Evaluating Open and Proprietary Large Language Models in Law Interpretation: The Case of the EU VAT Directive

Areti Karamanou[1] [0000−0003−2357−9169], Evangelos Kalampokis[1] [0000−0003−4416−8764], Fotios Fitsilis[2] [0000−0003−1531−4128], Georgios Theodorakopoulos[3], and Konstantinos Tarabanis[1] [0000−0002−4663−2113]

[1] Information Systems Lab, Department of Business Administration, University of Macedonia, Thessaloniki, Greece {akarm,ekal,kat}@uom.edu.gr
[2] Department for Scientific Documentation and Supervision, Hellenic Parliament, Athens, Greece fitsilisf@parliament.gr
[3] Hellenic State Legal Council, Athens, Greece
g.theodorakopoulos@nsk.gr

Pre-print Notice: This is a pre-print version of the following paper: Karamanou, A., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., Tarabanis, K. (2025). Evaluating Open and Proprietary Large Language Models in Law Interpretation: The Case of the EU VAT Directive. In: Janssen, M., et al. *Electronic Government. EGOV 2025*. Lecture Notes in Computer Science. Springer, Cham.

Abstract. Large Language Models (LLMs) are expected to exert unprecedented impact on the public sector. Specifically in the field of legal interpretation, the ability of LLMs to utilize the wealth of public sector information, such as legal texts and regulations, enables more efficient and accurate legal analysis and improves decision-making. However, when involving LLMs in such critical domains, it is important to ensure that they are trustworthy and, hence, they produce accurate responses. This study aims to explore and evaluate the trustworthiness of LLMs in interpreting law. Towards this direction, an exploratory case study is presented that engages nine proprietary and open (either open-weight or open-source) LLMs from four families, namely Claude, GPT, Mistral, and Llama, in answering a set of questions related to the European Union's VAT directive that have been selected by a legal professional. The questions, all of them of legal nature and varying complexity, are provided as prompts to the LLMs. Their responses are evaluated based on their legal precision. The results show significant insights, contributing to the development of more trustworthy and responsible AI systems, and ensuring their safe and effective use in critical areas such as law and public policy.

Keywords: Large Language Models · Trustworthiness · Law Interpretation · Public Sector.

1 Introduction

It is estimated that generative Artificial Intelligence (AI) could elevate the global Gross Domestic Product (GPD) by 7% over the next decade and, at the same time, bring forth automation to approximately 300 million jobs worldwide, contributing to the global economy $2.6 trillion to $4.4 trillion annually [6]. AI has long found its way into the public sector [33], yet the impact of utilizing Large Language Models (LLMs) will likely be unprecedented. LLMs utilize the wealth of information provided by the public sector, from documents across different departments, ministries, and local authorities to Open Government Data portals to enable, for instance, the deployment of chatbots and virtual assistants [4]; the analysis of documents for identifying key information in complex documents such as legal contracts [21]; the summarization of large volumes of text; and the assistance in decision-making by generating reports and evaluating applications and grants [14]. The introduction of LLMs can also lead to more proactive and data-driven public interventions, improving outcomes in areas such as public health, urban planning, and disaster response [24, 26].

Existing literature reveals both significant progress and persistent challenges in applying LLMs to legal interpretation tasks [25]. While general capabilities have advanced considerably, domain-specific applications—particularly for complex regulatory frameworks remain an active area of research.

Our work builds upon this foundation by examining the comparative performance of LLMs in the specific context of the European Union's Value Added Tax (VAT) Directive interpretation, addressing a gap in the literature regarding model efficacy in specialized legal domains. Towards this direction, an exploratory case study is presented and analyzed that involves nine proprietary and open (either open-weight or open-source) LLMs from four major families, namely Claude, GPT, Mistral, and Llama. The LLMs are tasked with answering a set of questions related to the EU VAT Directive. The questions, which are of varying complexity and of legal nature, are provided as prompts to the LLMs, and their responses were evaluated based on their legal precision.

The rest of the article is structured in the following sections. In the background section (Section 2) an examination of foundational concepts related to LLMs is presented (Section 2.1) along with an exploration of the LLM trustworthiness (Section 2.2), and a brief presentation of the EU VAT directive (Section 2.3). Section 3 presents the related work with recent research in the field. The Research Approach (Section 4) presents the methodology employed to assess the LLM truthfulness and presents an overview of the set of questions related to the EU VAT direction that were used to assess the trustworthiness of LLMs. The evaluation results are then presented in (Section 5). Finally, the Section 6, concludes this work providing key findings, contributions to the field, and directions for future research.

2 Background

2.1 Large Language Models

Foundation models refer to large, pre-trained models that serve as a starting point for other models. They are trained on massive datasets for extended periods of time, something that requires vast amounts of computational resources, and results in state-of-the-art performance.

In the field of natural language processing, ever since the invention of the revolutionary transformer architecture, many LLMs have been created and released. These include, for example, OpenAI's GPT-3.5 [37] and, the more recent, GPT-4 [2], LLMs from the Llama family of Meta [31] and Llama 2 [32], Mistral AI's Mistral [11] and Mixtral 8x7B [12], and Google's LaMDA [30]. LLMs vary in terms of openness and user access. Proprietary models such as Claude and ChatGPT are closed systems as neither their internal architecture nor their training data or model weights are publicly disclosed. They are only accessible through controlled interfaces, typically via API. In contrast, open-source models, including some from the Mistral family, provide both the source code and model weights under permissive licenses, allowing users to inspect, adapt, and deploy the models independently. A third category, represented by the Llama series, includes open-weight models, which make the trained parameters publicly available while withholding access to training data and full source code.

LLMs have proven to be extremely capable but also tremendously costly to train. As a result, efficient methods to harness their power have been explored including prompt engineering, prompt tuning, and fine tuning. Prompt engineering, uses techniques like chain of thought reasoning [35] and in-context learning [5] to design effective prompts in order to elicit desired responses from the LLM. It is the only technique that doesn't change the weights of the LLM. Prompt tuning [16], adds trainable parts to the input layer of the LLM and trains them in order to have them act as conceptual prompts for the model by guiding its predictive abilities to a certain direction. Finally, fine-tuning can be adapter or full parameter. Adapter fine-tuning [10] adds trainable components to the inner network, allowing for greater adaptation of the initial model to new tasks. Regarding full-parameter fine tuning, supervised fine tuning is commonly employed, while Reinforcement Learning (RL) methods also exist, such as RL from Human Feedback (RLHF) [27] and Reinforcement Learning from AI Feedback (RLAIF).

Even though LLMs hold considerable amounts of world knowledge, thanks to their initial training, they still lack the ability to be factually correct in all their responses, which often results in "hallucinations" [40], i.e., factually incorrect responses presented by the LLM as correct ones. In order to tackle this weakness, several methods have been developed, mainly utilizing prompt engineering. Among those methods is the supply of factually correct context along with the input. The main, state of the art, proposed architecture supporting this is Retrieval Augmented Generation (RAG). In RAG, specialized components retrieve information relevant to the original input from corpora and supply it to the model along with it. This has been shown to boost the factual capacities of the models significantly.

2.2 Trustworthiness in Large Language Models

The trustworthiness of Large Language Models (LLMs) has been subject of many previous published scientific works. In their effort to understand and evaluate the trustworthiness of LLMs, these works usually result in synthesizing holistic frameworks with definitions for all the different dimensions of trustworthiness as well as benchmarks that can be used to assess proprietary and\or open-weight LLMs for trustworthiness. Examples include the TrustLLM [28], HELM [18], DecodingTrust [34], Halueval [17], and other frameworks and benchmarks.

LLM trustworthiness frameworks and benchmarks identify various aspects of trustworthiness, including truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, and accountability [28]. *Truthfulness*, the most common and obvious aspect of LLM trustworthiness, is included in all related frameworks and benchmarks, and is used to assess whether or how often the LLMs respond with factually correct information in users' prompts. *Safety* assesses whether the LLM produces unsafe responses (e.g., illegal answers). In addition, *Fairness* assesses whether the responses provided by the LLM are fair, impartial, and not affected by specific race, gender, political, and other ideologies. *Robustness*, evaluates how well an LLM performs on edge cases (e.g., when user prompts include ambiguous or misleading texts). In addition, *Privacy Transparency* assesses the ability of the LLM to provide references and data sources in its responses. Finally, *explainability* evaluates to what extend is the LLM capable of providing reasoning that justifies its responses.

This work focuses on assessing the trustworthiness of LLMs based on the most common dimension, namely truthfulness.

2.3 The European Union's Value Added Tax Directive

The European Union's Value Added Tax (VAT) Directive (Council Directive 2006/112/EC) [1] establishes the framework for harmonizing the national VAT laws across the European Union. However, it allows some flexibility for the member states. The directive defines the main concepts that are related to the VAT (e.g., taxable persons, related transactions, and exceptional cases). At the same time, it also sets out rules for the VAT rates that should be utilized across the EU, as well as for the VAT reporting and collection processes. Among the main objectives of the directive is to minimize tax distortions in cross-border trading transactions and, at the same time, ensure a fair competition across the EU member states' markets.

The Directive outlines the place of supply regulations, which defines where and which goods and services are subject to VAT charges in the EU. This is very important when it comes to cross-border trade in Europe. The Directive also allows, in specific cases, for the adoption of the reverse charge mechanism, which reduces the possibility of tax fraud by transferring the burden of paying VAT from the seller to the buyer. Additionally, the Directive provides an overview of particular programs designed for specific sectors (e.g., digital services, small enterprises, and second-hand products trade). Although the Directive provides a a single legal framework for the VAT, it also provides the member states with some freedom in applying VAT rates and exemptions. As a result, there are some differences in the national implementations across the members of the European Union.

The Directive was introduced in November 2006 and, since then, has undergone numerous amendments. The most recent amendment was made by Council Directive (EU) in 2022 and aimed at updating the VAT rate provisions. In its latest consolidated form, the Directive consists of over 400 articles structured into 14 titles.

3 Related Work

LLMs are being applied across various public sector functions, demonstrating their versatility and potential to enhance administrative processes, and there is a growing body of research attempting to highlight various aspects of LLM implementation and application in the public sector.

For instance, one study examines the implementation of Intelligent Public Sector Automation (IPSA) based on LLMs for digital innovation in Korean public administration, focusing on critical challenges and their solutions [39]. Other research investigates the use and impact of LLMs on the management of Swedish and Croatian public records [29]. Further work explores the initial results of using a chatbot based on a LLM to address user queries about the Management and Performance Program, a Brazilian federal government initiative aimed at enhancing public sector efficiency by modernizing workforce management practices established in 2020 [15].

Due to the special responsibility of parliaments, governments, and administrations as the organizational instances of society, and through the inherent legitimation by society itself, there is a necessity to examine the implications of the use of generative AI within these institutions and traditional structures as well as their influence on political system logic [20]. Evidently, generative AI has emerged as a key technology in the parliamentary workspace, with 37% of the recorded use cases in 2024 involving such systems [8]. These systems are being used for tasks such as document summarization, legislative drafting, and multilingual translation. In the Judicial branch, structured multi-LLM setup can significantly improve decision-making accuracy, particularly in ambiguous situations, by harnessing the synergistic effects of diverse LLM arguments [13].

As an emerging field, the application of LLMs in the legal field is still in its early stages, with multiple challenges that need to be addressed. However, it has been found that the performance of models grounded in the Generative Pre-trained Transformer (GPT) architecture has consistently improved across various legal domains, including contract review, legal document summarization, and case outcome prediction [25]. Still several challenges exist related to the quality of legal data used for fine-tuning and training, as well as ensuring consistency in legal reasoning. High-quality, well-curated data is essential for successful LLM adoption in the legal domain. Researchers have collected a large amount of legal domain data and combined it with general domain data, using GPT-4 Turbo to build a high-quality legal dataset [36]. This again necessitates meticulous data preparation protocols. Beyond preparation, effective data governance is equally important, and developing and maintaining rigorous management strategies within Public Data Ecosystems (PDE) ensures alignment with societal, regulatory, and technical requirements in the digital era [23].

In law, reasoning is of particular significance and relevant benchmarks are being developed such as LegalBench, which was built through an interdisciplinary process, in which its makers defined tasks designed and hand-crafted by legal professionals [9]. Further research is also exploring moral and ethical aspects of LLM-driven legal reasoning [3]. However, in legal drafting, a European Commission study found that while tools like GPT-4 offer significant advantages in supporting the law-making process, many related functionalities can be effectively implemented using traditional AI techniques without relying solely on LLMs [7].

4 Research Approach

In order to explore the truthfulness of LLMs in interpreting law to answer legal questions, we utilized an exploratory case study [38]. Since LLMs are emerging technologies across various fields, including the legal one, research on their trustworthiness is still developing. An exploratory case study is particularly suitable, as a research approach, for the in-depth exploration of the field in order to enable gaining insights, explore new ideas, and identify unknown patterns.

The study focuses on nine LLMs from four families, namely Claude, ChatGPT, Mistral, and Llama. Each LLM represents a distinct case for analysis. The LLMs are selected based on their relevance in legal research, availability, and varying architectures. The majority of the selected LLMs are proprietary (five out of nine) and the rest of them are either open-weight or open-source models. Regarding the size of open LLMs, 8B, 70B, and 405B were selected for the Llama family in order to test their behavior regarding trustworthiness. The Mistral model's size is not known. An overview of the nine LLMs their access type, and specifications is presented in Table 1.

A legal expert from the Hellenic State Legal Council, highly skilled in EU tax laws, created a set of 19 progressively complex questions covering key aspects of the EU VAT Directive. The set of questions is presented in Table 2.

Each question was given as a prompt to each of the nine LLMs in order to be answered following a zero-shot approach, i.e., using its existing knowledge. To this end, inferences from LLM were obtained via SageMaker or Bedrock Amazon Web Services (AWS). Each LLM was configured with a temperature of 0, a top-p value of 0.9, and a maximum output token limit of 512, based on its level of parameterization. This combination of low temperature and increased top-

Table 1. Overview of access types and model specifications for the investigated Large Language Models (LLMs).

Access Type	Company	Model Series	Model Name	Size (B)	Release Date
Proprietary	Anthropic	Claude	Claude v3 Opus	Unknown	2024-Feb-29
Proprietary	Anthropic	Claude	Claude v3 Sonnet	Unknown	2024-Feb-29
Proprietary	OpenAI	GPT	GPT 4o	Unknown	2024-May-13
Proprietary	OpenAI	GPT	GPT 4 Turbo	Unknown	2024-Apr-09
Proprietary	OpenAI	GPT	GPT 3.5 Turbo	Unknown	2024-Jan-25
Open source	Mistral AI	Mistral	Mistral Large-2402	Unknown	2024-Feb
Open weight	Meta	Llama 3.1	Llama 3.1 8b Instruct	8	2024-Jul-23
Open weight	Meta	Llama 3.1	Llama 3.1 70b Instruct	70	2024-Jul-23
Open weight	Meta	Llama 3.1	Llama 3.1 405b Instruct	405	2024-Jul-23

P allows for creativity and, at the same time, increased reproducibility of the responses [22]. The responses were then anonymized to prevent biases based on the LLMs.

While accuracy is the most commonly used metric for assessing the truthfulness of LLM responses, involving humans in the evaluation process has also been recognized in literature [19]. In this context, this study involved the legal expert Hellenic State Legal Council to evaluate the anonymized LLM responses with regards to truthfulness. Specifically, the expert assigned to each LLM response a score ranging from one to ten, with one representing the lowest accuracy and ten the highest accuracy. All scores were then statistically analyzed so as to understand the performance of each LLM with regards to trustworthiness. In order to compare the LLMs, the Cohen's Kappa coefficient was calculated to measure agreement between their responses.

5 Results

The set of questions (Table 2) were given as prompts to the nine LLMs. Their responses were then evaluated by the legal expert to assess whether the LLM provides factually correct information based on the EU VAT Directive.

The average truthfulness score of each LLM is presented in Figure 1. The heatmap shows a generally high evaluation of responses across different LLMs with average truthfulness scores falling between 7.4 and 8.6, indicating overall strong performance. The top five LLMs related to truthfulness core are in descending order GPT 4 Turbo (8.63, 95% CI: 8.12, 9.15), GPT 4o (8.58, 95% CI: 8.14, 9.01), Llama 3.1 405B (8.45, 95% CI: 7.9, 9), Claude v3 Sonnet (8.16, 95% CI: 7.73, 8.9), and Llama 3.1 70b (8.21, 95% CI: 7.66, 8.76). Llama 3.1 8B (7.37, 95% CI: 6.43, 8.3) and Mistral Large (7.95, 95% CI: 7.3 , 8.6) achieved the lowest scores. Among the evaluated LLMs with open weights, Llama 3.1 with the 405 billion parameters provided the most accurate answers.

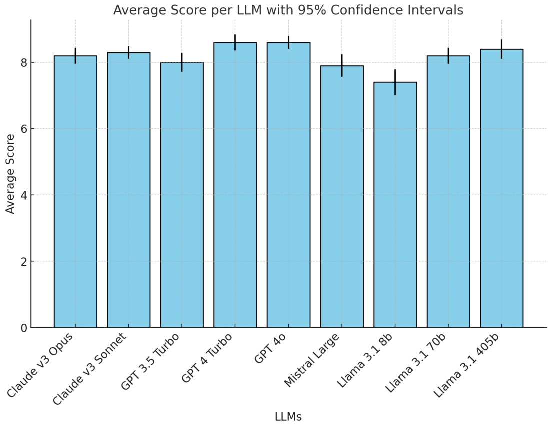

Fig. 1. Truthfulness scores for all LLMs (95% confidence interval).

The individual truthfulness scores for all LLM responses are presented in Figure 2. According to the heatmap, some LLMs seem to be more stable in their truthfulness scores, while others show more variability. Specifically, the GPT 4 LLMs are the most stable related to the trustworthiness of their responses, having a few low truthfulness scores in their responses. The Claude-3 variants have in general a good performance in truthfulness, although they have some inconsistencies (e.g., Opus was rated with a 5 on Q13). Llama 3.1 8B faces an increased variation across different questions. However, its responses in some questions achieved very good scores. For example, it was evaluated with truthfulness score of 10 for questions Q3 - Q5, and with 9 for questions Q7 and Q8. This indicates that there is a potential for improvement for the specific model, for example, by fine-tuning it using domain related data and, hence, enabling it to provide more accurate responses. At the same time, the large LLM of the family (Llama 405b) performs similar to GPT-4 LLMs, having clearly improved and more stable responses regarding trustworthiness. This indicates that scaling significantly improves results' accuracy.

If we place the focus to the questions, certain questions (e.g., Q2, Q6, Q10, and Q13) show more disagreement across LLMs, suggesting that these may be harder or more ambiguous to answer. Conversely, it can be observed that there are two questions (i.e., Q3 and Q4) that achieved the highest score (10) across all LLMs. This may be indicate that these questions are straightforward and, hence, less complex for LLMs to answer. Finally, Q6, which refers to the VAT treatment of farmers, is directly related to a specific exemption case and not all not align well on niche tax treatment. This emphasizes the need to go deeper and assess the specific aspects of trustworthiness identified in literature (see Section 2.2) including robustness, i.e., how well the LLM responses on questions regarding exceptional cases, which is extremely important in tasks like law interpretation.

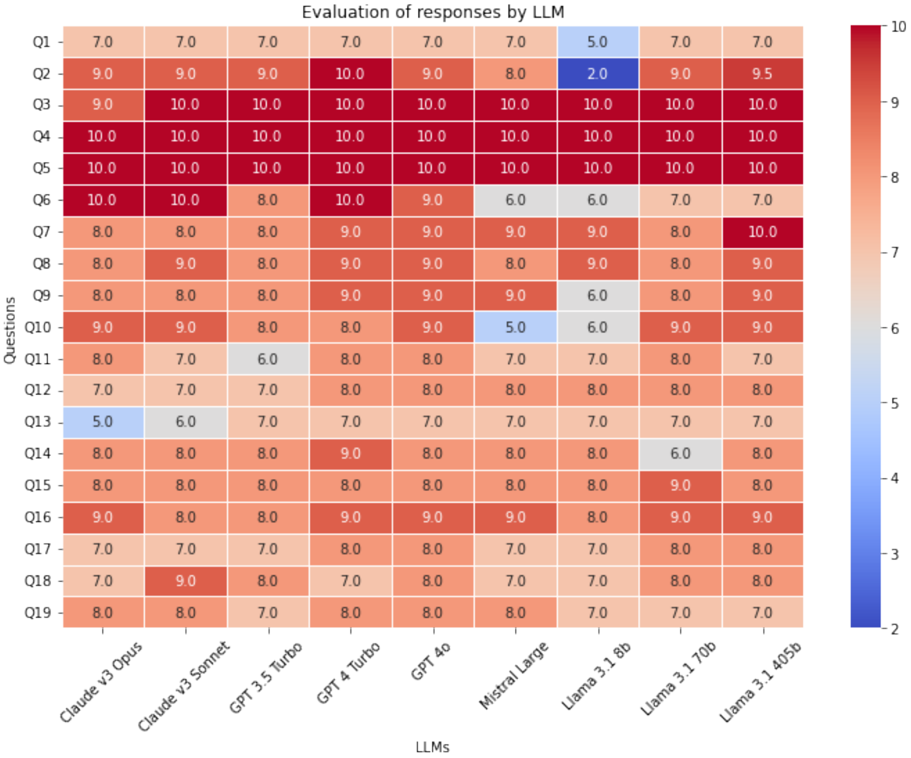

Fig. 2. Evaluation of the truthfulness of LLM responses.

Thereafter, the agreement of the different LLMs regarding the evaluation of their responses was evaluated responses' using the Cohen's Kappa coefficient agreement scores (Figure 3). The score is a value ranging from 0 to 1. Higher values of the Cohen's Kappa coefficient indicate strong agreement between LLMs, while lower values suggest weaker agreement (or, in some cases, even no agreement). Based on the results, the highest score achieved was 0.65. This highest agreement score is observed between GP4o and Llama 3.1 405B, meaning that these two LLMs have a similar behavior in producing accurate responses. This result can be translated as that Llama 3.1 405B has made a lot of progress related to the smaller LLMs of the family, and has a good improvement potential. In addition, based on the same results, LLMs from the same family have increased agreement scores when compared to the agreement between LLMs from different families. For example, the second highest agreement score is 0.64 between

GPT 4 Turbo and GPT 4o, which is followed by the agreement between the two Anhropic's LLMs (Claude v3 Opus and Claude v3 Sonnet). All remaining scores were lower indicating a higher degree of variability in the answers each model selected.

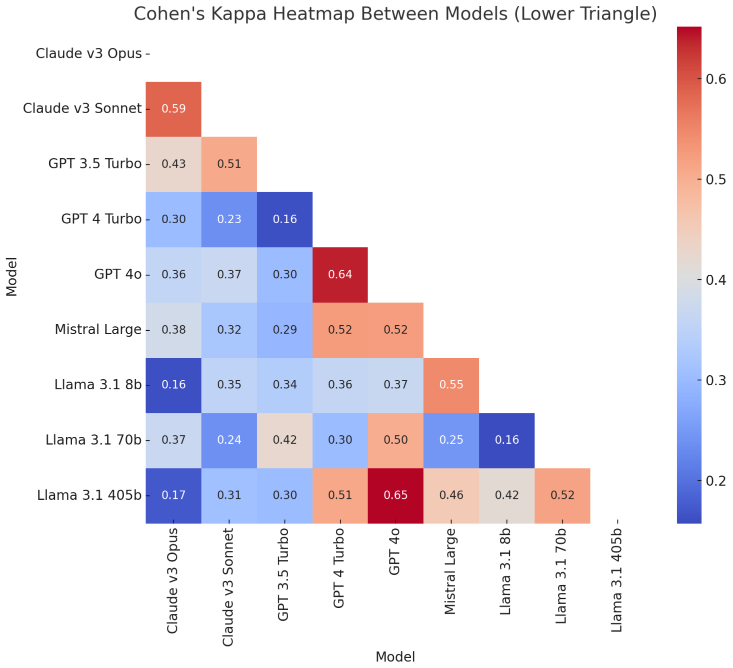

Fig. 3. Evaluation of the correlation of LLMs using Cohen's Kappa coefficient agreement scores.

The higher agreement between the LLMs from the same families possibly mirrors the anticipated impact of common architectures, training methodologies, and\or similar optimization processes. However, when comparing models from different families, such as, for example Claude v3 Opus and GPT 4 Turbo or Claude v3 Opus and Mistral Large, the agreement scores drop significantly (0.30 and 0.38 respectively). This indicates that different these LLMs are applying distinct ways for producing when responses, likely due to variations in training data or optimization goals. Finally, smaller models (e.g., Llama 3.1 8b) show lower alignment with mainstream models, underscoring the impact of model size and training diversity on evaluation consistency.

The above findings emphasize the need for creating and using standardized evaluation benchmarks across various LLM families. Since different LLMs are producing varying responses, human oversight and domain-specific fine-tuning remain crucial in applications where consistency is essential, such as the legal domain.

6 Conclusion

LLMs are increasingly integrated into governmental operations worldwide, significantly impacting public administration and service delivery. In this context, the legal reasoning and accuracy of generated text are crucial for providing precise, tailored, and, hence trustworthy responses to highly specialized legal or administrative inquiries.

In this context, this exploratory case study engaged nine proprietary and open (either open-weight or open-source) LLMs for evaluating their truthfulness in interpreting laws. A set of 19 progressively complex questions related to key aspects of the European Union's Value Added Tax (VAT) Directive were selected by a legal expert, highly skilled in EU tax laws. Each question was provided to the nine LLMs and their responses were evaluated by a legal expert with respect to truthfulness, i.e., how accurate is the response.

The best truthfulness score was achieved by the two GPT 4 variants (GPT 4 Turbo and GPT 4o). Interestingly, the third highest score was achieved by Llama 3.1 405B, which performed significantly better than Llama 3.1 8B indicating that scaling significantly improves results' accuracy. Although Llama 3.1 8B, the smallest model of the Llama family, achieved the lowest score, it had a good performance on certain questions, indicating its potential for improvement with methods like fine-tuning using well known benchmarks of trustworthiness. The evaluation of the agreement of LLMs showed an increased agreement between GP4o and Llama 3.1 405B, meaning that the LLMs have similar behavior in producing accurate responses. Finally, when it comes to questions that represent exceptional cases in the law, LLMs do not align well on providing the correct response.

A promising direction for future research is to enhance open LLMs by applying methods like fine-tuning or retrieval-augmented generation (RAG), making them more effective and better suited to legal use cases. Despite the main limitation of the present study that uses a single score to assess trustworthiness, it still provides significant insights. Next steps will include performing a more detailed evaluation of the LLMs, based on multiple aspects of trustworthiness identified in literature.

Acknowledgments. This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers" (Project Number:2412).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Council directive 2006/112/ec of 28 november 2006 on the common system of value added tax. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32006L0112 (2006), official Journal of the European Union, L 347, 11.12.2006, p. 1–118
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Almeida, G.F., Nunes, J.L., Engelmann, N., Wiegmann, A., de Araújo, M.: Exploring the psychology of llms' moral and legal reasoning. Artificial Intelligence **333**, 104145 (2024)
4. Androutsopoulou, A., Karacapilidis, N., Loukis, E., Charalabidis, Y.: Transforming the communication between citizens and government through ai-guided chatbots. Government Information Quarterly **36**(2), 358–367 (2019). https://doi.org/https://doi.org/10.1016/j.giq.2018.10.001, https://www.sciencedirect.com/science/article/pii/S0740624X17304008
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K.: The economic potential of generative ai (2023)
7. European Commission: Directorate-General for Digital Services, Fitsilis, F., Mikros, G.: Ai-based solutions for legislative drafting in the eu: Summary report. Tech. rep., Publications Office of the European Union, Luxembourg (2024)
8. Fitsilis, F.: Aspects of artificial intelligence in parliamentary governance. In: Fernandes, J., Martínez-Cantó, J. (eds.) Democracy at the Crossroads: Challenges for Governance and Representation (Essays in Honour of Thomas Saalfeld). Routledge, London (2025)
9. Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al.: Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems **36**, 44123–44279 (2023)
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
11. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 **10** (2023)
12. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
13. Jung, S., Jung, J.: Courtroom-llm: A legal-inspired multi-llm framework for resolving ambiguous text classifications. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 7367–7385 (2025)
14. Kalampokis, E., Karacapilidis, N., Tsakalidis, D., Tarabanis, K.: Understanding the use of emerging technologies in the public sector: A review of horizon 2020 projects. Digit. Gov.: Res. Pract. **4**(1) (Apr 2023). https://doi.org/10.1145/3580603, https://doi.org/10.1145/3580603

15. Kleiman, F., Barbosa, M.M.: Management and performance program chatbot: A use case of large language model in the federal public sector in brazil. Digital Government: Research and Practice (2024)
16. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
17. Li, J., Cheng, X., Zhao, X., Nie, J.Y., Wen, J.R.: Halueval: A large-scale hallucination evaluation benchmark for large language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing
18. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
19. Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3214–3252 (2022)
20. Lucke, J.V., Frank, S.: A few thoughts on the use of chatgpt, gpt 3.5, gpt-4 and llms in parliaments: Reflecting on the results of experimenting with llms in the parliamentarian context. Digital Government: Research and Practice (2024)
21. Mamalis, M.E., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., Tarabanis, K.: A large language model agent based legal assistant for governance applications. In: International Conference on Electronic Government. pp. 286–301. Springer (2024)
22. Mehandru, N., Miao, B.Y., Almaraz, E.R., Sushil, M., Butte, A.J., Alaa, A.: Evaluating large language models as agents in the clinic. NPJ digital medicine **7**(1), 84 (2024)
23. Nikiforova, A., Lnenicka, M., Milić, P., Luterek, M., Rodríguez Bolívar, M.P.: From the evolution of public data ecosystems to the evolving horizons of the forward-looking intelligent public data ecosystem empowered by emerging technologies. In: International Conference on Electronic Government. pp. 402–418. Springer (2024)
24. Pulapaka, S., Godavarthi, S., Ding, S.: Empowering the public sector with generative ai (2024)
25. Siino, M., Falco, M., Croce, D., Rosso, P.: Exploring llms applications in law: A literature review on current legal nlp approaches. IEEE Access (2025)
26. Sirait, E., Zuiderwijk, A., Janssen, M.: The readiness of the public sector to implement ai: A government-specific framework. In: Janssen, M., Crompvoets, J., Gil-Garcia, J.R., Lee, H., Lindgren, I., Nikiforova, A., Viale Pereira, G. (eds.) Electronic Government. pp. 302–316. Springer Nature Switzerland, Cham (2024)
27. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. Advances in neural information processing systems **33**, 3008–3021 (2020)
28. Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., et al.: Trustllm: Trustworthiness in large language models. arXiv preprint arXiv:2401.05561 **3** (2024)
29. Svard, P., Seljan, S.: The use of language models (llms) in the public sector and the impact on public records: A case of sweden and croatia. Atlanti **34**(2), 53–72 (2024)
30. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
31. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

32. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
33. van Noordt, C., Misuraca, G.: Artificial intelligence for the public sector: results of landscaping the use of ai in government across the european union. Government Information Quarterly **39**(3), 101714 (2022). https://doi.org/10.1016/j.giq.2022.101714
34. Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al.: Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In: NeurIPS (2023)
35. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
36. Yao, S., Ke, Q., Wang, Q., Li, K., Hu, J.: Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In: Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering. pp. 108–112 (2024)
37. Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al.: A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. arXiv preprint arXiv:2303.10420 (2023)
38. Yin, R.K.: Case study research: Design and methods, vol. 5. sage (2009)
39. Yoon, S., et al.: Digital innovation in public administration through intelligent public sector automation (ipsa): Strategies and challenges. Journal of Multimedia Information System **11**(4), 249–260 (2024)
40. Zhou, C., Neubig, G., Gu, J., Diab, M., Guzman, P., Zettlemoyer, L., Ghazvininejad, M.: Detecting hallucinated content in conditional neural sequence generation. arXiv preprint arXiv:2011.02593 (2020)

Table 2. The set of 19 questions covering key aspects of the EU VAT Directive that were provided as prompts to the LLMs

	Question
Q1	Can Member States adopt practices setting limits as regards to exercising VAT deduction, according to the EU VAT Directive? Justify the answer taking into account existing EU case law and the Advocate General's opinion.
Q2	Can a taxable person deduct VAT paid for purchasing goods or services, in the case this person exercises both for economic and non-economic activities, according to the EU VAT Directive?
Q3	Can tax fraud, tax evasion or other illegal practices influence the exercise of the right to deduct VAT, according to the EU VAT Directive? Justify the answer taking into account existing EU case law and the Advocate General's opinion.
Q4	Clarify the VAT tax obligations for taxable persons (both natural and legal) as outlined in the EU VAT Directive. Organize these obligations into categories.
Q5	Generate the content a model invoice relying on the elements outlined in the EU VAT Directive.
Q6	How are farmers treated by the EU VAT Directive? Identify deviations compared to other taxable persons.
Q7	How is fiscal neutrality interpreted by the EU case law and the Advocate General's opinion, and on which legal provisions of the EU VAT Directive is it based on?
Q8	Identify areas of public interest in the context of the EU VAT Directive and clarify their influence on the VAT implementation.
Q9	Identify the different treatments of small-sized enterprises for VAT reasons according to the EU VAT Directive.
Q10	Identify the place of supply of goods for the purposes of applying VAT in accordance with the EU VAT Directive. Classify the place of supply based on the criteria defined in the Directive for each of the categories of goods.
Q11	Identify the transitional VAT provisions or regimes according to the EU VAT Directive: a) specific to each Member State and b) applicable regardless of a specific Member State.
Q12	On which specific occasions does the EU VAT Directive grant Member States the discretion to establish their own deviation for VAT regulations? Identify the occasions associated with specific articles of the EU VAT Directive and provide all the relevant requirements for each one.
Q13	Provide a definition of the term 'legal certainty', exclusively based on the EU case law and the Advocate General's opinion interpreting the EU VAT Directive.
Q14	Provide specific circumstances under which taxable persons can deduct VAT that they have already paid on goods or services they have supplied, according to the EU VAT Directive.
Q15	Provide the basic principles of the EU common VAT system.
Q16	What is the impact of EU customs legislation on VAT legislation, according to the EU VAT Directive?
Q17	Which are the transactions that fall into the scope of the EU VAT Directive?
Q18	Which persons can be considered taxable taking into account the criteria and requirements defined for each category of natural or legal persons in the context of the EU VAT Directive? Identify the specific criteria for each one of the categories of taxable persons in this Directive.
Q19	Which transactions remain out of the scope of the EU VAT Directive?