Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Synthetic Legal Data Generation for Saudi Family Law Courts Using Multi-Role Simulation

Zarah Shibli

zshibli0002@stu.kau.edu.sa

King Abdulaziz University

Areej Alhothali

King Abdulaziz University

Lulwah Alharigy

King Abdulaziz University

Article

Keywords:

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9887183/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Synthetic Legal Data Generation for Saudi Family Law Courts Using Multi-Role Simulation

Zarah Shibli[1,*], **Areej Alhothali**[1], **and Lulwah Al-Harigy**[1]

[1]King Abdulaziz University, Department of Computer Science, Jeddah, Kingdom of Saudi Arabia
[*]zshibli0002@stu.kau.edu.sa

ABSTRACT

The growth of Legal Natural Language Processing is restricted by a critical lack of structured, high-quality corpora, particularly in specialized jurisdictions such as Saudi Family Law. This limitation is further compounded by strict data privacy and the sensitive nature of family cases, which limit access to real judicial records. To address these limitations, this paper presents a synthetic generation framework using Large Language Models (LLMs) and a multi-role simulation prompting strategy. Coordinated multi-turn interactions between role-specific were conducted to synthesize a dataset of 200 comprehensive litigation records focused on proof-of-divorce cases. These records capture the full judicial process—including claims, answers, pleadings, judicial reasoning, and final judgment—strictly in accordance with Saudi law while reflecting diverse factual scenarios. The dataset was evaluated through independent assessments by legal experts, achieving an exceptionally high inter-rater adjacent agreement of 94.6%. While the initial evaluation identified minor deviations in formal legal phrasing and conciseness, the integrated Human-in-the-Loop (HITL) validation successfully resolved these limitations, proving that multi-role simulation with HITL is a highly effective methodology for generating specialized standard legal datasets.

Introduction

Judicial proceedings are distinguished by complex procedural sequencing, adversarial exchanges, and multi-phase decision-making stages that require coordinated participation from judges, litigants, legal representatives, and court clerks. Effective frameworks are therefore essential to ensure procedural integrity, evidentiary compliance, and coherent adjudication across all phases of litigation. However, the inherent complexity of these interactions often leads to court congestion—a pervasive concern for judicial systems across comparative jurisdictions—as increasing caseloads continuously pressure adjudicative capacity and affect the timeliness of dispute resolution. In the Kingdom of Saudi Arabia, this issue is being handled as part of a larger structural vision. Rather than addressing congestion strictly as a quantitative capacity issue, the Kingdom has prioritized a qualitative transformation through legal codification.

A cornerstone of this transformation is the 2022 publication of the Family Law[1] (Personal Status Law) and the Law of Evidence[2]. By codifying substantive family law norms and evidentiary standards previously implemented through essentially uncodified Sharia-based judicial reasoning, the Saudi legal system has moved toward a clearer, more structured statutory framework. This codification improves legal certainty, promotes doctrinal consistency, increases the predictability of judgments, and enhances procedural efficiency.

Recent advancements in Artificial Intelligence (AI), particularly the emergence of Large Language Models (LLMs), have demonstrated profound potential for integration into the legal domain. LLMs are increasingly utilized across several legal tasks, ranging from the automated drafting of contracts and legal memos to intelligent case retrieval and the summarization of court transcripts. These capabilities extend beyond mere administrative efficiency to include complex semantic reasoning, enabling systems to interpret statutory texts, construct logical arguments, and predict judicial outcomes, which is known as Legal Judgment Prediction (LJP).

Despite these advancements, the development of LJP systems is often hindered by the scarcity of comprehensive case records. To address this data gap, the task of "Court Simulation" was proposed. In this context, court simulation serves as a generative mechanism where an LLMs is prompted to adopt multi-role-specific personas sequentially, creating the necessary full-case context—including the plaintiff's claim, the defendant's response, and the judge's legal reasoning and judgment. Furthermore, unlike "black-box" prediction models that output only a judgment, this stepwise generation enforces a Chain of Thought (CoT) process. This explicit reasoning aligns with statutory requirements in the Saudi legal system, where judges must

[1]Bureau of Experts of Council of Ministers, "Family Law", https://laws.boe.gov.sa/boelaws/laws/lawdetails/4d72d829-947b-45d5-b9b5-ae5800d6bac2/1 (Accessed: 2026-05-05).

[2]Bureau of Experts of Council of Ministers, "Law of Evidence", https://laws.boe.gov.sa/BoeLaws/Laws/LawDetails/2716057c-c097-4bad-8e1e-ae1400c678d5/1 (Accessed: 2026-05-05).

provide a rationale or causation for each decision. By automating this mandatory reasoning phase, Court Simulation transforms LJP from a simple predictive model into a robust judicial support tool, enhancing transparency, interpretability, and the practical drafting of judicial decisions.

Under the Family Law, specifically Article 77, divorce is the dissolution of marriage contingent upon the husband's will. While the husband is legally obligated to formally document the divorce, practical realities often lead to delays or intentional non-compliance. Such reluctance may arise from concerns over potential legal claims, including the wife's entitlement to deferred dowry or financial maintenance, or from deliberate attempts to exert control by withholding official documentation. In some cases, this behavior functions as a form of coercion, effectively preventing the wife from exercising her legal rights, including claims related to custody and financial support.

As a result, the wife may be placed in a state of legal uncertainty, unable to assert her rights without judicial intervention. To resolve this, she must initiate a lawsuit to establish the occurrence of divorce "Establish Divorce" (Ithbat Talaq) under Article 90 of the same law. This process often leads to a complex evidentiary dispute, requiring the court to assess conflicting claims and reconstruct the factual and legal circumstances surrounding the alleged divorce. Consequently, such cases demand rigorous legal analysis to ensure procedural fairness and the protection of the rights of both parties.

This research focuses specifically on this category—claims to establish divorce—as the primary case study. This domain was selected for two reasons. First, this category alone encompasses a wide range of legally significant scenarios and constitutes a substantial, procedurally complex category in which the interdependence of the husband's perspective and the wife's claim creates a clear structure suitable for simulation. Second, other categories, such as Khula (divorce at the wife's instance) or child custody disputes, were excluded. These cases often required external expert intervention, such as financial experts to assess compensation in dowry return cases in khula or medical or psychological professionals to evaluate harm or the best interests of the child in annulment and custody-related claims, which may introduce interdisciplinary complexities beyond the scope of this legal-focused study.

Despite AI's potential, there are two significant gaps in the area's study. First, Data Scarcity and Privacy: The Saudi Ministry of Justice enforces strict confidentiality regarding family cases, limiting access to real-world data and posing a significant challenge for model training. Second, Linguistic and Legal Specificity: Most LLMs are trained on high-resource languages (e.g., English, Chinese) or non-Arabic low-resource languages (e.g., Portuguese, Indonesian) and lack adequate exposure to Arabic legal texts, particularly the intersection of Sharia principles and Saudi statutory law, often resulting in inaccurate or hallucinated legal interpretations.

To address these gaps, this paper presents a framework for court simulation in Saudi Family Law. The primary contributions of this work are:

1. **A Specialized Synthetic Dataset:** To overcome privacy restrictions, a dataset comprising 200 *Ithbat Talaq* (establishing divorce) cases is introduced. These cases were generated via stratified sampling from over 70,000 combinatorial scenarios, ensuring a balanced and privacy-compliant representation of legal variables.

2. **Court Simulation Framework:** The study utilizes *Gemini 3.1 Pro* in a zero-shot prompt engineering configuration with law files. By injecting the full text of the Family Law and relevant articles from the Law of Evidence and the Civil Procedure Law, the framework simulates a multi-role workflow while strictly adhering to the law.

3. **Expert-Validated Evaluation:** The simulated data was evaluated by human legal experts against 8 distinct criteria—including Legal Reasoning, Pronouncement of the Judgment, Legal Language, and Judicial Procedures—demonstrating the system's efficacy in court simulation in adherence with Saudi law.

Literature Review

The existing research on LLMs has expanded across multiple application areas, particularly those requiring complex reasoning and structured decision-making. Within this context, the relevant literature can be organized into three strands: (i) LLM-based simulation across various domains, which demonstrates the effectiveness of LLMs in modeling complex system behavior, (ii) LLM-based court simulation in the legal context, which proposes to support judicial analysis and case simulation, and (iii) Arabic and Saudi legal judgment prediction.

LLM-Based Simulation Across Several Domains

The integration of LLMs into simulation frameworks has yielded promising results for modeling context-dependent, iterative interactions across multiple domains, including healthcare, education, and the social sciences. In recommendation systems, RecoWorld[1] develops a dual-agent setting in which user and recommender agents interact in iterative feedback loops, allowing for the rigorous evaluation of agentic recommendation strategies. To enhance patient safety and diagnostic precision, frameworks such as MATRIX[2] and MedAgentSim[3] model doctor-patient interactions. These applications indicate how agent-based methods

enhance the reliability of clinical communication. Educational simulations such as FACET[4] leverage teacher–learner agent collaboration to design personalized instructional content that reflects both cognitive needs and motivational profiles. Across these domains, LLM frameworks enable simulation of sophisticated reasoning patterns and interaction dynamics.

Court Simulation with LLM-Based

Courtroom simulations utilizing LLM-based systems indicate the potential to replicate both procedural and cognitive aspects of judicial decision-making. In adversarial frameworks such as AgentCourt[5], AdvEvol is an adversarial evolutionary mechanism that enables lawyer and judge agents to simulate over one thousand civil cases, demonstrating improvements in agents' legal reasoning and argumentation skills. Role-based simulations, exemplified by SimCourt[6], emulate Chinese trial procedures by multi-agent interactions. The system operates across five core judicial stages (Trial Preparation, Trial Investigation, Evidence Presentation, Trial Debate, and Final Statement) and incorporates five courtroom roles (Judge, Prosecutor, Attorney, Defendant, and Stenographer). By integrating agent planning and memory mechanisms to producing predictive judgments. AgentsBench[7] employs a deliberative multi-agent framework in which multiple LLM-judge agents engage in debate and consensus formation, achieving greater fairness and consistency than single-agent prediction models. Overall, these studies indicate that LLM-driven courtroom simulations can emulate the procedural rigor, reasoning depth, and ethical deliberation characteristic of human judicial processes.

Arabic and Saudi Legal Judgment Prediction

Within the Arabic legal domain, current LLM-based efforts in judgment prediction focus on advancing decision accuracy and enhancing the interpretability of model-generated outcomes. This is due to the Arabic language's unique linguistic characteristics and the complexities of its legal systems, which present substantial challenges for model development. Existing work has primarily focused on Saudi commercial courts, with data extracted from the Ministry of Justice's open judicial database. Kmainasi et al.[8] show that LLaMA-based models—tested under zero-shot, one-shot, and LoRA fine-tuning regimes—can capture core decision patterns with notable robustness. Almalki et al.[9] propose a binary judgment prediction model for Saudi commercial cases based on AraBERT, employing sentence-level classification with aggregation across Facts, Reasoning, and Decision sections to align predictions with judicial document structure. The study by Abu Shairah et al.[10] introduces ALARB, an extensive benchmark designed to assess Arabic legal reasoning within the context of Saudi commercial court judgments. The dataset contains structured legal cases, each annotated with four key parts: facts, reasoning, verdict, and cited legal articles. The authors used LLMs for named entity recognition (NER) to perform text cleaning, article extraction, and restructuring, and to link each case to the relevant statutes, drawn from eight major Saudi laws, such as the Sharia Procedure Law and the Company Law. To assess reasoning capacity, they define a suite of tasks that include verdict prediction, argument completion, and article identification. Similarly, Ammar et al.[11] compared GPT-3.5-turbo, JAIS-13b-chat, and LLaMA-7b on a dataset of Saudi commercial court cases, across zero-shot, one-shot, and customized fine-tuning, using efficient approaches such as LoRA.

However, to the best of our knowledge, no prior research has focused on simulating court proceedings in family law, particularly in cases concerning the establishment of divorce.

Methodology

This study utilizes a simulation-based approach to represent the proceedings of Saudi courts from a multi-role perspective. The objective is to present courtroom interactions and reflect the procedural structure, linguistic forms, and reasoning documented in real judicial records, ensuring that relevant legal articles are correctly identified and applied to reach a judgment. This research focuses on establishing divorce as its primary case study.

The constructed simulation represents a full judicial session, including the roles of the judge, the plaintiff (wife), the defendant (husband), witnesses where applicable, and the court clerk. The generation of claims and the formulation of case scenarios are achieved through a language model that operates using legally structured prompts and role-specific contextual instructions derived from established Saudi procedural norms. This includes the sequencing of litigation, the preparation of legal documents, the drafting of pleadings, the presentation and assessment of evidence, and the scope of judicial inquiry.

Divorce proceedings can often be complex, involving various legal, emotional, and procedural elements. In jurisdictions like Saudi Arabia, understanding the complexity of family law is crucial for navigating these challenges effectively. This simulation framework aims to replicate and analyze divorce scenarios within the context of Saudi Family Law, allowing for a deeper understanding of the judicial decision-making process.

The proposed methodology is divided into three main phases: Scenario Dataset Generation, Court Session Simulation Prompt Template, and Expert Review and Legal Validation. Figure 1 presents an overview of the proposed methodology.

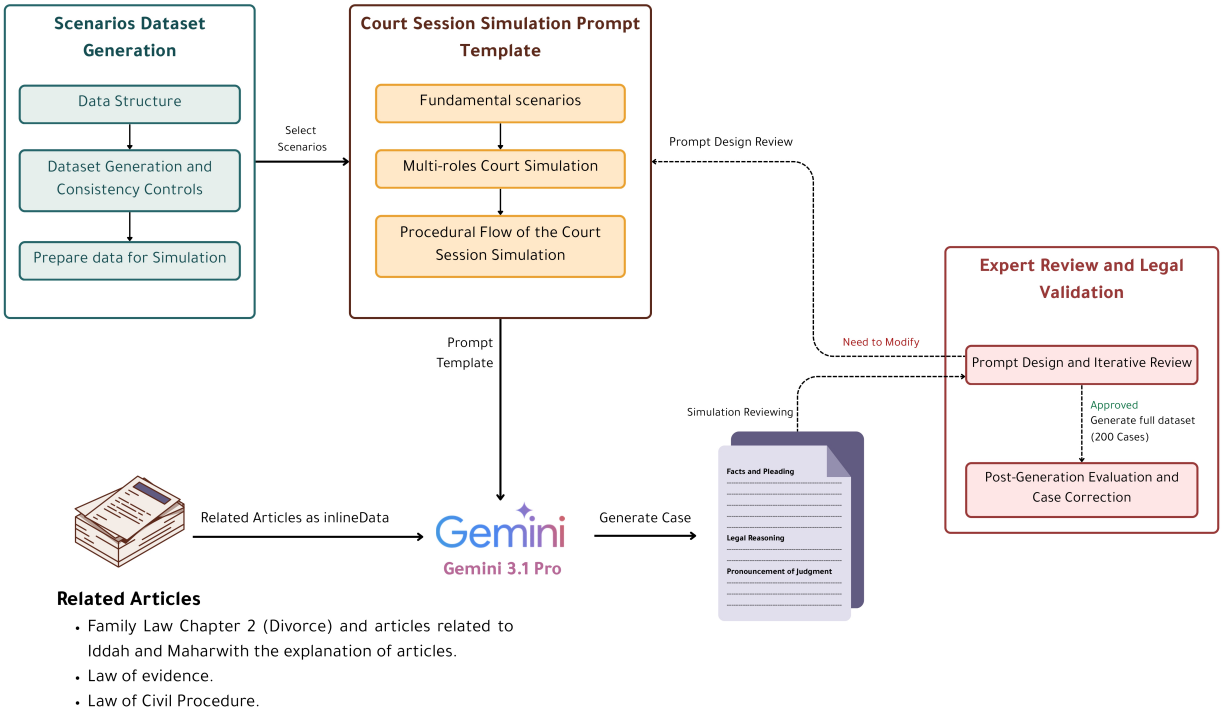


Figure 1. Overview of the proposed methodology

1. Scenarios Dataset Generation

The simulation framework is based on a dataset of modeled claims that establish divorce scenarios, providing the factual basis for virtual court proceedings. The dataset is designed to align with Chapter II (Divorce) of the Saudi Family Law, and each record includes the substantive and procedural factors necessary to reproduce judicial reasoning in cases claiming to establish divorce. The following sections detail the dataset's structure, generation methodology, and validation rules that ensure consistency and legal compliance for each scenario generated in the simulations.

Divorce-Related Terminology and Conceptual Definitions

This study employs a set of domain-specific legal terms derived from Saudi family law. Because several of these concepts carry specific meanings that differ from their ordinary linguistic usage, precise definitions are necessary to ensure methodological consistency. The following terminology provides the conceptual foundation upon which the dataset construction and court simulation framework are based.

- **Explicit Divorce:** A clear expression that can only be understood as divorce, typically using the word "divorce" or its direct derivatives.
- **Implicit Divorce:** An expression that may indicate divorce or another meaning according to custom, such as "Go to your family" or "You are not my wife." Divorce takes effect only if the husband intends these words to divorce.
- **Conditional Divorce (Suspended):** is divorce that is conditional on a specific action occurring or not occurring, such as "You are divorced if you leave the house." The divorce becomes effective when the condition is met, unless the husband is only aiming to encourage, discourage, prove, or disprove such an act, and not to affect the divorce.
- **Condition Satisfied:** a variable that reflects the respective perspectives of the spouses in cases of conditional divorce. It denotes that the legal effect of divorce is contingent upon the realization of the specified condition articulated by the husband. Consequently, the occurrence of divorce depends not only on the formulation of the conditional statement, but on the realization of the stipulated event and the acknowledgment of its occurrence by the parties. For instance, in the declaration "If you leave the house, you are divorced", the divorce becomes legally effective only if the wife leaves the house and the occurrence of that condition is established and recognized.

- **Valid Seclusion** (*Khalwa*): shall mean a married couple privately meeting at a place where no discerning person is present.

Scenarios Dataset Structure and Identification of Legal Variables

In order to support procedurally accurate, realistic, and consistent simulations, each record is uniquely identified and organized into two sets of attributes, representing the perspectives of the two litigants. The variables associated with each record are classified by perspective, with Table 1 detailing the wife's perspective and Table 2 detailing the husband's perspective.

Variable	Arabic Label	Description
Divorce form	صيغة الطلاق (صريح / كناية / معلق)	Linguistic form of the divorce as claimed by the wife (explicit, implicit, or conditional).
Intercourse State	حالة الدخول (لم يدخل ولم يخلو بها / تم الدخول والخلوة)	The divorce occurred prior to consummation or khalwa (consummation and khalwa occurred, or no consummation or khalwa occurred).
Divorce Phrasing	لفظ الطلاق	The specific phrasing was reportedly used by the husband, according to the wife's statement.
Divorce Number	رقم الطلقة (الأولى /الثانية /الثالثة)	The divorce number as claimed by the wife in the lawsuit (first, second, third).
Divorce Delivery Method	طريقة إيقاع الطلاق (نطق / كتابة / إشارة)	The method used to inform the wife of the divorce (verbal, written, or discernible sign).
Condition Satisfied	حالة تحقق الشرط (تحقق الشرط / لا يوجد شرط)	Since the wife is the one who brought the lawsuit, the wife asserts that the specified condition has occurred in conditional divorce scenarios and that no conditional element exists in explicit or implicit divorce.
Wife State	حالة الزوجة (طهر وطئ فيه / حيض / نفاس / اليأس من المحيض / حمل / طهر لم يطأ فيه)	The wife's physiological state at the time of divorce (period of purity during which intercourse occurred, menstruation, postpartum period, pregnancy, menopause, or purity without intercourse).
Evidence	البينة	The availability of evidentiary material substantiating the allegation, such as documents, audio messages, or testimonial witnesses, including a description if present.

Table 1. Divorce Case from the Wife's Perspective

Dataset Generation and Consistency Controls

The scenarios dataset was generated by enumerating all possible combinations of the identified variables. Once the combinations were established, a set of validation rules derived from statutory provisions and jurisprudential principles was applied to eliminate scenarios that were legally invalid or internally inconsistent. These rules ensured that all generated scenario complied with legal requirements and constraints. The following sections describe the rules applied to filter the combinations dataset:

1. **Wife-Related Rules**

   The wife's variables reflect her perspective on the divorce case, particularly regarding any circumstances that might lead to or affect the legal effect of the divorce, and the evidence presented to support her claim. To ensure the scenarios accurately reflect the wife's role and legal implications, four primary control rules were implemented:

   (a) **Suspended Divorce**

   If the divorce is suspended, it indicates that the divorce is conditional and will only take legal effect upon satisfaction of the specified condition. The type of the condition must be explicitly documented, including categories such as

Variable	Arabic Label	Description
Husband Position	موقف الزوج (إقرار / إنكار التلفظ بالطلاق إقرار مع اختلاف في التفاصيل)	The husband's stance regarding the divorce (Admission, Denying the pronouncement of divorce, Admission with divergence).
Husband State	حالة الزوج (عاقل / زال عقله اختيارًا غير مختار /غضب شديد)	The husband's mental or volitional condition (Sound mind, Mentally unsound, Acting under compulsion, Extreme anger impairing control).
Husband Intention	نية الزوج (يقصد الطلاق / لا يقصد الطلاق	The husband's intent behind the phrase (intended to influence the divorce, without intending an actual divorce).
Condition Satisfied	موقف الزوج من تحقق الشرط (تحقق الشرط لم يتحقق الشرط لا يوجد شرط)	Husband's position regarding the occurrence of the condition in conditional divorce (acknowledged, denied, or no condition).
Type of Conditional Divorce	نوع التعليق في الطلاق المعلق (التصديق / التكذيب / المنع / الحث / يمين الحرام	The classification of the conditional clause (Confirmation / Denial / Prohibition / Encouragement / Oath-based prohibition [yameen al-haram]).
Knowledge of Wife's State	علم الزوج بحال الزوجة (يعلم / لا يعلم)	The husband's awareness of the wife's physiological state at the relevant time (Aware / Unaware).

Table 2. Divorce Case from the Husband's Perspective

Confirmation, Denial, Prohibition, Encouragement, or Oath-based Prohibition (Yameen al-Haram). In instances where the divorce is not suspended, all relevant fields are recorded as "No condition", indicating that the divorce is unconditional. Any scenario that does not adhere to these specifications is rejected.

(b) **Evidence Validation**

The evidence type must match the wife's declaration: "None" if no evidence, otherwise it must not be "none." Descriptions must accurately reflect the evidence; for instance, witness testimony requires confirmation of at least one witness.

(c) **Evidence Constraints by Delivery Method**

The type of evidence admissible depends on the method used to inform the wife of the divorce:

- A verbal declaration cannot be supported by documentary evidence.
- A written declaration cannot be supported by witness testimony or audio/video recordings.
- A discernible sign cannot be corroborated by documentary evidence or audio recordings.

(d) **Physiological State and Intercourse Status**

The wife's physiological state constrains the intercourse state:

- The scenario is rejected if the wife is pregnant, in a state of purity during which marital intercourse has occurred, or postpartum and marital intercourse has not occurred.
- The scenario is rejected if the wife is menopausal or in a state of purity during which no marital intercourse has occurred, and marital intercourse has occurred.

2. **Husband-Related Rules**

Husband-related variables capture the husband's perspective, including his intention, acknowledgment, and mental or volitional state. These variables ensure that the scenarios accurately represent the husband's role. To prevent logical contradictions between these variables, additional consistency checks were applied:

(a) **Husband Position and Intention**

The husband cannot deny the pronouncement of divorce while claiming that his statements were intended to effect an actual divorce, nor can he admit the occurrence of divorce while asserting an absence of intent to effectuate the divorce.

(b) **Mental and Volitional State**

The husband's mental or volitional state affects admissibility; if he was acting involuntarily or was mentally impaired, he cannot be recorded as acknowledging the divorce or possessing the intent to effect it.

The resulting dataset from this stage contained over 70,000 distinct claims of divorce establishment scenarios. This exhaustive systematic generation process ensures comprehensive coverage of the legally possible scenario space. From this collection, a subset of 200 scenarios was selected to form the final dataset. A stratified sampling strategy was employed to ensure balanced representation across different forms of divorce. Specifically, the selection was designed to equally cover explicit, implicit, and conditional divorce cases, resulting in 66 explicit, 66 implicit, and 68 conditional cases. This balanced distribution ensures that each legal category is sufficiently represented, enabling robust evaluation across the full spectrum of divorce establishment scenarios. This sample size was determined based on practical constraints associated with expert refinements, including time and cost considerations, as well as the systematic generation process was designed to exhaustively capture the combinations of legally relevant factors affecting the occurrence and validity of divorce, the selected 200 cases already encompass the full spectrum of material legal situations. Beyond its role in the present study, this resulting subset constitutes a structured, expert-validated corpus that can serve as a standardized benchmark for downstream tasks, including legal judgment prediction (LJP) and other complex legal reasoning applications.

Prepare data for Simulation

For each simulation, a single case scenario is selected from the comprehensive dataset records, each representing a legally coherent dual-perspective account of the divorce dispute. The selected record is structured as text, preserving all critical variables, and is used as the factual input context for the LLM to generate the complete procedural narrative across all stages of the court session. An illustrative template of the record is shown in Table 3.

2. Court Session Simulation Prompt Template

In order to replicate court proceedings in Saudi family court, particularly in cases involving claims to establish the occurrence of divorce, this study utilized a multi-role agent simulation framework. The approach combines sequential procedural, legal-context conditioning, and structured role-based modeling to guarantee that simulated proceedings conform to Saudi family law, evidentiary regulations, and courtroom procedures.

Fundamental scenarios

Each case simulation is based on a selected factual scenario from the dataset that represents the divorce event, as shown in Table 3. The scenario serves as the foundation for the simulation, guiding the model in generating the supplementary components required to construct a complete case file—such as the marriage contract, child-related information, relevant dates, contextual factors surrounding the divorce, witness testimony, evidentiary materials, and the husband's legal position—while allowing controlled factual expansion during the interaction, reflecting how additional details are elicited through judicial questioning and party responses in real court proceedings.

Multi-roles Court Simulation

The simulation employs role-specific prompts corresponding to judicial actors typically involved in Saudi family court proceedings, such as the judge, clerk, defendant, and plaintiff. Each role defines clear functional boundaries, specifying the responsibilities and permissible actions of the role with respect to the shared factual scenario. Each of the four roles in the simulation represents a real procedural actor in Saudi family courts:

A. **Judge**

The judge executes judicial behavior consistent with procedural and substantive responsibilities as stipulated in Family Law Chapter 2 (Divorce) and articles related to Iddah and Mahar with the explanation of Article[3], the Law of Evidence, and the Civil Procedure Law[4] (selected the articles that affected the divorce case only, Articles from 41 to 74). The primary responsibilities include:

[3]Saudi Ministry of Justice, "Explanation Of Family Law", legal portal, https://laws-gateway.moj.gov.sa/apis/legislations/v1/ExportDocument/pdf?Serial=4DC3ARZ00YqnoqpraZMxCQ (Accessed: 2026-05-05).

[4]Bureau of Experts of Council of Ministers, "Civil Procedure Law", https://laws.boe.gov.sa/boelaws/laws/lawdetails/f0eaae46-9f84-40ee-815e-a9a700f268b3/1 (Accessed: 2026-05-05).

Arabic	English
=== منظور الزوجة === حالة الدخول والخلوة: لم يدخل ولم يخلو بها لفظ الطلاق: أنتِ طالق رقم الطلقة: ٣ حالة الزوجة وقت الواقعة: حيض وصف الدليل: تسجيل مكالمة هاتفية بين الزوجين تثبت الطلاق === منظور الزوج === موقف الزوج من الطلاق: إقرار حالة الزوج وقت الواقعة: عاقل ومختار أدركت معنى كل كلمة قلتها أثناء الطلاق النية وتحقق الشرط: يقصد الطلاق علم الزوج بحالة الزوجة: لا يعلم	=== Wife's Perspective === Consummation and Seclusion: He did not consummate the marriage nor was he alone with her (no consummation or khalwa occurred) Pronouncement of Divorce: "You are divorced". Number of Divorces: 3 Wife's State at the Time of the Divorce: Menstruating Description of Evidence: Recording of a phone call between the spouses proving the divorce. === Husband's Perspective === Husband's Stance on the Divorce: Acknowledgment Husband's State at the Time of the Incident: Sane and acting of his own free will – I understood the meaning of every word I said during the divorce Intention and Satisfaction of Condition: Intended divorce Husband's Knowledge of Wife's State: Unaware
=== منظور الزوجة === حالة الدخول والخلوة: لم يدخل ولم يخلو بها لفظ الطلاق: اذهبي إلى بيت أهلك رقم الطلقة: ١ حالة الزوجة وقت الواقعة: طهر لم يُطأ فيه وصف الدليل: شهادة شاهد بسماع لفظ الطلاق === منظور الزوج === موقف الزوج من الطلاق: إنكار التلفظ بالطلاق	=== Wife's Perspective === Consummation and Seclusion: He did not consummate the marriage nor was he alone with her Pronouncement of Divorce: "Go to your family's house" Number of Divorce: 1 Wife's State at the Time of the Incident: A period of purity during which no intercourse had taken place Description of Evidence: Testimony of a witness who heard the pronouncement of divorce === Husband's Perspective === Husband's Stance on the Divorce: Denies the pronouncement of divorce

Table 3. Samples from the dataset

- Initiating court sessions and confirming the presence of the litigants.
- Managing the presentation and exchange of claims and defenses.
- Conducting the assessment and approval of evidence.
- Determining the need for additional proof, clarification, or witnesses.
- Limiting inquiries to issues relevant to the divorce's validity.
- Providing a reasoned legal basis for the final judgment.

To ensure the LLM executes its adjudicative duties with logical rigor and strict fidelity to statutory guidelines, the prompt architecture incorporates a Sequential Implicit Thinking Framework. Before generating visible outputs, the judge persona is instructed to process the case internally through the following eight-step cognitive sequence:

1. **Comprehensive Review:** Examining all statements provided by the parties alongside the submitted evidence.
2. **Core Element Identification:** Isolating the fundamental elements of the dispute, including the exact phrasing of the divorce, the wife's physical state (e.g., menstruation), the fulfillment of statutory conditions, consummation/privacy (Khalwah), and available documents or witnesses.
3. **Sufficiency Assessment:** Evaluating whether the available information is sufficient to issue a binding judgment or if it necessitates further clarification.

4. **Directed Inquiry:** Requesting necessary clarifications by directing specific questions, demanding further evidence, or summoning witnesses to resolve ambiguous or disputed points.
5. **Dialogue Management:** Organizing the exchange of responses to ensure each party has the opportunity to address legally material claims while actively preventing immaterial or repetitive dialogue.
6. **Evidentiary Analysis:** Weighing the strength of evidence and witnesses against the case facts.
7. **Legal Balancing & Weighing (Tarjeeh):** Comparing all facts and evidence, giving due consideration to the husband's intent and the wife's state, to definitively determine the statutory validity and classification of the divorce.
8. **Final Adjudication:** If the facts and evidence are conclusive, closing the pleadings and issuing the final judgment. If incomplete, mandating further evidentiary requests prior to adjudication.

B. **Clerk**

The clerk records the proceedings in a formal case record that reflects Saudi court-documentation conventions without analysis or commentary. The documentation comprises:

- Recording session openings and litigant identification.
- Sequentially documenting statements, questions, objections, and testimonies.
- Recording adjournments and scheduling with justification.
- Formalizing the session's closure.

C. **Plaintiff (Wife)**

The plaintiff's responsibility includes:

- Presenting a claim for establishing the divorce, detailing the marital and children's circumstances leading to the divorce, and the legal request for proof of divorce.
- Supplying supporting documentation and requesting hearing witnesses or additional evidence when required.
- Responding to judicial inquiries based on factual and case-specific information.
- Addressing the defendant's statements and counter-arguments.

D. **Defendant (Husband)**

The defendant is responsible for responding to the plaintiff's claim:

- Specifying points of agreement and disagreement with the plaintiff's claim.
- Providing defense against the plaintiff's claim.

Procedural Flow of the Court Session Simulation

To clarify how the simulated proceedings reflect the typical workflow of Saudi family courts, the process progresses through a series of structured phases, culminating in a complete and coherent case output, as illustrated in Figure 2. The simulation advances through the following stages:

1. The session begins.
2. Presentation of the Plaintiff's claim.
3. Presentation of the Defendant's answer.
4. Iterative exchanges of pleadings, including judicial questioning, responses, and examination of witnesses.
5. Conclusion of the pleading phase once all factual and evidentiary elements have been addressed.
6. Presentation of the judge's legal reasons that lead to the final judicial determination.

7. Rendering of the final judicial decision.

The simulation results in a single, structured output that mirrors an authentic judicial case file and includes:

- A detailed transcript of the session recorded by the clerk, representing all facts.
- The reasons form the basis of the judicial decision.
- The formal pronouncement of the judgment.

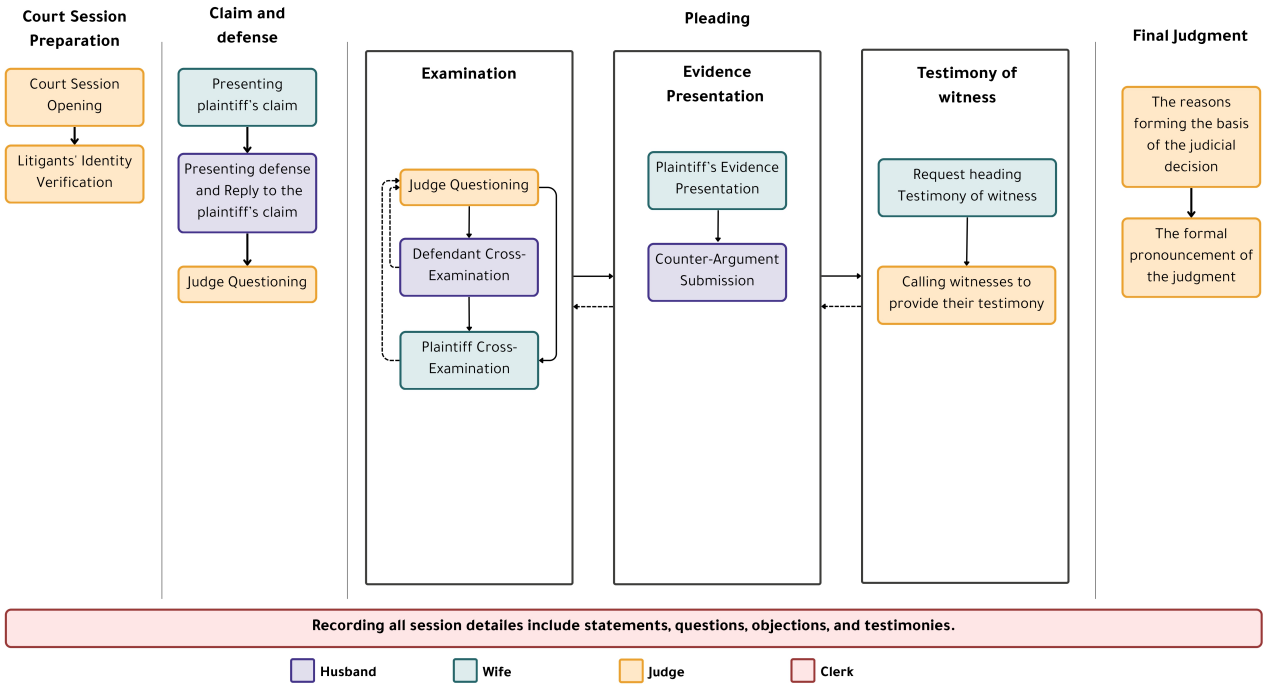


Figure 2. Schematic representation of the procedural flow in the simulated court session.

3. Expert Review and Legal Validation

To ensure the reliability and legal soundness of the simulated proceedings, the framework incorporates a Human-in-the-Loop (HITL) review process conducted by domain specialists.

Expert Selection

The evaluation panel comprised six legal experts specializing in Saudi Family Law (Personal Status Law), evidentiary requirements in divorce proceedings, and the Saudi Law of Evidence. This panel was deliberately convened to ensure comprehensive coverage of judicial practices pertaining to family disputes, with a specific emphasis on divorce litigation, evidentiary assessment, and procedural validation within the Saudi Personal Status Courts.

The evaluators' expertise encompasses navigating procedural frameworks, drafting statements of claim and defense memorandums, and assessing the admissibility and probative value of evidence. Furthermore, their professional backgrounds include specialized training and engagement with the Law of Evidence at the Saudi Judicial Training Center, ensuring a rigorous understanding of the statutory standards governing burden of proof, oral testimony, admissions, and documentary evidence. This targeted selection strategy guaranteed that all evaluators possessed direct expertise in Saudi family disputes, thereby bolstering the legal validity, methodological reliability, and jurisdictional appropriateness of the assessment process.

Two-Stage Review and Refinement Process

To integrate human legal reasoning with the LLM generation pipeline, a two-stage expert intervention was employed.

Stage 1: Prompt Design and Iterative Review

During the initial generation phase, a legal expert reviewed intermediate simulation outputs to continuously refine prompt instructions, system role boundaries, and factual grounding mechanisms. This iterative feedback loop corrected systemic generation errors and enhanced the model's adherence to Saudi judicial procedures and Sharia-compliant legal reasoning prior to scaling to full-case generation.

Stage 2: Post-Generation Evaluation and Case Correction

Following the full case generation ($N = 200$), expert feedback was collected using a highly structured, dual-purpose evaluation scheme. The criteria for this evaluation were systematically selected based on preparatory legal consultation with subject matter experts and a comprehensive review of previous studies in judicial NLP.

First, the evaluators utilized a quantitative numerical rating scale from 1 to 5 across eight critical dimensions: factual completeness, legal reasoning, judgment pronouncement, legal language, judicial procedures, clarity and conciseness, relevance, and consistency, as detailed in Table 4.

Second, alongside the numerical grading, experts provided qualitative, targeted feedback. Reviewers identified latent inconsistencies, unsupported assumptions, deviations from the scenario context, and deficiencies in evidentiary articulation under Law of Evidence. Based on these expert annotations, cases were manually revised and corrected as necessary. This crucial HITL refinement ensures that the final dataset outputs are not only linguistically coherent, but strictly legally valid, internally consistent, and suitable as a gold-standard corpus for advanced downstream AI tasks, such as Legal Judgment Prediction (LJP) within the Saudi legal context.

Evaluation Criterion	1	2	3	4	5
C1: Factual	Missing or incorrect	Major gaps	Clear with Minor gaps	Minor non-impactful omission	Complete & accurate
C2: Legal Reasoning	No linkage	Weak	Partial	Minor issues	Clear legal linking to applicable legal articles
C3: Pronouncement of the Judgment	Invalid/missing	Affecting determination of rights	Partly incomplete	Correct ruling with minor issues	Clear, specific, and enforceable ruling setting out legal effects
C4: Legal Language	Non-legal language	Frequent mistakes	Minor mistakes	Proper terminology	Precise
C5: Judicial Procedures	Non-compliant	Procedural defects affecting case integrity	Correct procedures with some missing steps	Correct procedures with non-impactful omissions	Correct and sequential procedures
C6: Clarity & Conciseness	Excessive padding	Disorganized and repetitive	Some ambiguity or repetition without material impact	Clear and slight verbosity	Organized and clear presentation
C7: Relevance	Unrelated to the case subject	Most core elements outside the case subject	Some irrelevant	Minor peripheral elements	Fully relevant to the case subject
C8: Consistency	Fundamental contradiction	Material conflict	Limited conflict	Clear with minor observations	Full coherence among facts, reasoning, and ruling

Table 4. Evaluation Rubric

Experiments

Experimental Setup

To systematically evaluate the efficacy and legal reasoning capabilities of the proposed framework, subsets from the generated dataset were deployed across two primary experimental phases. Specifically, Phase 1 aims to establish a foundational baseline and select the optimal LLMs by testing intrinsic legal knowledge in a zero-shot setting. Subsequently, Phase 2 aims to validate the robustness and accuracy of the fully proposed methodology across 200 diverse scenarios.

- **Phase 1: Pilot Study** ($N = 5$)**:** To establish a baseline and select the optimal foundational model, six LLMs were evaluated: three state-of-the-art multilingual models: Command A[12], GPT-5[13], Gemini 3.1 pro[14], and three Arabic-specific models: Command R7b Arabic[15], Fanar[16], ALLaM-7B-Instruct-preview[17]. A pilot study consisting of five carefully selected cases was conducted to select the optimal foundational model to use it in the proposed court simulation framework. These cases were intentionally chosen to capture diverse important legal configurations, rather than to reflect random sampling. Specifically, these cases cover key differences in: (i) marital state (consummated and, non-consummated marriage), (ii) type and form of divorce, (iii) divorce number, (iv) evidence (presence vs. absence of evidence), and (v) legal complexity arising from different physiological states (e.g., menstruation and postpartum conditions) and alongside variations in the husband's state at the time of pronouncement (e.g., sound mind, anger, mentally unsound). This phase assessed the intrinsic legal capabilities of each model without external knowledge, retrieval mechanisms, fine-tuning, or advanced prompting strategies. Using a standardized zero-shot prompt, the models were evaluated on their ability to accurately extract facts, identify and cite relevant legal articles, apply legal reasoning, pronounce judicial decisions, and generate formal legal language. To ensure a fair comparison, all experimental conditions were controlled across models. Specifically, the same five court cases, evaluation criteria, prompting template, and generation settings were used for all models. The temperature and sampling parameters were fixed at $T = 0$ and $top_p = 0.2$, respectively, to measure deterministic legal knowledge and minimize creative extrapolation. The output length was set to $max_tokens = 65K$ to avoid truncation of generated judicial records. For models with architectural limitations, like Fanar and ALLaM-7B-Instruct-Preview, the maximum context window (4K tokens) was used.

- **Phase 2: Full Court Simulation** ($N = 200$): The proposed court simulation methodology utilizing the advanced prompt template equipped with the Implicit Thinking and Legal Reference modules, was deployed across 200 scenarios. To empower the selected Gemini 3.1 Pro model to simulate a dynamic, human-like judicial persona, the parameters were set to $T = 1$ and $top_p = 0.95$, and $max_tokens = 65K$. Crucially, to accommodate the complex multi-step logic required by the prompt's sequential implicit thinking framework, the model's internal reasoning effort parameter was explicitly set to "High", maximizing the computational allocation for logical deduction prior to final text generation. The prompt was supplemented with three legal documents: Family Law Chapter II (Divorce) and articles related to Iddah and Mahar, the Law of Evidence, and the Law of Civil Procedure. Prompts were designed to describe the role and responsibilities of each participant in the courtroom, including the Judge, Plaintiff, Defendant, and Clerk.

Evaluation Metrics

The generated court simulation cases were evaluated using the criteria described in Table 4. To provide a comprehensive assessment of performance, both descriptive statistics and agreement-based metrics were employed.

- **Descriptive Statistics:** For each evaluation criterion, the mean, minimum, maximum, and mode scores were computed across all expert ratings. These statistics provide insights into the central tendency, score distribution, and consistency of the generated court simulation cases.

- **Success Rate:** In addition to descriptive statistics, a Success Rate metric was used to quantify the proportion of court simulation cases that received a perfect score of 5.0 for each evaluation criterion. The Success Rate is calculated as:

$$\text{Success Rate } (\%) = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100 \tag{1}$$

  where (N_{success}) denotes the number of court simulation cases that received a perfect score of 5.0 for the criterion under consideration, and (N_{total}) denotes the total number of evaluated court simulation cases.

- **Inter-Annotator Agreement:** To assess the reliability of expert judgments, two agreement measures were calculated. First, the Percentage Agreement (Exact Agreement) measures the proportion of evaluation decisions for which both experts assigned identical scores:

$$\text{Exact Agreement } (\%) = \frac{N_{\text{agreement}}}{N_{\text{total}}} \times 100 \tag{2}$$

  where ($N_{\text{agreement}}$) denotes the number of evaluation decisions receiving identical scores from both experts, and (N_{total}) denotes the total number of evaluated court simulation cases.

Second, Adjacent Agreement (± 1) was calculated to account for minor discrepancies between expert ratings. Under this measure, two evaluations are considered to be in agreement if their assigned scores differ by no more than one point on the five-point Likert scale.

$$\text{Adjacent Agreement } (\%) = \frac{N \pm 1}{N_{\text{total}}} \times 100 \tag{3}$$

where $(N \pm 1)$ denotes the number of evaluation decisions for which the experts' ratings differed by at most one point.

Results and Discussion

Phase 1: Model Selection (Pilot Study)

The results from the pilot evaluation, detailed in Table 5, indicate that Gemini 3.1 Pro outperformed both the multilingual models and the Arabic LLMs across all intrinsic capability metrics.

Evaluation Criterion	Command A	GPT-5	Gemini 3.1 pro	Command R7b Arabic	Fanar	Allam
C1: Factual	3.0	3.6	**4.8**	2.4	2.6	2.8
C2: Legal Reasoning	1.6	2.4	**4.2**	1.4	1.2	1.8
C3: Judgment Pronouncement	2.2	2.4	**4.2**	1.8	1.4	1.6
C4: Legal Language	3.2	4.2	**5.0**	2.6	3.2	3.8
C5: Judicial Procedures	3.0	4.2	**4.8**	2.0	2.6	2.8
C6: Clarity & Conciseness	3.0	3.6	**4.8**	2.8	3.2	3.0
C7: Relevance	2.8	3.4	**5.0**	2.6	2.0	1.6
C8: Consistency	2.0	2.6	**4.6**	2.0	1.8	1.6
Overall Average	2.6	3.3	**4.68**	2.2	2.25	2.38

Table 5. Phase 1 pilot study baseline evaluation of foundational LLMs. The table presents the intrinsic legal reasoning capabilities of six state-of-the-art models in a zero-shot setting, evaluated on a 5.0-point scale. Models were tested strictly on their foundational knowledge without the provision of external law references or advanced prompting guardrails.

Because this phase tested the models using a zero-shot prompt without external data retrieval, it exposed the foundational limitations of most LLMs in specialized jurisprudence. The most striking performance gap occurred in Legal Reasoning (C2) and Judgment Pronouncement (C3). While Gemini 3.1 Pro maintained robust scores (4.2 in both criteria), the next highest model, GPT-5, scored only 2.4. Furthermore, the Arabic models (Command R7b Arabic, Fanar, and Allam) suffered severe degradation in these categories, scoring between 1.2 and 1.8.

A qualitative inspection of the raw generated outputs confirms this statistical disparity. For example, in a test scenario involving a divorce pronounced while the wife was menstruating, Gemini 3.1 Pro was the only model to successfully reason through the scenario without external prompting. It correctly identified the act as a Bid'i (innovated) divorce and ruled it legally invalid.

In stark contrast, competing models failed entirely. Command R7b Arabic utilized uncharacteristically conversational vernacular and issued a legally invalid ruling validating the divorce. Similarly, ALLaM-7B hallucinated entirely irrelevant statutes (such as Article 117) to justify an incorrect ruling.

While Gemini 3.1 Pro exhibited superior baseline reasoning and perfect expert scores (5.0) in Legal Language (C4), it did exhibit minor inaccuracies in its intrinsic statutory recall (e.g., misidentifying Clause 4 as Clause 2). This specific limitation proved highly valuable; it justified Gemini's selection as the foundational architecture due to its superior cognitive reasoning, while simultaneously proving the absolute necessity of the proposed Legal Reference Module to explicitly ground the model with exact article texts in the final pipeline.

Validation of the Proposed Methodology

Before scaling the proposed methodology to the full dataset, it was imperative to validate its efficacy. To do this, the advanced court simulation pipeline was tested on the exact same 5 pilot scenarios used in Phase 1. The results, detailed in Table 6, compare the baseline zero-shot model against the proposed court simulation, both Before and After Human-in-the-Loop (HITL) expert refinement. Crucially, to prevent evaluation bias, the legal expert assigned to evaluating these five scenarios was entirely independent from the expert who conducted the qualitative HITL refinement.

Evaluation Criterion	Baseline (Gemini 3.1 pro)	Court Simulation (Before Expert Refinement)	Court Simulation (After Expert Refinement)
C1: Factual	4.8	4.8 (0.0)	5.0 (0.2)
C2: Legal Reasoning	4.2	4.9 (0.7)	5.0 (0.8)
C3: Judgment Pronouncement	4.2	4.8 (0.6)	5.0 (0.8)
C4: Legal Language	5.0	4.3 (-0.7)	5.0 (0.0)
C5: Judicial Procedures	4.8	4.9 (0.1)	5.0 (0.2)
C6: Clarity & Conciseness	4.8	4.3 (-0.5)	5.0 (0.2)
C7: Relevance	5.0	5.0 (0.0)	5.0 (0.0)
C8: Consistency	4.6	4.9 (0.3)	5.0 (0.4)
Overall Average	4.68	4.74 (0.0625)	5.0 (0.325)

Table 6. Validation of the proposed methodology on the 5-case pilot subset. The table compares the intrinsic legal performance of the baseline zero-shot model (Gemini 3.1 Pro) against the advanced court simulation pipeline, both prior to and following Human-in-the-Loop (HITL) intervention. Values in parentheses denote the performance delta (Δ) relative to the baseline.

Advancements in Legal Processing. The quantitative comparison proves that the proposed court simulation prompt architecture successfully resolves the core weaknesses of the zero-shot baseline. The most significant improvements occurred in substantive legal metrics: Legal Reasoning (C2) surged from 4.2 to 4.9 ($\Delta + 0.7$), and Judgment Pronouncement (C3) improved from 4.2 to 4.8 ($\Delta + 0.6$). Furthermore, Consistency (C8) increased by +0.3. This demonstrates that by forcing the model to implicitly reason and strictly cite external statutes, the pipeline virtually eliminates the risk of logical derailment and incorrect judicial rulings observed in unconstrained LLMs.

The Cognitive Trade-off in LLMs. Interestingly, the data reveals a documented LLM phenomenon: a cognitive trade-off between complex reasoning and linguistic fluency. While the unconstrained baseline scored a perfect 5.0 in Legal Language (C4) and 4.8 in Clarity & Conciseness (C6), the automated pipeline ("Before Refinement") saw these scores drop to 4.3 ($\Delta - 0.7$ and $\Delta - 0.5$, respectively). By mandating strict adherence to procedural, law citations, and multi-step reasoning, the model's output became overly formulaic, occasionally leading to verbosity that detracted from the natural flow of traditional judicial drafting.

The Necessity of HITL Expert Refinement. This stylistic degradation highlights the absolute necessity of the final Human-in-the-Loop intervention. As shown in the "After Expert Refinement" column, the manual expert review specifically targeted and corrected these linguistic inefficiencies. The human refinement successfully restored both Legal Language and Clarity to perfect 5.0 scores, while simultaneously pushing the reasoning metrics to absolute perfection. Ultimately, the methodology achieved a flawless overall average of 5.0 ($\Delta + 0.325$ over the baseline), proving that the combination of AI-driven legal reasoning and human refinement produces a high-quality, realistic court simulations.

Phase 2: Full Court Simulation

To evaluate the robustness of the proposed methodology, the evaluation pipeline was deployed across 200 diverse scenarios, generating 400 independent evaluations (two experts per case).

Overall Performance and Reliability

As illustrated in Table 7, the model's performance across the assessment criteria was exceptionally strong and consistent. The court simulation pipeline achieved a high overall mean score of 4.79, with a consistent mode of 5 across all categories.

Notably, the proposed methodology achieved an overall success rate of 70.8%, meaning that in over two-thirds of all evaluations, the generated judicial output was deemed entirely flawless (scoring a perfect 5) by the legal experts.

Evaluation Criterion	Mean	Min	Max	Mode	Success Rate Based on '5' Scores
C1: Factual	4.83	3.0	5	5	74.5%
C2: Legal Reasoning	4.80	2.5	5	5	71.0%
C3: Judgment Pronouncement	4.78	3.0	5	5	64.5%
C4: Legal Language	4.46	3.0	5	5	40.5%
C5: Judicial Procedures	4.86	3.5	5	5	80.5%
C6: Clarity & Conciseness	4.74	3.5	5	5	61.5%
C7: Relevance	4.93	**4.0**	5	5	**89.5%**
C8: Consistency	**4.88**	3.0	5	5	84.5%
Overall Average	**4.79**	**3.2**	**5**	**5**	**70.8%**

Table 7. Statistical summary of expert evaluation results for the full court simulation. Based on 200 highly diverse divorce scenarios, the table details the mean, minimum, maximum, and mode scores across eight judicial criteria on a 5.0-point Likert scale.

Because legal text evaluation is inherently subjective, analyzing the Inter-Annotator Agreement (IAA) is crucial for validating these high scores. The evaluation demonstrated exceptionally high consensus, achieving an overall exact agreement of 73.3% and a remarkable adjacent agreement (± 1 point) of 94.6%, as presented in Table 8. This indicates that the AI-generated rulings were consistently high in quality, and any divergence in expert scoring was virtually always limited to minor, single-point variances rather than fundamental disagreements on legal validity.

Evaluation Criterion	Percentage Agreement (Exact Agreement)	Adjacent Agreement (± 1 between experts)
C1: Factual Allegations	75.5%	**98.0%**
C2: Legal Reasoning	72.0%	96.0%
C3: Judgment Pronouncement	67.0%	96.5%
C4: Legal Language	50.5%	84.5%
C5: Judicial Procedures	83.0%	95.5%
C6: Conciseness & Clarity	63.5%	94.5%
C7: Relevance	**89.5%**	**96.5%**
C8: Coherence & Consistency	85.5%	95.5%
Overall Average	**73.3%**	**94.6%**

Table 8. Inter-Annotator Agreement (IAA) analysis across 400 independent expert evaluations. The table illustrates the consensus between two legal experts evaluating the 200 simulated cases. The metrics detail both exact agreement and adjacent agreement (± 1 variance on the 5.0-point scale).

Strengths in Objective Legal Criteria

A granular analysis reveals that the model excels in highly objective, structural criteria. Relevance (C7) achieved the maximum success rate (89.5% of scores were perfect 5s) and the highest mean (4.93). Notably, the minimum score for Relevance never dropped below 4.0, proving the model's exceptional reliability in staying on topic and preventing hallucination. Similarly, Consistency (C8) and Judicial Procedures (C5) achieved high success rates of 84.5% and 80.5%, respectively. The high exact agreement between experts in these categories (83%–88.5%) confirms that the integration of the Implicit Thinking module effectively grounds the model, allowing it to process massive amounts of contextual case data while strictly adhering to

Saudi courtroom sequencing without deviation.

Stylistic Variances vs. Substantive Errors

In contrast, Legal Language (C4) and Clarity & Conciseness (C6) showed comparatively lower performance, with success rates of 40.5% and 61.5%, and mean scores of 4.46 and 4.74, respectively. While these scores indicate occasional verbosity, stylistic inconsistency, or minor deviations from formal judicial phrasing, the IAA results provides critical context. Despite a lower exact agreement in Legal Language (50.5%), the experts maintained a high adjacent agreement (84.5%). This statistical variance accurately reflects the subjective nature of legal drafting; while the experts occasionally differed on whether the AI's specific lexicon deserved a "perfect 5" or an "excellent 4," they consistently agreed the text was structurally sound.

Finally, while the overall means for Legal Reasoning (C2) and Judgment Pronouncement (C3) remained high (4.80 and 4.78), the presence of lower minimum scores (2.5 and 3.0, respectively) highlights a critical limitation. These minimums suggest that while the methodology succeeds in the vast majority of scenarios, isolated cases—particularly highly complex or emotionally nuanced edge-case scenarios—remain susceptible to substantive legal errors. This observation reinforces the necessity of the final Human-in-the-Loop (HITL) review stage to correct these rare edge-case deviations prior to formal downstream deployment.

Phase 3: Component Analysis (Ablation Study)

To quantify the relative contribution of the proposed court simulation core components, an ablation study was conducted on a representative subset of 10 scenarios. For the fully optimized state (After Expert Refinement), generation parameters remained at $T = 1.0$ and $top_p = 0.95$. However, for the ablated ("w/o") models, parameters were strictly lowered to $T = 0.0$ and $top_p = 0.2$. This reduction is scientifically necessary; removing structural guardrails at high temperatures induces uncontrollable hallucinations. Lowering the temperature successfully isolates the true deterministic impact of the missing component. The statistical impact of removing specific components is detailed in Table 9.

Evaluation Criterion	Court Simulation (After Expert Refinement)	w/o Expert Refinement	w/o Legal Law	w/o Implicit thinking
C1: Factual	5.0	4.75 (-0.25)	4.80 (-0.20)	4.64 (-0.36)
C2: Legal Reasoning	4.7	4.95 (+0.25)	3.80 (-0.90)	4.45 (-0.25)
C3: Judgment Pronouncement	4.8	4.90 (+0.10)	4.60 (-0.20)	4.36 (-0.44)
C4: Legal Language	5.0	4.40 (-0.60)	4.80 (-0.20)	4.91 (-0.09)
C5: Judicial Procedures	5.0	4.85 (-0.15)	4.50 (-0.50)	4.73 (-0.27)
C6: Clarity & Conciseness	4.9	4.45 (-0.45)	4.80 (-0.10)	4.73 (-0.17)
C7: Relevance	5.0	4.90 (-0.10)	4.80 (-0.20)	4.82 (-0.18)
C8: Consistency	5.0	4.95 (-0.05)	4.40 (-0.60)	4.73 (-0.27)
Overall Average	4.925	4.769 (-0.156)	4.563 (-0.362)	4.670 (-0.255)

Table 9. Component analysis and ablation study results. Quantitative evaluation of the fully optimized court simulation pipeline (After Expert Refinement) compared against three ablated configurations. Scores are based on expert evaluations using a 5.0-point Likert scale across 10 representative scenarios. Values in parentheses denote the performance delta (Δ) relative to the fully refined baseline, highlighting the specific degradation caused by the removal of each architectural component.

The Impact of HITL Expert Validation. The ablation of the Human-in-the-Loop intervention ("w/o Expert Refinement") reveals a highly complex dynamic in LLM behavior. When the expert refinement is removed, the model suffers massive drops in Legal Language ($\Delta - 0.60$) and Clarity & Conciseness ($\Delta - 0.45$). However, its raw Legal Reasoning (C2) actually scores slightly higher ($\Delta + 0.25$) without the human. This indicates that the proposed methodology generates overly exhaustive, highly theoretical legal reasoning. While this exhaustive approach is technically accurate (resulting in a high reasoning score), it creates verbosity that is unnatural and inappropriate for a real courtroom. The human validation step is essential not because the AI lacks legal logic, but because the experts simplify the text. By cutting out unnecessary AI over-explanation, the expert sacrifices a fraction of the exhaustive reasoning to massively elevate the procedural tone, conciseness, and linguistic fluency to the standard expected in formal Saudi judicial rulings.

The Necessity of Legal Referencing. External law grounding proved to be the most critical component. Its ablation ("w/o Legal Law") caused the sharpest overall performance decline ($\Delta - 0.362$). Stripped of this component, the Legal Reasoning (C2) metric suffered a catastrophic degradation ($\Delta - 0.90$), dropping to a study-low of 3.80. Because Saudi Personal Status

law is strictly governed by codified guidelines, the absence of jurisdictional referencing leaves the model unable to generate defensible legal arguments. This inability to anchor arguments to law subsequently causes a cascading failure in the model's internal logic, reflected by a severe drop in Consistency ($\Delta - 0.60$).

The Role of Implicit Thinking. Ablating the implicit thinking component resulted in the second-largest overall degradation ($\Delta - 0.255$). Without multi-step internal cognitive processing, the model's ability to extract and retain Factual (C1) dropped significantly ($\Delta - 0.36$). More importantly, its Judgment Pronouncement (C3) capability suffered notably ($\Delta - 0.44$). Qualitative observation confirmed that in complex scenarios—such as evaluating the husband's state of mind during the pronouncement, parsing explicit versus implicit phrasing, or assessing statutory impediments like menstruation *(Hayd)*—the implicit reasoning component acts as the necessary bridge connecting raw case facts to accurate, logical judicial outcomes.

Conclusion

In this study, we introduced a simulation-based framework for modeling Saudi divorce establishment (*Ithbat Talaq*) proceedings. Utilizing Gemini 3.1 Pro augmented with *Implicit Thinking* and *Legal Reference* modules, the system successfully compelled the LLM to internalize the judicial process and generate highly accurate court records. Evaluated at scale across 200 diverse scenarios, the pipeline achieved flawless expert scores in over 70% of cases, supported by a remarkable 94.6% inter-rater adjacent agreement.

However, critical evaluation revealed notable cognitive limitations. The model occasionally exhibited stylistic verbosity, scoring lower in formal Legal Language. More substantively, the AI struggled with jurisprudential (*Fiqh*) edge cases—such as the divorce of a menstruating woman prior to consummation. Because the model relies strictly on codified texts, it lacks the normative guidance to infer these doctrinal exceptions independently. These findings definitively establish the necessity of Human-in-the-Loop (HITL) expert validation, confirming that current LLMs function optimally as human-supervised drafting tools rather than autonomous judicial arbiters.

The study focused on claims for establishing divorce to avoid interdisciplinary requirements present in other areas, such as *khul*, annulment, or custody. Future work will explore a multi-agent framework, decomposing judicial functions into specialized agents for fact analysis, legal reasoning, and judgment. This approach is expected to improve reasoning accuracy, reduce stylistic errors, and enable expansion to broader family law domains.

References

1. Liu, F. *et al.* Recoworld: Building simulated environments for agentic recommender systems. *arXiv preprint arXiv:2509.10397* (2025).
2. Lim, E. *et al.* Matrix: Multi-agent simulation framework for safe interactions and contextual clinical conversational evaluation. *arXiv preprint arXiv:2508.19163* (2025).
3. Almansoori, M., Kumar, K. & Cholakkal, H. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678* (2025).
4. Gonnermann-Müller, J., Haase, J., Fackeldey, K. & Pokutta, S. Facet: Teacher-centred llm-based multi-agent systems-towards personalized educational worksheets. *arXiv preprint arXiv:2508.11401* (2025).
5. Chen, G. *et al.* Agentcourt: Simulating court with adversarial evolvable lawyer agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, 5850–5865 (2025).
6. Zhang, K. *et al.* Chinese court simulation with llm-based agent system. *arXiv preprint arXiv:2508.17322* (2025).
7. Liu, X. *et al.* Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688* (2023).
8. Kmainasi, M. B., Shahroor, A. E. & Al-Ghraibah, A. Can large language models predict the outcome of judicial decisions? *arXiv preprint arXiv:2501.09768* (2025).
9. Almalki, A., Alsafari, S. & Alotaibi, N. M. Legal judgment prediction in the saudi arabian commercial court. *Futur. Internet* **17**, 439 (2025).
10. Shairah, H. A. *et al.* Alarb: An arabic legal argument reasoning benchmark. In *Proceedings of The Third Arabic Natural Language Processing Conference*, 389–406 (2025).
11. Ammar, A., Koubaa, A., Benjdira, B., Nacar, O. & Sibaee, S. Prediction of arabic legal rulings using large language models. *Electronics* **13**, 764 (2024).
12. Ahmadian, A. *et al.* Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698* (2025).
13. OpenAI. GPT-5 (2025). Accessed: 2026-05-05.

14. Google. Gemini 3.1 Pro (2026). Accessed: 2026-05-05.

15. Alnumay, Y. *et al.* Command r7b arabic: a small, enterprise-focused, multilingual, and culturally aware arabic llm. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, 126–135 (2025).

16. Team, F. *et al.* Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944* (2025).

17. Bari, M. S. *et al.* Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390* (2024).

Author contributions statement

Zarah Shibli: conceptualization, methodology, experiments, analysis, figures and visualization, writing the first draft, revising and finalizing the manuscript. Areej Alhothali and Lulwah Al-Harigy: conceptualization, revising the final version, supervision.

Data availability

The datasets generated during the current study are publicly available in Zenodo:
Shibli, Z. (2026). *Synthetic Saudi Divorce Court Simulation Dataset* [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.20736304

Declarations

Competing interests

The authors declare no competing interests.

Funding

The project was funded by KAU Endowment (WAQF) at king Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SyntheticLegalDataGenerationforSaudiFamilyLawCourtsUsingMultiRoleSimulationSupplementary.pdf