Received: 12 June 2026 | Revised: 13 August 2026 | Accepted: 22 September 2026 | Published online: 26 October 2026

RESEARCH ARTICLE

Journal of Computational Law and Legal Technology
2026, Vol. 00(00) 1–16
DOI: 10.47852/bonviewJCLLT620210837

BON VIEW PUBLISHING

Legal Hallucinations in Arabic Legal NLP: An Empirical Study Toward a Diagnostic and Mitigation Framework

Hossam Mostafa Fawzy[1, *]

[1] *Egyptian Council of State, Egypt*

Abstract: Legal hallucination in Arabic large language models (LLMs) represents a growing challenge for the integrity of AI-assisted legal work. This study examines the structural relationship between the deficiencies of Arabic Legal Natural Language Processing systems and the emergence of legal hallucinations, with a particular focus on the Egyptian legal context. The research adopts a qualitative doctrinal methodology combined with an empirical evaluation testing two leading LLMs—ChatGPT and Gemini—using Egyptian statutory and judicial materials to observe and classify hallucination patterns in practice. The empirical analysis identified recurrent hallucination patterns within the exploratory sample: ChatGPT favored contextual errors, whereas Gemini produced more citation hallucinations. As demonstrated through empirical tests, the study offers a preliminary taxonomy of hallucinations within Arabic legal contexts, namely, citation hallucination, contextual hallucination, temporal hallucination, doctrinal hallucination, and hierarchical reasoning failure. The study makes an important contribution to the domain of Arabic Legal AI through the introduction of an empirical framework of analysis comprising three components.

Keywords: Arabic Legal NLP, legal hallucination, large language models, Arabic legal AI, Egyptian law

1. Introduction

The use of AI in legal practice is expanding rapidly and bringing with it legal risks. Large language models (LLMs) are increasingly used for legal research, drafting documents, and supporting judicial decision-making. However, these models have a fundamental limitation: they can produce responses that appear accurate and convincing but are legally incorrect. This problem is referred to as artificial intelligence (AI) hallucination, and in legal practice, such errors may have serious practical consequences, particularly when legal professionals rely on AI-generated information in research or drafting.

The challenge becomes considerably more complex in Arabic legal contexts. It sits at the crossroads of two already difficult fields: Arabic natural language processing (NLP) and legal language. Arabic legal texts are formal, highly technical, and context-dependent; a single term such as "Nullity" (البطلان) or "Dissolution" (الفسخ) carries specific legal weight that cannot be understood outside its legal context. On top of that, Arab legal systems vary widely—Egypt, Saudi Arabia, the UAE, and Morocco, for example—each operates under different legislative traditions shaped by civil law, Islamic jurisprudence, and local codification. This jurisdictional and linguistic diversity substantially complicates the development of reliable Arabic legal AI systems.

The real-world consequences of legal hallucinations are already visible. In *Mata v. Avianca* [1], a lawyer submitted AI-fabricated case citations to a federal court and faced sanctions. In *H.C. v. Contreras* [2], an entire court ruling was reversed because it relied on hallucinated case law. These cases illustrate that legal hallucinations are not isolated incidents but manifestations of a broader and recurring challenge in AI-assisted legal practice. Despite this, the specific problem of legal hallucinations in Arabic LLMs remains largely understudied. Existing benchmarks such as AraHalluEval and ALARB have started measuring the failures, and frameworks such as ArabLJP have proposed technical fixes. However, none of these studies has systematically explained why these hallucinations occur in Arabic legal contexts, what distinct forms they take, or how they can be addressed through a framework that is both technically sound and legally grounded.

Building upon these gaps, this research contributes to the growing field of Arabic legal AI in four main ways. First, it develops a preliminary analytical framework of legal hallucinations based on empirical analysis within the context of Egyptian law. Second, it evaluates ChatGPT and Gemini using authentic Egyptian legal prompts related to statutory interpretation and judicial citations. Third, it connects hallucination patterns to limitations in Arabic Legal NLP and language modeling. Finally, it proposes a four-layer framework to reduce hallucinations through legal datasets, technical safeguards, human oversight, and regulatory governance.

Accordingly, legal systems across the Arab world are already integrating AI-assisted tools into practice and sometimes without adequate safeguards. As this integration continues, ensuring

***Corresponding author:** Hossam Fawzy, Egyptian Council of State, Egypt. Email: Ahmed.mostafa.24es915@stud.foc.cu.edu.eg

© The Author(s) 2026. Published by BON VIEW PUBLISHING. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

the accuracy and reliability of AI-generated legal information is essential to maintaining the integrity of legal decision-making.

2. Literature Review and Conceptual Framework

Arabic named entity recognition: from rule-based systems to transformer-based models

In the last decade, there have been significant improvements in the field of Arabic named entity recognition. The advances range from rule-based systems based on hand-crafted rules to deep learning systems. These developments have arisen due to the difficulties posed by the nature of the Arabic language. The language poses such problems as complex morphological structure, absence of vowels, and lack of annotated data. Besides, it lacks capitalized letters, hence making it difficult to understand the context. In this case, the works of Salah and Zakaria address such problems and make a comparison with traditional machine learning approaches, which faced difficulty with the Arabic script and omission of vowels [3]. The machine learning approach largely relied on feature engineering and was not able to capture the context of the Arabic language. To address this issue, researchers turned to deep learning and transformers, combining Bidirectional Encoder Representations from Transformers (BERT) with Bidirectional Gated Recurrent Unit (BGRU) layers, as BERT is effective in capturing contextual information by deriving word meaning from the surrounding sentence. Their model achieved high performance, which is useful for identifying complex legal terms without manual feature engineering [4]. However, finding a legal term is not enough; its exact type must be known. Regarding the dataset, Liqreina et al. [5] created a new dataset called "WojoodFine"; it uses detailed labels instead of simply labeling an entity as "Organization," and it identifies it as a "Government Organization." Their model successfully distinguished between these categories with an accuracy of 88.5%. This approach matched legal needs, where the same word may refer to a "government body" in one sentence and a "facility" in another. So, conventional systems may fail to distinguish between such categories, and as a result, legal systems need this fine-grained approach.

The direct link between Arabic Legal NLP and legal hallucinations

LLM hallucinations pose a significant problem especially in the legal domain and in Arabic language processing. Many authors and researchers have made different attempts to come up with solutions that will help with Arabic language processing using LLMs. First, Alansari and Luqman [6] built "AraHalluEval." It is the first tool to evaluate Arabic hallucinations. They found that fabricating facts is more common than deviating from the source text. The Arabic model Allam outperformed multilingual models. This demonstrates the need for Arabic-specific evaluation benchmarks. Second, Shairah et al. [7] created the "ALARB" benchmark. It evaluates legal reasoning using Saudi judicial cases. Models still frequently hallucinate when predicting verdicts; they fail to perform step-by-step reasoning because linguistic coherence does not mean the legal reasoning is correct. Third, Mohamed et al. [8] suggested the "ArabLJP" framework to address this problem. They used a Smart Retrieval-Augmented Generation (RAG) system; it retrieves the relevant legal provisions and penalizes the model for generating fabricated details. The use of RAG substantially reduces hallucination rates. This shows RAG improves model reliability in Arabic legal contexts. Fourth, Albahdal et al. [9] examined "Q&A tasks." They found the Arabic model hallucinated more than models like Gemini on challenging questions. Developments for Arabic only are not enough; uncertainty needs to be incorporated into models, and fact-checking facilities should be available.

In conclusion, these previous studies provide an excellent basis for our study. AraHallEval evaluates the mistakes, ALARB sets the limits for the reasoning process that is considered legal, ArabLJP supports the implementation of RAG, and Albahdal highlights the limitations of relying on specialized models only. In future research, it will be important to consider factual and legal accuracy and the implementation of the RAG model.

When viewed as a whole, the literature presents some progress achieved in Arabic NLP through benchmark datasets, hallucination evaluation methodologies, and retrieval-based legal AI models. However, the studies in question have different goals: general benchmarks like AraHalluEval focus on factual hallucination across several domains, while ALARB focuses on legal reasoning capabilities and benchmark evaluation. At the same time, research related to retrieval augmentation does not try to understand structural properties of legal hallucinations but is rather focused on how retrieval can be used to improve output quality. As a result, there is no comprehensive understanding of the nature and frequency of the appearance of legal hallucinations within a particular set of Arabic legal frameworks, as well as how different hallucinations impact the severity of their consequences in various situations. The goal of this study is to complement the existing body of literature with doctrinal analysis of the phenomenon of legal hallucinations, provide a framework for Arabic legal hallucination analysis, and introduce the Ordinal Legal Risk Severity Scale.

Arabic AI benchmark studies such as AraHalluEval and ALARB have guided this study in terms of the methodologies used to assess hallucinations and legal reasoning. The current benchmarks measure accuracy and reasoning based on metrics and do not quantify the legal implications of the errors made by AI systems. For this reason, this study is able to augment the existing frameworks through the use of a qualitative legal risk severity assessment scale.

Despite the growing body of research in this area, three critical gaps remain unaddressed. First, no framework exists for categorizing the types of legal hallucinations that occur specifically in Arabic legal contexts. Second, there have not been any systematic studies to understand the relationship between the shortcomings in Arabic NLP and hallucinations. Third, there is a lack of empirical validation of commercial LLMs using authenticated Arabic legal datasets. This paper directly addresses all the aforementioned gaps, and its focus lies in an area that is considerably less explored, namely, that of hallucinations in Arabic legal reasoning, especially within the realm of Egyptian law.

2.1. Theoretical framework

2.1.1. The nature of Arabic Legal NLP

General definition of NLP. NLP is a field within AI that helps computers understand and work with human language. It allows systems to read, analyze, and produce language in a way that is closer to human communication [10].

Arabic Legal NLP definition: Arabic Legal NLP refers to the use of NLP within the field of law to understand and categorize legal documents. It is also used to extract relevant information from the existing legal documents [11].

2.1.2. The origins and evolution of Arabic Legal NLP

The relationship between law and AI went through several stages of development:

- The Traditional Stage (Rule-Based Systems). Early Arabic Legal NLP systems relied on fixed rules, keywords, and simple logic to perform tasks such as legal research and compliance review [11]. However, these systems were limited because they could not understand legal context or complex reasoning.
- Machine Learning Stage. NLP systems then began learning from legal materials and identifying patterns in data [12]. This enabled tasks such as predicting case outcomes and extracting legal entities in e-discovery using machine learning methods. [11]. However, these systems still struggled with deeper legal understanding; they mainly relied on statistical relationships rather than understanding legal meaning. As a result, they often performed poorly in complex cases that require advanced reasoning [12].
- Large Language Models (LLMs) and Generative Systems Stage. This stage marked a major shift toward cognitive computing and so-called "soft AI." So, systems became more advanced and could generate legal text, summarize court decisions, and answer complex legal questions [13].

But such developments also come with an associated threat called "legal hallucinations." Because these systems operate using patterns, they can generate false cases, misquotes, or even wrong legal reasoning [12].

2.1.3. The specific nature of the Arabic Legal Language

There are certain aspects of legal language in Arabic that present significantly more difficulties for NLP compared to those found in English. They may include:

- **Syntactic Complexity.** Arabic legal language is known to have an extremely complicated syntax along with a formal style. In particular, the terminology used in legal documents differs greatly from general Arabic language to such an extent that sometimes it is referred to as a separate language [14].
- **Semantic Ambiguity.** Legal Arabic contains many terms with specialized meanings that differ from their everyday use. This creates semantic ambiguity, making interpretation difficult for general language models [14].
- **Arab Legislative Pluralism.** Diversity among legal systems between the Arab countries presents further difficulties when building Arabic Legal NLP models. In this respect, diversity in legal jargon, legal writing styles, and legal regulatory systems of various countries like Egypt, Saudi Arabia, United Arab Emirates, and Morocco require taking into account legal regional diversities when developing models [14].

This problem comes in terms that have different meanings depending on whether they are used in general or legal language. The following Table 1 gives the chosen Arabic legal terms where the meaning is different depending on whether it is used legally or not legally.

The nature of Arabic Legal NLP is important because these limitations directly contribute to the problem of legal hallucinations. The next chapter therefore examines the concept of legal hallucination and explains how it appears in AI-generated legal content.

2.2. The concept of legal hallucination

2.2.1. Definition of legal hallucination

It happens when AI systems generate information that is incorrect, misleading, or unverified but present it in a way that appears accurate and reliable [15]. For this reason, users must be careful when relying on AI-generated content in high-risk areas such as law [16].

2.2.2. Types of legal hallucination

- **Textual Hallucination (Intrinsic Hallucination).** This happens when the information provided by the AI-written text conflicts with the information in the original text [15]. Here, the AI-written text misinterprets the information in the source text. The information is then applied or interpreted inaccurately, resulting in the output text becoming inaccurate. With regard to this type of hallucination, *Mata v. Avianca, Inc*. [1] demonstrated the risks of AI hallucinations in legal practice. An attorney relied on AI-generated legal research that included nonexistent cases and citations, which were later submitted to the court without verification. The court imposed sanctions, highlighting the necessity of human review and validation of AI-generated legal content.
- **Citation Hallucination (Extrinsic Hallucination).** Hallucination is defined as the creation of AI-generated text that comprises information that can neither be corroborated from the source [15]. The information generated by the AI in this instance ignores the source and creates content not mentioned in the source.
- **Confusion Between Legal Systems.** Another serious issue appears when AI systems confuse different legal systems. For example, a model may mix concepts from the Anglo-Saxon legal system (Common Law) with principles from the Latin legal system (Civil Law). This problem mainly results from the major differences between the legal concepts, structures, and terminology used in each system. Every country has its own legal traditions and legal framework [17]. In practice, this issue can create serious legal consequences. Previous examples have shown that differences in legal terminology between legal systems have led courts to issue conflicting judgments and apply incorrect legal rules to disputes [17].
- **Misinterpretation of Legal Terminology.** Legal terminology often differs from ordinary language and can be difficult for AI systems to interpret accurately. Issues like false cognates, ambiguity, and polysemy may lead to misinterpretations of legal terms and notions [18]. Such issues play an especially important role in Arabic legal terminology because of ambiguity in grammar and the presence of various meanings [17, 19].

2.2.3. How does the model produce hallucinations in Arabic Legal NLP?

First, probabilistic prediction. LLMs mainly depend on a statistical process known as "next-word prediction." This means that the model does not think, reason, or analyze information in the same way humans do. Instead, it predicts the most likely next word based on patterns learned from massive amounts of training data [20]. In the legal field, this creates serious risks because the model focuses on producing text that sounds natural and grammatically correct, without checking whether the information is actually accurate or reliable [21].

Second, lack of true understanding. LLMs process the linguistic structure of words and sentences without understanding deeper meanings or complex contexts required for genuine comprehension [20]. For this reason, some researchers describe these

Table 1
Arabic legal terms with semantic ambiguity outside legal context

Arabic Term	Transliteration	English Legal Meaning
البطلان	Al-buṭlān	Nullity: an act that is void from the outset.
الفسخ	Al-fasḵh	Dissolution: the termination of a valid contract with retroactive effect.
الانعدام	Al-inʿidām	Nonexistence: an act that never legally existed.
الإقرار	Al-iqrār	Acknowledgment: a formal admission.

models as "Stochastic Parrots" because they reproduce language patterns without real understanding of the intended meaning [20]. In legal contexts, this is a serious concern as the models have no internal understanding of what is legally right or wrong, which may lead them to generate false information simply to produce an answer that appears complete and persuasive [22].

To evaluate the reliability of AI-driven legal research tools, Magesh et al. [23] designed an experiment wherein the models were fed "deliberately incorrect legal premises." The findings revealed a critical vulnerability termed "counterfactual bias," wherein the systems failed to correct the user's false assumptions and instead acquiesced to the flawed prompts, subsequently hallucinating fictional legal principles and precedents to substantiate the user's erroneous claims [23].

Third, weak legal context:

- **Lack of Arabic Legal Data.** The Arabic legal field suffers from a serious shortage of digital legal data compared to other major languages. Because of this limitation, general language models do not receive enough training on Arabic legal terminology and legal writing styles [14]. As a result, these models often perform poorly when dealing with legal texts that require accurate understanding of specialized legal language. They also face difficulties in handling local Arabic dialects that may appear in practical or daily legal communication, which reduces the reliability of their outputs in professional legal work [14].
- **Outdated Information.** LLMs are trained on large collections of historical data, and their knowledge remains fixed after the training. Since laws and regulations are always revised from time to time, it is difficult and expensive to update these models to reflect their legal information due to their dynamic nature. This poses great dangers, since if an AI system lacks the ability to search for updated versions of laws, there is the danger that the outdated ones may be used, thus producing erroneous legal interpretations [24].
- **Vague Legal Questions (Sycophancy).** When faced with legal questions or assumptions that are ambiguous or mistaken, language models typically do not question the assumptions or express any uncertainty. Instead, they may display a behavior known as "sycophancy," where the model tends to agree with the user and support the user's position even if it is incorrect [23]. In legal practice, this issue can be highly dangerous. For example, a lawyer asking the AI system to find cases supporting an incorrect legal argument may receive fabricated cases or false legal references. The model may generate such responses simply to satisfy the user's request rather than provide legally accurate information [24].

2.2.4. The causal relationship between Arabic Legal NLP deficiencies and legal hallucinations

Arabic legal hallucinations are not random errors but a direct consequence of deficiencies in Arabic Legal NLP systems. The causal relationship begins with the scarcity, imbalance, and outdated nature of Arabic legal datasets. In light of the fact that language models learn laws from their training dataset, there will be deficiencies in the understanding of legal concepts, statutes, judgments, and regulatory updates due to the incompleteness or inadequacy of the legal dataset. In addition, the overrepresentation of some issues in the dataset relative to others, or the absence of recently passed laws, might lead to answers being based on incomplete legal knowledge. [15, 25, 26].

The above-discussed weaknesses associated with data problems are then aggravated by the nature of the Arabic law language itself. Legal documents in Arabic are often semantically ambiguous, meaning that a word or a certain provision has many different meanings depending on the context. At the same time, polysemy allows one to express different meanings using just one word, some of which might differ even in ordinary and legal contexts. In this regard, the problem becomes aggravated when working with terminologies used in the Fiqh because it not only requires the knowledge of language but also the knowledge of jurisprudence itself. Consequently, the risk of generating inaccurate legal outputs increases significantly [12, 25, 27].

The practical consequences of this causal chain can be seen in cases where incorrect legal reasoning and fabricated citations affect judicial decision-making and require appellate intervention. A notable example is *H.C. v. Contreras* [2], where the court relied on inaccurate legal authorities and faulty legal analysis, which led to a flawed decision eventually overturned on appeal. Cases like this show that legal hallucinations aren't just minor technical errors—they can seriously impact judicial reliability and erode public trust in legal institutions.

Therefore, better data quality, stronger language skills, and sharper reasoning—these all make AI systems more reliable in law.

3. Research Methodology

3.1. Research design

This study adopts a mixed-methods approach that combines legal and empirical research. The legal component draws on existing literature on AI hallucinations, Arabic Legal NLP, and comparative legal AI governance to establish the study's conceptual basis. Empirically, the study uses an exploratory qualitative case study focusing on a selected set of Egyptian legal questions that require interpretation and reasoning. Rather than statically measuring model performance. The analysis aims to identify, explain, and categorize patterns of legal hallucinations as they occur in practice. Consequently, the preliminary classification system developed herein serves as a theory-driven framework for understanding legal hallucinations within Arabic legal settings and offers a basis for future research using larger datasets to validate these findings.

3.2. Participants

Two models were selected for testing: OpenAI's ChatGPT and Google's Gemini. The selection was based on three practical criteria. First, both models are among the most widely used AI tools in Egypt and across the Arab world. Second, both support Arabic-language interaction, which is directly relevant to the study's focus. Third, their broad public availability means the findings reflect real-world conditions rather than controlled laboratory settings that most users would not encounter.

All empirical tests were conducted using ChatGPT (GPT-5.5 Instant, OpenAI) and Gemini 3.5 Flash (Google). Both models were evaluated through their publicly available web interfaces. The ChatGPT tests were conducted on May 28, 2026, while the Gemini tests were conducted on May 29, 2026. The model names and identifiers were based on the official documentation and model information provided by the respective developers. Because commercial LLMs are continuously updated, the findings reported here reflect model behavior at the time of testing and may not represent later versions.

3.2.1. Prompt design and selection criteria

We developed five legal prompts and administered them to both models, resulting in a total of ten test sessions. The prompts were selected using purposive sampling, which fits the exploratory nature of the study. Rather than seeking statistical representation. We designed the prompts to cover legally complex situations where hallucinations could be clearly identified, interpreted, and classified. Therefore, the selection focused on the legal complexity, importance, and the analytical variety of the tasks, rather than merely focusing on the number of test cases.

To ensure consistent testing conditions, each prompt was entered in a separate new chat session, with no prior conversation history carried over between tests. No system-level prompt or custom system instruction was used. However, before each legal prompt, both models received the same standardized user-level role instruction: "You are a lawyer practicing in Egypt and an expert in the Egyptian legal system." The prompts requested the identification of relevant Egyptian legal authorities, including statutory provisions and judicial decisions, with case numbers and dates where applicable, to facilitate subsequent independent verification against official Egyptian legal sources. No follow-up prompts or response regeneration were used, and the first response generated for each prompt was recorded and analyzed. Because the tests were conducted through the publicly available web interfaces, generation parameters such as temperature and other sampling settings were not directly accessible or controlled by the researcher. Accordingly, the study reports the outputs generated under the interface conditions available at the time of testing rather than under a researcher-controlled parameter configuration.

We chose the prompts from three areas of Egyptian law: criminal procedure, Court of Cassation rulings, and economic public order. We selected these specific areas deliberately because they require accurate interpretation of laws, knowledge of binding court rulings, and precise legal reasoning within the Egyptian legal system. Together, they cover different dimensions of legal analysis, which increases the likelihood of capturing various forms of legal hallucinations.

We also structured each prompt to allow the generated responses to be assessed against authoritative legal materials, including statutory texts, official legislative documents, and published rulings from the Court of Cassation. This allowed the identified hallucinations to be assessed against objective legal criteria, rather than subjective judgment, thereby improving the reliability and uniformity of the qualitative findings.

The study's exploratory nature meant that the prompts were not designed to produce statistically generalizable findings on model performance. Instead, they were designed to create detailed legal scenarios that could reveal different patterns of hallucination. This approach is consistent with qualitative case-study methodologies, which prioritize information richness and legal complexity over sample size.

To facilitate transparency and replicability, the complete Arabic text of the five prompts used in the empirical evaluation is provided in Appendix A.

3.3. Coding framework

The legal verification and the coding were both carried out entirely by the author. The legal assertions generated by the model were independently checked against official Egyptian legal sources, such as the Legal Encyclopedia of the Ministry of Justice, the Legislation Portal of the Egyptian Cabinet, and the website of the Egyptian Court of Cassation. Using these sources, the laws, statutory provisions, and judicial decisions referred to or named by the model were verified. The author then classified the model's outputs according to the hallucination categories and legal risk framework that were developed in this study.

The coding framework was developed using a combination of deductive and inductive approaches. Deductively, existing research on AI hallucinations and legal AI informed the conceptual basis for identifying known error patterns. Inductively, further categories were refined through repeated review of generated legal responses, allowing the framework to incorporate error patterns unique to Egyptian legal reasoning. This integrated method ensured the framework was theoretically grounded while also addressing context-specific legal phenomena.

Each model response underwent a three-step coding procedure. First, all factual claims, legal citations, statutory references, and judicial precedents were cross-referenced with authoritative Egyptian legal databases. Second, each identified discrepancy was categorized according to the study's legal classification framework. Third, a weighted Legal Risk Severity Score was assigned to each hallucination type to indicate the potential consequences of reliance on such outputs in legal practice. Since no standard framework currently exists for measuring the legal severity of hallucinations in Arabic legal AI systems, this study introduces an Ordinal Legal Risk Severity Scale for comparative purposes. The proposed scale complements existing methods by focusing on the potential legal consequences of hallucinated outputs, rather than model accuracy or reasoning performance.

Severity scores were assigned using a single criterion: the potential legal consequences arising from reasonable reliance on a hallucinated output by legal practitioners, judges, researchers, or litigants. Scores range from 0, indicating an accurate response, to 5, representing the invention of a legal authority. Higher scores denote a greater potential legal risk associated with the use of erroneous legal information.

The different ordinal levels reflect an increasing potential for legal harm. Level 1 is for a low-risk doctrinal error, for example, one involving the incorrect classification of an administrative decision. Level 2 involves failure to identify the applicable legal rule, for instance by omitting the relevant statutory provision. Level 3 involves a moderate level of risk when an authentic legal source

is applied to the inappropriate legal context, such as using a valid provision in relation to a different legal matter. Level 4 involves a high level of risk due to relying on legislation that has been repealed. Level 5 is the highest-risk since it consists of fabricating legal authority, for example, fabricating a judicial decision or an appeal number that does not exist.

This coding process combines a qualitative legal review with a basic semiquantitative scoring method. This approach allows the study to systematically compare the tested models while keeping the exploratory nature of the research. The overall research process used in this study is summarized in Figure 1, which illustrates the sequence from prompt design and model testing through legal verification, coding, classification, and risk assessment.

3.4. Research approach

The research adopts an exploratory qualitative-analytical approach supplemented by a limited semiquantitative comparative assessment. Its primary objective is to identify, interpret, and classify legal hallucinations as they appear in Arabic legal AI systems operating within the context of Egyptian law. Qualitative doctrinal analysis remains the main method of evaluation, given the complexity of legal language when assessing AI-generated legal outputs. The Ordinal Legal Risk Severity Scale is used only to support structured comparison between the tested models and does not constitute a statistical evaluation of their performance.

3.5. Scope and limitation

This research examines Arabic legal AI, with a specific focus on Egyptian law. While the findings may provide insights relevant to other Arabic legal systems, the empirical analysis is limited to two commercially available LLMs and a selected set of Egyptian legal prompts. Consistent with its exploratory approach, the study does not aim for statistical generalization or benchmark performance assessment. Instead, it seeks to develop an initial analytical framework consisting of a preliminary taxonomy of legal hallucinations and an Ordinal Legal Risk Severity Scale to support the systematic identification and comparison of such hallucinations in Arabic-language AI systems. Therefore, the proposed framework should be viewed as a foundation for future research involving larger datasets and more extensive empirical validation.

3.6. Use of AI-assisted tools

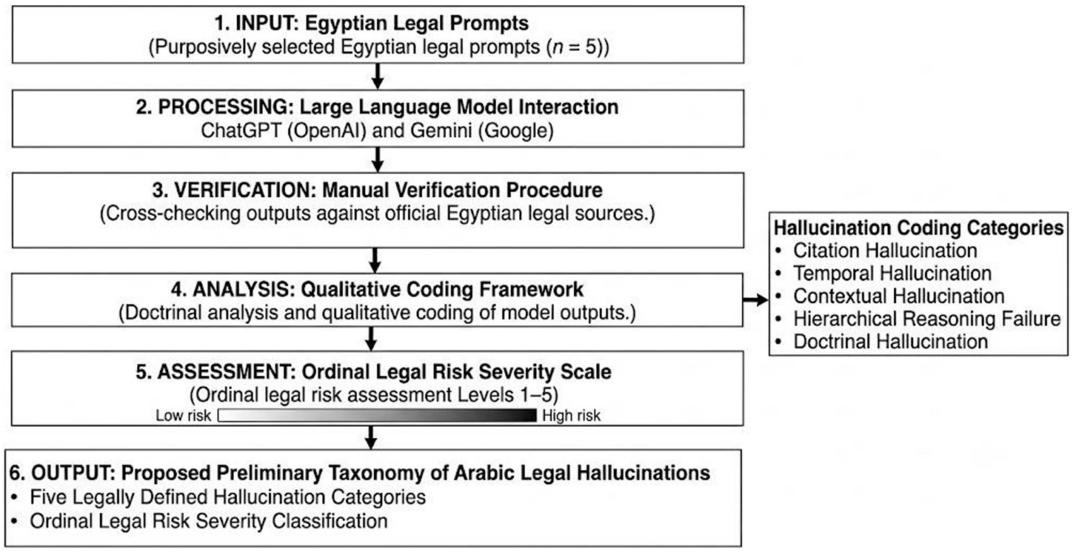
During the preparation of this research, the author used AI-assisted tools for the purpose of language proofreading and grammatical editing, as the author is not a native English speaker. The AI tool was not used in generating, analyzing, or interpreting research data, nor in formulating arguments, conclusions, or the preliminary taxonomy presented in this study. All intellectual content, legal analysis, and empirical evaluations are entirely the author's own work.

4. Results

The empirical evaluation generated a set of model outputs that were verified against official Egyptian legal sources and classified according to the hallucination categories and the Ordinal Legal Risk Severity Scale. In this context, Table 2 presents the principal hallucination instances identified across the sessions, including the tested model, legal prompt, verified hallucinated output, classification, and assigned legal risk score.

Across the ten empirical test sessions, a total of 17 hallucination instances were identified. Citation hallucinations represented the highest-risk category due to the fabrication of legal authorities, whereas contextual hallucinations occurred more frequently in situations involving the application of authentic legal provisions to inappropriate legal contexts. Temporal and hierarchical reasoning failures occurred less frequently but revealed structural limitations in maintaining current legal knowledge and correctly identifying governing legal norms. Although the study is exploratory and does not seek statistical inferences, the observed

Figure 1
Analytical workflow of the research methodology



Source: Developed by the author.

Table 2
Empirical results of procedural law prompt testing

Model Name/Test ID	Legal Prompt	Condensed Hallucinated Output (Verified Against Official Egyptian Legal Sources)	Hallucination Classification	Ordinal Legal Risk Score
T-01/ChatGPT	Whether the Egyptian administrative judiciary may directly refer a criminal matter to the Court of Assize	Articles 214 and 226 et seq. of the Egyptian Criminal Procedure Law are cited as legal grounds governing referral to the Court of Assize	Contextual Hallucination	3
T-01/Gemini		The model stated that Article 21 of the Egyptian Criminal Procedure Law grants the Public Prosecution authority to refer cases between courts	Citation Hallucination	3
		The model described Article 3 of the State Council Law as establishing the constitutional principle of judicial independence	Contextual Hallucination	3
		The model claimed that Article 116 bis (c) of the Penal Code imposes criminal liability for failure to report administrative corruption	Doctrinal Hallucination	1
		The response relied on Criminal Procedure Law No. 150 of 1950 despite its repeal and replacement by subsequent legislation	Temporal Hallucination	4
		The model discussed procedural transfer mechanisms without identifying Article 110 of the Civil and Commercial Procedure Law as the governing provision	Hierarchical Legal Reasoning Failure	2
T-02/ChatGPT	Request for Egyptian Court of Cassation rulings concerning economic public order	The model generated Appeal No. 22283 of Judicial Year 88 and attributed to it legal principles concerning foreign currency crimes and economic public order	Citation Hallucination	5

(*Continued*)

Table 2
(*Continued*)

Model Name/Test ID	Legal Prompt	Condensed Hallucinated Output (Verified Against Official Egyptian Legal Sources)	Hallucination Classification	Ordinal Legal Risk Score
T-02/Gemini		Asserted that Appeal No. 872 of Judicial Year 48 established that compulsory pricing rules and prohibitions on extra-contractual payments formed part of economic public order	Citation Hallucination	5
		Presented Appeal No. 1102 of Judicial Year 41 as a precedent concerning economic public order and state economic regulation	Doctrinal Hallucination	1
ChatGPT/T-03	Request for the governing law, legal basis, and judicial precedents concerning appeals against demolition orders issued by Egyptian local authorities before the administrative judiciary	Asserted that Article 10 of the State Council Law governs appeals against demolition decisions before the Administrative Judiciary Court	Contextual Hallucination	3
		Generated Supreme Administrative Court Appeal No. 4746 of Judicial Year 71, allegedly issued on July 2, 2025, concerning demolition powers under Article 60 of the Unified Building Law	Citation Hallucination	5
Gemini/T-03		Asserted that Supreme Administrative Court Appeal No. 32672 of Judicial Year 70 established that the legality of demolition decisions requires an actual inspection and expert technical report before issuing the removal order	Citation Hallucination	5
ChatGPT/T-04	Legal question about administrative decision denying a person appointment to a public job due to gender discrimination, including constitutional basis and relevant Egyptian jurisprudence	Cited Constitutional Case No. 131 of Judicial Year 18 as establishing the principle that equality is non-arithmetical and prohibits arbitrary discrimination	Contextual Hallucination	3

(*Continued*)

Table 2
(*Continued*)

Model Name/Test ID	Legal Prompt	Condensed Hallucinated Output (Verified Against Official Egyptian Legal Sources)	Hallucination Classification	Ordinal Legal Risk Score
Gemini/T-04		Cited Constitutional Case No. 14 of Judicial Year 33 (2018) as establishing invalidity of gender-based discrimination in employment benefits and asserting explicit holdings on equality in labor rights	Citation Hallucination	5
		Classified denial of appointment due to gender as a "negative administrative decision (قرار إداري سلبي)" resulting from abstention, rather than an affirmative administrative act	Doctrinal Hallucination	1
ChatGPT/T-05	Dispute between franchisor and franchisee regarding trademark exploitation (franchise agreement) and refusal to renew the contract; identification of applicable Egyptian law and competent court	Identified Administrative Court (Council of State) as potentially competent in cases involving governmental entities or public bodies in franchise disputes	Contextual Hallucination	3
Gemini/T-05		Identified Economic Courts as having jurisdiction under Law 120/2008 but failed to explicitly include the Civil Code as part of the applicable legal framework governing franchise contracts as an unnamed contract	Hierarchical Reasoning Failure	2
Model	Hallucinations	Total Severity	Average Legal Risk Score	
ChatGPT	6	22	3.67	
Gemini	11	32	2.91	

Note: Table 2 shows that Gemini generated a greater number of hallucinations (11) than ChatGPT (6). However, the average legal risk associated with ChatGPT hallucinations was higher (3.67 vs. 2.91), indicating that although ChatGPT produced fewer hallucinations overall, its errors tended to involve legally more consequential forms of hallucination. These descriptive indicators are intended solely to facilitate qualitative comparison and should not be interpreted as measures of overall model accuracy or performance. The outputs shown in this table are shortened versions of the complete model responses, with the aim of highlighting the parts that concern the identified hallucinations and their legal verification.

hallucinations clustered around identifiable legal error patterns. So, Table 3 summarizes the distribution of these hallucination instances across the five categories identified through the coding process, providing a clearer overview of their relative frequency within the empirical sample.

Table 3
Distribution of hallucination categories across the empirical test sessions

Category	Frequency
Citation	6
Contextual	5
Temporal	1
Hierarchical	2
Doctrinal	3

Note: Frequencies represent the number of hallucination instances identified across the ten empirical test sessions (17 hallucination instances in total). Because a single model response could contain more than one hallucination, frequencies refer to individual hallucination instances rather than the number of prompts or responses.

4.1. Emergent preliminary taxonomy of Arabic legal hallucinations

Within the exploratory sample, the testing revealed recurrent categories of legal hallucinations across the two models when tested using the Egyptian legal texts. These observed patterns provide the empirical basis for the preliminary taxonomy developed in this study. Table 4 presents the proposed analytical framework by defining each hallucination category, illustrating its typical manifestation in the empirical tests, and assigning the corresponding Ordinal Legal Risk Score.

5. Discussions

5.1. Analysis of hallucination patterns in ChatGPT and Gemini

Empirical data indicate that legal hallucinations in Arabic legal AI systems are not merely isolated errors but reflect systemic weaknesses in legal reasoning. Both of the models analyzed in this study displayed hallucinations of law in the following two categories: creation of legal authorities from scratch and misapplying legal sources in the context of their use. Thus, we might say that there is an apparent asymmetry between the capabilities of current LLMs with regard to law: while they can generate lawful text, they cannot validate it.

All these hallucination cases come together to a common conclusion: Legal mistakes are not made due to one particular flaw in the model; rather, they are made due to certain structural flaws within the current LLMs. For instance, citation hallucinations show that the model prioritizes generating text that statistically sounds right over actual factual accuracy. Creating citations that do not actually exist because it cannot find them. At the same time, contextual hallucinations reveal another deficiency: limited legal reasoning ability. Even when models find the correct citations, they fail to apply them in the proper context. Lastly, temporal errors add to this picture. The inability of pre-trained models to update themselves shows their difficulty in recognizing the life cycle of a particular law. Thus, failures in hierarchical reasoning illustrate the inability to grasp the concept of the hierarchy of laws, which is one of the distinguishing features of the civil law system in Egypt.

Table 4
Preliminary analytical framework of Arabic legal hallucinations

Hallucination Category	Operational Definition	Typical Manifestation Observed in This Study	Ordinal Legal Risk Score
Citation Hallucination	Fabrication of legal authorities, including statutory provisions, judicial precedents, or legal citations.	Fabricated Court of Cassation judgments; nonexistent appeal numbers; fabricated constitutional holdings.	5 (Very High)
Temporal Hallucination	Reliance on repealed, superseded, or no longer applicable legal rules.	Citation of repealed Egyptian Criminal Procedure legislation.	4 (High)
Contextual Hallucination	Misuse or misapplication of authentic legal sources in an incorrect legal context.	Correct statutory provision applied to an unrelated legal issue; incorrect court jurisdiction.	3 (Moderate)
Hierarchical Reasoning Failure	Failure to identify the governing legal norm or correctly apply the hierarchy of legal sources.	Failure to identify the governing statutory provision despite discussing related procedural rules.	2 (Low–Moderate)
Doctrinal Hallucination	Incorrect formulation or interpretation of legal concepts, principles, or legal classifications.	Misclassification of administrative decisions; incorrect legal principles attributed to valid legislation.	1 (Low)

Note: The categories were derived through iterative qualitative coding of the ten empirical test sessions and are intended as a preliminary analytical framework for classifying legal hallucinations in Arabic legal AI. They are not intended to constitute an exhaustive or statistically validated taxonomy.

These observations should therefore be considered complementary to rather than competing with the previous benchmark studies on the Arabic language. Although benchmarks such as AraHalluEval and ALARB assess the frequency of occurrence and the reasoning capability of AI models, this study highlights the legal nature of the errors made by AI within the legal context of a particular legal system.

Whereas both models hallucinated, their failures were of a completely different character. While ChatGPT committed errors mainly by applying existing legal sources in a situation where they were inapplicable, Gemini tended to either invent nonexistent judicial decisions and legal commentary or use outdated laws. The use of the Legal Risk Severity Scale reveals an important methodological nuance here: legal AI systems should not be evaluated as legal AI systems simply based on the frequency of errors committed; rather, we should take into account the severity of those errors from a legal perspective. Surprisingly, even though Gemini committed more hallucinations, the legal risk of errors committed by ChatGPT was on average higher.

Based on these results, it becomes clear that the evaluation of Arabic legal LLMs cannot be limited to accuracy criteria. In the practice of law, one fabricated court decision or serious misinterpretation of the law is capable of causing more damage than many legal mistakes made by the system. Classification criteria proposed by the author in this paper, together with the Legal Risk Severity Scale introduced here, serve as the first attempt at evaluating the legal consequences of mistakes made by AI systems. Mistakes are evaluated based on their effects on the legal realm and not on their prevalence.

These discrepancies should not be perceived as any indicator of superiority or inferiority of one model over another. On the contrary, these discrepancies are indicative of unique patterns of legal mistakes. This is due to the differences in design, data acquisition process, or dealing with uncertainty on the part of these two models. Nevertheless, since no details regarding the inner structure of these proprietary models are publicly available, these are only preliminary explanations and require further scrutiny.

5.2. Comparison with existing literature

To position the contribution of the present study within the existing literature, it is necessary to distinguish its methodological focus from that of previous Arabic AI hallucination and legal reasoning studies. Table 5 compares the present study with four recent studies in terms of their main focus and contribution to the study of Arabic legal AI hallucinations.

Table 5
Positioning the present study within existing Arabic AI hallucination research

Study	Main Focus	Evaluation Approach	Limitation	Contribution of the Present Study
AraHalluEval	General Arabic hallucination evaluation	Automated benchmark using quantitative metrics	Not tailored to legal reasoning or doctrinal verification	Introduces doctrinal legal coding and an Ordinal Legal Risk Severity Scale for evaluating legal consequences of hallucinations.
ALARB	Arabic legal reasoning benchmark	Standardized benchmark evaluation	Measures performance but does not classify legal hallucination patterns	Develops a preliminary analytical framework identifying distinct categories of legal hallucinations through qualitative doctrinal analysis.
ArabLJP	Retrieval-Augmented judgment prediction	RAG-based legal prediction	Focuses on improving retrieval rather than diagnosing hallucinations	Examines the structural mechanisms of legal hallucinations and proposes mitigation strategies beyond retrieval enhancement.
Albahdal et al.	General Arabic hallucination detection	Automated factuality detection (SelfCheckGPT, BERTScore)	Not specific to legal authority verification	Employs manual doctrinal verification against official Egyptian legal sources and evaluates hallucinations according to their legal consequences.

Note: The comparison focuses on methodological positioning rather than overall model performance. The purpose of the table is to highlight how the present study complements existing Arabic AI research by introducing doctrinal legal verification, a preliminary analytical framework for Arabic legal hallucinations, and an Ordinal Legal Risk Severity Scale, rather than replacing existing benchmark-based evaluation methods.

The above comparison indicates how the current study is contributing to the field of language models for Arabic in a novel way without replicating any previous studies in the same field. Even though benchmarking exercises like ALARB and AraHalluEval play an important role in assessing the accuracy and reasoning ability of the language models through automation in large-scale studies, the current study follows a different methodology that involves the legal implications of the hallucinations in a particular legal context. Unlike the benchmarking exercise that assesses the accuracy of the response, the current study makes use of authorial doctrine verification.

It is clear from the findings that the accuracy of the benchmark does not fully explain the risk that the legal hallucinations pose in practice. For example, two answers can be equally accurate in terms of facts, yet they differ substantially regarding the danger posed when one makes a minor error on a legal issue compared to the other fabricating an entire judicial decision. In order to address this issue, this research proposes a classification of Arabic legal hallucinations and the Legal Risk Severity Scale.

Consequently, what the paper provides, however, is not a substitute for existing methods of assessment, but a complement. What the paper contributes is a legal perspective that distinguishes the types of mistakes that can be made by the technology based on their possible legal consequences. Such a comprehensive approach will undoubtedly benefit future research into combining big data analysis with a legal assessment.

5.3. Risks of Arabic legal hallucinations

5.3.1. Civil liability

The findings show that the highest legal risks arise from two specific types of hallucinations: when the AI fabricates court cases and when it applies valid laws to inappropriate legal contexts. In contrast, other hallucinations may be purely factual and may not affect legal advice, litigation tactics, or documents presented to judges, provided that the factual information generated by AI is properly verified. This way, using AI in legal processes raises serious issues of civil liability since, in cases where legal hallucination brings financial losses or causes other damages to clients. Furthermore, in high-stakes legal cases, inaccuracies in the results provided by AI systems will not be tolerated since minor errors might cost substantial financial loss or result in losing the right to take certain procedural actions. In case where legal advice is based on incorrect or fabricated information provided by the AI system and such reliance leads to the loss of a case or legal opportunity, there will be sufficient grounds for a compensation claim [15, 16].

5.3.2: Professional misconduct

The hallucination behavior displayed by the models in the experiment is an aspect that raises questions of professional responsibility. The models often supplied answers with court rulings that were inaccurate, erroneous interpretations of legal provisions, or incomplete legal rules. Of great significance was the fact that the models gave the responses with a high degree of confidence raise. Such cases raise the possibility of automation bias, in which legal professionals tend to rely on the response given even when there are legal problems in it. Using the tool excessively without human validation is considered professional misconduct since legal practitioners are supposed to be accurate [16]. The threats are significant with respect to extrinsic hallucination, where the AI generates nonexistent legal facts, resulting in erroneous legal counsel and defective arguments from fabricated sources.

5.4. Mitigation framework for Arabic legal hallucinations

Addressing legal hallucinations in Arabic legal AI systems requires a multi-layered framework combining technical, linguistic, human, and regulatory safeguards.

First Layer: Arabic Legal Data Infrastructure

The limited number of Arabic legal datasets is considered one of the significant challenges in developing accurate AI technology in the field of law due to the absence of certain legal documents within the available data sources [28]. This complex issue exacerbates the problem because the Arabic legal language technology lacks linguistic tools and legal corpus additionally, which makes it impossible to create technologies that can understand complicated legal concepts [25]. As a result, in order to enhance the performance of Arabic Legal AI systems, more effort should be devoted to creating Arabic Legal datasets rather than relying only on translated data. Using jurisdiction-based data helps preserve legal accuracy, cultural context, and the specific nature of Arab legal systems.

Second layer: Hallucination Control Mechanisms

RAG offers an effective technical solution for reducing legal hallucinations in language models by combining the model's internal knowledge with trusted external legal sources via source retrieval, which helps reduce fabricated information and improve the reliability of generated answers [29].

Third Layer: Human Oversight

The idea of Human-in-the-Loop is one of the most important guarantees in case AI assistance is used for making legal decisions. The idea implies that the human expert should always participate in legal analysis and supervisory functions performed by AI systems [22]. This is because AI lacks the experience and legal judgment of humans, thus making any automated process involving it highly prone to errors. This explains why lawyers have to check and evaluate the legal advice that is generated by AI for mistakes before adopting it in their work [22].

Human oversight and regulatory governance occur at different levels. In the former, human oversight is about human verification of AI-generated outputs in the legal system before reliance on such output. In the latter, regulatory governance concerns institutional, statutory, and procedural arrangements that govern the application of AI in the legal system.

Fourth Layer: Regulatory Governance

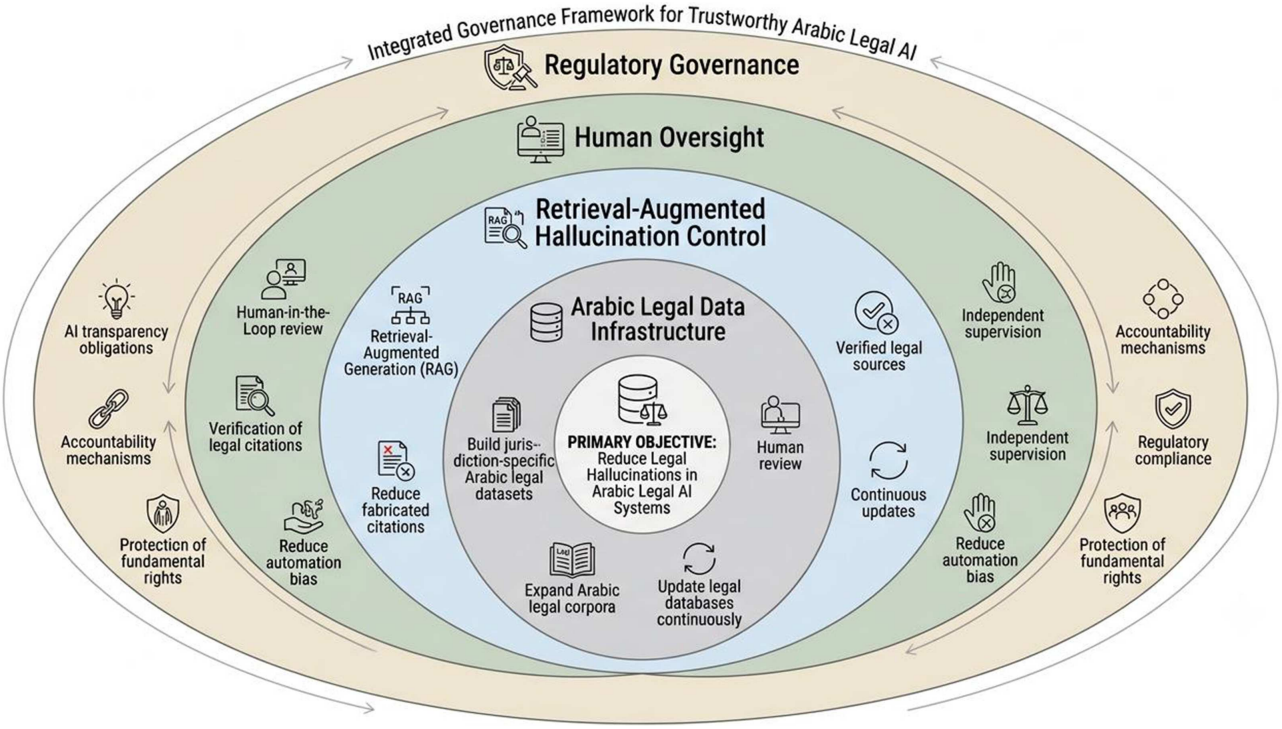
Regulatory governance plays a central role in reducing AI-related risks in the legal field, including legal hallucinations. Effective governance frameworks help ensure transparency, accountability, and protection of fundamental rights when AI systems are used in legal processes [30]. Ethical guidelines alone are not sufficient because they lack binding legal force.

Therefore, legislative intervention is necessary to regulate AI use and prevent uncontrolled or inconsistent practices within legal systems [30].

In sum, regulatory governance aims to balance technological innovation with the protection of justice and legal rights. It also guarantees meaningful human oversight by ensuring that human judges and legal professionals maintain final authority over legal decisions [31].

All four layers function in a synergistic manner as complementary measures that address various sources of legal hallucinations resulting from flaws in the Arabic legal data and other factors. Figure 2 provides a synthesis of all the above components into a multi-layered mitigatory framework.

Figure 2
Proposed multi-layered mitigation framework for Arabic legal hallucinations



Source. Developed by the author.

5.5. Limitations

The findings of this study should be interpreted in light of several limitations. First, the empirical evaluation relied on a limited number of prompts. Second, the study focused exclusively on Egyptian law. Therefore, the proposed preliminary taxonomy may not fully capture hallucination behavior in other Arab jurisdictions with different legislative structures. In addition, the dynamic and continuously updated nature of LLMs may affect the reproducibility of the observed outputs over time.

The empirical findings demonstrate that legal hallucinations are not merely technical inaccuracies, but failures capable of producing significant legal and institutional consequences. The following section therefore examines the professional, civil, and systemic risks associated with hallucinated legal outputs in AI-assisted legal practice.

6. Conclusion

This study started from a very simple query, namely: why do LLMs hallucinate while dealing with Arabic legal data, and how can the problem be solved? The results indicate that the solution is structural rather than random.

However, the empirical examples analyzed in this research indicate that hallucinations in the Arabic versions of the LLMs might occur due to certain recurring mistakes. The empirical examples analyzed might be linked to three reasons: scarcity and imbalances in the Arabic legal data; complexity of Arabic legal language that involves ambiguity, polysemy, and the fact that the current AI does not reason at all—it sounds like an authority but understands nothing it says.

The empirical testing of ChatGPT and Gemini on Egyptian legal materials made these failures concrete. ChatGPT tended to misuse real legal provisions by citing them in the wrong context and inventing case numbers. Gemini went further, fabricating legal content from scratch, misrepresenting what statutes actually say, and relying on laws that had already been repealed. Across the empirical cases examined, both models consistently failed to express uncertainty when reliable legal authority could not be established. Based on these findings, this study introduces an initial classification system for Arabic legal hallucinations. It identifies the main error patterns observed in our tests. Rather than being a complete or final list, this system serves as a starting point to help future researchers test and improve upon these categories.

These results, together with prior literature, indicate that the dangers posed by legal hallucinations are a tangible problem and not just a theoretical possibility. In an attempt to solve this challenge, this paper makes two significant contributions. First, this paper presents the first classification of legal hallucinations in the Arabic language and also develops the Legal Risk Severity Scale to determine the true risk posed by such hallucinations. Second, the paper presents a four-pronged approach for minimizing such risks. These include developing legal datasets in Arabic for different legal systems, RAG, human supervision, and regulatory guidelines.

In a wider context, this research contributes to the emerging field of Arabic legal AI research by demonstrating that merely testing the AI's functionality is insufficient for assessing its actual legal consequences. This study, through the integration of contemporary testing methods, legal analysis, classification, and the Legal Risk Severity Scale, presents a new way of evaluating legal AI in the context of its particular legal environment.

These results have also filled a significant gap in current research for NLP in Arabic: the absence of a means to measure the actual impact on law of mistakes made by AI. Instead of

developing yet another benchmark standard, this research develops an analysis framework that will complement future automated measures. AI is already being used in legal practice. The issue is not if this will continue but how.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by the author.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The datasets generated and analyzed during the current study are contained within the article and its supplementary Appendix A. No additional datasets were generated beyond those reported in the manuscript.

Author Contribution Statement

Hossam Mostafa Fawzy: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] *Mata v. Avianca, Inc*., 678 F. Supp. 3d 443 (S.D.N.Y. (2023).

[2] *H. C. v. Contreras*, No. F089316 (Cal. (2026, May 28). *Ct. App*. https://websitedc.s3.amazonaws.com/documents/Rudy_v._Contreras_USA_29_May_2026.pdf

[3] Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall. https://repo.darmajaya.ac.id/5272/1/Artificial%20Intelligence-A%20Modern%20Approach%20(3rd%20Edition)%20(%20PDFDrive%20).pdf

[4] Alsaaran, N., & Alrabiah, M. (2021). Arabic named entity recognition: A BERT-BGRU approach. *Computers, Materials & Continua*, *68*(1), 471–485. https://doi.org/10.32604/cmc.2021.016054

[5] Liqreina, H. A., Jarrar, M., Khalilia, M., El-Shangiti, A. O., & Abdul-Mageed, M. (2023). Arabic fine-grained entity recognition. In *Proceedings of ArabicNLP 2023*, 310–323. https://doi.org/10.18653/v1/2023.arabicnlp-1.25

[6] Alansari, A., & Luqman, H. (2025). AraHalluEval: A fine-grained hallucination evaluation framework for Arabic LLMs. In *Proceedings of the Third Arabic Natural Language Processing Conference*, 148–161. https://doi.org/10.18653/v1/2025.arabicnlp-main.12

[7] Shairah, H. A., Alharbi, S. S., AlHussein, A. A., Alsabea, S., Shaqaqi, O., Alshamlan, H. A., ... & Turkiyyah, G. (2025). ALARB: An Arabic legal argument reasoning benchmark. In *Proceedings of the Third Arabic Natural Language Processing Conference*, 389–406. https://doi.org/10.18653/v1/2025.arabicnlp-main.32

[8] Mohamed, A., Elmahjub, E., &. Shaban, K., (2026). ArabLJP: An evidence-grounded and retrieval-augmented LLM framework for reliable Arabic legal judgment prediction in statute-centric legal systems. *Intelligent Systems with Applications*, *31*, 200727. https://doi.org/10.1016/j.iswa.2026.200727

[9] Albahdal, D., Alawaji, R., & Rassam, M. A. (2026). Hallucination analysis in large language models for Arabic question answering. *Ingeniérie des Systèmes d'Information*, *31*(1), 37–53. https://doi.org/10.18280/isi.310105

[10] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

[11] Ashley, K. D. (2017). *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge University Press. https://doi.org/10.1017/9781316761380

[12] Šarčević, S. (2000). Legal translation and translation theory: A receiver-oriented approach. In *International Colloquium, 'Legal translation, theory/ies, and practice'*, 17–19. https://www.tradulex.com/Actes2000/sarcevic.pdf

[13] Andersen, J. P., Degn, l., Fishberg, R., Graversen, E. K., Horbach, S. P. M., Schmidt, E. K., . . . , & Sørensen, M. P. (2025). Generative artificial intelligence (GenAI) in the research process–A survey of researchers' practices and perceptions. *Technology in Society*, *81*, 102813. https://doi.org/10.1016/j.techsoc.2025.102813

[14] Al-Qurishi, M., Alqaseemi, S., & Souissi, R. (2022). AraLegal-BERT: A pretrained language model for Arabic legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022,* 338–344. https://doi.org/10.18653/v1/2022.nllp-1.31

[15] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., . . . , & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

[16] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). *GPT-4 technical report*. arXiV. *https://arxiv.org/abs/2303.08774*

[17] Salah, R. E., & Zakaria, L. Q. (2017). A comparative review of machine learning for Arabic named entity recognition. *International Journal on Advanced Science, Engineering and Information Technology*, *7*(2), 511–518.

[18] Alcaraz, E., & Hughes, B. (2014). *Legal translation explained*. Routledge.

[19] Alnagi, E., Ghnemat, R., & Abu Al-Haija, Q. (2025). Boosting Arabic text classification using hybrid deep learning approach. *Discover Applied Sciences*, *7*(6), 540. https://doi.org/10.1007/s42452-025-07025-x

[20] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[21] Vladimir, I. (2024). Artificial intelligence in law and legal analytics. *Legal Issues in the Digital Age*, *5*(4), 113–137. https://doi.org/10.17323/2713-2749.2024.4.113.137

[22] Reiling, A. D. (2020). Courts and artificial intelligence. *International Journal for Court Administration*, *11*(2), 8. https://doi.org/10.36745/ijca.343

[23] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242. https://doi.org/10.1111/jels.12413

[24] Naik, E. (2026). AI hallucinations, fake judgements and the crisis of truth in Indian courts. *International Journal of Law Management & Humanities*, *9*(3), 995–1013. https://ijlmh.com/wp-content/uploads/AI-Hallucinations-Fake-Judgements-and-the-Crisis-of-Truth-in-Indian-Courts.pdf

[25] Alayba, A. M. (2025). Arabic natural language processing (NLP): A comprehensive review of challenges, techniques, and emerging trends. *Computers*, *14*(11), 497. https://doi.org/10.3390/computers14110497

[26] Matrane, Y., Benabbou, F., & Sael, N. (2023). A systematic literature review of Arabic dialect sentiment analysis. *Journal of King Saud University – Computer and Information Sciences*, *35*(6), 101570. https://doi.org/10.1016/j.jksuci.2023.101570

[27] Goldfarb, R. L. (1964). Mellinkoff: The language of the law. *Michigan Law Review*, *63*(1), 180–186. https://repository.law.umich.edu/mlr/vol63/iss1/11

[28] Alkhowaiter, M., Alshahrani, N., Alshahrani, S., Masoud, R. I., Alzahrani, A., Alnuhait, D., . . . , & Almubarak, K. (2025). Mind the gap: A review of Arabic post-training datasets and their limitations. In *Proceedings of the Third Arabic Natural Language Processing Conference*, 323–337. https://doi.org/10.18653/v1/2025.arabicnlp-main.26

[29] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. https://doi.org/10.18653/v1/2021.findings-emnlp.320

[30] European Parliament, & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC). (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. Official Journal of the European Union. No 300/2008, (EU) No 167/2013, (EU) No 168/2013. http://data.europa.eu/eli/reg/2024/1689/oj

[31] Calo, R. (2018). Artificial intelligence policy: A primer and roadmap. *University of Bologna Law Review*, *3*(2), 180–218.

[32] Darwish, K., & Magdy, W. (2014). Arabic information retrieval. *Foundations and Trends in Information Retrieval*, *7*(4), 239–342. https://doi.org/10.1561/1500000031

How to Cite: Fawzy, H. M. (2026). Legal Hallucinations in Arabic Legal NLP: An Empirical Study Toward a Diagnostic and Mitigation Framework. *Journal of Computational Law and Legal Technology*. https://doi.org/10.47852/bonviewJCLLT620210837

Appendix A

Arabic Prompts Used in the Empirical Evaluation

Prompt	Arabic Text
T01	هل يجوز للقضاء الإداري أن يحيل مسألة جنائية مباشرة إلى محاكم الجنايات.
T02	ابحث عن أحكام نقض في النظام العام الاقتصادي.
T03	ابحث عن القانون الحاكم، الأساس القانوني والأحكام القضائية المتعلقة بالطعن أمام القضاء الإداري على قرارات الإزالة الصادرة عن الجهات الإدارية.
T04	ما الأساس الدستوري والأحكام القضائية المتعلقة برفض تعيين شخص في وظيفة عامة بسبب التمييز على أساس الجنس
T05	هناك نزاع بين مانح الامتياز وصاحب الامتياز بشأن استغلال العلامة التجارية (عقد الامتياز التجاري) ورفض تجديد العقد، مع تحديد القانون المصري الواجب التطبيق والمحكمة المختصة.