

Automated Generation of Arabic Procedural Court Orders in the Moroccan Legal Context

Saliha Yassine, Mohammed Essoujaa, Mustapha Esghir, Ouafaa Ibrihich

Faculty of Sciences, Mohammed V University in Rabat, Laboratory MIA-SI, Rabat, Morocco

Abstract—Automating the drafting of legal documents can reduce repetitive judicial workload and support procedural efficiency in e-Justice systems. However, Arabic legal text generation remains underexplored, particularly for Moroccan procedural court orders, which are characterized by highly structured language and recurring legal formulations. This study compares retrieval-based and transformer-based approaches for generating Moroccan procedural court orders using a corpus of 4,098 authentic orders collected from 2020 to 2025. Three approaches are evaluated: a TF-IDF-based Template Baseline, AraBART, and AraT5. A random stratified split and a stricter template-disjoint evaluation are used to assess both performance and generalization. Under the random split, the Template Baseline achieves the highest BLEU score (87.39), followed by AraBART (77.33) and AraT5 (51.88). In contrast, under the template-disjoint setting, AraBART outperforms the Template Baseline, achieving BLEU scores of 68.00 and 63.67, respectively, with a statistically significant difference ($p = 0.004$). These findings demonstrate that retrieval-based generation can benefit from repeated legal templates, whereas AraBART achieves better performance under the template-disjoint setting when exact templates are absent. The results highlight the importance of evaluation design and corpus structure when selecting generation approaches for standardized legal documents. A conceptual hybrid strategy combining retrieval, neural generation, and human validation is proposed for future e-Justice applications.

Keywords—*Arabic legal NLP; legal text generation; moroccan procedural court orders; retrieval-based generation; AraBART; AraT5; e-Justice*

I. INTRODUCTION

The rapid digitalization of judicial systems has increased interest in Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies for supporting judicial document drafting and administrative procedures. In Morocco, procedural court orders constitute an essential component of judicial administration, supporting routine legal procedures such as payment orders, notifications, deposits, and other administrative decisions. Unlike judicial judgments, these documents are generally concise, highly structured, and frequently follow standardized legal formulations. Automating their drafting has the potential to improve document consistency, accelerate routine judicial procedures, and support the development of intelligent e-Justice services while allowing legal professionals to focus on more complex judicial tasks [24].

Recent advances in transformer-based models and large language models have significantly improved automatic text generation across numerous domains, including legal NLP [1]. Arabic generative models such as AraBART and AraT5 have

demonstrated promising performance by leveraging large-scale pretraining and contextual language representations. However, applying these models to procedural legal documents presents a unique challenge. Because many procedural court orders contain recurring legal expressions and standardized templates, it remains unclear whether computationally intensive neural generation models consistently provide advantages over simpler retrieval-based approaches. Consequently, selecting the most appropriate generation paradigm depends not only on the model architecture but also on the intrinsic characteristics of the legal corpus.

Existing research on legal text generation has primarily focused on judgment summarization, legal question answering, statute retrieval, and general legal document drafting, with most studies conducted for English- or Chinese-language legal systems. Arabic legal NLP remains comparatively underexplored, particularly for generative tasks involving authentic judicial documents. Although recent Arabic transformer models, including AraBART and AraT5, have achieved encouraging results across several natural language generation tasks, their effectiveness for generating procedural court orders has received little attention. Furthermore, previous studies rarely examine how corpus repetitiveness influences the comparative performance of retrieval-based and neural generation methods in highly standardized legal domains.

This observation reveals an important research gap. To the best of our knowledge, no prior study has reported a systematic comparison of retrieval-based and transformer-based generation for Moroccan procedural court orders while explicitly analyzing the influence of corpus repetition on model selection. Addressing this question is particularly important because highly repetitive legal corpora may favor retrieval-based methods when relevant templates are available in the retrieval database, whereas neural generation models may achieve better performance when exact decision templates are absent from the training data. Understanding this trade-off is essential for designing efficient, reliable, and practically deployable legal text generation systems for Moroccan courts and other legal environments characterized by standardized procedural documents.

To address this research gap, this study presents a comprehensive comparative study of retrieval-based and transformer-based approaches for generating Moroccan procedural court orders. Rather than assuming that increasingly complex neural models always produce superior results, this work investigates how corpus characteristics, particularly the degree of textual repetition, influence the effectiveness of different generation paradigms.

The main contributions of this study are summarized as follows:

- Use and systematic characterization of a corpus comprising 4,098 authentic Moroccan procedural court orders from 2020 to 2025, representing a systematically described dataset for Arabic legal text generation in this domain.
- Comprehensive comparative evaluation of a TF-IDF-based Template Baseline, AraBART, and AraT5 using complementary automatic evaluation metrics, enabling a systematic assessment of retrieval-based and transformer-based generation methods.
- Corpus repetition analysis demonstrating how the degree of textual repetition affects the relative performance of retrieval-based and transformer-based generation across different evaluation settings.
- A conceptual hybrid generation strategy derived from the experimental findings, recommending retrieval-based generation for highly repetitive cases and transformer models for more diverse drafting scenarios under human supervision, providing guidance for future intelligent e-Justice systems.

The remainder of this study is organized as follows. Section II reviews related work on Arabic legal NLP and legal text generation. Section III describes the proposed methodology, including corpus construction, preprocessing, generation models, and evaluation metrics. Section IV presents and discusses the experimental results. Finally, Section V concludes the study and outlines directions for future research.

II. RELATED WORK

Legal Natural Language Processing (Legal NLP) has rapidly evolved from rule-based systems to transformer-based architectures, enabling significant advances in legal document classification, information extraction, judgment prediction, and text generation. Among these tasks, legal text generation has recently attracted increasing attention because of its potential to support judicial drafting, reduce repetitive workload, and improve the efficiency of e-Justice systems.

Domain-specific pre-training has proven effective across various legal NLP applications. LEGAL-BERT [2], which builds upon the BERT architecture [20], achieves strong performance for European court judgment prediction, while Zhong et al. [3] examined the application of NLP methods to several tasks in the Chinese legal system. Large-scale legal datasets such as CAIL2018 [16] have further supported research on judgment prediction. In legal text generation, Feijo et al. [4] demonstrated the feasibility of GPT-2 for automatic contract generation in Brazilian law. In the Arabic NLP domain, AraBERT [5] established a strong baseline for Arabic language understanding, whereas AraBART [6] and AraT5 [7] introduced transformer-based sequence-to-sequence architectures that have shown promising performance in Arabic text generation and summarization tasks. AraBART is grounded in the BART architecture [17] and AraT5 in the T5 text-to-text framework [18], both building on the foundational Transformer architecture [21]. Bidirectional Arabic pretraining models such as ARBERT

and MARBERT [22] and comprehensive Arabic NLP resources [25] have further supported the development of these models.

Retrieval-based and case-based approaches have also been explored as alternatives to fully generative systems, particularly in domains where documents exhibit strong structural regularities. Such approaches can preserve established domain-specific formulations by reusing previously validated outputs. However, their effectiveness for Moroccan procedural court orders in Arabic has not been systematically compared with neural generation models.

More recently, large language models and retrieval-enhanced generation have further expanded the potential of legal text generation. General-purpose LLMs such as GPT-3 [23] and Llama 2 [8] and retrieval-augmented generation frameworks [9] illustrate the growing trend toward combining knowledge retrieval with neural generation. Dedicated benchmarks such as LegalBench-RAG [10] and Arabic-specific models such as AceGPT [11] further demonstrate the breadth of this research landscape. Although these approaches are promising, their application to highly standardized Arabic procedural court orders remains largely unexplored. Table I summarizes representative studies in Legal NLP and Arabic text generation. Existing research has mainly focused on legal judgment prediction, legal document classification, contract generation, and general-purpose Arabic text generation. However, the automatic generation of Moroccan procedural court orders has not yet been systematically investigated. Moreover, previous studies have not examined how corpus repetitiveness influences the relative effectiveness of retrieval-based and transformer-based generation approaches. To address this gap, the present work investigates how corpus repetitiveness influences the comparative performance of retrieval-based and transformer-based generation, providing practical guidance for designing efficient Arabic legal text generation systems for highly structured legal domains.

TABLE I. REPRESENTATIVE STUDIES IN LEGAL NLP AND ARABIC TEXT GENERATION

Study	Domain	Task	Main Method
Chalkidis et al. [2]	EU courts	Judgment prediction	LEGAL-BERT
Zhong et al. [3]	Chinese legal system	Multiple legal NLP tasks	Legal NLP methods
Feijo et al. [4]	Brazilian law	Contract generation	GPT-2
Antoun et al. [5]	Arabic NLP	Text classification	AraBERT
Kamal et al. [6]	Arabic NLP	Text generation	AraBART
Nagoudi et al. [7]	Arabic NLP	Text generation	AraT5
Sulea et al. [12]	French law	Legal classification	SVM + TF-IDF
Medvedeva et al. [13]	ECHR	Decision prediction	Linear SVM + textual n-grams

Although transformer-based models have demonstrated strong performance in Arabic text generation, their application to highly repetitive procedural legal documents raises a methodological question: whether generative flexibility provides a genuine advantage when valid judicial formulations

are already available in historical records. Retrieval-based approaches preserve established legal formulations and can reduce the risk of generative deviations at low computational cost, but they are limited in their ability to generate formulations unseen in the training corpus. Transformer-based approaches offer greater generative flexibility and can adapt to novel inputs, but remain exposed to hallucination and semantic deviation. This trade-off motivates the use of TF-IDF character n-gram retrieval, rather than a purely generative approach, as a baseline for highly standardized procedural court orders.

III. METHOD

A. Dataset

We compiled a corpus of 4,098 authentic Moroccan procedural court orders issued by first-instance courts between 2020 and 2025. The corpus was previously used for a classification study of Moroccan procedural court orders [26]; the present work addresses a distinct research question focused on automatic text generation and the comparative analysis of retrieval-based and transformer-based approaches. Each document consists of two main sections: *faits* (procedural facts describing the legal request) and *décision* (the judicial ruling issued by the court). In this study, the *faits* section serves as the input to all generation models, while the corresponding *décision* is used as the reference output during training and evaluation. Table II summarizes the main characteristics of the corpus. The corpus covers four procedural categories representing the most frequent types of procedural court orders handled by Moroccan first-instance courts.

TABLE II. CORPUS STATISTICS

Statistic	Value
Total documents	4,098
Training set	3,278 (80%)
Validation set	410 (10%)
Test set	410 (10%)
Average facts length (characters)	323
Average decision length (characters)	159
Number of procedural categories	4
Global repetition rate (training set)	64.0%
Random seed	42

To ensure a fair, reproducible, and representative evaluation, the corpus was partitioned using a stratified 80/10/10 split, resulting in 3,278 training documents, 410 validation documents, and 410 test documents. Stratification preserves the distribution of procedural categories across all subsets, while a fixed random seed (42) ensures the reproducibility of the data partition. Because the corpus contains repeated judicial formulations, the random stratified split may place identical decision formulations in both the training and test sets. This setting reflects the intended retrieval scenario but may favor retrieval-based methods on highly standardized cases. To address this limitation, an additional template-disjoint evaluation was conducted. Decision texts were grouped by their exact string content after normalizing whitespace, and a group-

based split (scikit-learn's GroupShuffleSplit) was used to ensure that all documents sharing an identical decision formulation were assigned entirely to the same subset (training, validation, or test), thereby preventing any decision string from appearing in more than one split.

This evaluation provides a more stringent assessment of generalization beyond previously observed judicial templates. The resulting template-disjoint partition contained 3,364 training documents, 354 validation documents, and 380 test documents.

B. Corpus Analysis: Repetition Structure

A preliminary analysis of the training corpus reveals a global repetition rate of 64.0%, corresponding to 1,180 unique judicial decisions among 3,278 training documents. In this study, a repeated decision refers to an identical judicial ruling appearing in multiple procedural court orders after whitespace normalization. The repetition rate was computed by comparing unique decision texts within each procedural category, providing a quantitative measure of the degree of standardization in judicial drafting.

Fig. 1 illustrates the distribution of procedural categories in the training set, while Table III summarizes the corresponding repetition rates, showing that rejection orders exhibit the highest repetition (89.9%), followed by validation-of-warning orders (75.9%). In contrast, payment orders show considerably greater textual diversity, with a repetition rate of only 27.0%.

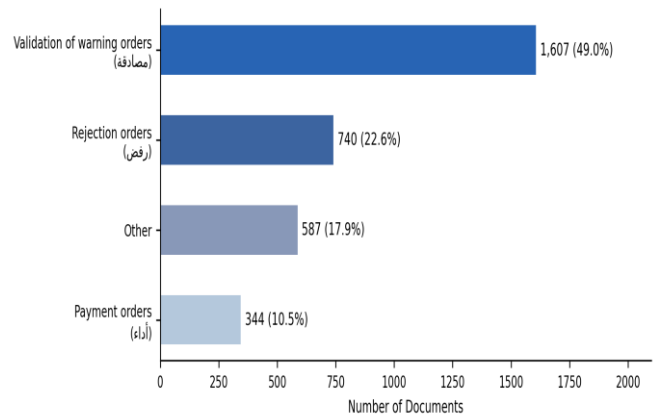


Fig. 1. Distribution of procedural court order categories in the training set (n = 3,278).

TABLE III. REPETITION ANALYSIS OF THE TRAINING CORPUS BY PROCEDURAL CATEGORY

Category	Documents	Unique Decisions	Repetition Rate (%)
Validation of warning orders (مصادقة)	1,607	388	75.9%
Rejection orders (رفض)	740	75	89.9%
Payment orders (أداء)	344	251	27.0%
Other	587	466	20.6%
Total (training set)	3,278	1,180	64.0%

These observations indicate that corpus repetition is a key factor in selecting an appropriate generation paradigm and

motivate the inclusion of a retrieval-based Template Baseline in the comparative evaluation.

Rejection orders exhibit the highest repetition rate (89.9%), reflecting the highly formulaic nature of procedural refusals under Moroccan procedural law. Validation-of-warning orders (75.9%) follow a similarly standardized structure. This observation explains the strong performance of the Template Baseline.

C. Systems

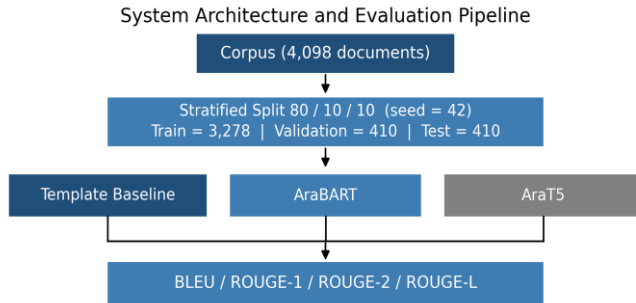


Fig. 2. Experimental framework for retrieval-based and transformer-based generation of Moroccan procedural court orders.

Fig. 2 illustrates the main experimental framework used for the random-split evaluation. An additional template-disjoint evaluation was conducted as described in Section III-A. Three complementary approaches were evaluated: a retrieval-based Template Baseline, AraBART, and AraT5. The retrieval-based system represents a non-neural approach that exploits the high repetition rate of procedural court orders, whereas AraBART and AraT5 represent neural sequence-to-sequence architectures capable of generating more flexible judicial texts.

1) *Template baseline*: For each test document, the most similar training document is retrieved using TF-IDF with character n-grams (3–5) and cosine similarity. The retrieved document’s judicial decision is then returned as the generated output. Character n-grams were selected because they are more robust than word-based representations for Arabic text, effectively handling morphological variations, affixes, spelling differences, and tokenization inconsistencies while preserving lexical similarity between highly repetitive legal documents. This approach requires no neural training and directly exploits the standardized nature of procedural court orders.

2) *AraBART*: We fine-tuned moussaKam/AraBART [6], a BART-based sequence-to-sequence model [17] pre-trained on approximately 70 GB of Arabic text. AraBART was selected because it provides an Arabic sequence-to-sequence architecture designed for text generation and has demonstrated strong performance on Arabic generation tasks. The facts section is provided as input, while the corresponding decision is generated as output, with both input and output sequences limited to 128 tokens.

3) *AraT5*: We fine-tuned UBC-NLP/AraT5-base-title-generation [7], an Arabic adaptation of the T5 text-to-text framework [18]. AraT5 was included as a complementary Arabic text-to-text baseline for comparison with AraBART.

Although this checkpoint was originally developed for title-generation tasks, it was included here as a complementary Arabic T5 baseline to assess its transferability to procedural legal text generation. The same input-output configuration and maximum sequence length were used to ensure a fair comparison with AraBART. A preliminary experiment with mT5-small [19] did not produce coherent Arabic judicial text and was therefore excluded from the main comparison; this result should not be interpreted as a general limitation of the mT5 architecture.

D. Hyperparameters

The hyperparameters used for fine-tuning AraBART and AraT5 are summarized in Table IV. To ensure a fair comparison between both transformer models, identical training settings were adopted whenever applicable while maintaining computational feasibility on the available hardware.

TABLE IV. TRAINING HYPERPARAMETERS

Parameter	AraBART	AraT5
Pre-trained model	moussaKam/AraBART	UBC-NLP/AraT5-base-title-generation
Training epochs	3	3
Batch size	8	4
Max input/output length (tokens)	128 / 128	128 / 128
Learning rate	5×10^{-5}	5×10^{-5}
Warmup steps	100	100
Precision	fp16	fp16
Hardware	Kaggle T4 GPU	Kaggle T4 GPU
Optimizer	AdamW	AdamW

E. Evaluation Metrics

The generated judicial decisions were evaluated using the BLEU [14] and ROUGE [15] metrics, which are widely adopted for automatic text generation evaluation. BLEU measures modified n-gram precision with a brevity penalty by comparing the generated text with the reference decision, whereas ROUGE evaluates recall-based lexical overlap. Specifically, ROUGE-1, ROUGE-2, and ROUGE-L were computed to assess unigram, bigram, and longest common subsequence similarity, respectively.

Because retrieval-based and neural generation models exhibit different generation behaviors, multiple evaluation metrics were employed to provide a comprehensive assessment. BLEU is particularly informative for highly repetitive legal documents where exact lexical matching is expected, while ROUGE captures partial lexical overlap when generated texts differ from the reference, complementing BLEU in cases where exact n-gram matching is limited.

Because Arabic is morphologically rich, BLEU and ROUGE may underestimate semantic similarity when equivalent legal expressions differ only in their morphological or orthographic forms. Therefore, these metrics are interpreted jointly with the qualitative analysis presented in Section IV-C to provide a more comprehensive evaluation of generation quality.

IV. RESULTS AND DISCUSSION

A. Quantitative Comparison of Generation Models

TABLE V. PERFORMANCE COMPARISON ON 410 TEST DOCUMENTS

Method	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Template Baseline	87.39	21.30	15.29	21.14
AraBART	77.33	20.01	13.92	20.01
AraT5	51.88	19.00	12.81	19.00

Table V summarizes the quantitative performance of the three evaluated generation approaches on the 410-document test set. The retrieval-based Template Baseline achieves the highest overall performance, obtaining a BLEU score of 87.39, followed by AraBART (77.33) and AraT5 (51.88). Similar performance trends are observed across the ROUGE metrics, where the Template Baseline consistently outperforms both transformer-based models.

Statistical significance analysis: To assess whether the observed differences between the generation approaches were statistically significant, a paired bootstrap analysis was conducted over the 410 test documents using 2,000 bootstrap replications. The analysis was performed on document-level BLEU scores, with each comparison using the same test documents for both methods. The Template Baseline significantly outperformed AraBART, with a mean BLEU difference of 10.07 points (95% CI: [7.74, 12.34], $p < 0.001$). Similarly, the Template Baseline significantly outperformed AraT5 by 35.51 BLEU points (95% CI: [32.59, 38.38], $p < 0.001$). AraBART also significantly outperformed AraT5, with a difference of 25.45 BLEU points (95% CI: [23.03, 27.66], $p < 0.001$). These results confirm that the observed performance differences are statistically significant under the present test-set configuration.

The results under the random split initially show a strong advantage for the Template Baseline, which is consistent with the highly repetitive nature of Moroccan procedural court orders. However, the template-disjoint evaluation provides a more stringent test of generalization and reverses this ranking, with AraBART achieving a significantly higher BLEU score than the Template Baseline. This contrast indicates that the apparent superiority of retrieval-based generation under the random split is partly associated with the availability of repeated decision formulations across the training and test sets. Since many judicial decisions follow standardized legal formulations, retrieving the most similar precedent frequently produces an output that closely matches the reference decision. Consequently, retrieval-based generation can be particularly effective for highly standardized procedural court orders when sufficiently similar validated precedents are available in the retrieval corpus.

Among the evaluated checkpoints, AraBART achieves higher BLEU, ROUGE-1, and ROUGE-L scores than the selected AraT5 checkpoint (UBC-NLP/AraT5-base-title-generation), while AraT5 obtains a slightly higher ROUGE-2 score than AraBART under the template-disjoint setting; this

discrepancy may partly reflect checkpoint mismatch rather than a general architectural disadvantage of AraT5.

Under the random split, the Template Baseline achieves the highest scores, while the relatively small differences in ROUGE-1 (21.30 vs. 20.01) and ROUGE-L (21.14 vs. 20.01) indicate that transformer models may generate expressions that differ lexically from the reference text. This observation highlights the complementary roles of BLEU and ROUGE in evaluating legal text generation, particularly when comparing retrieval-based and neural generation approaches.

B. Template-Disjoint Evaluation

To assess the generalization of the evaluated approaches beyond repeated decision templates, an additional template-disjoint evaluation was conducted. In contrast to the initial random stratified split, the template-disjoint protocol prevents documents sharing an identical decision string from being assigned to different subsets. The resulting test set contains 380 documents and provides a more stringent assessment of model generalization to decision formulations not represented in the training set. The AraBART and AraT5 models were re-trained using the corresponding template-disjoint training set, while the Template Baseline retrieved candidate decisions only from that same training set.

The results reveal a substantially different performance pattern from that obtained under the random split. As shown in Table VI, AraBART achieved the highest BLEU score (68.00), followed by the Template Baseline (63.67) and AraT5 (39.89). AraBART also obtained higher ROUGE-1, ROUGE-2, and ROUGE-L scores than the Template Baseline, indicating better lexical overlap with the reference decisions under the template-disjoint setting.

TABLE VI. TEMPLATE-DISJOINT EVALUATION ON 380 TEST DOCUMENTS

Method	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Template Baseline	63.67	19.37	8.99	19.20
AraBART	68.00	26.32	16.54	26.32
AraT5	39.89	25.61	17.09	25.61

The difference between the two evaluation settings is particularly informative. Under the random split, the Template Baseline achieved a BLEU score of 87.39 compared with 77.33 for AraBART. Under the template-disjoint split, however, AraBART outperformed the Template Baseline by 4.33 BLEU points. This contrast indicates that retrieval-based generation benefits substantially from the presence of repeated decision formulations across the training and test sets, whereas AraBART achieves better performance when exact decision templates are excluded from the training data.

To determine whether this difference was statistically significant, a paired bootstrap analysis with 2,000 replications was performed on document-level BLEU scores using the same 380 test documents for each method. AraBART significantly outperformed the Template Baseline, with a BLEU difference of 4.33 points (95% CI: [1.60, 6.95], $p = 0.004$). The Template Baseline also significantly outperformed AraT5 by 23.78 BLEU

points ($p < 0.001$), while AraBART significantly outperformed AraT5 by 28.11 BLEU points ($p < 0.001$). These results confirm that the advantage of AraBART over the retrieval-based baseline under the template-disjoint setting is statistically significant.

Overall, the template-disjoint evaluation provides complementary evidence to the random-split results. Rather than indicating that one generation paradigm is universally superior, the two-evaluation settings show that the relative performance of retrieval-based and generative approaches is strongly influenced by the degree of template overlap between training and test data.

C. Qualitative Analysis of Generated Judicial Decisions

TABLE VII. QUALITATIVE EXAMPLES — RANDOM-SPLIT TEST SET (EXCERPTS)

Facts (excerpt)	Reference	Template	AraBART	Observation
طلب مقدم بتاريخ 06 يناير 2022 يلتبس تبليغ إنذار بأداء...	نوافق على الطلب و نأمر بتبليغ الإنذار المرفق بالمقال للمطلوب. We approve the request and order the notification of the warning attached to the petition to the respondent (fem.).	نوافق على الطلب و نأمر بتبليغ الإنذار للمطلوب. We approve the request and order the notification of the warning attached to the petition to the respondent (masc.).	نوافق على الطلب و نأمر بتبليغ الإنذار بالأداء المرفق بالمقال. We approve the request and order the notification of the payment warning attached to the petition.	Minor morphological variation
مقال مسجل بتاريخ 07 نونبر 2023 يلتبس الإنذار بتبليغ بالأداء...	نوافق على الطلب ونأمر بتبليغ الإنذار بالأداء المرفق بالمقال للمطلوب. We approve the request and order the notification of the payment warning attached to the petition to the respondent.	نوافق على الطلب و نأمر بتبليغ الإنذار المرفق بالمقال للمطلوب. [Identical to Reference]	نوافق على الطلب ونقرر المصادقة على الإنذار المرفق بالطلب. We approve the request and decide to validate the warning attached to the application.	Semantic deviation (different judicial action)
مقال مقدم بتاريخ 28 نونبر 2023 يلتبس تبليغ إنذار بأداء دين...	نوافق على الطلب و نأمر بتبليغ الإنذار بالأداء المرفق بالمقال للمطلوب. We approve the request and order the notification of the payment warning attached to the petition to the respondent.	نوافق على الطلب و نأمر بتبليغ الإنذار بالمقال للمطلوب. [Identical to Reference]	نوافق على الطلب و نأمر بتبليغ الإنذار المرفق للجهة المطلوبة. We approve the request and order the notification of the warning attached to the petition to the respondent party.	Alternative wording preserved legal meaning

Table VII focuses on the Template Baseline and AraBART, the two approaches selected for qualitative comparison, to illustrate the contrast between retrieval-based and neural generation. AraT5 is not included because it obtained lower BLEU, ROUGE-1, and ROUGE-L scores than AraBART across the two evaluation settings, although it slightly outperformed AraBART in ROUGE-2 under the template-disjoint setting. Table VII presents representative examples illustrating the

qualitative behavior of the retrieval-based Template Baseline and AraBART on held-out test documents. While automatic metrics quantify overall performance, these examples illustrate the different behaviors of retrieval-based and transformer-based generation.

The Template Baseline frequently retrieves judicial decisions that are identical or nearly identical to the reference text, reflecting the high level of standardization observed in Moroccan procedural court orders. In the second and third examples, the retrieved decisions exactly match the reference output. In the first example, only a minor morphological agreement variation, المطلوب instead of المطلوبة, is observed, while the underlying procedural action remains unchanged.

In contrast, AraBART generally generates fluent and grammatically well-formed judicial text, but it may introduce semantic deviations. As illustrated in the first and third examples, the generated decisions preserve the overall legal intent while using alternative legal expressions. However, the second example shows that AraBART may also generate a different judicial action (validation of the warning) instead of the expected notification of the warning, leading to a legally inappropriate output despite its linguistic fluency.

Overall, the qualitative analysis complements the quantitative results reported in Table V and illustrates the behavior of the two approaches under the random-split setting. Retrieval-based generation can produce highly faithful outputs when sufficiently similar precedents are available, whereas transformer-based generation provides greater linguistic flexibility but may remain susceptible to semantic deviations in legally sensitive contexts.

To complement the illustrative examples in Table VII, all 410 test documents were automatically classified into seven output categories by comparing each prediction with its reference decision, using text similarity, judicial-action keywords, and monetary-amount comparison. This automatic, similarity-based classification is not equivalent to expert legal review, but it provides a systematic, corpus-wide complement to the qualitative examples above. Table VIII summarizes the results for AraBART and the Template Baseline.

TABLE VIII. SYSTEMATIC ERROR ANALYSIS ACROSS THE RANDOM-SPLIT TEST SET

Output Category	AraBART	Template	Description
Exact match	30.5%	60.7%	Output identical to the reference
Lexical variation	26.1%	8.8%	Alternative phrasing, meaning preserved
Morphological variation	23.4%	10.7%	Minor difference (agreement, punctuation)
Numerical substitution	9.8%	13.2%	Amount (and sometimes identity) differs from reference
Procedural omission	2.4%	1.2%	Expected amount or element missing
Semantic deviation	0.2%	0.5%	Different judicial action (e.g., validation instead of notification)
Other/major deviation	7.6%	4.9%	Differences not classified above

A closer analysis of AraBART's outputs further reveals that 9.8% of test documents were automatically categorized as involving numerical or beneficiary-identity substitution. In several cases, recurring names or amounts were observed across different inputs, suggesting that the model may sometimes rely on frequently observed patterns rather than fully conditioning its output on the input facts. This observation requires further expert validation and should therefore be interpreted as an indication rather than a definitive causal explanation. Although automatic metrics such as BLEU and ROUGE do not necessarily distinguish this type of error from a minor lexical variation, it may constitute a materially more serious risk in a judicial deployment context, reinforcing the need for systematic human validation before any generated decision is adopted in practice.

D. Practical Implications and Recommended Generation Strategy

The combined experimental results demonstrate that no single generation approach is universally optimal for Moroccan procedural court orders; performance depends substantially on the availability of repeated templates and the evaluation setting.

For highly standardized procedural categories, where sufficiently similar and validated precedents are available, retrieval-based generation can be particularly suitable. By retrieving an existing precedent, the Template Baseline preserves the original judicial wording and can achieve strong quantitative performance without requiring computationally intensive neural inference. However, the template-disjoint evaluation shows that this advantage does not necessarily extend to cases where exact decision templates are absent from the training data.

For procedural categories exhibiting greater textual diversity, transformer-based models offer increased linguistic flexibility by generating decisions from the procedural facts; however, their outputs require legal validation, as semantic deviations may occur. Among the evaluated transformer models, AraBART achieves the highest BLEU, ROUGE-1, and ROUGE-L scores in the template-disjoint evaluation, while AraT5 obtains a slightly higher ROUGE-2 score than AraBART, making AraBART the preferred neural alternative when retrieval-based generation is less appropriate.

Based on the complementary results obtained under the random and template-disjoint evaluation settings, we propose a conceptual hybrid generation strategy for future practical e-Justice applications, illustrated in Fig. 3. Retrieval-based generation should be employed when a sufficiently similar precedent is identified with high confidence. In contrast, a transformer model such as AraBART may be preferable when exact templates are unavailable. In all cases involving low-confidence retrievals or legally sensitive decisions, human validation remains essential before the generated text is adopted in judicial practice.

The present study is limited to Moroccan procedural court orders and therefore does not establish cross-jurisdictional generalizability. Differences in legal terminology, procedural rules, document structure, and drafting conventions may substantially affect model performance in other jurisdictions.

Future studies should evaluate the proposed methodology on procedural legal corpora from other Arabic-speaking jurisdictions. Due to the legal and privacy constraints associated with judicial documents, the full corpus cannot be publicly released.

A further limitation concerns the absence of systematic human evaluation by legal professionals. The qualitative analysis presented in Table VII and the automatic error categorization in Table VIII provide complementary evidence, but they cannot replace expert assessment of legal correctness, semantic fidelity, and procedural appropriateness. A future study should involve legal experts and report inter-annotator agreement to provide a more rigorous human-centered evaluation.

The study also does not provide a systematic quantitative analysis of computational cost, including training time, inference latency, and memory consumption. Such measurements should be incorporated in future experiments to provide a more complete assessment of deployment feasibility.

Although the present study focuses on Moroccan procedural court orders, the proposed strategy may also be applicable to other legal domains characterized by varying levels of textual standardization. Future work may investigate adaptive systems capable of automatically selecting the most appropriate generation paradigm according to corpus characteristics and retrieval confidence.

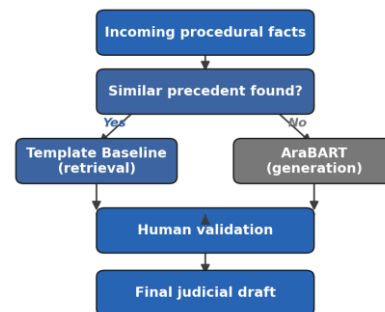


Fig. 3. Conceptual hybrid generation strategy recommended for Moroccan procedural court orders.

V. CONCLUSION

This study presented a comparative study of retrieval-based and transformer-based generation for Moroccan procedural court orders. Using a corpus of 4,098 authentic Moroccan procedural court orders, we compared a retrieval-based Template Baseline with two Arabic sequence-to-sequence models, AraBART and AraT5, under complementary evaluation settings.

Under the random stratified split, the Template Baseline achieved the highest BLEU score (87.39), reflecting the strong repetition present in the corpus. However, the additional template-disjoint evaluation produced a different result: AraBART achieved a BLEU score of 68.00 compared with 63.67 for the Template Baseline, and the 4.33-point difference was statistically significant ($p = 0.004$). This contrast demonstrates that retrieval-based performance is strongly influenced by the presence of repeated templates across training

and test data, whereas AraBART achieves better performance under the template-disjoint setting.

These findings show that no single generation paradigm is universally optimal. Retrieval-based generation is advantageous when reliable and sufficiently similar precedents are available, while transformer-based generation may be advantageous when the required formulations are not directly represented in the retrieval corpus. The evaluation protocol itself therefore plays an important role in determining the apparent effectiveness of each approach.

Beyond the quantitative comparison, this study shows that the effectiveness of a generation approach depends on the degree of textual repetition within the target corpus. Retrieval-based generation is particularly suitable when sufficiently similar and validated precedents are available, whereas transformer-based models provide greater flexibility and may be advantageous when exact legal formulations are not represented in the retrieval corpus. Based on these findings, we propose a conceptual hybrid generation strategy, intended as a practical guideline that combines retrieval-based generation, neural text generation, and human validation to support future e-Justice applications; this strategy has not yet been implemented or empirically evaluated as an integrated system.

Future work will investigate retrieval-confidence estimation, automatic selection between retrieval-based and neural generation, and human-in-the-loop validation. Recent developments in legal retrieval-augmented generation [10] and Arabic large language models such as AceGPT [11] also provide promising directions for extending the proposed strategy. The proposed strategy will additionally be evaluated on other categories of Moroccan legal documents to assess its generalizability.

VI. DECLARATION ON GENERATIVE AI

Generative AI tools were used for language refinement and editorial assistance during manuscript preparation. The authors critically reviewed and validated the resulting text and remain fully responsible for the scientific content, methodology, results, and conclusions of this work.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA), and the CNRST of Morocco, grant number Alkhawarizmi/2020/25.

REFERENCES

- [1] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, and M. J. Bommarito, "Natural Language Processing in the Legal Domain," arXiv preprint arXiv:2302.12039, 2023.
- [2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," in Findings of EMNLP, Online, Nov. 2020, pp. 2898–2904, doi: 10.18653/v1/2020.findings-emnlp.261.
- [3] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, "How Does NLP Benefit Legal System," in Proc. 58th Annu. Meeting ACL, Online, Jul. 2020, pp. 5218–5230, doi: 10.18653/v1/2020.acl-main.466.
- [4] D. Feijo and V. P. Moreira, "Automatic Contract Generation with GPT-2," in Proc. Brazilian Conf. Intell. Syst. (BRACIS), Nov. 2021, pp. 532–546, doi: 10.1007/978-3-030-91699-2_36.
- [5] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in Proc. 4th Workshop Open-Source Arabic Corpora (OSACT4), Marseille, France, May 2020, pp. 9–15.
- [6] I. Kamal and M. Abdul-Mageed, "AraBART: A Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization," in Proc. 6th Arabic Natural Language Processing Workshop, Punta Cana, Nov. 2021, pp. 231–236.
- [7] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-Text Transformers for Arabic Language Generation," in Proc. 60th Annu. Meeting ACL, Dublin, Ireland, May 2022, pp. 628–647, doi: 10.18653/v1/2022.acl-long.47.
- [8] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.
- [9] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.
- [10] N. Pipitone and G. Houir Alami, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," arXiv preprint arXiv:2408.10343, 2024.
- [11] H. Huang et al., "AceGPT, Localizing Large Language Models in Arabic," in Proc. 2024 Conf. North Am. Chapter ACL (NAACL-HLT), Mexico City, Mexico, Jun. 2024, pp. 8139–8163, doi: 10.18653/v1/2024.naacl-long.450.
- [12] O. Sulea, M. Zampieri, M. Vela, and J. van Genabith, "Exploring the Use of Text Classification in the Legal Domain," in Proc. Workshop Automated Semantic Analysis Information Law (ASAIL), London, UK, Jun. 2017, pp. 1–9.
- [13] M. Medvedeva, M. Vols, and M. Wieling, "Using Machine Learning to Predict Decisions of the European Court of Human Rights," Artif. Intell. Law, vol. 28, no. 2, pp. 237–266, Jun. 2020, doi: 10.1007/s10506-019-09255-y.
- [14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in Proc. 40th Annu. Meeting ACL, Philadelphia, PA, USA, Jul. 2002, pp. 311–318, doi: 10.3115/1073083.1073135.
- [15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Proc. Workshop Text Summarization Branches Out, Barcelona, Spain, Jul. 2004, pp. 74–81.
- [16] C. Xiao et al., "CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction," arXiv:1807.02478 [cs.AI], Jul. 2018.
- [17] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proc. 58th Annu. Meeting ACL, Online, Jul. 2020, pp. 7871–7880, doi: 10.18653/v1/2020.acl-main.703.
- [18] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020.
- [19] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in Proc. 2021 Conf. North Am. Chapter ACL (NAACL-HLT), Online, Jun. 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.
- [20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [21] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.
- [22] M. Abdul-Mageed et al., "ARBERT and MARBERT: Deep Bidirectional Transformers for Arabic," in Proc. 59th Annu. Meeting ACL, Online, Aug. 2021, pp. 7088–7105, doi: 10.18653/v1/2021.acl-long.551.
- [23] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, Online, Dec. 2020, pp. 1877–1901.
- [24] K. D. Ashley, Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge, UK: Cambridge Univ. Press, 2017.
- [25] N. Habash, Introduction to Arabic Natural Language Processing. San Rafael, CA, USA: Morgan & Claypool Publishers, 2010, doi: 10.2200/S00277ED1V01Y201008HLT010.
- [26] S. Yassine, O. Ibrhich, M. Essouja, and M. Esghir, "Automated Classification of Moroccan Procedural Court Orders Using CAMEL-BERT," International Journal of Advanced Computer Science and Applications, vol. 17, no. 7, 2026, doi: 10.14569/IJACSA.2026.0170719.