SAVIA: Artificial Intelligence in support of the lawmaking process

Michele **Visciarelli**[1,*,†], Giovanni **Guidi**[1,†], Laura **Morselli**[1,†], Domitilla **Brandoni**[1], Giuseppe **Fiameni**[2], Luisa **Monti**[3], Stefano **Bianchini**[3] and Cosimo **Tommasi**[3]

[1]*CINECA, via Magnanelli 6/3, Casalecchio di Reno (B0), 40033, Italy*

[2]*NVIDIA AI Technology Center, Milan , Italy*

[3]*Assemblea Legislativa Emilia Romagna, viale Aldo Moro 50, Bologna, 41127, Italy*

Abstract

We explore the use of open-source Large Language Models (LLMs) to support legal professionals, lawmakers, and citizens in accessing information on the current and past legislation of the Emilia-Romagna region. We develop a generative AI tool based on the Retrieval-Augmented Generation (RAG) technique to answer questions related to regional laws and their implementing acts, retrieving relevant information from the Emilia-Romagna law corpus. To adapt pre-trained LLMs to this downstream task, we follow a multi-step approach. First, we use the QLoRa technique to quantize and adapt the pre-trained LLMs to the regional legal text dataset. Next, we fine-tune the domain-adapted models using an "ad-hoc" instruction-based dataset. We then implement a module to retrieve relevant contextual information from the legal documents dataset. Finally, we align the models with domain-specific instructions using RAG-based prompting. We evaluate the performance of the domain-adapted models using the perplexity metric, and the results of the final fine-tuned models are assessed by domain experts, focusing on the quality of the generated text and the relevance of the answers. Our results show that domain adaptation on domain-specific text is a crucial step for enhancing the quality of the generated text in expert domains, such as legal texts, which contain a vast amount of specialized vocabulary and expressions. This approach leads to higher performance compared to models fine-tuned only on small Question-Answer datasets. Additionally, our findings highlight the importance of the retrieval module, which must be able to reliably find the most relevant documents to provide useful and up-to-date insights to lawmakers and citizens.

Keywords
Generative AI, LLM, Legal AI, NLP

1. Introduction

In the last years the interest for Generative Artificial Intelligence (Generative AI) applications has grown importance among the research and industry community, thanks to the introduction of Foundation Models in different AI domains, such as text generation (GPT-series [1, 2, 3], LLaMA-series [4, 5, 6], MEGATRON [7]), image generation (DALL·E [8, 9, 10], Stable Diffusion [11]), and video generation (Sora [12]). The progress in Deep Learning modelling has been fostered by important advancements in the neural network (NN) research, such as the introduction of the Transformer architecture in Nat-

ural Language Processing (NLP) [13], by improvements in hardware acceleration for linear algebra, expanding model's size up to several billions of parameters, by the introduction of quantization techniques allowing training of large NNs even on consumer GPUs [14, 15], and by the release of large high-quality and open-source datasets [16].

Large Language Models (LLMs) for text generation have achieved remarkable performance and great interest even outside the research and industry professional community, in particular after the release of ChatGPT to the public [17]. Despite their success, the use of LLMs on domain-specific Question-Answer (QA) tasks still face several challenges, that hinder their spread beyond the research community, especially when applied to tasks for which explainability and high quality responses are of paramount importance [18].

Some of the challenges that LLMs are still facing are the following:

- difficulty of maintaining up-to-date knowledge
- costs of training and inference of large models, costs and difficulty to collect large amount of high-quality domain-specific data
- hallucinated answers, i.e. answers that provided false information without warning

- out-of-date or generic answers, even when the user expects a specific, current response

Retrieval-Augmented Generation or RAG has recently emerged as a paradigm to address such challenges [19]. In particular, RAG combines a language model with an information retrieval system to dynamically fetch relevant external information to enhance the model's responses, by encoding the user's question into a dense representation, and retrieving passages relevant to the question from an indexed data source, adding this information to the LLM prompt. Different studies have shown that RAG enhances the quality of the generation process, leading to higher accuracy, better robustness, reduced hallucinations, higher interpretability, and even the possibility to perform open-domain QA just by updating the knowledge-base [20, 21]. RAG also offers a balanced approach in terms of customization and resource requirements, being more flexible and cost-effective than full fine-tuning, although still requiring labeled data and a supervised training phase.

In this work we present SAVIA, a project developed by CINECA and Assemblea Legislativa of Emilia-Romagna. This project, started in Autumn 2023 and expected to end in march 2025, has the goal of creating a model capable of answering questions on the Region's law and their respective implementing acts, as well as on related "ex-ante" and "ex-post" reports of the laws' impact. In Section 2 we present the data used for this project and the workflow that has been adopted. In Section 3, we describe the procedure and the details of the experiments and tests conducted, and in Section 4 we show the obtained results. Our conclusion are then presented in Section 5.

2. Methodology

To obtain a model capable of understanding Italian language in the law domain and responding to questions related to laws enacted in the Emilia-Romagna region, we followed a multi-step approach. We started from an open-source LLM and adapted it to the legal language through unsupervised domain adaptation (Section 2.1). The resulting domain-adapted model was then fine-tuned for question-answering (Q&A) on an instruction-based dataset prepared by domain experts for this purpose (Section 2.2). Finally, we implemented a domain-adapted retrieval model (Section 2.3) to enrich the answers with relevant information from the law corpus.

The full workflow was reproduced starting from different open-source LLMs:

- **LLaMAntino-2-7b-hf-ITA**: a 7B model, based on LLama-2, specifically fine-tuned for the Italian language [22].

- **Mistral-7B-v0.1**: a 7B model, that implements grouped-query and sliding window attention, Rotary Position Embedding, that can handle context of arbitrary size [23].
- **Mixtral-8x7B-Instruct-v0.1**: a 46.7B mixture of experts model, trained on instructions in English, French, Italian, German and Spanish, with maximum context length of 32k [24].

Domain experts qualitatively evaluated the performances of the final models obtained from the different pre-trained LLMs.

2.1. Unsupervised Domain-Adaptation

The first step in the procedure was the domain-adaptation of the model on legal text. We collected the PDFs of the regional laws of Emilia-Romagna, as well as the relative implementing acts at the regional level (e.g. "atto del dirigente" and "atto di giunta") and the available reports on the expected and measured impact of a given law (e.g. "clausola valutativa", "ex-ante and "ex-post" reports). We split the legal documents in chunks, and we implemented a cleaning pipeline to remove typos, bad characters, and irrelevant parts of the documents such as headers and footers. We also added mii-llm/gazzetta-ufficiale [25] in the training dataset, given the affinity of this dataset to our application, both in language, semantics and type of documents. We did not perform domain-adaptive tokenization [26], using instead the pre-trained models native tokenizers to tokenize the legal corpus.

Not all the three model under investigation underwent domain adaptation. LLaMAntino-2-7b-hf-ITA and Mistral-7B-v0.1 were adapted, while Mixtral-8x7B-Instruct-v0.1, after tests regarding its native capabilities of producing adequate Italian legal text, has not been domain adapted.

2.2. Model Alignment on Instruction-Based dataset

With the support of domain experts, we generated an Q&A dataset mimicking different levels of domain language proficiency, ranging from questions that could by written by non-expert users, to the ones that may be asked by experts in the legal domain. We developed a semi-automatic procedure to further enrich this Q&A dataset, using legal documents metadata. The following is an example included in the instruction-based dataset:

- Q: "Da quando è stata istituita la regione, quali normative sono state adottate per incentivare la partecipazione?"
- A: "La prima legge regionale riguardante la partecipazione ad essere stata approvata è la legge numero 3 del 2010. In seguito, la legge numero 3

del 2010 è stata abolita e sostituita con la legge regionale numero 15 del 2018.",

To fine-tune the domain-adapted LLMs, we used the instruction-based dataset prepared by the domain experts. For the loss function computation, we removed the portion of the text containing the prompt, as in many cases the prompt added by the RAG module can account for up to 50% of the total text length. This approach helped to optimize the training process more effectively.

2.3. Domain-Adapted Retrieval Model

To enrich the user's question with relevant information from the legal documents database, we develop a retrieval module based on Semantic Similarity search technique. We used a Sentence-BERT model [27] to populate a vector store with embedding generated from the legal documents text chunks. The content similar to a user's question is retrieved using the semantic search library FAISS [28, 29].

3. Experiment

The project has been carried out exploiting the computational resources of the supercomputer LEONARDO, hosted by CINECA. Each node in the booster partition is equipped with four NVidia A100 SXM6 64GB GPUs and a single 32-cores Intel Ice Lake CPU.

For all models, only data parallelism has been employed, given the fact that all these models could adequately fit in the VRAM of the GPUs at our disposal. For the same reason, LLaMAntino-2-7b-hf-ITA and Mistral-7B-v0.1 have not been quantized during domain adaptation and instruction fine-tuning, opting to preserve the weights' precision. Mixtral-8x7B-Instruct-v0.1 underwent 4-bit quantization instead [30], due to its size. For domain adaptation and instruction fine-tuning, we applied LoRA adapters on Q, K, V layers of the models [15]. The training procedure for the models under study were the following:

- pre-trained LLMs causal Language Modelling on the legal text chunks. This has been performed on LLaMAntino-2-7b-hf-ITA and Mistral-7B-v0.1 and needed for each model, on average, 400 GPU hours (approximately 4 days on a single LEONARDO booster node) training for four epochs;
- model alignment on domain adapted LLaMAntino-2-7b-hf-ITA and Mistral-7B-v0.1, and on base pre-trained Mixtral-8x7B-Instruct-v0.1, on the QA dataset. This step required approximately 96 GPU hours, or 24 node hours (4 GPUs per node), to complete a 12 epochs training run, on a single LEONARDO node.

Table 1
Perplexity for base and domain-adapted models under study.

Model	Perplexity
LLaMAntino-2-7b-hf-ITA	**5.2447**
LLaMAntino-2-7b-hf-ITA domain-adapted	**1.1789**
Mistral-7B-v0.1	**5.2254**
Mistral-7B-v0.1 domain-adapted	**2.3618**

4. Results

To evaluate the quality and proceed to select the candidate for instruction-based fine-tuning, all domain-adapted models were evaluated using the perplexity metric (PPL) on an held-out evaluation dataset based on laws. The metric is reported in Table 1.

Three different domain experts (lawmakers of Assemblea Legislativa) were asked to evaluate the answers generated by the final instruction fine-tuned models to a set of 25 questions. The qualitative analysis of the experts reported that, in general, the answers provided by the LLaMAntino-based model were considered too short and dry. The answers provided by the Mixtral-based model were considered the most complete, clear and satisfactory in terms of quality of the used specific words. Below we report an example of answers provided by the different final models on a given question. For context, we also include the answer of chatGPT (3.5) to the same question.

- Question: Sul tema della partecipazione, quali leggi sono state fatte in Emilia-Romagna?
- *Answer of Mixtral-8x7B-Instruct-v0.1 fine-tuned*: La prima legge regionale approvata in tema di partecipazione è la legge regionale 9 febbraio 2010, n. 3, intitolata "NORME PER LA DEFINIZIONE, RIORDINO E PROMOZIONE DELLE PROCEDURE DI CONSULTAZIONE E PARTECIPAZIONE ALLA ELABORAZIONE DELLE POLITICHE REGIONALI E LOCALI". Questa legge è stata successivamente abrogata e sostituita con la legge regionale 22 ottobre 2018, n. 15, denominata "LEGGE SULLA PARTECIPAZIONE ALL'ELABORAZIONE DELLE POLITICHE PUBBLICHE. ABROGAZIONE DELLA LEGGE REGIONALE 9 FEBBRAIO 2010, N. 3". La legge regionale attualmente in vigore in materia di partecipazione è la legge regionale 22 ottobre 2018, n. 15.
- *Answer of LLaMAntino-2-7b-hf-ITA domain-adapted + instruction fine-tuned*: - Legge Regionale 9 febbraio 2010, n. 3 - Legge Regionale 22 ottobre 2018, n. 15
- *Answer of Mistral-7B-v0.1 domain-adapted + instruction fine-tuned*: In relazione alla partecipazione, le leggi regionali sono state le seguenti:

- Legge Regionale 3 del 2010 - Legge Regionale 15 del 2018

- *Answer of chatGPT-3.5*: Non ho accesso a informazioni in tempo reale o a dati specifici riguardanti leggi promulgate dopo gennaio 2022. Tuttavia, posso fornirti un'informazione generale sul tema della partecipazione politica in Italia. In Italia, molte regioni hanno promulgato leggi e normative per favorire la partecipazione dei cittadini alla vita politica e amministrativa. Queste leggi possono riguardare vari aspetti, come la consultazione pubblica su decisioni amministrative, la promozione della trasparenza nelle istituzioni, o l'istituzione di forme di democrazia partecipativa a livello locale. Per quanto riguarda l'Emilia-Romagna, una delle regioni più attive nel promuovere la partecipazione cittadina, potrebbero essere state adottate diverse leggi e iniziative in questo ambito. Alcune di queste potrebbero riguardare la partecipazione ai processi decisionali locali, la promozione della trasparenza e dell'accesso alle informazioni pubbliche, o la creazione di strumenti e piattaforme per coinvolgere attivamente i cittadini nelle decisioni che li riguardano. Per ottenere informazioni specifiche sulle leggi relative alla partecipazione in Emilia-Romagna dopo il 2022, ti consiglio di consultare le fonti ufficiali della Regione Emilia-Romagna, come il sito web istituzionale o i comunicati stampa delle autorità regionali. In alternativa, potresti contattare direttamente gli uffici regionali competenti per ottenere informazioni aggiornate sulle leggi e le iniziative in materia di partecipazione politica e amministrativa.

5. Conclusions

We explored different approaches to adapt open-source LLMs for question-answering on the Emilia-Romagna law corpus. We adapted the different LLMs on a corpus composed of the Emilia-Romagna regional laws and the relative implementing acts, and we further refined the domain-adapted models on a custom QA dataset provided by domain experts. Finally, we exploited RAG to enrich the user's question with relevant contextual information extracted from the law database.

We experimented with different open-source LLMs, such as Mistral-7B-v0.1, LLaMAntino-2-7b-hf-ITA, Mixtral-8x7B-Instruct-v0.1. Our results show that domain-adapted LLMs that are able to answer specific domain questions can be a helpful tool to support decision-making in specialized fields such as the legal domain, that often need to retrieve exact, concise and easy-to-understand information from large and unstructured data sources.

Given the scope and the length of the project, several improvements to the workflow are foreseen in the near future, as well as the possibility to test with more pre-trained open source models, for example new Italian-native models that will be developed in the near future and domain adaptation of Mixture-of-Experts models (such as Mixtral-8x7B-v0.1). Our future work will also focus on further improving the retrieval module with better embedding models, and on applying more powerful techniques to train the LLMs, such as Direct Preference Optimization (DPO, [31]) and Reinforcement Learning from Human Feedback (RLHF, [32]).

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., Language Models are Few-Shot Learners, arXiv e-prints (2020) arXiv:2005.14165`arXiv:2005.14165`, `doi:10.48550/arXiv.2005.14165`.

[2] M. Chen, J. Tworek, H. Jun, Q. Yuan, et l., Evaluating Large Language Models Trained on Code, arXiv e-prints (2021) arXiv:2107.03374`arXiv:2107.03374`, `doi:10.48550/arXiv.2107.03374`.

[3] OpenAI, J. Achiam, S. Adler, S. Agarwal, et al., GPT-4 Technical Report, arXiv e-prints (2023) arXiv:2303.08774`arXiv:2303.08774`, `doi:10.48550/arXiv.2303.08774`.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, et al., LLaMA: Open and Efficient Foundation Language Models, arXiv e-prints (2023) arXiv:2302.13971`arXiv:2302.13971`, `doi:10.48550/arXiv.2302.13971`.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, et al., Llama 2: Open Foundation and

Fine-Tuned Chat Models, arXiv e-prints (2023) arXiv:2307.09288`arXiv:2307.09288`, `doi:10.48550/arXiv.2307.09288`.

[6] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, et al., Code Llama: Open Foundation Models for Code, arXiv e-prints (2023) arXiv:2308.12950`arXiv:2308.12950`, `doi:10.48550/arXiv.2308.12950`.

[7] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, et al., Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM, arXiv e-prints (2021) arXiv:2104.04473`arXiv:2104.04473`, `doi:10.48550/arXiv.2104.04473`.

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, arXiv e-prints (2021) arXiv:2102.12092`arXiv:2102.12092`, `doi:10.48550/arXiv.2102.12092`.

[9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv e-prints (2022) arXiv:2204.06125`arXiv:2204.06125`, `doi:10.48550/arXiv.2204.06125`.

[10] Z. Shi, X. Zhou, X. Qiu, X. Zhu, Improving Image Captioning with Better Use of Captions, arXiv e-prints (2020) arXiv:2006.11807`arXiv:2006.11807`, `doi:10.48550/arXiv.2006.11807`.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, arXiv e-prints (2021) arXiv:2112.10752`arXiv:2112.10752`, `doi:10.48550/arXiv.2112.10752`.

[12] OpenAI, Video generation models as world simulators, Tech. rep., OpenAI (2024).

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, arXiv e-prints (2017) arXiv:1706.03762`arXiv:1706.03762`, `doi:10.48550/arXiv.1706.03762`.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv e-prints (2021) arXiv:2106.09685`arXiv:2106.09685`, `doi:10.48550/arXiv.2106.09685`.

[15] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, arXiv e-prints (2023) arXiv:2305.14314`arXiv:2305.14314`, `doi:10.48550/arXiv.2305.14314`.

[16] H. Face, Hugging face datasets (2016). URL https://huggingface.co/datasets

[17] OPENAI, Introducing chatgpt, Tech. rep., OpenAI (2022).

[18] L. Huang, W. Yu, W. Ma, W. Zhong, et al., A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, arXiv e-prints (2023) arXiv:2311.05232`arXiv:2311.05232`, `doi:10.48550/arXiv.2311.05232`.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv e-prints (2020) arXiv:2005.11401`arXiv:2005.11401`, `doi:10.48550/arXiv.2005.11401`.

[20] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, et al., Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering, arXiv e-prints (2022) arXiv:2210.02627`arXiv:2210.02627`, `doi:10.48550/arXiv.2210.02627`.

[21] P. Zhao, H. Zhang, Q. Yu, Z. Wang, et al., Retrieval-Augmented Generation for AI-Generated Content: A Survey, arXiv e-prints (2024) arXiv:2402.19473`arXiv:2402.19473`, `doi:10.48550/arXiv.2402.19473`.

[22] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, LLaMAntino: LLaMA 2 Models for Effective Text Generation in Italian Language, arXiv e-prints (2023) arXiv:2312.09993`arXiv:2312.09993`, `doi:10.48550/arXiv.2312.09993`.

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, et al., Mistral 7B, arXiv e-prints (2023) arXiv:2310.06825`arXiv:2310.06825`, `doi:10.48550/arXiv.2310.06825`.

[24] M. A. team, Mixtral of experts, Tech. rep., Mistral AI (2023).

[25] E. Federici, M. Ferraretto, N. Landro, Gazzetta Ufficiale: A dataset of legislative texts, public and private acts (2024). URL https://huggingface.co/datasets/mii-llm/gazzetta-ufficiale

[26] M. Liu, T.-D. Ene, R. Kirby, C. Cheng, et al., ChipNeMo: Domain-Adapted LLMs for Chip Design, arXiv e-prints (2023) arXiv:2311.00176`arXiv:2311.00176`, `doi:10.48550/arXiv.2311.00176`.

[27] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv e-prints (2019) arXiv:1908.10084`arXiv:1908.10084`, `doi:10.48550/arXiv.1908.10084`.

[28] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (3) (2019) 535−547.

[29] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). `arXiv:2401.08281`.

[30] H. Face, bitsandbytes, Tech. rep., Hugging Face (2023).

[31] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct Preference Optimization: Your Language Model is Secretly a Reward Model, arXiv e-prints (2023) arXiv:2305.18290`arXiv:2305.18290`, `doi:10.48550/arXiv.2305.18290`.

[32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, et al., Training language models to follow instructions with human feedback, arXiv e-prints (2022) arXiv:2203.02155`arXiv:2203.02155`, `doi:10.48550/arXiv.2203.02155`.