electronics

MDPI

Article

Prediction of Arabic Legal Rulings Using Large Language Models

Adel Ammar *, Anis Koubaa, Bilel Benjdira, Omer Nacar and Serry Sibaee

Robotics and Internet-of-Things Lab, Prince Sultan University, Riyadh 12435, Saudi Arabia; akoubaa@psu.edu.sa (A.K.); bbenjdira@psu.edu.sa (B.B.); onajar@psu.edu.sa (O.N.); ssibaee@psu.edu.sa (S.S.)
* Correspondence: aammar@psu.edu.sa

Abstract: In the intricate field of legal studies, the analysis of court decisions is a cornerstone for the effective functioning of the judicial system. The ability to predict court outcomes helps judges during the decision-making process and equips lawyers with invaluable insights, enhancing their strategic approaches to cases. Despite its significance, the domain of Arabic court analysis remains under-explored. This paper pioneers a comprehensive predictive analysis of Arabic court decisions on a dataset of 10,813 commercial court real cases, leveraging the advanced capabilities of the current state-of-the-art large language models. Through a systematic exploration, we evaluate three prevalent foundational models (LLaMA-7b, JAIS-13b, and GPT-3.5-turbo) and three training paradigms: zero-shot, one-shot, and tailored fine-tuning. In addition, we assess the benefit of summarizing and/or translating the original Arabic input texts. This leads to a spectrum of 14 model variants, for which we offer a granular performance assessment with a series of different metrics (human assessment, GPT evaluation, ROUGE, and BLEU scores). We show that all variants of LLaMA models yield limited performance, whereas GPT-3.5-based models outperform all other models by a wide margin, surpassing the average score of the dedicated Arabic-centric JAIS model by 50%. Furthermore, we show that all scores except human evaluation are inconsistent and unreliable for assessing the performance of large language models on court decision predictions. This study paves the way for future research, bridging the gap between computational linguistics and Arabic legal analytics.

Keywords: large language models; Arabic court analysis; foundation models; natural language processing; transformers

Citation: Ammar, A.; Koubaa, A.; Benjdira, B.; Nacar, O.; Sibaee, S. Prediction of Arabic Legal Rulings Using Large Language Models. *Electronics* **2024**, *13*, 764. https://doi.org/10.3390/electronics13040764

Academic Editor: Alberto Fernandez Hilario

Received: 15 January 2024
Revised: 7 February 2024
Accepted: 9 February 2024
Published: 15 February 2024

1. Introduction

The fusion of law, artificial intelligence (AI), and natural language processing (NLP) stands as a groundbreaking frontier in contemporary research. The legal domain, with its intricate statutes, precedents, and interpretations, offers a unique challenge for computational models. Yet, the potential implications of successfully navigating this domain are profound. If legal decisions can be predicted with high precision using machine learning models, a set of invaluable insights would be given to the judicial system. Such advancements could advance legal research, case preparation, and help judges and lawyers with deeper insights that they may not take into consideration during the cases' analysis.

Predicting court decisions is challenging, especially for under-represented languages in NLP studies, such as Arabic [1]. The inherent complexity of case description texts, combined with the nuances of the Arabic language, compounds the difficulty. Arabic, with its rich morphological structure and myriad dialects, has been a challenging landscape for NLP tasks [2,3]. Moreover, case description texts in Arabic are characterized by their detailed rhetoric, extensive use of precedents, and domain-specific terminologies [4].

1.1. Context

The effectiveness of language model pretraining has been demonstrated in enhancing various tasks within the realm of natural language processing. This approach has proven

successful in elevating the performance of a wide range of tasks related to processing and understanding human language [5,6].

In the area of AI applied to the legal domain, there have been significant advancements [7]. The rise of machine learning has brought in a new wave of research, with scholars exploring the potential of statistical models for legal prediction [8]. The recent advancements in large language models, especially transformer-based models, have further expanded the horizons in this domain [9]. These models have demonstrated exceptional capabilities in a range of NLP tasks from machine translation [10,11] to sentiment analysis [12–14], making their application to legal area an exciting avenue of exploration.

This paper embarks on an exploration of predicting Arabic court decisions using large language models (LLMs). By leveraging the latest in NLP and deep learning, we aim to test different approaches to using LLMs to maximize the predictive capability.

1.2. Related Works

Language models (LMs) serve as the basis for various language technologies, but an understanding of their capabilities, limitations, and risks is still lacking. Several benchmarks were built to bridge this gap. The objective of a benchmark is to set a standard by which the performance of systems can be evaluated across a variety of tasks. Kumar et al. [15] introduced the Holistic Evaluation of Language Models (HELM) to enhance the transparency of language models. Initially, they created a taxonomy to categorize the wide range of possible scenarios and metrics relevant to language models. Subsequently, a comprehensive subset of scenarios and metrics was selected based on coverage and feasibility, while also identifying any gaps or underrepresentation. Finally, a multi-metric approach was adopted to evaluate language models.

The Beyond the Imitation Game benchmark (BIG-bench) was introduced by Srivastava et al. [16], featuring 204 tasks contributed by 444 authors from 132 institutions. These tasks covered diverse topics and aimed to test the limits of current language models. The performance of various model architectures, including OpenAI's GPT models and Google's dense and sparse transformers, was evaluated on BIG-bench across a wide range of model sizes. Human expert raters also participated to establish a strong baseline. The findings revealed that model performance and calibration improved with larger model sizes, but they still fell short compared to human performance. Interestingly, performance was similar across different model classes, with some advantages observed for sparse transformers. Tasks that showed gradual improvement often required extensive knowledge or memorization, while tasks with breakthrough behavior involved multiple steps or components. In settings with ambiguous context, social bias tended to increase with scale, but it could be mitigated through prompting techniques.

Elmadany et al. [17] presented ORCA, which is an openly accessible benchmark aimed at evaluating Arabic language comprehension. ORCA was meticulously developed to encompass various Arabic dialects and a wide range of complex comprehension tasks. It leveraged 60 distinct datasets across seven clusters of natural language understanding (NLU) tasks. To assess the current advancements in Arabic NLU, ORCA was employed to conduct a thorough comparison of 18 multilingual and Arabic language models. Furthermore, a public leaderboard was provided, featuring a unified evaluation metric (ORCA score). This score represents the macro-average of the individual scores across all tasks and task clusters.

Abdelali et al. [18] evaluated the performance of Foundation Models (FMs) in various text and speech tasks related to Modern Standard Arabic (MSA) and Dialectal Arabic (DA), including sequence tagging and content classification, across different domains. ChatGPT (OpenAI's GPT-3.5-turbo), Whisper (OpenAI) [19], and USM (Google) [20] were used to conduct zero-shot learning and address 33 distinct tasks using 59 publicly available datasets, resulting in 96 test setups. They found out that LLMs performed worse compared to state-of-the-art (SOTA) models across most tasks, dialects, and domains, although they achieved comparable or superior performance in a few specific tasks. The study emphasized the

importance of prompt strategies and post-processing for enhancing the performance of FMs and provided in-depth insights and findings.

On the other hand, the field of prompt engineering [21] has gained prominence in developing and refining inputs for language models. It provides a user-friendly and intuitive interface for human interaction with LLMs. Given the sensitivity of models to even minor changes in input, prompt engineering focuses on creating tools and techniques to identify robust prompts that yield high-performance outcomes. Various automatic optimization approaches [22,23] have been suggested to determine the optimal prompt for a particular task or a range of tasks. These methods aim to find the most suitable prompt that yields the best performance outcome.

More specifically, numerous studies have ventured into predicting court decisions across different jurisdictions. In the US, machine learning has been used to anticipate the outcomes of Supreme Court decisions [24]. In Europe, deep learning models have been employed to predict the decisions of the European Court of Human Rights [25]. Concerning the Arabic legal domain, a pioneering model named AraLegal-BERT [26], inspired by the English-based LEGAL-BERT [27], was proposed. It is a bidirectional encoder transformer-based model (BERT [28]) fine-tuned for the Arabic legal domain. The model was evaluated against three BERT variations for Arabic across three natural language understanding (NLU) tasks, showcasing superior accuracy over the general and original BERT models on legal text. This work exemplifies how domain-specific customization can significantly improve language model performance in narrow domains, advancing the field's understanding of model adaptation for specialized use-cases. However, the tasks targeted in the study were specifically legal text classification tasks, named entity recognition tasks, and keyword extraction tasks. These tasks are different from the scope of our paper targeting the prediction of Arabic legal rulings, which is more challenging and complex. Moreover, AraLegal-BERT is trained from scratch on specific Arabic datasets. This approach is different from the current study, where we tried first to profit from the LLMs advanced linguistic capabilities. Then, we tried to enhance the eliciting performance of these LLMs on Arabic legal ruling prediction using zero-shot and few-shot learning. To our knowledge, the application of the aforementioned approach to the Arabic legal system remains a new field, with an attractive potential. This paper aims to bridge this gap by presenting a systematic investigation into predictive analysis of Arabic court decisions via an array of cutting-edge large language models tested on a dataset of real commercial cases.

1.3. Contributions

Given the aforementioned context and the gap identified in Arabic legal system analysis, our research offers the following novel contributions:

- Comprehensive Model Evaluation: This study conducted a predictive analysis of Arabic court decisions by leveraging three prominent large language models, LLaMA-7b, JAIS-13b, and GPT-3.5-turbo, applied to a dataset comprising 10,813 real commercial court cases.
- Significance of Text Preprocessing: The study thoroughly investigated the potential benefits derived from summarizing and translating the original Arabic input texts, culminating in the creation of 14 distinct model variations.
- Highlighting LLaMA's Limitations: LLaMA models have been touted as almost equivalent to GPT models [29]. Nevertheless, the findings of this paper reveal the intrinsic reduced performance of all LLaMA model variants compared to JAIS and GPT-3.5 on our dataset of Arabic court decisions.
- Insights into Evaluation Metrics: The paper offers a detailed evaluation of model performance using diverse metrics, namely human assessment; GPT evaluation; ROUGE (1, 2, and L); and BLEU scores. Importantly, the research underscored the unreliability of all metrics, barring human assessment.

- Bridging Research Domains : This pivotal study bridges the gap between computational linguistics and Arabic legal analytics, laying a foundation for future scholarly endeavors in this interdisciplinary realm.

2. Materials and Methods

2.1. Base Large Language Models

LLaMA-7b [29] (designed by Meta AI), JAIS-13b-chat [30] (MBZUAI University), and GPT-3.5-turbo [31–34] (OpenAI) are three recent representatives of a frontier of advancements in large language model (LLM) technology, each hailing from different origins with distinct architectural innovations. LLaMA-7b, an open-source LLM emanating from Meta AI, showcases a unique architectural approach with a range of models tailored for various applications. On the other hand, JAIS-13b-chat, with its focus on bilingual (Arabic and English) capabilities, offers a novel solution to Arabic-centric language processing tasks. GPT-3.5-turbo, a product of OpenAI, stands out for its optimization for chat-based applications, demonstrating a balance between performance and cost-effectiveness. Table 1 summarizes the main characteristics of these three models, providing a comparative glimpse into their architectural underpinnings, language and domain proficiency, training data, and use cases. Only JAIS was trained on a sizeable proportion (29%) of Arabic texts. In contrast, Arabic language represented 0.03% of GPT3's training dataset by word and character count, and 0.01% by document count [35]. Similar figures are assumed for GPT-3.5-turbo. Meta AI did not disclose the proportion of tokens per language in LLaMA models' training datasets, but the description of the sources of their pretraining datasets reveals that it is overwhelmingly in English [29].

Another important element in a large language model is the tokenizer. Tokenization consists in subdiving words into sub-word tokens in order to learn vocabulary that encompasses sub-word units such as prefixes, suffixes, and root components, enabling effective handling of diverse word morphologies. Each of the three base models considered use tailored pretrained tokenizers that are based on byte-pair encoding (BPE). BPE is a data compression algorithm initially designed to reduce the size of files by replacing frequent sequences of bytes with shorter representations [36]. In recent years, it has been adopted in NLP to tokenize text into subwords or characters in a way that strikes a balance between the flexibility of character-level representations and the efficiency of word-level representations [37]. Nevertheless, we noticed that most common tokenizers used in LLMs are not adapted to Arabic language, as can be seen in an example in Figure 1. In this example, the LLaMA tokenizer segments a word into individual characters which do not have any independent meaning. The same occurs with GPT's Tiktoken tokenizer. By contrast, JAIS tokenizes the same word in this example into a single token, which conserves the meaning.

Input

sentence = 'مرحبا' # means 'Hello'
tokens = tokenizer.tokenize(sentence)
tokens

Output
['ا', 'ب', 'ح', 'ر', 'م', '_']

Figure 1. Over-Segmented example by LLaMA Tokenizer. The Arabic word means 'Hello'. The tokenizer segments into individual characters.

Table 1. Theoretical comparison of LLaMA-7b, JAIS-13b-chat, and GPT-3.5-turbo base models.

Characteristic	LLaMA-7b	JAIS-13b-chat	GPT-3.5-turbo
Model Size and Architecture	Total of 7B parameters, SwiGLU activation, Rotary positional embeddings.	Total of 13B parameters, transformer-based decoder-only (GPT-3) architecture, SwiGLU non-linearity.	Total of 175B parameters, GPT architecture.
Language and Domain Proficiency	Outperforms on many benchmarks including reasoning, coding, proficiency, and knowledge tests.	Bilingual (Arabic and English), state-of-the-art Arabic-centric performance.	Optimized for chat, capable of understanding and generating natural language or code.
Training Data and Open-source Availability	Trained on 1.4 trillion tokens from publicly available datasets, overwhemingly in English.	Total of 395B tokens (116B Arabic tokens), pretrained with an additional 10M instruction/response pairs.	Total of 300B tokens from various sources, overwhemingly in English.
Tokenizer	LLaMA tokenizer (BPE model based on SentencePiece [38]).	JAIS tokenizer (BPE custom-built tokenizer that weighs both languages equally)	Tiktoken (fast optimized BPE).
Use Cases and Performance	Fine-tuned for dialogue, optimized versions for chat (Llama-2-Chat).	Bilingual tasks, outperforms existing open Arabic/multilingual chatbots.	Optimized for chat-based applications, human-like responses in conversations.

In contrast to LLaMA and GPT-3.5, the JAIS model initially refused to generate predictions concerning court decisions. This refusal reveals the type of precautionary measures incorporated into the model during its reinforcement learning from human feedback (RHLF) phase. Nevertheless, we successfully elicited predictions from the model by including an explicit instruction, stating that these are experiments that are intended solely for educational and research purposes.

We employed multiple configurations of the three aforementioned foundational models. These configurations encompass the zero-shot, single-shot, and fine-tuning training paradigms. Furthermore, they are implemented on either the original Arabic dataset or on pre-processed texts that have undergone summarization and/or translation. Cumulatively, these diverse configurations result in 14 distinctive model variants. A comprehensive description of these variants is provided in Section 2.4.

2.2. Fine-Tuning Using LLM-Adapters

Engaging in complete fine-tuning has the potential to result in catastrophic forgetting [39,40], given that it involves altering all parameters within the model. In contrast, parameter efficient fine-tuning (PEFT), by exclusively modifying a limited subset of parameters, as opposed to full-parameter fine tuning, demonstrates greater resilience against the detrimental impacts of catastrophic forgetting [41]. In this context, LLM adapters offer a simple and efficient approach to PEFT in large language models [42]. LoRA (low-rank adaptation) is a method that can significantly reduce the number of trainable parameters required for fine-tuning large language models. It is a type of LLM adapter that is integrated into the LLM-adapters framework and supports fine-tuning of LLaMA models among others [42]. As LoRA has significant motivations for successfully lowering the number of trainable factors without sacrificing performance, applying it to the LLaMA model aims to achieve high performance while minimizing computational costs.

With this approach, LoRA follows a strategy that reduces the number of parameters to be trained during fine-tuning by freezing all of the original model parameters and then inserting a pair of rank decomposition matrices alongside the original weights. Additionally, LoRA utilizes the adapter method in such a way of adding a subset of parameters, enabling a few low-intrinsic adapters in parallel with the attention module without increasing

inference latency. In this work, we carry out the fine-tuning of the LLaMA-7b base model using LoRA approach on Arabic texts, following the implementation of [43]. In fact, LoRA's design allows for more flexibility in adding adapters [44], making it efficient for scaling up to large language models for improved performance on custom datasets and tasks. Nevertheless, we did not manage to fine-tune the larger JAIS-13b and GPT-3.5-turbo base models due to resource constraints.

Figure 2 illustrates the integrated mechanism of the LoRa adapter within the LLM module of the transformer, highlighting the modified forward pass in the network, and the weight adjustment mechanism. The LoRa method enhances the fine-tuning of large language models (LLMs) by decomposing the weight update matrix into a lower-rank representation instead of updating the original weight matrix directly, leading to fewer parameters during adaptation. This results in faster training and potentially reduced computational needs without losing vital information. In conventional fine-tuning, weight changes are computed via backpropagation based on the loss gradient. LoRa, instead, decomposes these changes into two smaller, lower-dimensional matrices. Then, it trains these smaller matrices, enabling effective representation in a lower-dimensional space and reducing the parameter space.

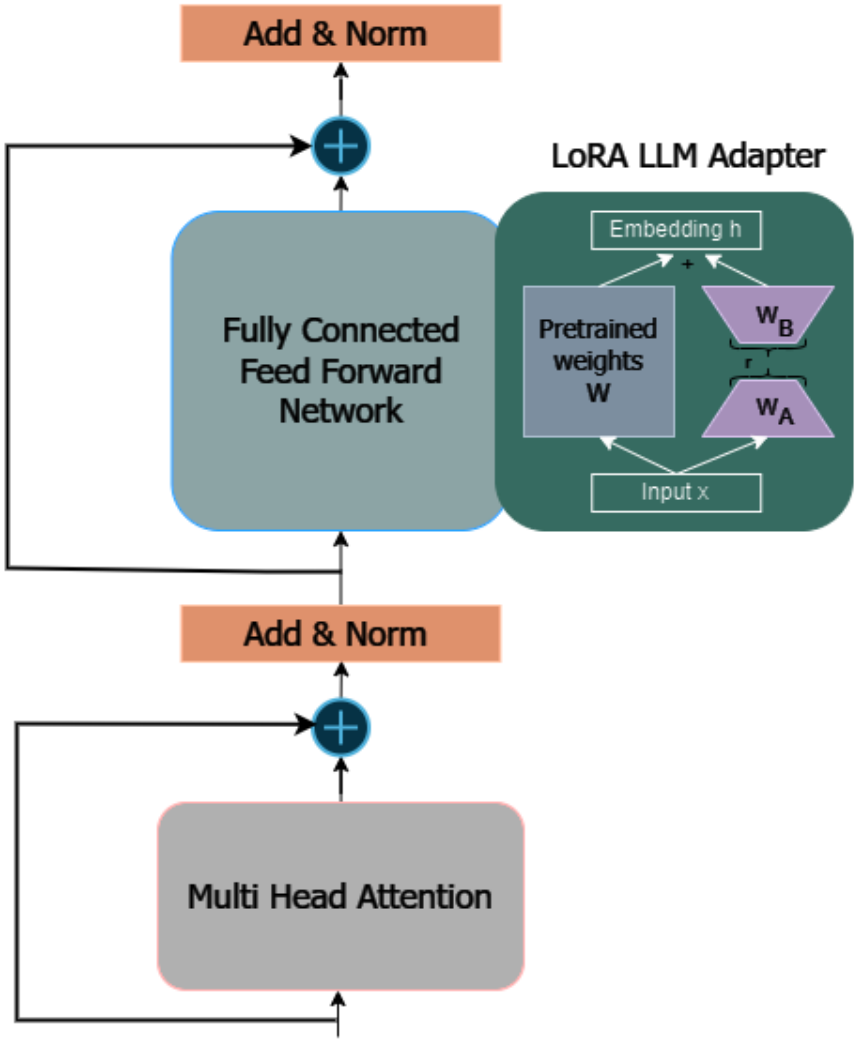

Figure 2. Operational schematic of LoRa adapters within the transformer.

In the LoRA method, the decomposition of the weight update matrix ΔW into two matrices W_A and W_B is given by:

$$\Delta W = W_A W_B. \quad (1)$$

Assuming W_A and W_B are of dimensions $m \times r$ and $r \times n$, respectively, where r is the rank, and m and n are the dimensions of the original matrix ΔW, the total number of parameters to be learned reduces from $m \times n$ to $m \times r + r \times n$.

Further, if X is the input to a layer and Y is the output, the modified forward pass in LoRA can be represented as:

$$Y = (W + W_A W_B)X + b, \tag{2}$$

where b is the bias vector.

The error E in approximation can also be analyzed. It is given by

$$E = ||\Delta W - W_A W_B||_F, \tag{3}$$

where $||.||_F$ denotes the Frobenius norm (aka Hilbert–Schmidt norm), which is defined as the square root of the sum of squares of all the matrix entries.

This mathematical formulation elucidates the reduction in computational complexity and the preservation of essential information for task adaptation achieved by LoRA. This method preserves the essential information required for task adaptation while reducing the computational burden, showcasing a trade-off between model complexity and adaptation capacity.

The implementation of LoRa is relatively straightforward, as seen in Figure 2. A modified forward pass in the network is applied, adjusting the magnitude of weight updates to balance pretrained knowledge with new task-specific adaptation.

2.3. Dataset

We retrieved the Saudi Ministry of Justice dataset (SMOJ) through a web scraping from the Saudi Justice Portal (SJP) website [45], focusing on the category of commercial courts, which contains a series of court decisions about financial and commercial disputes, all in Arabic language. To facilitate the data retrieval, we used Selenium Python library [46], which enables programmatic interactions with web pages, essentially simulating user actions to access and gather data.

The data collection process for SMOJ was structured and systematic, starting with the iteration through a range of page numbers. This range spans from page 1 to 60,000, a scope determined based on the expected volume of data available on the website. Before any data extraction occurs, each page's availability is verified by checking for the presence of the text 'Page not found.' This precautionary measure ensures that only existing pages are processed, minimizing potential errors and preventing unnecessary resource consumption.

Once page availability is confirmed, the data extraction process is initiated using the Beautiful Soup Python package [47] which is tailored for HTML and XML parsing and is employed to dissect the HTML structure of the web pages. This allows for the extraction of specific elements, focusing on critical legal information contained within the SJP website. The data extraction process focuses on three primary categories: case description, justification, and court decision. We used the case description as the input (prompt) to the LLM models and the court decision as output (completion). There is no strictly pre-defined format or ordering for the case description and court decision, which complicates the data processing by the LLMs, because of the greater challenge represented by unstructured texts. We decided to ignore the justification field and not include it in the input, seeing that it often unveils the inclination of the court decision.

After removing duplicates and excessively long cases (more than 4096 tokens), we randomly subdivided the SMOJ dataset into a training dataset containing 10,713 cases and a testing dataset containing 100 cases. We opted for a reduced testing dataset to be able to manually evaluate the outputs of each LLM model. In fact, we will show in Section 3 that all other automatic evaluations were unreliable and inconsistent.

Figure 3 depicts the histogram of the number of words in the prompts (case descriptions) and completions (court decisions) in the SMOJ training dataset. The total number

of words in the training set is 5M, and the average number of words in the prompts and completions are 422 and 52, respectively. The large size of the prompts is a real challenge, which motivated us to test LLM models on summarized prompts, as will be detailed in Section 2.4.

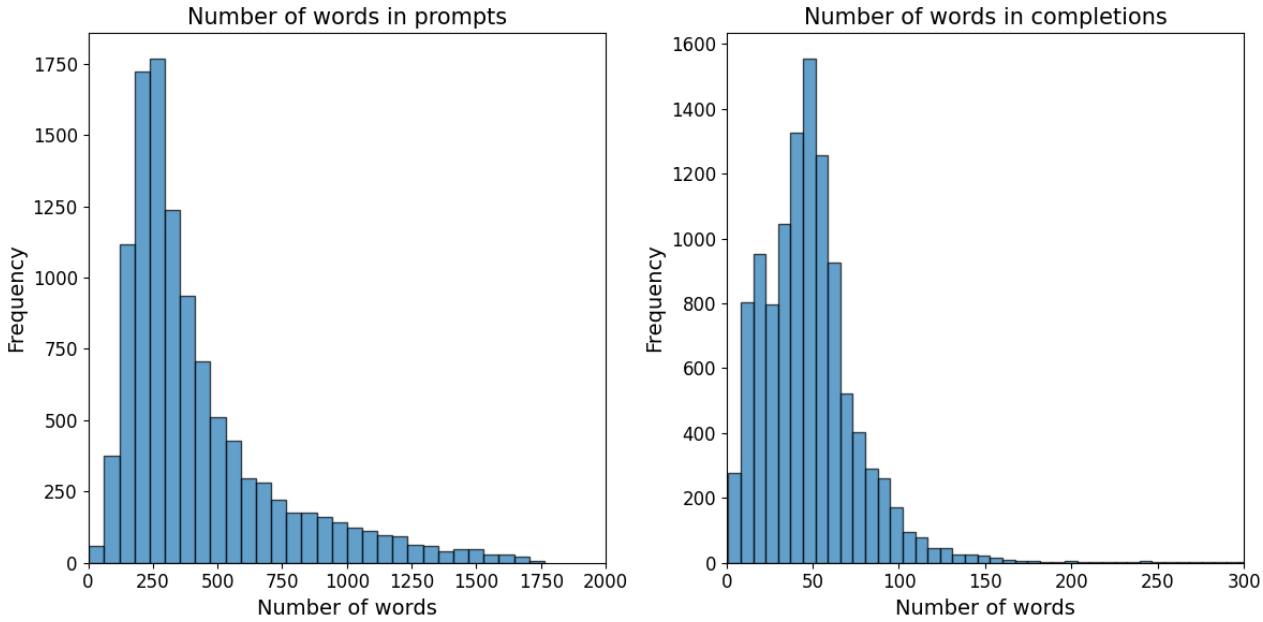

Figure 3. Histogram of the number of words in the prompts (case descriptions) and completions (court decisions) in the SMOJ training dataset.

2.4. LLM Model Variants

For each of the three base pretrained models described in Section 2.1 (LLaMA-7b, GPT-3.5-turbo, and JAIS-13b-chat), we implemented different variants. Figure 4 illustrates the main steps for evaluating the 8 LLaMA variant models on the SMOJ dataset. The base pretrained LLaMA-7b model is the core of all these models. They differ by the inclusion or not of single-shot or fine-tuning learning and the addition or not of summarizing and/or translation steps:

- Model L0 is a zero-shot model. It is the mere application of the base pretrained LLaMA-7b model on each prompt of the Arabic testing dataset without any pre-processing or learning steps.
- Model L1 is a single-shot variant, where a single prompt/completion pair from the original Arabic training dataset is provided in the instruction to act as an example to follow.
- Model LT0 is a zero-shot model applied to an English testing dataset. This dataset was obtained using the Google Translate API through Python translators package [48]. The translation of the original Arabic dataset into English can be beneficial to enhance the prediction for LLaMA and GPT-3.5 models, since they are overwhelmingly pre-trained on English texts, as explained in Section 2.1. The assessment of the usefulness of this pre-processing step will be discussed in Section 3.2.
- Model LT1 is a single-shot model applied to the translated English testing dataset. It includes a single translated prompt/completion pair from the training dataset in its instructions.
- Model LF is obtained by fine-tuning the base model on the original Arabic training dataset for 200 epochs.
- Model LFT is obtained by fine-tuning the base model on the translated English training dataset for 200 epochs.

- Model LFS is obtained by fine-tuning the base model on a subset of the Arabic training dataset after summarizing the prompts through GPT-3.5-turbo API. We selected only a subset of 1000 prompt and completion pairs due to budget limitations, since requests to the GPT API are costly.
- Model LFST is obtained by fine-tuning the base model on a subset of 1000 prompt and completion pairs from the Arabic training dataset after summarizing and translating them through GPT-3.5-turbo API.

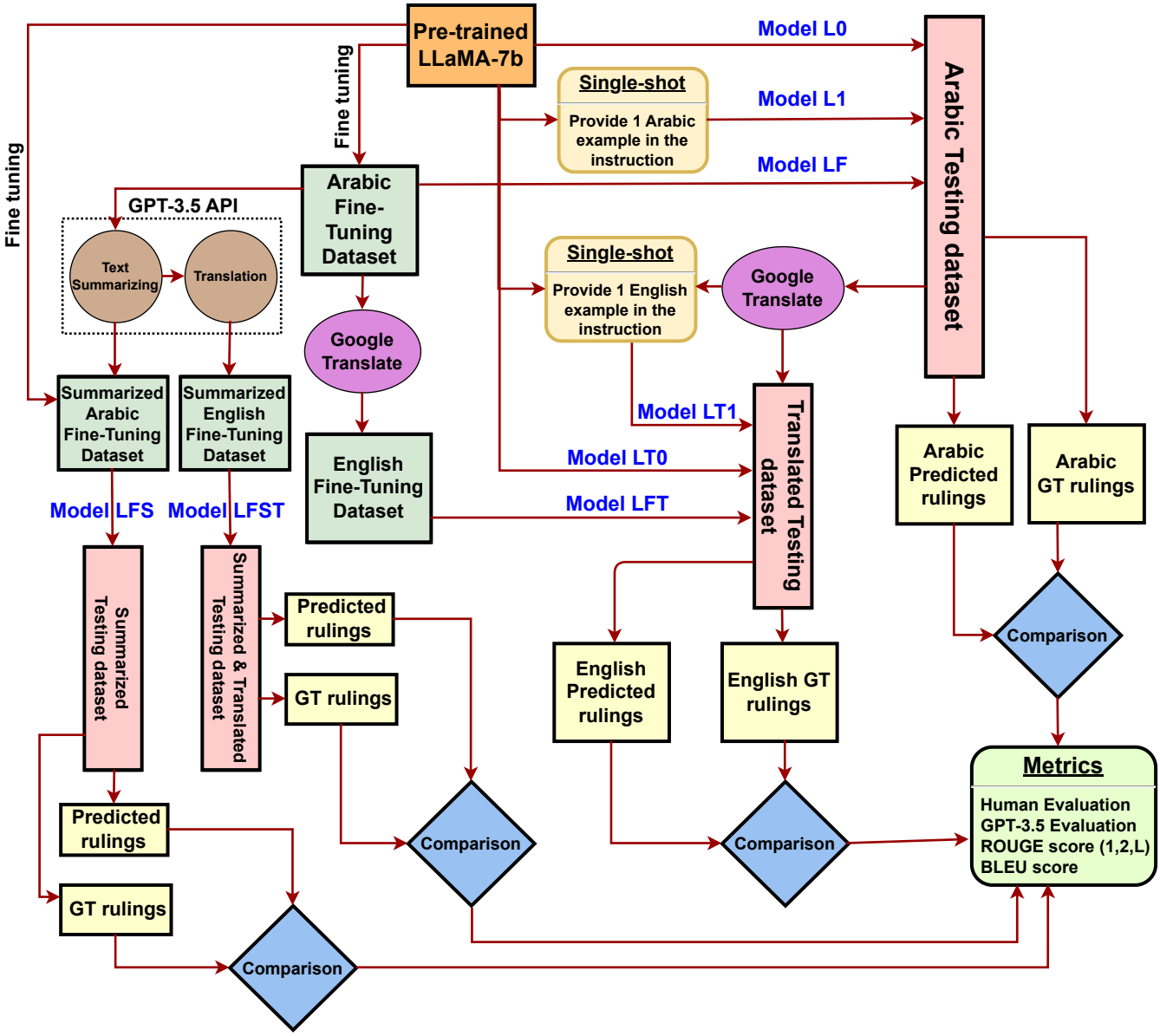

Figure 4. Diagram of the main steps for evaluating the 8 LLaMA variant models on the SMOJ dataset.

Figure 5 showcases the instructions fed to GPT-3.5-turbo API to summarize and/or translate the original SMOJ dataset, for fine-tuning the LFS and LFST models.

Similarly, models G0, G1, GT0, and GT1 are obtained in the same way as L0, L1, LT0, and LT1, respectively, but using GPT-3.5-turbo as a base model, instead of LLaMA-7b. Likewise, J0 and J1 are zero-shot and single-shot variants of JAIS-13b-chat model.

Task	Instruction to GPT-3.5
Arabic Summarization	لخص هذا النص العربي في فقرة مترابطة في أقل من 100 كلمة. *[Summarize this Arabic text in a coherent paragraph in less than 100 words].*
Translation from Arabic to English and summarization	`Give a summarized translation into English of the following Arabic text in less than 100 words, as a paragraph without bullets.`

Figure 5. Instructions used to summarize and translate the SMOJ dataset through GPT-3.5-turbo API. The Arabic summarization instruction is translated between square brackets.

Due to limited resources, we could not fine-tune the GPT and JAIS models in the same way that we did for the LLaMA models. In addition, it is pointless to apply JAIS on translated text, since it was pretrained with a special focus on the Arabic language. Furthermore, multi-shot variants were not examined in our research. This is due to the extensive input size present in our dataset and the inherent limited context length associated with the models (4096 tokens).

For each model, we employ a suite of metrics, as detailed in Section 2.5, to evaluate their performance by comparing the predicted rulings to the suitable version of the ground-truth (GT) rulings from the test dataset. Specifically:

- Models L0, L1, LF, J0, J1, G0, and G1 are evaluated against the original Arabic version of the GT rulings.
- Model LFS is gauged against the summarized Arabic version.
- Models LT1, LT0, LFT, GT0, and GT1 are assessed based on the translated form of the GT rulings.
- Model LFST is measured against the GT rulings that have been both summarized and translated.

2.5. Metrics

The following metrics were applied to evaluate each of the LLM models described in Section 2.4:

- Human score: A human evaluator was tasked with assessing the accuracy of the predicted rulings generated by each model in relation to the ground-truth rulings of the test dataset. This evaluation was conducted on a scale ranging from 0 to 5. A score of 0 indicated that the predicted ruling was either nonsensical or wholly incorrect, while a score of 5 signified a flawless prediction, mirroring the decisions encapsulated in the ground-truth ruling, regardless of the actual wording. If the model predicts a decision close to the ground truth (e.g., 'The defendant should pay an amount of X to the plaintiff') but does not guess the exact amount to be paid, the human score will be strictly between 0 and 5, and its value will depend on how close the predicted amount is to the real amount. To ensure a uniform evaluation standard and minimize variability in scoring, all model outputs were reviewed by the same evaluator, with a background on Arabic legal decisions.
- GPT score: We used GPT-3.5-turbo API to automatically and systematically compare all the predicted rulings generated by each model to the ground-truth rulings of the test dataset. To guide this assessment, we provided the GPT model with the following instruction: "Compare the following two court decisions (predicted: 'Decision (predicted)' and ground-truth: 'Decision (GT)') and assign a score from 0 to 5 to the predicted decision. 0 means: Non sense. 5: means perfect answer. Format the response as: Score; Justification. For example: 0; Non sense."
- BLEU score: BLEU [49,50], an acronym for bilingual evaluation understudy, was designed as a metric for assessing the quality of machine-translated text between two natural languages. The BLEU score is computed using a weighted geometric mean

of modified n-gram precision. This is further adjusted by the brevity penalty, which diminishes the score if the machine translation is notably shorter than the reference translation. The utilization of the weighted geometric mean ensures a preference for translations that consistently perform well across different n-gram precision levels. More specifically, the BLEU score is given by

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4}$$

where:

- p_n is the n-gram precision.
- w_n are the weights for each precision (typically $w_1 = w_2 = w_3 = w_4 = 0.25$ for BLEU-4).
- BP is the brevity penalty:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases} \tag{5}$$

with c as the predicted output length and r as the ground-truth length.

We calculated the BLEU metric using the nltk.translate.bleu_score Python module [51].

- ROUGE score : ROUGE [52] is an acronym for recall-oriented understudy for gisting evaluation. Its primary purpose is to evaluate the performance of automatic summarization tools and machine translation systems within the realm of natural language processing (NLP). The fundamental idea behind ROUGE is to juxtapose an algorithmically generated summary or translation with one or multiple human-crafted reference summaries or translations. This comparison helps to determine how well the machine-generated output aligns with the human standard. We apply it here to the comparison between predicted and GT rulings in the SMOJ testing dataset. We specifically used three variants of the ROUGE score:
  - ROUGE-1 : This metric gauges the overlap of unigrams (individual words) between the predicted output and the GT ruling. By examining the matching single words between both texts, ROUGE-1 provides insights into the basic lexical similarity.
  - ROUGE-2 : Stepping beyond individual words, ROUGE-2 considers bigrams (pairs of adjacent words). By comparing the overlap of these word pairs between the predicted and GT outputs, ROUGE-2 offers a deeper understanding of the phrasal and structural alignment. The general formula for the ROUGE-N score is

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{GT}} \sum_{\text{N-gram} \in s} \text{Count}_{\text{match}}(\text{N-gram})}{\sum_{s \in \text{GT}} \sum_{\text{N-gram} \in s} \text{Count}(\text{N-gram})} \tag{6}$$

    where
    - $\text{Count}_{\text{match}}(\text{N-gram})$ is the maximum number of times an n-gram is found in both the predicted and GT outputs.
    - $\text{Count}(\text{N-gram})$ is the count of the n-gram in the GT output.
  - ROUGE-L : This metric employs the concept of the longest common subsequence (LCS). LCS is the maximum sequence of tokens that appear in both the machine-produced and reference texts. This metric offers a more holistic perspective on similarity as it naturally considers sentence-level structures and automatically identifies the co-occurring n-gram sequences.

For each of these three metrics, we compute the precision (P), recall (R), and F1-score (F):

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{8}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

We calculated the ROUGE metrics using the ROUGE 1.0.1 Python library [53].

While the BLEU and ROUGE metrics were primarily conceived for tasks related to translation and summarization, they can potentially serve as indicative tools for evaluating the alignment between predicted and GT rulings in the SMOJ dataset. We will assess the correctness of this hypothesis in Section 3.

3. Results

Within this section, we undertake a comprehensive evaluation of the implemented models, encompassing both qualitative and quantitative assessments. In Section 3.1, we present a qualitative comparison between human and GPT scores on a small sample, exemplifying scenarios where predictions align with or deviate from expectations. We also discuss the challenges and nuances of employing GPT-3.5 as an evaluation metric. In Section 3.2, we delve into the performance evaluation of the 14 models using diverse metrics, shedding light on the impact of zero-shot, single-shot, and fine-tuning approaches, as well as the prompt summarization and/or translation pre-processing steps. We further discuss the reliability of GPT, BLEU, and ROUGE scores. This holistic evaluation provides insights into the strengths and limitations of LLMs for the prediction of court decisions.

3.1. Qualitative Evaluation

Figure 6 provides a qualitative comparison between human and GPT scores on a small sample of predicted and GT rulings from the testing dataset. This sample is representative of most of the encountered cases. The first row shows an example in Arabic. The predicted output contains a correct decision briefly expressed with implicit reference to the amount mentioned in the input (case description), while the GT ruling explicitly mentions the names of the plaintiff and defendant and the amount of money that the latter should pay to the former. Because of the difference in formulation, the GPT API gave the prediction a score of only 2/5, whereas the human evaluator took into account the semantic matching and assigned a higher score of 4/5.

In the second example, the LLM model issues a perfect ruling matching the same amount to be paid by the plaintiff as in the GT decision. Even though the identities of the plaintiff and defendant are not explicitly mentioned, this is not important, since they are already mentioned in the case description. In this case, both the human evaluator and GPT-3.5 assigned a perfect score of 5/5.

In the third example, the predicted output is a series of nonsensical words and symbols. This happens often with LLaMA models, especially when the input size is large. As expected, the human score in this case is 0. However, GPT-3.5 oddly assigns a score of 2/5 to this prediction. This example also reveals the poor Google translation in the GT output, especially for the last sentence where the Arabic word Al-hādī ('guide') was mistaken for its homonym: 'pacific'. Such translation shortcomings can affect the quality of LLM training.

The fourth prediction example in Figure 6 is similar in terms of meaningless predicted output and poor GT translation, but in this case, both the human and GPT scores are rightly equal to 0.

In the fifth and last example, the LLM model just rehashed the instruction and part of the input that was fed to it, without adding any prediction. This also often happens with LLaMA models. As expected, the human score in this case is 0. However, GPT-3.5

surprisingly assigns a score of 4/5 to this prediction, which showcases its unreliability as en evaluation metric. This will be discussed more precisely using quantitative analysis in the next section.

Predicted ruling	GT ruling	GPT score	Human score
تقرر استحقاق المدعى عليه لرد رأس المال المذكور في الشراكة مع المدعى الممثل بالمدعي وكالة. قررت الدائرة رفع الجلسة للمداولة.	حكمت الدائرة بإثبات الصلح المتفق عليه بين الأطراف، والقاضي بإلزام المدعى عليه/ جلال بن صالح بن منصور الخميري -هوية وطنية رقم (...)- بأن يدفع للمدعية/ زينب بنت أحمد بن إبراهيم عبيدى -هوية وطنية رقم (...)-، مبلغ وقدره (٢٥٠،٠٠٠) مائتان وخمسون ألف ريال بتاريخ (٣٠/ ٣/ ١٤٤٤هـ)، وبالله التوفيق.	2	4
The court should order the defendant to pay the remaining rental fee of SAR 39,360 to the plaintiff.	The court ruled that the defendant, Mazaaya Al-Mawared Construction Company, with commercial registration number (…), must pay an amount of SAR 39,360.00 (thirty-nine thousand, three hundred and sixty Saudi riyals) to Layl and Nahar Construction and Equipment Rental Company, with commercial registration number (…).	5	5
Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good Ple Pleottom Good Ple Ple Ple Ple Ple Ple Ple Ple Ple Ple Ple Good Ple Ple Ple Good Good Good Good Good Good Good Good Good Good Good Good Good Ple Ple Ple Ple Good Good Good Good Good Good Good Good Good Good Good Good Good Good Good / Ple Ple Ple Ple/# / / / / / / // // // // / / / / / / / / / / / / / / /########<s><s><s><s><s><s><s><s><s>	The text of the ruling: The department ruled to compel the defendant, Salah bin Hassan bin Muhammad Naqi, Saudi, according to the national identity No.: (...) The owner of the modern men's sew Commercial No. (...) an amount of (67,698) sixty -seven thousand and six hundred and ninety -eight riyals. God bless and Pacific to the straight path.	2	0
<s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s> <s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s> <s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s> <s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s> <s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s><s>	The text of the ruling: The department ruled to obligate/ Muhammad bin Abdul Rahman bin Saleh Al -Dailaj, a national identity No. (...), to pay a growing company for trade, an amount of (2.200,000) riyals, two million and two hundred thousand riyals, and refused except for requests , God bless.	0	0
You are an assistant at a Saudi justice court. You examine the case, and you give a final judgement ### Input: Facts Agency No. (441693313), as well as/ Mazen Mishaal bin Humaid Al -Harbi National ID No. (...) was attended as an agent of the defendant according to the Agency No. (435016242), then the prosecutor's agent was asked about his client's lawsuit and he wrote it with what he stipulated: [Facts:	The text of the ruling: obligatory/ Fatima bint Ayed Zaid Al -Omari National ID No. (...) The owner of the first development institution for trade with commercial registry No. (...), that you pay to/ Ahmed Marzouq Abdullah Al -Raddadi National ID No. (...) , An amount of (80,000) eighty thousand riyals, and he refused otherwise; When it is shown in the reasons, and may God's prayers and peace be upon our Prophet Muhammad, his family and companions, all of the members, Ahmed Abdullah Muhammad Al -Sa'diya, the second member, Yahya Muhammad Al -Suhaibran, head of the Judicial Department, Hassan Nayef met Al -Harbi	4	0

Figure 6. Human- and GPT-assigned scores on sample predicted and GT rulings from the testing dataset.

3.2. Quantitative Evaluation

Table 2 presents the performance evaluation of the eight LLaMA-7b variant models on the test datasets, measured using different metrics. The LFST model, which underwent both summarization and translation processes using GPT-3.5-turbo, consistently delivered the top performance across nearly all evaluation metrics. In stark contrast, the other LLaMA variants displayed significantly subpar results. Their human evaluation scores were under 0.5 out of 5, and their GPT scores did not surpass 2 out of 5. Furthermore, both the ROUGE and BLEU scores for all models were notably low. This underperformance was particularly pronounced for Arabic language models. L1 receives a score of 0 for all metrics except GPT score. This performance degradation compared to L0 can be explained by the increase in the input size due to adding an example of prompt/completion in the instruction, which often makes the total input size exceed the model's maximum context length.

Table 2. Results of the evaluation of the 8 LLaMA-7b variant models on the testing datasets using various metrics.

LLaMA-7b Model		Zero-Shot		Single Shot		Fine-Tuned			
		Arabic (L0)	Translated (LT0)	Arabic (L1)	Translated (LT1)	Arabic Original (LF)	Arabic Summarized (LFS)	Translated (LFT)	Summarized and Translated (LFST)
Fine-tuning epochs		0	0	0	0	200	200	200	200
Human score		0.24	0.38	0	0.40	0.10	0.021	0.062	**1.2**
GPT score		1.3	0.91	0.60	0.93	1.5	1.35	1.8	**2.8**
BLEU score		0.0016	0.035	0	0.050	0.058	0.053	0.22	**0.25**
ROUGE-1	**R**	0.00032	0.064	0	0.075	0.018	0.018	0.19	**0.27**
	P	0.0010	0.32	0	0.29	0.029	0.036	**0.23**	0.20
	F	0.00049	0.097	0	0.096	0.020	0.021	0.17	**0.21**
ROUGE-2	**R**	0.0	0.0060	0	0.0077	0.0044	0.0052	0.044	**0.065**
	P	0.0	0.015	0	0.0072	0.0084	0.0084	0.043	**0.048**
	F	0.0	0.008	0	0.0066	0.0055	0.0061	0.038	**0.048**
ROUGE-L	**R**	0.00032	0.062	0	0.070	0.018	0.017	0.17	**0.25**
	P	0.0010	0.31	0	0.29	0.029	0.033	**0.21**	0.18
	F	0.00049	0.094	0	0.092	0.019	0.019	0.15	**0.19**

The primary reason for summarizing prompts in the LFS and LFST models stems from our observation that the LLaMA models frequently produce low-quality responses to longer prompts. This observation finds some validation in Figure 7, which showcases scatter plots correlating input size (measured by word count) with human evaluation scores for the LT1 and LFST models. Notably, for the LT1 model, prompts exceeding 1000 words invariably receive a score of zero. However, the overall correlation remains relatively weak, at –0.3. In contrast, upon summarizing the prompts for the LFST model, the correlation between input size and evaluation score vanishes. This suggests that the modified LLaMA model can handle moderately sized inputs in an equal manner.

Table 3 presents the outcomes of applying the same metrics to the four GPT-3.5-turbo variant models. Both G0 and GT1 demonstrate closely aligned performance when evaluated using human scores. This suggests that the integration of translation and single-shot training did not significantly enhance performance for GPT-based models. However, when we consider the BLEU and ROUGE scores, the translated models, GT0 and GT1, consistently outperform their counterparts. Interestingly, there is a noticeable discrepancy between the GPT score and the human judgment. A more detailed examination of specific prediction instances confirms that the GPT score can be unreliable in several scenarios.

Table 4 shows the performance of the two JAIS-13b-chat models. Only Arabic-based models were tested in this scenario since the JAIS base model is specifically tailored for Arabic language. We observe a slight improvement when moving form zero-shot (J0) to single-shot (J1) according to all metrics, except for the GPT score. This further highlights the unreliability of the GPT score for this task. Even though JAIS was pretrained with special focus on Arabic language, it falls short in comparison with all GPT-based models (Table 3). This confirms the superiority of GPT-based models for a wide range of tasks even for under-represented languages in its learning dataset, such as Arabic.

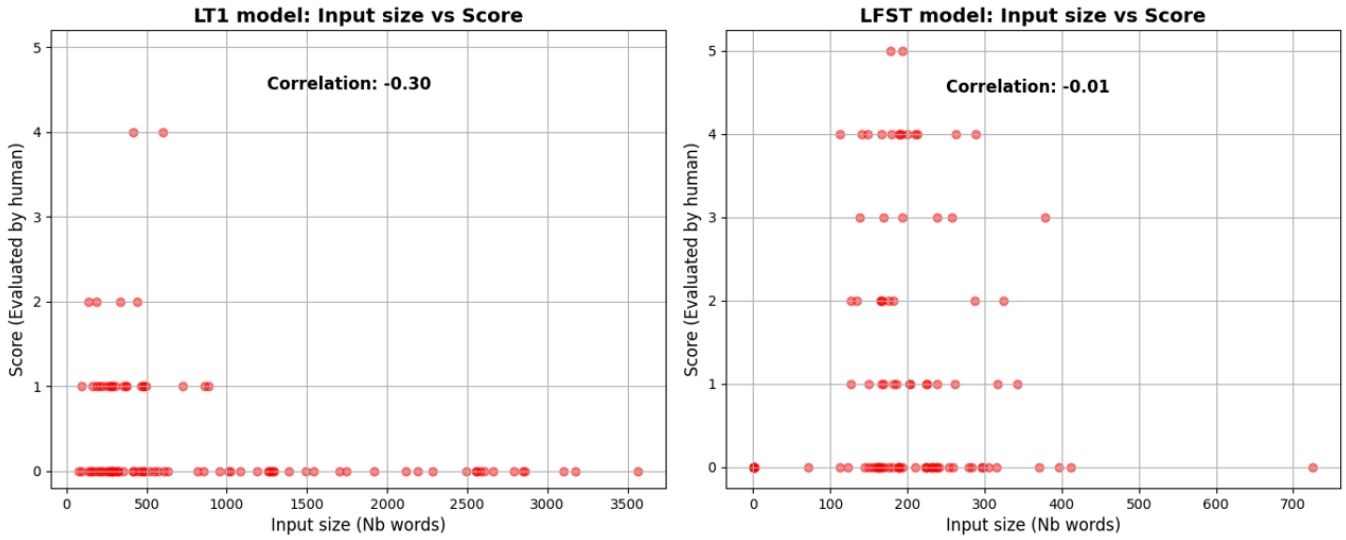

Figure 7. Scatter plot between the input size (in terms of number of words) and the human score obtained for LT1 (**left**) and LFST (**right**) models.

Table 3. Results of the evaluation of the four GPT-3.5-turbo variant models on the testing datasets, using various metrics.

GPT-3.5-turbo Model		Original Arabic		Translated	
		Zero Shot (G0)	Single Shot (G1)	Zero Shot (GT0)	Single Shot (GT1)
Human score		**2.4**	1.9	2.0	**2.4**
GPT score		**2.7**	**2.7**	2.3	2.1
BLEU score		0.19	0.21	0.18	**0.25**
ROUGE-1	**R**	0.13	0.13	**0.30**	0.26
	P	0.081	0.095	0.13	**0.18**
	F	0.086	0.096	0.16	**0.19**
ROUGE-2	**R**	0.031	0.036	**0.083**	0.069
	P	0.025	0.035	0.033	**0.050**
	F	0.025	0.032	0.043	**0.051**
ROUGE-L	**R**	0.12	0.12	**0.27**	0.23
	P	0.075	0.091	0.12	**0.16**
	F	0.081	0.093	0.15	**0.17**

Figure 8 maps out the 14 implemented models in the (human score, GPT score) space. This visualization underscores the dominance of the GPT-based models and the underperformance of the LLaMA-based counterparts. Among the LLaMA models, only the LFST variant comes close to the performance of JAIS and GPT models in terms of human evaluation. Notably, LFST is not a pure LLaMA model as it leverages the summarizing and translation capabilities of GPT-3.5. On the other hand, while JAIS models outpace LLaMA models, they lag behind the GPT models. A striking feature of Figure 8 is the evident discrepancy between GPT and human scores. For instance, despite LFST achieving the highest GPT score across all models, it secures a merely moderate human score. In a similar vein, LFT showcases a higher GPT score than both LT0 and LT1, even though the latter pair surpass it in human evaluations. This incongruence is especially pronounced in English-based models.

Table 4. Results of the evaluation of the two JAIS variant models on the testing datasets using various metrics.

JAIS-13b-chat		Original Arabic	
		Zero Shot (J0)	**Single Shot (J1)**
Fine-tuning epochs		0	0
Human score		1.4	**1.6**
GPT score		**2.1**	2.0
BLEU score		0.16	**0.2**
ROUGE-1	**R**	0.072	**0.11**
	P	0.077	**0.086**
	F	0.057	**0.081**
ROUGE-2	**R**	0.016	**0.028**
	P	0.013	**0.021**
	F	0.014	**0.020**
ROUGE-L	**R**	0.068	**0.097**
	P	0.074	**0.079**
	F	0.054	**0.074**

Figure 8. Human score versus GPT score for each tested large language model. Arabic-based models are represented as circles, while English-based models are represented as squares, with a different color code for each base model (LLaMA, JAIS, GPT-3.5).

This observation is further confirmed in Figure 9 where we notice that the correlation between the human score and GPT score is much higher for Arabic-based models (0.92) than for English-based models (0.60). A plausible explanation for this divergence is that the process of translating from Arabic to English may introduce errors, omission or misrepresentation of key details, which makes score evaluation by GPT-3.5 more challenging. This observation extends to the BLEU and ROUGE scores, which consistently display a lower alignment with human scores for English models. Most notably, ROUGE-1 precision and ROUGE-L precision exhibit negative correlations with human scores, standing at –0.79 and

–0.77, respectively. All these results suggest that the GPT, BLEU, and ROUGE scores are unreliable for performance evaluation in the considered task.

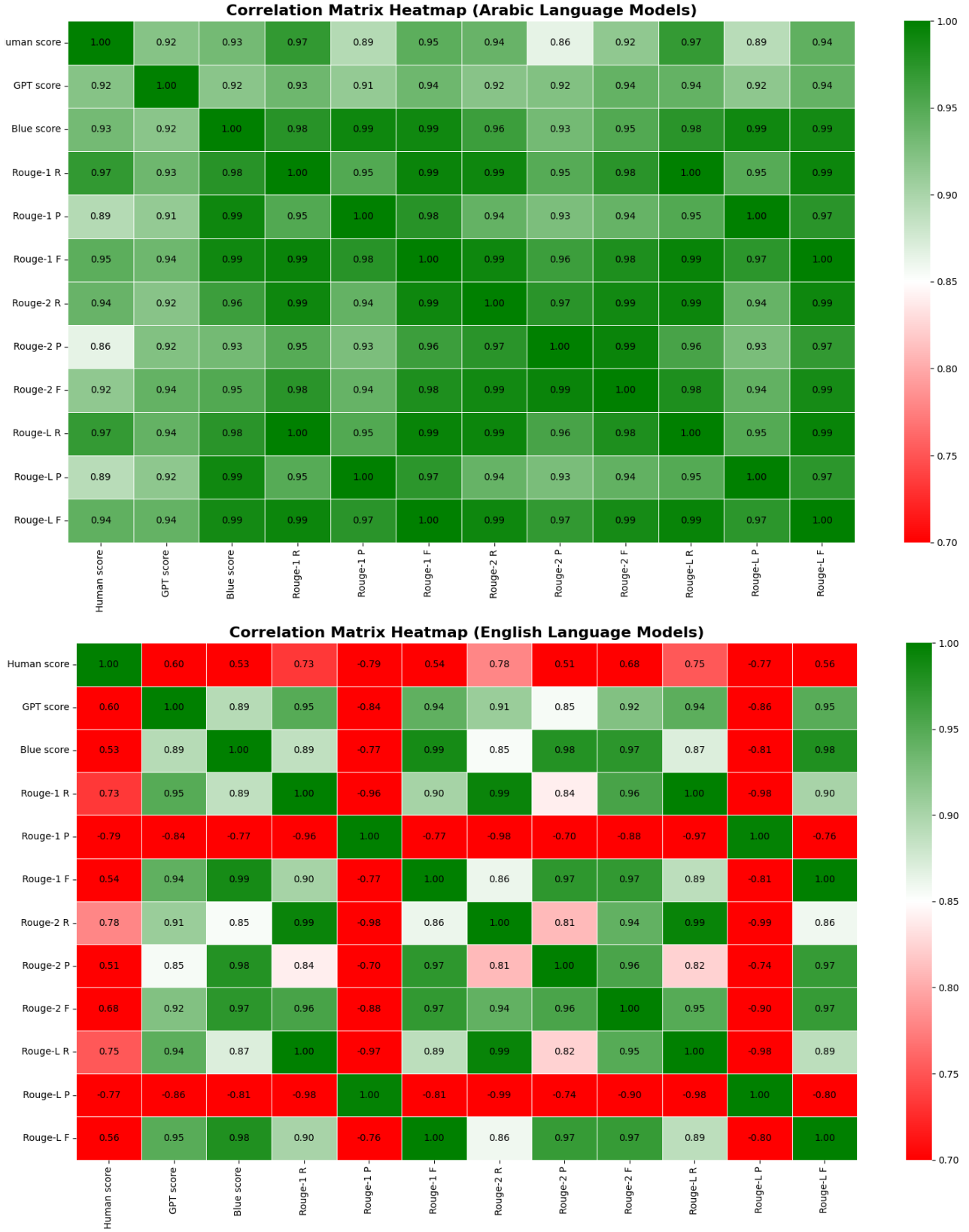

Figure 9. Heatmap of the correlation coefficient between the values of the metrics used for evaluating Arabic (**top**) and translated (**bottom**) language models.

3.3. Summary of the Results

The results presented in Section 3 provide several important insights in the context of legal ruling prediction using large language models:

- Performance of GPT-3.5-based Models: The GPT-3.5-based models outperform all other models by a wide margin, surpassing even the dedicated Arabic-centric JAIS model's average score by 50%. This is surprising since the proportion of Arabic language in JAIS's pretraining dataset is around 1000 times larger than in GPT3's pretraining dataset, and JAIS's tokenizer is a priori more adapted to Arabic than GPT's (see Section 2.1).
- Reliability of the Human Score: The human score serves as the gold standard, highlighting the nuanced comprehension humans have over automated metrics in assessing the quality of legal ruling prediction. This confirms the superiority of human skills over LLMs in certain domains that require careful reasoning [54].
- GPT Score Limitations: The GPT score, though indicative, showcases its limitations in several instances, rendering it potentially misleading. Moreover, the significant divergence between GPT scores and human evaluations, especially on translated datasets, underscores potential translation errors or inherent metric limitations.
- Inefficiency of ROUGE and BLEU: The ROUGE and BLEU scores, originally designed for translation and summarization tasks, exhibit their unsuitability for the task at hand. The low scores of these two metrics across all models can be attributed, in part, to the absence of stemming or punctuation filtering preprocessing steps applied to the data. Given the morphologically rich and highly inflectional nature of the Arabic language, exact matching—upon which the ROUGE and BLEU scores rely—is anticipated to yield lower values.

4. Conclusions

This study represents a pioneering effort in the realm of Arabic court decision analysis, shedding light on the efficacy of advanced language models in predicting legal outcomes. The findings underscore the remarkable out-performance of GPT-3.5-based models, surpassing even domain-specific models tailored for Arabic language. This unexpected outcome challenges conventional assumptions about the importance of domain specificity and dataset size in model performance. Nevertheless, in spite of the relative superiority of GPT-3.5-based models, their absolute performance on predicting Arabic legal rulings is still unsatisfactory, with an average human score of 2.4 out of 5. Better models fine-tuned on larger Arabic legal datasets need to be developed before LLMs can act as useful legal assistants.

However, it is crucial to acknowledge the limitations of this research. Firstly, the study may be constrained by the availability and quality of the dataset used, which could affect the generalizability of the findings. Additionally, the reliance on automated metrics such as GPT scores, ROUGE, and BLEU highlights the need for caution, as these metrics may not fully capture the complexity of legal language and decision-making processes. Moreover, while human evaluation serves as the gold standard, the subjectivity inherent in human judgment introduces its own set of challenges, potentially impacting the reliability of the evaluations conducted. Ideally, several human evaluators should grade the model outputs, and their evaluations should be compared to detect any possible bias or outlier scores.

On the other hand, it is important to acknowledge that the court decisions used in the study are not anonymized. Consequently, it is possible that some sensitive court decisions may not have been published on the SMOJ website, potentially introducing bias into the training dataset. This bias could impact the performance of prediction models, particularly if certain types of cases or courts are overrepresented or underrepresented in the dataset. To address this limitation and ensure the robustness of future research in this area, several potential solutions could be considered. Firstly, efforts could be made to obtain a more comprehensive and diverse dataset by collaborating with judicial authorities to access anonymized court decisions from a wider range of sources. Additionally, techniques such

as data augmentation or bias correction methods could be applied to mitigate the effects of any existing biases in the dataset. Finally, transparent reporting of dataset limitations and biases in research publications is essential for fostering a clear understanding of the study's scope and implications.

Furthermore, the study emphasizes the indispensable role of human evaluation as the gold standard for assessing the quality of legal ruling predictions. While automated metrics like GPT scores, ROUGE, and BLEU can provide valuable indications in some cases, they exhibit limitations in capturing the nuanced and context-dependent nature of legal language. The inefficacy of ROUGE and BLEU scores in this context underscores the need for tailored evaluation metrics when applying advanced language models to legal analysis tasks. Future research in this domain should focus on developing more contextually relevant evaluation measures to better reflect the accuracy and relevance of predictions in the legal context.

Overall, this study serves as a foundation for future research at the intersection of computational linguistics and Arabic legal analytics. It encourages further exploration into the potential of large language models in assisting legal professionals and policy-makers, ultimately contributing to the effective functioning of the judicial system and the enhancement of legal decision-making processes.

Author Contributions: Conceptualization, A.K. and A.A.; methodology, A.A. and A.K.; software, A.A. and O.N.; validation, A.A.; formal analysis, A.A.; investigation, A.A.; resources, A.K.; data curation, S.S. and A.A.; writing—original draft preparation, A.A., B.B., and O.N.; writing—review and editing, A.A. and B.B.; visualization, A.A. and O.N.; supervision, A.A. and A.K.; project administration, A.K.; funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Prince Sultan University grant number SEED-CCIS-2023-145. And the APC was funded by Prince Sultan University.

Data Availability Statement: Data are available on request.

Acknowledgments: The authors thank Prince Sultan University for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guellil, I.; Saâdane, H.; Azouaou, F.; Gueni, B.; Nouvel, D. Arabic natural language processing: An overview. *J. King Saud-Univ.-Comput. Inf. Sci.* **2021**, *33*, 497–507. [CrossRef]
2. Habash, N. *Introduction to Arabic Natural Language Processing*; Morgan & Claypool Publishers: Kentfield, CA, USA, 2010.
3. Shaalan, K.; Siddiqui, S.; Alkhatib, M.; Abdel Monem, A. Challenges in Arabic natural language processing. In *Computational Linguistics, Speech and Image Processing for Arabic Language*; World Scientific: Singapore, 2019; pp. 59–83.
4. Attia, M. Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. Ph.D. Thesis, University of Manchester, Manchester, UK, 2008.
5. Dai, A.M.; Le, Q.V. Semi-supervised sequence learning. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 3079–3087.
6. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. *arXiv* **2018**, arXiv:1801.06146.
7. Surden, H. Artificial intelligence and law: An overview. *Ga. State Univ. Law Rev.* **2019**, *35*, 19–22.
8. Katz, D.M.; Bommarito, M.J.; Blackman, J. A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE* **2017**, *12*, e0174698. [CrossRef] [PubMed]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
10. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *Proc. Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
11. Afzaal, M.; Imran, M.; Du, X.; Almusharraf, N. Automated and Human Interaction in Written Discourse: A Contrastive Parallel Corpus-Based Investigation of Metadiscourse Features in Machine-Human Translations. *Sage Open* **2022**, *12*. [CrossRef]
12. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011.
13. Khan, A. Improved multi-lingual sentiment analysis and recognition using deep learning. *J. Inf. Sci.* **2023**. [CrossRef]
14. Chaudhry, H.N.; Javed, Y.; Kulsoom, F.; Mehmood, Z.; Khan, Z.I.; Shoaib, U.; Janjua, S.H. Sentiment analysis of before and after elections: Twitter data of us election 2020. *Electronics* **2021**, *10*, 2082. [CrossRef]

15. 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* **2022**, arXiv:2211.09110.
16. Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A.A.M.; Abid, A.; Fisch, A.; Brown, A.R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv* **2022**, arXiv:2206.04615.
17. Elmadany, A.; Nagoudi, E.M.B.; Abdul-Mageed, M. ORCA: A Challenging Benchmark for Arabic Language Understanding. *arXiv* **2022**, arXiv:2212.10758.
18. Abdelali, A.; Mubarak, H.; Chowdhury, S.A.; Hasanain, M.; Mousi, B.; Boughorbel, S.; Kheir, Y.E.; Izham, D.; Dalvi, F.; Hawasly, M.; et al. Benchmarking Arabic AI with Large Language Models. *arXiv* **2023**, arXiv:2305.14982.
19. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. *arXiv* **2022**, arXiv:2212.04356.
20. Zhang, Y.; Han, W.; Qin, J.; Wang, Y.; Bapna, A.; Chen, Z.; Chen, N.; Li, B.; Axelrod, V.; Wang, G.; et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv* **2023**, arXiv:2303.01037.
21. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv* **2023**, arXiv:2302.11382.
22. Zhou, Y.; Muresanu, A.I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; Ba, J. Large language models are human-level prompt engineers. *arXiv* **2022**, arXiv:2211.01910.
23. Shin, T.; Razeghi, Y.; Logan IV, R.L.; Wallace, E.; Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv* **2020**, arXiv:2010.15980.
24. Lauderdale, B.E.; Clark, T.S. Scaling politically meaningful dimensions using texts and votes. *Am. J. Political Sci.* **2014**, *58*, 754–771. [CrossRef]
25. Medvedeva, M.; Vols, M.; Wieling, M. Using machine learning to predict decisions of the European Court of Human Rights. *Artif. Intell. Law* **2019**, *27*, 237–266. [CrossRef]
26. AL-Qurishi, M.; AlQaseemi, S.; Soussi, R. AraLegal-BERT: A pretrained language model for Arabic Legal text. *arXiv* **2022**, arXiv:2210.08284. [CrossRef]
27. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. *arXiv* **2020**, arXiv:2010.02559.
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
29. 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* **2023**, arXiv:2302.13971.
30. Sengupta, N.; Sahu, S.K.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Afzal, O.M.; Kamboj, S.; Pandit, O.; Pal, R.; et al. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv* **2023**, arXiv:2308.16149.
31. 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. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
32. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
33. Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. *Preprints* **2023**, 2023030422. [CrossRef]
34. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
35. OpenAI. GPT3 Dataset Language Statistics. Available online: https://github.com/openai/gpt-3/tree/master/dataset_statistics (accessed on 9 October 2023).
36. Shibata, Y.; Kida, T.; Fukamachi, S.; Takeda, M.; Shinohara, A.; Shinohara, T.; Arikawa, S. *Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching*; Technical Report DOI-TR-161; Department of Informatics, Kyushu University: Fukuoka, Japan, 1999.
37. Bostrom, K.; Durrett, G. Byte pair encoding is suboptimal for language model pretraining. *arXiv* **2020**, arXiv:2004.03720.
38. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
39. French, R.M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **1999**, *3*, 128–135. [CrossRef]
40. Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; Kanan, C. Measuring catastrophic forgetting in neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
41. Pu, G.; Jain, A.; Yin, J.; Kaplan, R. Empirical Analysis of the Strengths and Weaknesses of PEFT Techniques for LLMs. *arXiv* **2023**, arXiv:2304.14999.
42. Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.P.; Lee, R.K.W.; Bing, L.; Poria, S. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv* **2023**, arXiv:2304.01933.
43. AGI-Edgerunners. LLM-Adapters Github Repository. Available online: https://github.com/AGI-Edgerunners/LLM-Adapters (accessed on 9 October 2023).
44. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
45. SJP. Saudi Justice Portal. Available online: https://sjp.moj.gov.sa (accessed on 5 October 2023).
46. PyPI. Selenium Python Library. Available online: https://pypi.org/project/selenium (accessed on 5 October 2023).

47. PyPI. Beautiful Soup Python Package. Available online: https://pypi.org/project/bs4 (accessed on 4 October 2023).
48. PyPI. Translators Python Package. Available online: https://pypi.org/project/translators/ (accessed on 28 September 2023).
49. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
50. Chen, B.; Cherry, C. A systematic comparison of smoothing techniques for sentence-level BLEU. In Proceedings of the 9th Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 362–367.
51. NLTK. Bleu Python Package. Available online: https://www.nltk.org/api/nltk.translate.bleu_score.html (accessed on 4 October 2023).
52. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
53. PyPI. Rouge Python Package. Available online: https://pypi.org/project/rouge (accessed on 4 October 2023).
54. Koubaa, A.; Qureshi, B.; Ammar, A.; Khan, Z.; Boulila, W.; Ghouti, L. Humans are still better than chatgpt: Case of the ieeextreme competition. *arXiv* **2023**, arXiv:2305.06934.