization," in *Legal Knowledge and Information Systems*, 2023, pp. 311–316.

[14] D. Licari and G. Comandè, "Italian-legal-bert models for improving natural language processing tasks in the italian legal domain," *Computer Law & Security Review*, vol. 52, 2024.

[15] T. D. Pont, F. Galli, A. Loreggia, G. Pisano, R. Rovatti, and G. Sartor, "Legal summarisation through llms: The prodigit project," *arXiv preprint arXiv:2308.04416*, 2023.

[16] M. Shanahan, "Talking about large language models," *Commun. ACM*, vol. 67, no. 2, p. 68–79, jan 2024. [Online]. Available: https://doi.org/10.1145/3624724

[17] G. McGraw, H. Figueroa, K. McMahon, and R. Bonett, "An architectural risk analysis of large language models: Applied machine learning security," Berryville Inst. Mach. Learn. (BIML), Tech. Rep., 2024.

[18] M. Campbell and M. Jovanovic, "Disinfecting ai: Mitigating generative ai's top risks," *Computer*, vol. 57, no. 05, pp. 111–116, may 2024.

[19] S. B. Tete, "Threat modelling and risk analysis for large language model (llm)-powered applications," 2024. [Online]. Available: https://arxiv.org/abs/2406.11007

[20] G. McGraw, R. Bonett, H. Figueroa, and K. McMahon, "23 security risks in black-box large language model foundation models," *Computer*, vol. 57, no. 04, pp. 160–164, apr 2024.

[21] F. Schilder, "Legal expertise meets artificial intelligence: A critical analysis of large language models as intelligent assistance technology," in *Int.l Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace, LegalAIIA 2023*, 2023.

[22] I. Hupont, M. Micheli, B. Delipetrev, E. Gomez, and J. Garrido, "Documenting high-risk ai: A european regulatory perspective," *Computer*, vol. 56, no. 05, pp. 18–27, may 2023.

[23] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. D. Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, and P. Resnik, "The prompt report: A systematic survey of prompting techniques," 2024. [Online]. Available: https://arxiv.org/abs/2406.06608

[24] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[25] D. L. Paulhus, "Two-component models of socially desirable responding." *Journal of personality and social psychology*, vol. 46, no. 3, p. 598, 1984.

[26] G. M. Sullivan and A. R. Artino Jr, "Analyzing and interpreting data from likert-type scales," *Journal of graduate medical education*, vol. 5, no. 4, pp. 541–542, 2013.