LongDoc Summarization using Instruction-tuned Large Language Models for Food Safety Regulations

Guido Rocchietti[1,5,*], Cosimo Rulli[1], Korbinian Randl[2], Cristina Ioana Muntean[1], Franco Maria Nardini[1], Raffaele Perego[1], Salvatore Trani[1], Manos Karvounis[3] and Jakub Janostik[4]

[1]ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
[2]Department of Computer and Systems Sciences Stockholm University Postbox 7003, SE-164 07 Kista, Sweden
[3]Agroknow, Greece
[4]Digicomply, Romania
[5]Department of Computer Science, University of Pisa, Largo B. Pontecorvo, 3 56127 Pisa, Italy

Abstract

We design and implement a summarization pipeline for regulatory documents, focusing on two main objectives: creating two silver standard datasets using instruction-tuned large language models (LLMs) and finetuning smaller LLMs to perform summarization of regulatory text. In the first task, we employ state-of-the-art models, Cohere C4AI Command-R-4bit and Llama-3-8B, to generate summaries of regulatory documents. These generated summaries serve as ground-truth data for the second task, where we finetune three general-purpose LLMs to specialize in high-quality summary generation for specific documents while reducing the computational requirements. Specifically, we finetune two Google Flan-T5 models using datasets generated by Llama-3-8B and Cohere C4AI, and we create a quantized (4-bit) version of Google Gemma 2-B based on summaries from Cohere C4AI. Additionally, we initiated a pilot activity involving legal experts from SGS-Digicomply to validate the effectiveness of our summarization pipeline.

Keywords

Summarization, Large Language Models, Finetuning, Food Safety Regulations

1. Introduction

The legal industry is characterized by an overwhelming influx of textual data, encompassing case law, statutes, regulations, legal opinions, and contracts. Navigating these vast amounts of information can be laborious and time-consuming for legal professionals. Hence, there arises the need for efficient and accurate summarization tools to improve productivity and facilitate better decision-making.

IIR 24: Italian Information Retrieval Workshop, September 5-6, 2024, Udine, Italy
*Corresponding author.
guido.rocchietti@isti.cnr.it (G. Rocchietti); cosimo.rulli@isti.cnr.it (C. Rulli); korbinian.randl@dsv.su.se (K. Randl); cristina.muntean@isti.cnr.it (C. I. Muntean); francomaria.nardini@isti.cnr.it (F. M. Nardini); raffele.perego@isti.cnr.it (R. Perego); salvatore.trani@isti.cnr.it (S. Trani); manos.karvounis@agroknow.com (M. Karvounis); jakub.janostik@digicomply.com (J. Janostik)
0009-0004-9704-0662 (G. Rocchietti); 0000-0003-0194-361X (C. Rulli); 0000-0002-7938-2747 (K. Randl); 0000-0001-5265-1831 (C. I. Muntean); 0000-0003-3183-334X (F. M. Nardini); 0000-0001-7189-4724 (R. Perego); 0000-0001-6541-9409 (S. Trani); 0000-0003-3750-2066 (M. Karvounis)

Recent advancements in Natural Language Processing (NLP) and particularly Large Language Models (LLMs) have shown significant promise in automating text summarization tasks. The introduction of the transformer architecture by Vasvani et. al. [1] and models derived from it have improved the quality of generated summaries [2, 3]. Yet, summarizing legal texts presents unique challenges compared to different domains. On the one hand, legal texts present a high level of syntactic and semantic complexity combined with a highly domain-specific vocabulary. On the other hand, legal text summarization asks for a high level of accuracy and comprehension, as even minor errors in summarization can lead to significant misinterpretations.

Instruction-tuned Large Language Models (ILLMs) such as ChatGPT[1] or Perplexity,[2] have shown remarkable capabilities in various NLP tasks, including text summarization. However, these models are often large and computationally expensive for practical deployment. There is a pressing need to develop methods to reduce model size without compromising performance, especially for domain-specific applications like legal regulation summarization.

The present research is conducted within the Extreme Food Risk Analytics European (EFRA) project framework. The project's main goal is to develop an AI-driven approach to help and promote food risk prevention. In particular, we address the challenges of summarizing regulatory documents, offering insights that can be applied to other domain-specific applications. When considering the introduction of food safety regulations, it is important to remember that it is a complex procedure involving several steps. In fact, public authorities and regulators require an integrated decision framework that allows an automatic evaluation of both the regulatory aspect and the food and risk-related one. In this framework, our partner, SGS-Digicomply[3], plays a crucial role. They are a company specialized in "regulatory compliance and risk prediction with modern technology" with the leading software in the Food Safety market. For this research, they provide us with their extensive set of regulatory data.

In this paper, we develop and evaluate a method for summarizing regulatory food safety-related documents using instruction-tuned LLMs. Due to the fact there is little annotation available in this regard, we first create a dataset consisting of document summary pairs. The dataset captures the complexities and specificities of regulatory text summarization. To this end, we employ powerful — yet expensive — ILLMs to generate silver standard summaries of our collection of documents. This weak supervision method allows us to enhance the amount of data used for creating a fine-tuned summarization model, our second contribution. We then employ this dataset as a training set for *smaller* LLMs that are finetuned on the previously generated output of their larger counterparts. We aim to transfer the reliable knowledge of billion-sized LLM into smaller models, tearing down the summarization cost at the price of negligible degradation in the generated summaries. Finally, we provide a comprehensive evaluation of our models using a dataset of regulatory documents provided by SGS-Digicomply. Our results demonstrate the effectiveness of our approach in generating accurate and concise summaries.

The rest of the paper is organized as follows. In Section 2, we present the current state of the art available in the literature. In Section 3, we explain our research methods and the

[1] https://openai.com/chatgpt/
[2] https://www.perplexity.ai/
[3] https://www.digicomply.com

experimental setup. Finally, in Section 4, we present and comment on the results, followed by the conclusions in Section 5.

2. Related Work

Text summarization has been a significant area of research within natural language processing (NLP), with recent advancements driven by large language models (LLMs). This section reviews the most relevant contributions in the field.

Large language models have shown remarkable capabilities in generating coherent and contextually relevant summaries. Zhang *et al.* [4] explored the use of transformer-based architectures for summarization tasks, highlighting the superior performance of LLMs in handling long document contexts. Their work emphasizes the importance of model size and finetuning in achieving high-quality summaries, which aligns with our use of Llama-3-8B and Cohere Command-R-4bit models for initial summary generation. Many other applications of ILLMs and finetuned ones can be found in the literature. For instance, [5] shows that ILLMs have excellent rewriting capabilities in the context of query rewriting. [6] proposed a new framework to adapt LLMs to different domains by injecting legal information during a continual training stage. [7] introduce Legal Electra, a Language Model specialized in the legal domain. [8] and [9], explore new techniques to summarize documents in a low resources setting. The first use models such as BART [10] and GPT-2 [11] to summarize long documents, while the second investigate how to adapt models to the domain while keeping the resources low.

The concept of instruction-tuning, where models are finetuned with specific instructions to perform a task, has proven effective in various NLP applications. Wei *et al.* [12] demonstrated that instruction-tuning significantly enhances the performance of LLMs across multiple tasks, including summarization. [13] offer a good survey on the main techniques in the Natural Language Processing field, including summarization. Regarding model compression, [14] discussed the effectiveness of quantization in reducing model size and improving inference speed, which is critical for deploying models in resource-constrained environments.

The application of LLMs to regulatory text summarization poses unique challenges due to the complexity and specificity of regulatory documents. Our collaboration with SGS-Digicomply provides a practical setting for evaluating our summarization pipeline. By involving legal experts in the pilot phase, we ensure the generated summaries are concise and comply with regulatory standards and legal requirements. This practical application highlights the real-world relevance and effectiveness of our proposed methods. In summary, our work builds on the advancements in instruction-tuned LLMs, finetuning, and model compression to develop an energy-efficient summarization pipeline tailored to regulatory texts. This integration of state-of-the-art techniques enhances the summarization quality and addresses the practical constraints of deploying such models in resource-limited environments.

3. Experimental Setup

This section presents the methodology used to conduct the current research. As indicated in Section 1, our first objective is to generate a dataset of regulatory document summaries

exploiting the capabilities of ILLMs. Subsequently, we finetune several smaller models to learn how to summarize regulatory documents, with the purpose of distilling the summarization capabilities of ILLMs into more resource-efficient architectures.

Data Collection. The primary dataset for this study consists of regulatory documents provided by one of our industry partners, SGS-Digicomply. The dataset (SGS Dataset) they created for us, which cannot be made public for copyright reasons, is a large collection of HTML regulatory documents. It comprises items collected from websites identified by experts as pertinent to the food industry. For each selected website, a strategy was devised to identify the most relevant documents, which were then scraped using a proprietary framework built on top of the Scrapy Python library[4]. The source documents come in various formats, including HTML, PDFs, and Docx files. Each document undergoes processing and conversion into HTML and JSON formats suitable for machine learning applications. The original language of each document is detected, and non-English documents are translated into English.

The SGS-Digicomply dataset is intended to serve as a comprehensive collection of documents relevant to global markets detailing the regulatory landscape. It includes government publications, news articles, and scientific papers on legislative changes and food safety issues. For this research, we focus exclusively on the subset of data related to food regulatory frameworks. The version of the dataset used for this research comprises a total of 14,307 documents in 28 different languages. Most of these documents are in Italian, totaling 8,191, while English is the second most represented language, with 4,034 documents. As stated before, each document in a language different from English has a corresponding version in English, which we use for our experiments. All of these documents are provided with a summary. Most of them have a "scraped summary," while 44 have a manual summary, which human experts wrote. This set of summaries constitutes part of our test set, and we use it to evaluate our summaries. Finally, two different datasets must be created first to perform the finetuning phase.

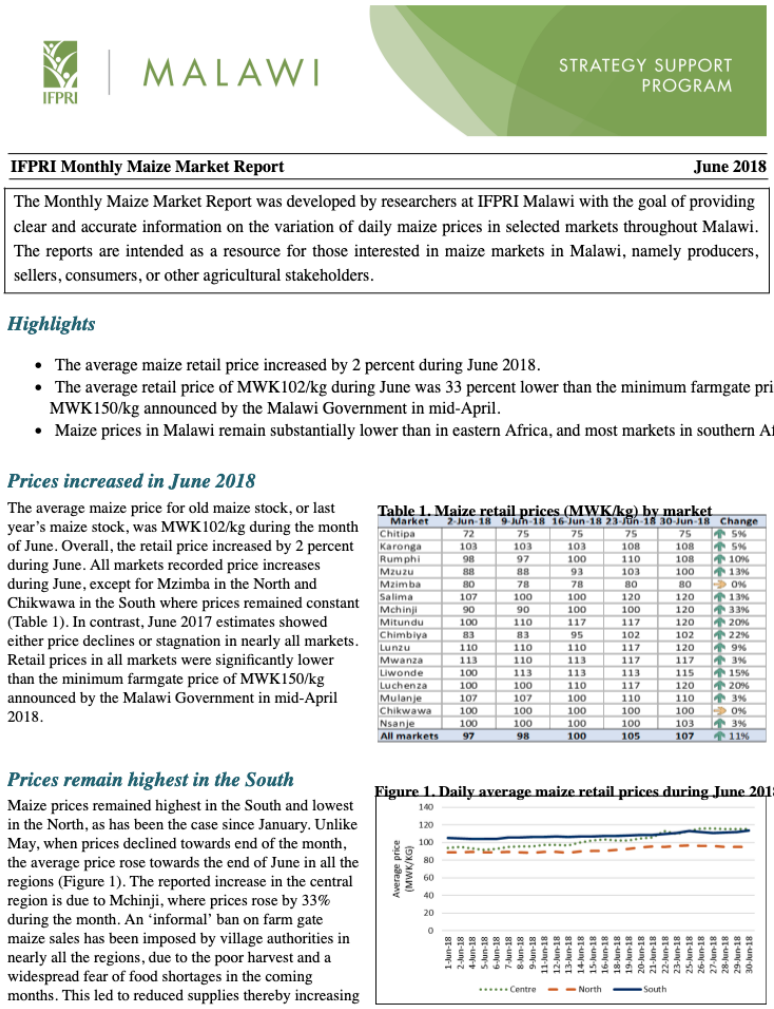
Data Preprocessing. We apply a simple pre-processing step to remove non-textual elements — metadata, footnotes, and references — as shown in Figure 1. We employ the BeautifulSoup[5] Python library to eliminate all the non-HTML elements.

To deal with the GPU memory limit, we cannot feed the entire textual input to the ILLMs that otherwise cause an Out-of-Memory GPU error. For this purpose, we create two datasets with different configurations of the same HTML content:

- The first dataset (SGS-Cut) is created using the first 40k characters of each cleaned document while eliminating the rest.
- The second dataset (SGS-Split) is created by splitting the cleaned HTML documents into chunks of 30k characters each. In this way, we produce multiple training samples for each text, notably increasing the total number of samples.

These two datasets are then used as input for the selected ILLMs to generate a new dataset of summaries that will be used for the finetuning phase.

[4] https://scrapy.org/
[5] https://www.crummy.com/software/BeautifulSoup/



IFPRI | MALAWI

STRATEGY SUPPORT PROGRAM

IFPRI Monthly Maize Market Report **June 2018**

The Monthly Maize Market Report was developed by researchers at IFPRI Malawi with the goal of providing clear and accurate information on the variation of daily maize prices in selected markets throughout Malawi. The reports are intended as a resource for those interested in maize markets in Malawi, namely producers, sellers, consumers, or other agricultural stakeholders.

Highlights

- The average maize retail price increased by 2 percent during June 2018.
- The average retail price of MWK102/kg during June was 33 percent lower than the minimum farmgate price of MWK150/kg announced by the Malawi Government in mid-April.
- Maize prices in Malawi remain substantially lower than in eastern Africa, and most markets in southern Africa.

Prices increased in June 2018

The average maize price for old maize stock, or last year's maize stock, was MWK102/kg during the month of June. Overall, the retail price increased by 2 percent during June. All markets recorded price increases during June, except for Mzimba in the North and Chikwawa in the South where prices remained constant (Table 1). In contrast, June 2017 estimates showed either price declines or stagnation in nearly all markets. Retail prices in all markets were significantly lower than the minimum farmgate price of MWK150/kg announced by the Malawi Government in mid-April 2018.

Table 1. Maize retail prices (MWK/kg) by market

Market	2-Jun-18	9-Jun-18	16-Jun-18	23-Jun-18	30-Jun-18	Change
Chitipa	72	75	75	75	75	5%
Karonga	103	103	103	108	108	5%
Rumphi	98	97	100	110	108	10%
Mzuzu	88	88	93	103	100	13%
Mzimba	80	78	78	80	80	0%
Salima	107	100	100	120	120	13%
Mchinji	90	90	100	100	120	33%
Mitundu	100	110	117	117	120	20%
Chimbiya	83	83	95	102	102	22%
Lunzu	110	110	110	117	120	9%
Mwanza	113	110	113	117	117	3%
Liwonde	100	113	113	113	115	15%
Luchenza	100	100	110	117	120	20%
Mulanje	107	107	100	110	110	3%
Chikwawa	100	100	100	100	100	0%
Nsanje	100	100	100	100	103	3%
All markets	**97**	**98**	**100**	**105**	**107**	11%

Prices remain highest in the South

Maize prices remained highest in the South and lowest in the North, as has been the case since January. Unlike May, when prices declined towards end of the month, the average price rose towards the end of June in all the regions (Figure 1). The reported increase in the central region is due to Mchinji, where prices rose by 33% during the month. An 'informal' ban on farm gate maize sales has been imposed by village authorities in nearly all the regions, due to the poor harvest and a widespread fear of food shortages in the coming months. This led to reduced supplies thereby increasing

Figure 1. Daily average maize retail prices during June 2018

Figure 1: Example of HTML document contained in the SGS Dataset

Summaries Generation. We use the ILLMs to generate the summaries for each dataset; in particular, we select Llama-3-8B-Instruct and CohereForAI/c4ai-command-r-v01-4bit, which were the top-performing open-source model on the HuggingFace Open LLMs Leaderboard [6] at the time of performing the present research. Llama is used for the SGS-Split dataset to generate summaries relative to each chunk of the documents, and Cohere for the SGS-Cut to generate a single summary per document. We provide the ILLMs with a prompt to input the data, asking them to summarize the regulatory documents, i.e., *"I want you to summarize the following legal document"*, followed by the document itself. This results in the creation of two different datasets: the Llama dataset, composed of a training set of 15,101 entries, and validation and test sets of 1,888 entries each. On the other hand, the Cohere dataset consists of a training set of 7,485 entries and a validation and test set of 935 entries each. Both datasets are then used to perform

[6] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

the finetuning phase.

Model Finetuning. We finetune several models from the HuggingFace repository. Models can belong to different architectures, and each architecture requires a specific input format. We rely on encoder-decoder and decoder-only transformers. The former requires clean text as input and provides a generated text as output. The latter creates a continuation of the input text token by token. Hence, we employ a separation token that indicates the end of the input, marking the start of the summary. At training time, we provide the model with the properly formatted input and, as a target, the summary of the relative document or chunk of it.

With our generated datasets in hand, the next step is to finetune several models to see if we could replicate or even improve upon the performance of the ILLMs, but with lower computational requirements. To pursue this goal, we experiment with three distinct finetuning paths. First, we use the summaries generated by Llama-3-8B-Instruct to finetune google/flan-T5-large. This involved training the model to understand and replicate the style and precision of Llama-3-8B-Instruct's summaries.

Similarly, we finetune another instance of google/flan-T5-large, using the summaries produced by CohereForAI/c4ai-command-r-v01-4bit. This allows us to compare the impact of different summary sources on the same base model. Lastly, we finetune a 4-bit quantized version of google/gemma-2B using the CohereForAI-generated summaries. The quantization significantly reduced the model size and computational load, making it more efficient while aiming for high-quality output. The finetuning process was conducted on a Nvidia V-100 80GB GPU to handle large models and extensive datasets effectively.

Newly generated summaries go to a final post-processing phase that eliminates all the noise and errors that the models might produce. For instance, in some cases, the generative models fit the maximum number of tokens to generate and create sentences that do not conclude. In those cases, we simply eliminate the latest generated sentence.

Evaluation Metrics. Summarization is not an easy task to evaluate. The most used metrics, such as ROUGE, use the lexical overlap to establish the similarity between the documents and the summaries, which poorly estimate the semantic overlapping between the two. To address this problem, we incorporate neural evaluation metrics, such as BERTScore and the newly released LongDocFactScore metric [15]. These metrics overcome the limits of the lexical-based approach, aiming at assessing the factual accuracy and consistency of the summaries, ensuring that the finetuned models not only generated concise summaries but also preserved the integrity and essential facts of the original legal content.

We list the metrics employed in our evaluation.

- ROUGE-1 (R1) [16] measures the overlap of unigrams (single words) between the summaries generated by our models and the reference summaries from the handmade dataset.

$$P_{\text{ROUGE-1}} = \frac{\text{number of overlapping words}}{\text{total words in generated summary}} \tag{1}$$

$$R_{\text{ROUGE-1}} = \frac{\text{number of overlapping words}}{\text{total words in reference summary}} \tag{2}$$

- ROUGE-L (RL) [16] evaluates the longest common subsequence (LCS), which is the longest sequence of words (not necessarily contiguous) present in both the generated summary

and the reference.

$$P_{\text{ROUGE-L}} = \frac{\text{number of words in LCS}}{\text{total words in generated summary}} \tag{3}$$

$$R_{\text{ROUGE-L}} = \frac{\text{number of words in LCS}}{\text{total words in reference summary}} \tag{4}$$

- BERTScore [17] uses a model based on BERT to compare the similarity between pairs of texts. It creates embeddings for both the automatically generated summaries (i.e., $\hat{\mathbf{x}}$) and the reference summaries (i.e., $\mathbf{x}$), then evaluates the similarity between these embeddings.

$$P_{\text{BERTScore}} = \frac{1}{|\hat{x}|} \sum_{\hat{\mathbf{x}}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^T \hat{\mathbf{x}}_j \qquad R_{\text{BERTScore}} = \frac{1}{|x|} \sum_{\mathbf{x}_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^T \hat{\mathbf{x}}_j \tag{5}$$

- LongDocFactScore (LDFS) [15]: Given the importance of factual accuracy in regulatory documents, we incorporated LongDocFactScore. This recently developed metric assesses both factual accuracy and consistency, ensuring that the summaries retained the key facts and logical flow of the originals.

Also, for R1, RL, and BERTScore, we calculate the F1-score as shown in Equation 6.

$$F1 = 2 * \frac{P * R}{P + R} \tag{6}$$

The metrics that we use to evaluate are the F1-measure for Rouge 1 (F1@R1), Rouge L (F1@RL), and BertScore (F1@BS), plus the newly introduced one LongDocFACTScore (LDFS). Although Rouge in the original formulation only represents recall, we argue that using F1 shows a more comprehensive picture.

4. Results

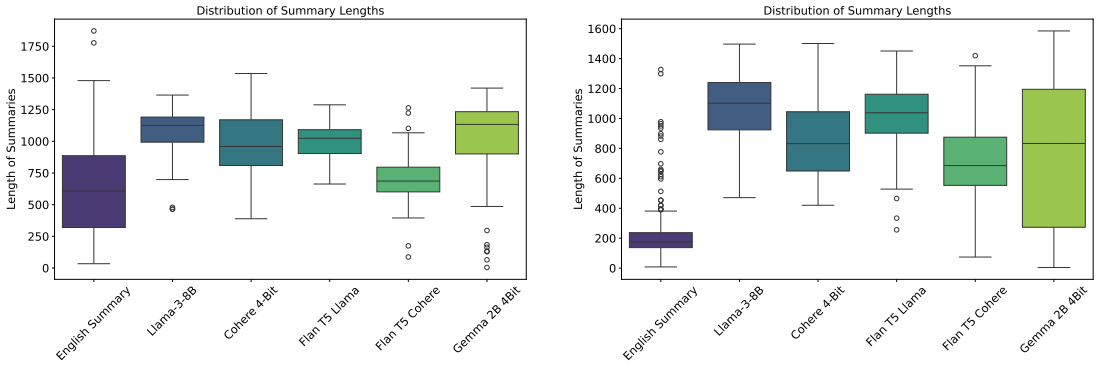
In this section, we present the results of the experiments after the finetuning phase. We compare five different LLMs, both instruction-tuned and finetuned. Llama-3-8B and Cohere 4-bit are the instructed ones that we also use to generate the training datasets for the finetuning phase (see Sec. 3). On the other hand, we use two finetuned versions of Flan-T5, one finetuned on Cohere-generated data and one on Llama. Furthermore, we evaluate the performance of a 4-bit quantized version of Gemma-2B finetuned on the data generated by Cohere.

In Figures 2a and 2b, we report the average length of the available summaries for the two datasets we use to evaluate. Figure 2a reports the boxplot indicating the length distribution of the summaries considering the subset of 44 manual summaries. On the other hand, Figure 2b reports the length distribution of the summaries in the test set provided by SGS-Digicomply.

As we can observe, in both cases, the summaries generated by the finetuned and ILLMs are, on average longer than the reference ones indicated by the *English Summary* label.

In Table 1, we report the results of the evaluation phase when comparing the generated summaries with the manual ones provided by SGS-Digicomply. In this case, we can observe that the best-performing model is Flan T5, finetuned with the dataset generated using Cohere.

In Table 2, we report the results of the chosen metrics calculated when comparing all of the summaries, including the 44 manual ones provided with the dataset, with the content of the



(a) (b)

Figure 2: Length distribution of the summaries. On the left (a) are the ones of the 44 entries subset containing the manual summaries, while on the right (b) the test set provided by SGS-Digicomply.

Table 1
Results of the evaluation phase calculated between all the summaries and the HTML documents generated by SGS-Digicomply. F1 indicates the f-measure, while R1, RL, and BS indicate Rouge 1, Rouge L, and BERTScore, respectively.

Metric	Silver Standard		Fine Tuned		
	Llama-3-8B	Cohere 4bit	Flan T5 Llama	Flan T5 Cohere	Gemma 2B 4bit
F1@R1	0.252	0.243	0.239	**0.264**	0.232
F1@RL	0.161	0.150	0.160	**0.173**	0.148
F1@BS	0.821	0.823	0.817	**0.829**	0.796
LDFS	-5.379	-4.914	-5.565	**-4.606**	-5.643

HTML documents. As we can observe, the summaries generated with Llama-3-8B are the ones that obtain the highest result for all the metrics, with a high gap with the manual ones. All generated summaries get higher scores than the ones obtained manually. This can be probably due to the length of the manual summaries, which has a significant influence when evaluating lexical overlap features and will be assessed in the next iterations of the research.

Table 2
Results of the evaluation phase calculated between all the summaries, including the manual and the HTML documents generated by SGS-Digicomply. F1 indicates the f-measure, while R1, RL, and BS indicate Rouge 1, Rouge L, and BERTScore, respectively.

Metric	Manual	Silver Standard		Fine Tuned		
		Llama-3-8B	Cohere 4bit	Flan T5 Llama	Flan T5 Cohere	Gemma 2B 4bit
F1@R1	0.099	**0.241**	0.192	0.213	0.226	0.167
F1@RL	0.065	**0.174**	0.129	0.146	0.170	0.120
F1@BS	0.791	**0.832**	0.825	0.805	0.827	0.828
LDFS	-4.633	**-3.019**	-3.397	-3.175	-3.071	-2.963

Finally, Tables 3 and 4 show the results obtained when evaluating the summaries on the external test set provided by SGS-Digicomply. In the first table, we can see the results of the generated summaries evaluated when compared with the content of the HTML document. In line with the evaluation shown in Table 2, we observe that the higher scores are achieved by the summaries generated using Llama-3-8B-Instructed, which seems the best model to grasp the content of the original HTML better. The only exception is the LongDocFACTScore metric, which indicates that Flan T5 trained on the Llama summaries is the best way to keep track of the facts in the original HTML.

When we consider the scraped summaries contained in the SGS-Digicomply test set, we can see that Flan T5 finetuned on the Cohere dataset achieves the best results when considering Rouge L, BertScore, and LongDocFACTScore, while Gemma 2B 4bit is the best performing one when considering Rouge 1. Also, in this case, we need to remember that a higher metric value might not involve the fact that the generated summaries are better than those with lower scores, as the current metrics for evaluating summarization retain little information regarding the content of the summaries.

Table 3
Results of the evaluation phase calculated between the generated summaries and the cleaned HTML documents.F1 indicates the f-measure, while R1, RL, and BS indicate Rouge 1, Rouge L, and BERTScore, respectively.

Metric	Silver Standard		Fine Tuned		
	Llama-3-8B	Cohere 4bit	Flan T5 Llama	Flan T5 Cohere	Gemma 2B 4bit
F1@R1	**0.383**	0.292	0.358	0.279	0.239
F1@RL	**0.285**	0.186	0.281	0.207	0.172
F1@BS	**0.870**	0.853	0.859	0.851	0.852
LDFS	-3.032	-3.483	**-3.012**	-3.042	-3.112

Table 4
Results of the evaluation phase calculated between the generated summaries and the ones automatically generated by SGS-Digicomply.F1 indicates the f-measure, while R1, RL, and BS indicate Rouge 1, Rouge L and BERTScore, respectively.

Metric	Silver Standard		Fine Tuned		
	Llama-3-8B	Cohere 4bit	Flan T5 Llama	Flan T5 Cohere	Gemma 2B 4bit
F1@R1	0.231	0.247	0.218	0.266	**0.268**
F1@RL	0.172	0.174	0.171	**0.208**	0.207
F1@BS	0.855	0.861	0.851	**0.863**	0.860
LDFS	-6.470	-5.612	-6.551	**-5.128**	-5.650

In conclusion, we can state that the ILLMs and the consequent finetuned models achieve good-quality summarization capabilities for the chosen metrics. Furthermore, when comparing with the content of the HTML documents, Llama-3-8B is the best performing one, in line with its size in terms of parameters. It is interesting to notice that Flan T5, finetuned on the summaries generated by Llama, achieves similar results to the ones obtained by Llama while reducing the

parameter number by approximately 10.2 times.

5. Conclusions

This paper presents a new approach to the automatic summarization of regulatory documents exploiting Istruction-tuned LLMs and finetuning conducted in the Extreme Food Risk Analytics (EFRA) European project framework. Thanks to our collaboration with SGS-Digicomply, we were provided with a large dataset of HTML documents containing legal text in the form of regulations, news, and laws. In this research phase, we exploit the content of these documents, appropriately cleaning off the noisy HTML tags, to generate two summary datasets exploiting ILLM to finetune smaller LLMs later. We created these two datasets using the instruction-tuned version Llama-3 with 8B parameters and CohereForAI/c4ai-command-r-v01-4bit using two approaches to input these models. We then use the newly generated summaries as targets for three distinct LLMs to teach them how to summarize the regulatory documents adequately. To do so, we finetuned two versions of Flan T5, one on the summaries generated by Llama and the other on the ones generated by the Cohere model. Finally, we finetuned a 4-bit quantized version of the Google model Gemma with 2B parameters.

As shown in Sections 4, the results achieved when evaluating with standard metrics for the summarization task achieve interesting scores. We achieved better scores for every model than those calculated using manually created summaries of the regulatory documents. At the same time, the scores achieved by the finetuned models are also comparable, if not better, than the ones achieved by the two ILLMs.

This leaves us with good hopes for future research steps. In the following research phase, we plan on using a pool of legal experts from the SGS Digicomply partner to manually evaluate and label the summaries generated by the different models on the test set they provided us. In this way, we plan to apply various techniques, such as knowledge distillation, to exploit the newly labeled data and finetune even better models while reducing their size. Simultaneously, we plan on applying all the state-of-the-art quantization techniques to further reduce the size of the models while maintaining a good summarization quality.

Ackowledgements.

Funding for this research has been provided by the European Union's Horizon Europe research and innovation program EFRA (Grant Agreement Number 101093026). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1073–1083. URL: https://aclanthology.org/P17-1099. doi:10.18653/v1/P17-1099.

[3] Y. Liu, M. Lapata, Text summarization with pretrained encoders, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, p. 3721.

[4] H. Zhang, X. Liu, J. Zhang, DiffuSum: Generation enhanced extractive summarization with diffusion, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13089–13100. URL: https://aclanthology.org/2023.findings-acl.828. doi:10.18653/v1/2023.findings-acl.828.

[5] E. Galimzhanova, C. I. Muntean, F. M. Nardini, R. Perego, G. Rocchietti, Rewriting Conversational Utterances with Instructed Large Language Models, in: 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2023, pp. 56–63. URL: https://ieeexplore.ieee.org/document/10350178. doi:10.1109/WI-IAT59888.2023.00014.

[6] Q. Huang, M. Tao, C. Zhang, Z. An, C. Jiang, Z. Chen, Z. Wu, Y. Feng, Lawyer LLaMA Technical Report, 2023. URL: http://arxiv.org/abs/2305.15062. doi:10.48550/arXiv.2305.15062, arXiv:2305.15062 [cs].

[7] W. Hua, Y. Zhang, Z. Chen, J. Li, M. Weber, Mixed-domain Language Modeling for Processing Long Legal Documents, in: D. Preo\textcommabelowtiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. Spanakis, N. Aletras (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2023, Association for Computational Linguistics, Singapore, 2023, pp. 51–61. URL: https://aclanthology.org/2023.nllp-1.7. doi:10.18653/v1/2023.nllp-1.7.

[8] A. Bajaj, P. Dangati, K. Krishna, P. Ashok Kumar, R. Uppaal, B. Windsor, E. Brenner, D. Dotterrer, R. Das, A. McCallum, Long Document Summarization in a Low Resource Setting using Pretrained Language Models, in: J. Kabbara, H. Lin, A. Paullada, J. Vamvas (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Online, 2021, pp. 71–80. URL: https://aclanthology.org/2021.acl-srw.7. doi:10.18653/v1/2021.acl-srw.7.

[9] T. Yu, Z. Liu, P. Fung, AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 5892–5904. URL: https://aclanthology.org/2021.naacl-main.471. doi:10.18653/v1/2021.naacl-main.471.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault

(Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[13] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Computing Surveys 56 (2023) 1–40.

[14] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, K. Keutzer, Q-bert: Hessian based ultra low precision quantization of bert, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8815–8821.

[15] J. A. Bishop, S. Ananiadou, Q. Xie, LongDocFACTScore: Evaluating the Factuality of Long Document Abstractive Summarisation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 10777–10789. URL: https://aclanthology.org/2024.lrec-main.941.

[16] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).