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.

Evaluating Artificial Intelligence Tools for Forensic-Linguistic Analysis of Arabic Judicial Texts: A Layer-Based Comparative Case Study of Saudi Judgments

Bunder Alshammari

Ban.Alshammari@uoh.edu.sa

University of Ha'il

Khaled Nasser Alfraidi

University of Ha'il

Mohammad Omar Aljudaiey

University of Ha'il

Walid Abdelhalim

Beni-Suef University

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10959403/v1

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

Additional Declarations: No competing interests reported.

Evaluating Artificial Intelligence Tools for Forensic-Linguistic Analysis of Arabic Judicial Texts: A Layer-Based Comparative Case Study of Saudi Judgments

Bunder Sebail Alshammari[1*], Khaled Nasser Alfraidi[1], Mohammad Omar Aljudaiey[1], and Walid Abdelhalim[2]

[1] Department of Arabic Language, College of Arts and Letters, University of Hail, Hail, Saudi Arabia

[2] Faculty of Education, Beni-Suef University, Beni-Suef, Egypt

Corresponding author: Bunder Sebail Alshammari, Department of Arabic Language, College of Arts and Letters, University of Hail, Hail, Saudi Arabia. Email: ban.alshammari@uoh.edu.sa. ORCID: 0009-0003-2111-2968.

Khaled Nasser Alfraidi. Email: k.alfraidi@uoh.edu.sa. ORCID: 0009-0002-5853-6142.

Mohammad Omar Aljudaiey. Email: m.aljedery@uoh.edu.sa. ORCID: 0009-0009-5618-5327.

Walid Abdelhalim. Email: walid0118@edu.bsu.edu.eg. ORCID: 0000-0001-9492-8245.

Abstract This study evaluates GPT-4o (Omni; version 2024-05-13), supported by Arabic preprocessing with CAMeL Tools, against expert forensic-linguistic annotation of Saudi judicial judgments. Using a comparative evaluation design, the study analyzed 10 Saudi criminal judgments from Majmūʿat al-Aḥkām al-Qaḍāʾiyya (Ministry of Justice, 1435 AH/2013-2014 CE), from which 24 representative segments involving psychiatric evidence, legal capacity, and criminal responsibility were extracted. Each segment was coded across the five P-DIM layers, producing 120 segmentby-layer classification decisions. Two expert coders established the adjudicated benchmark after pre-adjudication reliability testing (88% agreement; pooled Cohen's kappa =.82). The AI arm used deterministic prompting (temperature = 0.0), a system-level forensic-linguistic persona, the full P- DIM codebook, and three few-shot gold-standard examples per layer. Results showed strong AI performance for formulaic and structurally explicit layers, especially legal formulae and institutional weighting, but weak performance for pragmatic-doctrinal interpretation. Layer 4 produced the main pragmatic bottleneck (F1 =.41; kappa =.28). The corrected fold-level ANOVA used 20 F1 observations (four validation folds for each of the five P-DIM layers) and showed a significant layer effect, $F(4, 15) = 63.04$, $p = 3.367e-09$, $\eta^2 = .944$. Tukey HSD confirmed that L4 was significantly lower than all other layers, while also showing additional significant non-L4 contrasts. Overall AI-expert agreement was fair (kappa =.275, N = 120), reflecting aggregate instability caused mainly by Layer 4. The study therefore supports a human-in-the-loop verification protocol rather than autonomous AI interpretation in Arabic forensic-linguistic analysis. Keywords: Arabic forensic linguistics; Saudi judicial judgments; AI-assisted annotation; expert coding; legal NLP evaluation; pragmatic-doctrinal inference; diglossia

1. Introduction

Forensic linguistics applies linguistic theory and analytical methods to questions of legal evidence, interpretation, authorship, and institutional communication (Coulthard & Johnson, 2007; Gibbons, 2003). In recent years, the field has faced a practical and methodological pressure point: the growing scale of digitized legal records, electronically mediated evidence, and computationally searchable discourse has made purely manual close-reading approaches difficult to sustain for large corpora, while fully automated approaches often struggle with context-sensitive interpretation (Casey, 2011). As language technologies become increasingly embedded in legal and institutional workflows, forensic analysts must determine which components of linguistic

analysis can be delegated to computational tools and which require expert, theory-guided inference.

This problem is especially acute in Arabic judicial discourse. Arabic legal texts combine highly conventionalized formulae with dense institutional reasoning and intertextual references to doctrinal principles. In addition, Arabic diglossia complicates computational processing because judgments are typically drafted in Modern Standard Arabic, whereas quoted testimony and reported speech may contain dialectal features, code-switching, loan forms, and register shifts (Ferguson, 1959; Myers-Scotton, 1993; Poplack, 1980). These features matter for forensic analysis because they can index stance, alignment, credibility, hesitation, or pragmatic force in ways that are not reducible to keyword matching. Accordingly, Arabic forensic linguistics requires analytic models that integrate lexical and morphosyntactic cues with discourse-pragmatic interpretation and the institutional logic of legal reasoning (Tiersma, 1999; Wright & Picornell, 2024).

A particularly high-stakes domain concerns cases where psychiatric evidence intersects with criminal responsibility and legal capacity. In such cases, courts must evaluate how clinical reports, lay narratives, and textualized behavioral descriptions jointly support or weaken attributions of capacity and culpability (Appelbaum, 2006; Ferguson & Ogloff, 2011). The legal construction of responsibility and capacity, including concepts such as ahliyya, is rarely expressed through a single marker; rather, it is built through temporal framing, evidential weighting, stancetaking, and certainty grading across extended stretches of judicial reasoning. This makes the domain a demanding testbed for AI-supported analysis: tools that perform well on formulaic extraction may still fail when the interpretive task requires linking linguistic choices to argumentative aims, doctrinal warrants, or culturally embedded legal reasoning (Goźdź-Roszkowski, 2021, 2024; Mazzi, 2010).

At the same time, recent AI evaluation work has shown that legal and Arabic language tasks cannot be judged through a single global accuracy score. LegalBench separates types of legal reasoning rather than treating legal competence as monolithic (Guha et al., 2023), while LegalBench-RAG isolates retrieval quality in legal-domain RAG pipelines (Pipitone & Alami, 2024). Arabic-specific benchmarks similarly show that LLM performance is task-sensitive and language-sensitive, with LAraBench comparing LLMs against task-specific Arabic systems across many Arabic NLP and speech-processing tasks (Abdelali et al., 2024), ArabLegalEval focusing on Arabic legal knowledge (Hijazi et al., 2024), and ALARB extending evaluation to multistep Saudi legal argument reasoning (Abu Shairah et al., 2025). These studies justify a layer-based

design for the present work: the relevant question is not whether AI can process Arabic judicial text in general, but which levels of forensic-legal meaning it can classify reliably.

Against this backdrop, the present study addresses a focused and auditable problem: how AI-assisted classification can be evaluated across linguistically and doctrinally different layers of Arabic judicial discourse. Rather than treating AI accuracy as a single global score, the study operationalizes the Pragmatic-Doctrinal Inference Model (P-DIM) as a diagnostic evaluation protocol that separates comparatively surface-stable cues from higher-inference pragmatic and doctrinal judgments. This framing supports reproducible AI evaluation in legal-discourse settings while acknowledging that the corpus is a case-study corpus rather than a general benchmark for all Arabic legal NLP.

The study offers a calibrative and methodological contribution. Rather than proposing a new AI architecture or a domain-specific trained model, it provides a layer-based evaluation protocol for identifying where ready-to-use LLM outputs align with, or diverge from, expert forensic annotation in a Sharia-informed Arabic legal context. Its central value lies in showing how model performance changes when the same Arabic judicial passages are evaluated across explicitly defined layers and in specifying which layers require expert verification before AI-assisted labels can be used in forensic-linguistic interpretation.

2. Literature Review

The academic trajectory of forensic linguistics has evolved from the foundational study of "language in evidence" (Coulthard & Johnson, 2007) to a multidisciplinary examination of the justice system as a whole (Gibbons, 2003). Current scholarship emphasizes that forensic linguistics is not merely an auxiliary tool for legal inquiry but an epistemological framework for examining the complex intersections of discourse, evidence, authority, and institutional decision-making (Heydon, 2019). As legal processes increasingly rely on the analysis of legal discourse to interpret confessions, evidential weight, credibility, and institutional stance (Gallego Balcells, 2023), there is a pressing need for a semiotic perspective that unifies the language of the legal process with the language found in evidence (Wright & Picornell, 2024). This study builds upon these theoretical foundations by testing the efficacy of the Pragmatic-Doctrinal Inference Model (P-DIM) in the context of contemporary Arabic judicial digital transformation.

2.1 Forensic Linguistics: From "Language in Evidence" to Institutional Discourse Analysis

Forensic linguistics developed initially as a systematic application of linguistic analysis to evidential texts and legal problems, including authorship attribution, police interviewing, courtroom interaction, and the interpretation of contested meanings (Coulthard & Johnson, 2007; Gibbons, 2003). Contemporary work, however, increasingly treats the justice system itself as a discourse-producing institution in which authority is enacted through recurrent genres, including judgments, indictments, forensic reports, expert testimony, and written records of oral proceedings (Heydon, 2019). This shift is consequential because it reframes forensic linguistics from a set of techniques into an epistemological approach to institutional meaning-making, one that is attentive to evidential hierarchies, procedural constraints, credibility assessments, and the rhetorical management of doubt and certainty.

Within this broadened orientation, scholars have emphasized the need to align micro-level linguistic features with macro-level institutional functions, particularly in texts where confessions, denials, psychiatric evidence, and evidential weighting are central (Gallego Balcells, 2023). A key development in this respect is the semiotic perspective that explicitly links the language of the legal process, that is, how institutions formulate and record legal reasoning, with the language in evidence, namely the linguistic behavior attributed to parties, witnesses, defendants, and experts (Wright & Picornell, 2024). This distinction is directly relevant to AI-supported forensic analysis because it clarifies what counts as a "feature" in legal discourse: not merely tokens, phrases, or keywords, but also institutionalized forms of stance-taking, evaluative framing, evidential sequencing, and genre-bound formulaicity.

The expansion of digital and institutional discourse has further widened the evidential field of forensic linguistics. The key methodological problem across these domains is the gap between pattern recognition and meaning reconstruction: language technologies may identify recurring forms, but forensic interpretation requires determining how such forms function within social, institutional, and evidential contexts. This distinction is particularly important when computational outputs are used to support interpretation of judgments, expert reports, and records of attributed speech.

2.2 Arabic Judicial Discourse: Diglossia, Formulaicity, and Doctrinal Intertextuality

Arabic legal and judicial discourse presents a high-density environment for forensic analysis because it combines formulaic legal expressions, institutional evidential ranking, and

intertextual links to doctrinal principles. Work on legal language and psycholinguistic accessibility has long shown that legal texts maintain authority partly by relying on specialized formulae and stable constructions that distinguish professional legal discourse from lay understanding (Charrow & Charrow, 1979; Tiersma, 1999). In Arabic judicial writing, this separation is amplified by the co-presence of Modern Standard Arabic as the primary drafting register and dialectal, colloquial, or code-switched speech attributed to defendants, witnesses, or litigants, an arrangement classically described as diglossic (Ferguson, 1959). Because reported speech may be normalized, paraphrased, or re-entextualized in the written record, the stylistic imprint of speakers can be partially filtered through institutional drafting practices. As a result, forensic analysis must remain sensitive to the distinction between speaker-originating linguistic cues and institutionally produced cues introduced through judgment-writing conventions.

The relevance of code-switching and lexical variation is supported by broader studies of multilingual discourse. Alternation between languages, dialects, and registers may perform social, affective, and pragmatic functions rather than simply reflecting linguistic deficiency or casual mixing (Jake et al., 2002; Myers-Scotton, 1993; Poplack, 1980). In Arabic legal contexts, such variation is especially consequential because judicial records may contain traces of dialectal speech, institutional paraphrase, specialized legal Arabic, and translated or borrowed technical terminology. These features complicate automated processing because meaning may reside not only in the lexical item itself but in the relation between register, speaker identity, institutional mediation, and evidential function.

Arabic judicial discourse also frequently encodes reasoning through culturally and doctrinally saturated warrants, especially where Sharia-based maxims and legal principles function as implicit inferential bridges. This has been described as an "engineering of meaning," whereby judgments do not simply summarize evidence but construct a legally coherent narrative by aligning propositions with institutional and doctrinal standards (Saudi & Qabayli, 2023). From a pragmatic standpoint, judicial texts operate as argumentative systems: they justify legal outcomes by sequencing evidential claims, rebuttals, mitigations, and evaluations that manage competing versions of events (Hussein Hamadi, 2022). This matters directly for automated analysis because pragmatic functions such as mitigation, rebuttal, hedging, risk framing, and credibility grading are often realized through dispersed cues rather than explicit labels. Automated systems may therefore detect a relevant lexical signal while failing to identify the doctrinal warrant that gives that signal legal force.

A domain that illustrates these complexities with particular clarity is the intersection of mental-health evidence and criminal responsibility. In such cases, courts must integrate clinical expertise with legal concepts of capacity and culpability, producing judgments where psychiatric reports are assigned institutional weight and then translated into legally relevant assessments of responsibility (Appelbaum, 2006; Ferguson & Ogloff, 2011). The social implications of mentalillness discourse, especially where public safety, risk, and dangerousness are invoked, also shape how courts frame the defendant and justify social control, which is often reflected in evaluative lexis and narrative selection (Markowitz, 2011; Nee & Witt, 2013). Linguistically, this domain relies on temporal framing, stance-taking, and evidential grading: courts must determine whether a mental condition preceded the offence, coincided with it, or appeared afterward, and then assess whether that temporal relation is legally relevant to responsibility.

Work on judicial evaluation has shown that courts recurrently employ patterns of epistemic marking and evaluation to signal whether an inference is treated as robust, disputed, contingent, or insufficiently warranted (Goźdź-Roszkowski & Hunston, 2016; Mazzi, 2010; Szczyrbak, 2014). These insights motivate the need for models that do not treat forensic features as a single layer, but instead separate formulaicity, temporal reasoning, institutional weighting, pragmatic function, and evidential stance into analytically distinct dimensions.

2.3 AI and Forensic Feature Extraction: From Surface-Stable Cues to Context-Dependent Inference

The expansion of computational approaches to language analysis has created strong incentives to integrate AI into forensic workflows, particularly for large corpora and time-sensitive tasks. The human-machine framing emphasizes that emerging language technologies can process volumes of text beyond the reach of purely manual analysis, enabling rapid scanning for recurring formulae, evaluative patterns, and stylistic regularities (Sayers et al., 2021). At the same time, Arabic remains challenging for many NLP pipelines because of morphological complexity, orthographic variation, dialectal diversity, and resource scarcity, all of which complicate tokenization, disambiguation, normalization, and robust generalization across domains (Habash, 2010; Obeid et al., 2020). Consequently, AI performance in Arabic forensic settings depends not only on model capacity but also on preprocessing decisions, domain adaptation, dialect sensitivity, and the stability of the target cues.

Existing work suggests that automated systems can be effective when target features are surface-stable and strongly lexicalized, such as standardized formulae, recurring legal

constructions, and overt stance markers. Corpus-based approaches to legal discourse demonstrate that many evaluative and argumentative moves in judgments correlate with recurring lexicogrammatical configurations, which are well suited to computational discovery (Goźdź-Roszkowski, 2021, 2024). However, a key limitation remains the transition from extracting candidate features to producing valid interpretive inferences, especially where doctrinal warrants, institutional stance, and intertextual reasoning are required. In these contexts, a tool may correctly detect an epistemic verb, a legal formula, or a reference to silence while still failing to interpret its argumentative role or doctrinal implication, producing outputs that appear linguistically plausible yet are institutionally misleading.

This limitation is particularly important in light of recent legal-LLM evaluation research. LegalBench shows that legal reasoning should be decomposed into task families rather than treated as a single capability (Guha et al., 2023). LegalBench-RAG further demonstrates that legal AI evaluation may need to isolate retrieval quality from generation quality (Pipitone & Alami, 2024). For Arabic, LAraBench reports that LLMs do not uniformly outperform task-specific systems in zero-shot Arabic tasks and that few-shot prompting narrows, but does not eliminate, gaps with specialized approaches (Abdelali et al., 2024). ArabLegalEval and ALARB are especially relevant because they focus on Arabic/Saudi legal knowledge and argument reasoning, yet they evaluate tasks such as legal knowledge, article identification, verdict generation, and reasoning-chain completion rather than forensic-pragmatic annotation of judicial passages (Hijazi et al., 2024; Abu Shairah et al., 2025).

Accordingly, AI deployment in forensic linguistic analysis must be coupled with explicit verification practices rather than treated as a neutral automation of analysis. Recent work on legal hallucinations shows that LLMs can generate fluent but legally inaccurate outputs and that factuality remains a high-stakes challenge in legal question answering (Dahl et al., 2024; Hu et al., 2025). Within forensic linguistics, this means that AI-generated labels should be treated as provisional candidates that must be validated against pragmatic, doctrinal, and evidential criteria rather than accepted as determinations. The question is therefore not whether AI can identify recurring Arabic legal markers, but whether it can correctly interpret the institutional and doctrinal functions those markers perform in judicial reasoning.

To strengthen the AI-specific framing, the present study is situated in work on Arabic NLP resources, legal NLP evaluation, foundation models, and annotation transparency. Arabic preprocessing decisions are consequential because tokenization, morphological disambiguation,

dialectal variation, and orthographic normalization can change downstream labels (Habash, 2010; Obeid et al., 2020). Transformer-based Arabic models such as AraBERT and broader Arabic pretrained model evaluations have improved representational coverage, but their performance remains task- and domain-sensitive (Antoun et al., 2020; Inoue et al., 2021). Legal NLP benchmarks such as LexGLUE, LegalBench, LegalBench-RAG, ArabLegalEval, and ALARB show that legallanguage evaluation must report task definitions, units of analysis, prompt or model configuration, data limitations, and reproducible evidence rather than relying on broad claims about model competence (Chalkidis et al., 2022; Guha et al., 2023; Pipitone & Alami, 2024; Hijazi et al., 2024; Abu Shairah et al., 2025). The present study follows this evaluation logic by making the unit of analysis, codebook, prompt structure, fold-level values, and error categories explicit, while retaining human annotation as the validity anchor (Pustejovsky & Stubbs, 2012).

2.4 Textualized Demeanour and Silence: What Can Be Analyzed Computationally

An additional challenge for forensic analysis, and a frequent source of conceptual confusion in AI studies, concerns non-verbal and demeanour evidence. Computational tools cannot detect non-verbal behavior directly from written judgments; what they can analyze are textualized references to non-verbal conduct, such as the judge's written description of silence, gaze, posture, hesitation, emotional display, or failure to respond. Research in evidence and credibility assessment has shown that written judgments sometimes record such cues and treat them as relevant to evaluating credibility, demeanor, or stance (Denault, 2024; Denault & Bozin, 2024). From a semiotic standpoint, these textualizations function as part of the institutional meaningmaking system: they are not neutral observations but linguistically framed selections that can strengthen or weaken interpretations of testimony (Wright & Picornell, 2024).

In Sharia-informed adjudication, the interpretive status of silence can be doctrinally consequential, particularly where legal maxims are invoked to treat silence under specific conditions as communicatively meaningful (Al-Qahtani, 2021). For AI-supported analysis, this implies a clear boundary: systems can be trained or prompted to identify textual signals that reference silence or demeanor, but the legal force of these references depends on doctrinal context and argumentative positioning within the judgment. Thus, the computational task is best conceptualized as extracting candidate segments where demeanour is textualized and classifying their function under an explicitly defined analytic scheme, rather than inferring credibility directly.

This distinction also helps explain why general discourse-pattern studies are relevant but insufficient for forensic interpretation. A hesitation marker, silence reference, dialectal phrase, or

stance expression in a judgment is not merely a discourse feature; it is a textualized evidential object whose legal significance depends on procedural context, doctrinal warrant, and judicial uptake. This is why the present study evaluates AI at the level of layer-specific forensic function rather than only at the level of lexical detection.

2.5 Research Gap and Rationale for a Layer-Based Benchmark

Across the above strands, a consistent methodological gap emerges: studies often discuss AI effectiveness in broad terms without separating the distinct interpretive demands of legal language. In Arabic judicial discourse, some features are comparatively extractable, including stable legal formulae, institutional source labels, and overt epistemic markers, while others require context-sensitive inference, including pragmatic aims, doctrinal warrants, dialect-mediated stance, and intertextual legal reasoning. This distinction motivates the present study's adoption of the Pragmatic-Doctrinal Inference Model (P-DIM) as a multilayer coding framework. By benchmarking AI-assisted extraction against expert coding across separate layers—institutional weighting, temporal framing, formulaicity, pragmatic function, and evidential stance—the study aims to specify where AI aligns with forensic interpretation and why misalignment occurs.

The rationale for this layer-based approach is reinforced by three bodies of prior work. First, forensic and legal linguistic scholarship shows that legal meaning is produced through genre, stance, evaluation, and institutional authority rather than through lexical content alone (Coulthard & Johnson, 2007; Goźdź-Roszkowski, 2021, 2024; Tiersma, 1999; Wright & Picornell, 2024). Second, Arabic linguistic research shows that diglossia, code-switching, and register variation can affect how stance and identity are encoded and interpreted (Ferguson, 1959; Jake et al., 2002; Myers-Scotton, 1993; Poplack, 1980). Third, legal NLP and Arabic LLM evaluation literature shows that models must be assessed through transparent task definitions, reproducible annotation schemes, and explicit error analysis rather than through broad claims of accuracy or competence (Bender & Friedman, 2018; Chalkidis et al., 2022; Guha et al., 2023; Abdelali et al., 2024; Hijazi et al., 2024; Abu Shairah et al., 2025; Pustejovsky & Stubbs, 2012).

In doing so, the present study advances a practical, testable approach to integrating AI into Arabic forensic linguistics. AI can serve as a high-speed filter for surface-stable cues, but expert verification remains necessary for layers where meaning depends on doctrinal relevance, pragmatic argumentation, register-sensitive stance, and legally consequential inference. The purpose of P-DIM is therefore not to present a universal legal NLP benchmark, but to offer a

transparent, domain-specific evaluation protocol that can be adapted and externally validated in future studies across other Arabic legal corpora, legal systems, and institutional genres.

3. Theoretical Framework: The Pragmatic-Doctrinal Inference Model (P-DIM)

P-DIM functions in this study as an author-developed operational framework synthesized from forensic pragmatics, legal-discourse analysis, Arabic diglossia research, and Islamic legalepistemic reasoning. Its purpose is evaluative: to make explicit which types of Arabic judicial meaning are comparatively extractable by AI and which require expert doctrinal-pragmatic interpretation. The framework should therefore be read as a domain-specific annotation protocol that requires further external validation in other legal systems, genres, and Arabic varieties.

The Pragmatic-Doctrinal Inference Model (P-DIM) is the theoretical framework guiding this study. It is designed to explain how Saudi judicial texts construct legal reality through a layered interaction between linguistic evidence, institutional reasoning, and Sharia-based doctrinal warrants. Building on the notion of "engineering of meaning" in Arabic judgments (Saudi & Qabayli, 2023), P-DIM treats judicial writing as an institutional genre that does not merely report facts but ranks voices, stabilizes interpretations, and justifies outcomes through patterned evaluative and inferential moves. In this sense, the model aligns with semiotic approaches that connect "language of the legal process" (institutional drafting and reasoning) with "language in evidence" (reported speech, attributed stance, and narrated conduct) (Wright & Picornell, 2024).

Unlike approaches limited to surface linguistic description (e.g., lexical counts or syntactic patterns), P-DIM integrates pragmatic inference with Islamic legal epistemology. It is therefore particularly appropriate for cases where psychiatric evidence intersects with responsibility and capacity (masʾūliyya, ahliyya), because such cases require the court to reconcile clinical discourse with doctrinally relevant legal categories and principles (Al-Dulaimi, 2024; Appelbaum, 2006). The model is also compatible with an AI-assisted workflow: it specifies which aspects of judicial discourse are likely to be extractable computationally (e.g., formulae and overt stance markers) and which require expert verification because they depend on contextual, doctrinal, and argumentative integration.

3.1 Foundations of the P-DIM

P-DIM is anchored in four interlocking components

Speech Act Theory. P-DIM treats courtroom discourse as performative and institutionally consequential: utterances do not merely describe states of affairs, but can also carry legal force when they are interpreted within procedural and doctrinal frameworks. In the Saudi context, legally relevant meanings may be inferred not only from explicit statements but also from textualized absences, including silence recorded in judgments. Such silences can acquire illocutionary force when interpreted through established legal maxims and institutional expectations (Al-Qahtani, 2021). Accordingly, the model distinguishes between silence as a narrative gap and silence as a pragmatically and doctrinally meaningful act within legal reasoning.

Politeness, Facework, and Stance. P-DIM also incorporates facework-based reasoning to explain how deference, mitigation, self-presentation, and alignment are constructed in institutional settings. In written judgments, courts may textualize demeanour (e.g., hesitation, lowered head, emotional display) and treat it as evidence relevant to credibility or moral evaluation (Denault, 2024; Denault & Bozin, 2024). P-DIM therefore treats such descriptions as semiotically meaningful components of judicial stance rather than as peripheral narrative details.

Relevance Theory and inferential filtering. Because judicial reasoning often involves ambiguity, competing narratives, and uneven evidential quality, P-DIM models how judges select interpretations that best satisfy institutional needs for coherence and legal justification. This includes interpreting temporal framing (e.g., pre-offence chronicity versus an acute episode), aligning psychiatric claims with legal thresholds, and deciding which propositions are sufficiently supported to warrant doctrinal conclusions (Mazzi, 2010). In other words, P-DIM assumes that meaning in judgments emerges through inferential filtering, not through lexical triggers alone.

Islamic legal epistemology and doctrinal warrants. Finally, P-DIM explicitly integrates Sharia-based legal epistemology, including fiqh maxims and discretionary sentencing logic (ta'zir), as the warranting layer that authorizes inferences from language and conduct to legal responsibility. In this framework, determinations of ahliyya are not reducible to medical description; they are doctrinal inferences made by the court on the basis of evidence evaluation, stance grading, and legally recognized warrants (Al-Dulaimi, 2024; Yassin, 2002). This integration is essential for analyzing Saudi judgments because doctrinal relevance often explains why a feature matters and how it is transformed into an institutional conclusion.

3.2 The Multilayered Coding Scheme

To operationalize the framework for comparative analysis, P-DIM is implemented as a five-layer coding scheme that separates surface-stable features from context-dependent inference. The scheme functions as an "interpretive grid" against which AI outputs and expert coding can be compared transparently, layer by layer. The core assumption is that different layers impose different inferential burdens: some are largely extractable by pattern recognition, while others require doctrinal and pragmatic interpretation that must be verified by human experts.

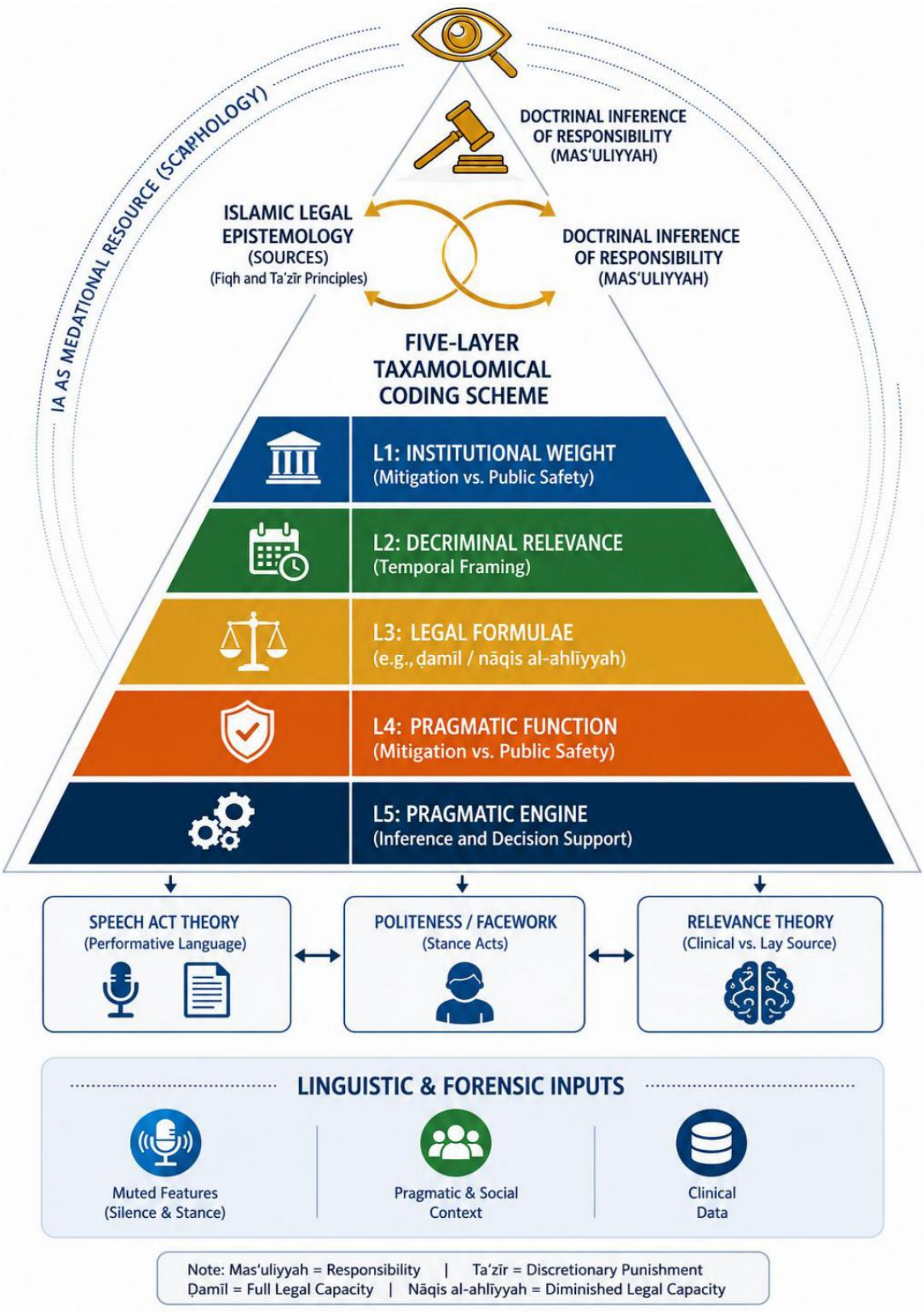

IA AS MEDATIONAL RESOURCE (SCAPHOLOGY)
DOCTRINAL INFERENCE OF RESPONSIBILITY (MAS'ULIYYAH)
ISLAMIC LEGAL EPISTEMOLOGY (SOURCES) (Fiqh and Ta'zīr Principles)
DOCTRINAL INFERENCE OF RESPONSIBILITY (MAS'ULIYYAH)
FIVE-LAYER TAXAMOLOMICAL CODING SCHEME
L1: INSTITUTIONAL WEIGHT (Mitigation vs. Public Safety)
L2: DECRIMINAL RELEVANCE (Temporal Framing)
L3: LEGAL FORMULAE (e.g., ḍamīl / nāqis al-ahlīyyah)
L4: PRAGMATIC FUNCTION (Mitigation vs. Public Safety)
L5: PRAGMATIC ENGINE (Inference and Decision Support)
SPEECH ACT THEORY (Performative Language)
POLITENESS / FACEWORK (Stance Acts)
RELEVANCE THEORY (Clinical vs. Lay Source)
LINGUISTIC & FORENSIC INPUTS
Muted Features (Silence & Stance)
Pragmatic & Social Context
Clinical Data
Note: Mas'uliyyah = Responsibility | Ta'zīr = Discretionary Punishment
Ḍamīl = Full Legal Capacity | Nāqis al-ahlīyyah = Diminished Legal Capacity

Figure 1. The Pragmatic-Doctrinal Inference Model (P-DIM): pragmatic engine, five-layer coding scheme, and

doctrinal oversight

Figure 1 summarizes the architecture of P-DIM, showing how the pragmatic engine feeds the five-layer coding scheme under doctrinal oversight. The integration of AI into Arabic forensic linguistics offers a practical response to long-standing constraints, particularly resource limitations, morphological complexity, and high levels of linguistic variation in Arabic NLP (Habash, 2010; Obeid et al., 2020; Abdelali et al., 2024). Yet the comparative logic underpinning P-DIM indicates that AI gains are uneven across interpretive layers. Automated tools are most dependable where judicial meaning is strongly lexicalized and genre-stable-especially in extracting standardized legal formulae (Layer 3) and identifying overt stance and certainty markers (Layer 5).

By contrast, AI reliability drops when analysis requires doctrinally warranted inference about why a segment is doing argumentative work (Layer 4), because pragmatic aims (e.g., mitigation, rebuttal, public-safety justification) are often realized through discourse positioning, intertextual warrants, and culturally specific legal reasoning rather than isolated keywords. For this reason, P-DIM formalizes a hybrid workflow in which AI supports scalable feature discovery while expert analysts retain responsibility for layer-specific verification and doctrinal synthesis.

This hybrid orientation is especially timely given the broader human-machine turn and the accelerating use of legal AI systems (Sayers et al., 2021; Guha et al., 2023; Pipitone & Alami, 2024). Manual discourse analysis remains essential for interpretive depth, but it cannot easily scale to expanding digital corpora without computational assistance. At the same time, the engineering of meaning in Arabic judgments depends on how courts integrate formulaic legal language with epistemic verbs, evaluative predicates, and Sharia-informed warrants to construct responsibility (mas'uliyya) and capacity (ahliyya) (Saudi & Qabayli, 2023; Al-Dulaimi, 2024; Yassin, 2002). P-DIM therefore treats AI not as an autonomous adjudicative interpreter but as a cognitive partner whose outputs must be routed through explicit verification routines, particularly for layers where doctrinal relevance and pragmatic intent are decisive. Within this framing, the contribution of the present study is to specify, empirically and procedurally, where AI meaningfully augments forensic work and where human interpretive authority remains indispensable. Research Questions (RQs) RQ1. How does AI-assisted analysis compare with expert manual coding in identifying and classifying forensic-linguistic features in Arabic judicial judgments? RQ2. How does AI accuracy

vary across the five P-DIM layers-institutional weight (L1), temporal framing (L2), legal formulae (L3), pragmatic function (L4), and stance/evidentiality (L5)? RQ3. To what extent do AI-extracted linguistic markers (e.g., formulaic expressions, evaluative lexis, epistemic verbs) align with the pragmatic aims and doctrinal warrants identified through manual qualitative analysis? RQ4. Which linguistic and doctrinal factors (e.g., diglossia, entextualized testimony, intertextual fiqh maxims) most constrain AI from producing a holistic, context-valid interpretation of mental-health framing in Sharia-based judgments? Research Hypotheses (Hs) H1. AI will perform most strongly on surface-stable features, particularly legal formulae (L3) and overt stance/evidentiality markers (L5). H2. AI performance will be significantly weaker for pragmatic-function classification (L4) than for other layers, especially in detecting discourse moves associated with mitigation, hedging, and public-safety justification in Sharia-informed reasoning. H3. When calibrated with domain-specific terminology and a fixed P-DIM codebook, GPT-4o outputs will show strong agreement with human coding for institutional weight (L1) and moderateto-strong agreement for temporal framing (L2), although temporal legal relevance may remain inference-sensitive. H4. Diglossia and doctrinal intertextuality (fiqh maxims functioning as warrants) will be primary drivers of AI-human divergence, reducing inferential alignment in higher-inference layers (especially L4). H5. The results are expected to indicate whether a hybrid workflow (AI extraction + expert verification for pragmatic/doctrinal inference) is methodologically necessary for forensic validity; this hypothesis is interpreted as an empirically motivated recommendation rather than as a directly tested third-arm comparison.

This study adopts a comparative case-study evaluation design. The purpose is not to develop a new AI architecture, but to test how a ready-to-use, instruction-tuned LLM-supported workflow behaves when its outputs are compared with expert coding across the five P-DIM layers. The design contains two parallel analytical arms applied to the same textual units: (a) expert manual coding, which provides the adjudicated benchmark, and (b) AI-assisted labeling, which produces provisional layer labels under constrained prompting. All comparisons are therefore made at the same analytical-unit level rather than at the whole-judgment level.

4.1 Design Overview

This study adopts a comparative case-study design to examine how GPT-4o-assisted analysis performs relative to expert manual coding when extracting and interpreting forensiclinguistic features in Arabic judicial judgments. Consistent with the P-DIM framework, the methodology separates surface-stable, genre-formulaic features from high-inference,

doctrinally warranted interpretations. The research is implemented as two parallel analytical arms, manual and AI-assisted, applied to the same corpus and evaluated through matched outcome metrics. The total sample size of N = 120 represents the number of unique classification decisions, derived from 24 distinct textual segments extracted from 10 judgments, with each segment independently coded across the five analytical layers of P-DIM (24 segments x 5 layers = 120 data points).

4.2 Materials: Corpus, Data Characteristics, and Case Selection

4.2.1 Data source

The corpus comprises criminal judgments drawn from Majmūʿat al-Aḥkām al-Qaḍāʾiyya (1435 AH), published by the Saudi Ministry of Justice (Wizārat al-ʿAdl, 1435 AH). This source was selected because it represents an institutionalized genre with recurrent formulae, explicit evidential ranking, and evaluative stance-taking typical of Saudi judicial reasoning.

The corpus consisted of 10 comprehensive criminal judgments, averaging approximately 1,500 words each. From these judgments, 24 representative segments ranging from approximately

50 to 150 words were purposively extracted on the basis of mental-health claims, capacity/responsibility markers, stance expressions, temporal illness framing, or psychiatricevidence references. Each of the 24 segments was evaluated across the five P-DIM layers, producing 120 unique classification points (24 x 5 = 120) for statistical comparison. The lexical screening procedure described in the appendices was used only to locate candidate judgments and passages; it was not itself the final sampling frame or the unit of statistical analysis.

Cases were included when the written judgment contained at least one of the following: (1) references to psychiatric assessment, such as a committee report or clinician statement; (2) narrative framing of mental state, including pre-offence chronicity, acute episode, or post-offence claims; (3) doctrinally relevant capacity markers, such as kamil al-ahliyya or naqis al-ahliyya; or (4) overt stance/evidentiality markers, such as tabayyana or yarjuh, used to grade certainty.

The final analytical unit was a bounded judicial passage: a clause, sentence, or short multisentence segment that expressed one coherent evidential, temporal, formulaic, pragmatic, or stance-related move. Each retained passage was evaluated across five P-DIM layers. Thus, N =

120 refers specifically to segment-by-layer coding decisions, not to 120 judgments, 120 pages, or

120 independent legal cases. This clarification separates three levels of data: 10 source judgments,

24 retained analytical segments, and 120 quantitative classification decisions.

For data accounting, the corpus comprised 10 source judgments, from which 24 analytical segments were retained. Because each segment was independently coded across the five P-DIM layers, the final dataset consisted of 120 segment-by-layer classification decisions.

4.2.3 Linguistic characteristics of the corpus

The corpus foregrounds a key methodological challenge: diglossia and entextualization. While judgments are drafted in Modern Standard Arabic, witness statements and narrated speech may reflect dialectal forms, paraphrase, or institutional reformulation, which can obscure "stylistic imprints" and complicate automated interpretation (Ferguson, 1959; Mahajna, 2019). This feature is analytically valuable because it stress-tests AI under realistic conditions of Arabic legal textuality.

4.3 Operational Framework: P-DIM Multilayered Coding Scheme

The study operationalizes the Pragmatic-Doctrinal Inference Model (P-DIM) as a fivelayer coding scheme. The goal is to ensure that both analytical arms (manual and AI) target the same units of analysis while allowing performance to be assessed layer-by-layer, rather than through undifferentiated "overall accuracy."

Layer 1: Institutional weight of psychiatric evidence. Codes the evidential source and its institutional authority (e.g., multi-member committee report vs. individual clinician statement vs. lay narrative), tracking how judgments rank psychiatric material (Ferguson & Ogloff, 2011).

Layer 2: Temporal framing and doctrinal relevance. Codes the temporal positioning of illness relative to the offense (pre-offence chronicity, acute episode during the act, post-offence claims), reflecting the legal relevance of temporal construal for culpability evaluation (Hart, 2013; Al-Dulaimi, 2024).

Layer 3: Legal formulae and capacity constructions. Extracts standardized legal expressions that encode capacity/responsibility determinations (e.g., kāmil al-ahliyya, nāqiṣ

alahliyya) and other recurrent formulaic anchors typical of judgment-writing (Goźdź-Roszkowski, 2021).

Layer 4: Pragmatic-argumentative function. Codes the argumentative aim of segments (e.g., mitigation, rebuttal, public-safety justification, hedging judicial commitment), following pragmatic analyses of Saudi judicial reasoning (Hussein Hamadi, 2022).

Layer 5: Stance, evidentiality, and certainty grading. Codes epistemic and evaluative markers used to grade certainty and warrant conclusions (e.g., tabayyana, yarjūḥ), including stance-taking and evidential patterns (Szczyrbak, 2014; Goźdź-Roszkowski, 2024).

Two trained forensic-linguistic coders conducted directed qualitative content analysis guided by the P-DIM codebook (Krippendorff, 2013; Pustejovsky & Stubbs, 2012). Coding was conducted in NVivo 14 as a CAQDAS environment for organizing passages, storing memos, maintaining an audit trail, and running matrix comparisons (Zamawe, 2015). The manual arm served as the adjudicated benchmark for evaluating AI outputs at the layer level.

Doctrinal interpretation was operationalized conservatively. A segment was not coded as doctrinally relevant merely because it contained religious or legal vocabulary; it had to connect a linguistic cue to a legal consequence, such as capacity, culpability, evidential sufficiency, mitigation, or risk. This rule was added to prevent the AI system from treating doctrinal terminology as a keyword trigger without verifying its argumentative role in the judgment.

The codebook specified clear decision rules for each layer. A passage was coded in Layer 1 when the text identified the evidential source and its institutional authority; in Layer 2 when it positioned illness relative to the act, investigation, or judgment; in Layer 3 when it contained a recurrent capacity/responsibility formula; in Layer 4 when the segment performed an argumentative function such as mitigation, rebuttal, risk framing, or warranting; and in Layer 5 when it graded certainty, evidential source, or stance. Borderline cases were adjudicated by asking whether the label depended primarily on lexical form (Layers 1, 3, and parts of 5), event construal (Layer 2), or doctrinal-pragmatic warranting (Layer 4).

Before adjudication, the two expert coders achieved an unadjusted agreement of 88% and a pooled Cohen's kappa of.82. Layer-specific human agreement was highest for L3 (Legal Formulae, kappa =.96) and lowest for L4 (Pragmatic Function, kappa =.68). This indicates that the pragmatic-doctrinal layer was inherently more interpretive even for trained human coders, but human agreement remained substantial and clearly higher than AI-expert agreement for the same layer.

4.4.1 Manual arm: expert qualitative coding (benchmark)

Reliability and adjudication. Human-human agreement was reported before adjudication, and AI-expert agreement was reported after the benchmark had been stabilized. Disagreements were most frequent in Layer 4 (Pragmatic Function) and were resolved through codebook-guided discussion and third-party consultation to establish the final benchmark dataset.

4.4.2 AI arm: AI-assisted feature extraction and classification

The AI arm used GPT-4o (Omni; version 2024-05-13) via the OpenAI API as a ready-touse, instruction-tuned model. No Arabic judicial model was trained, fine-tuned, or modified using the study corpus. The model was prompted using a system-level forensic-linguistic persona, the full P-DIM codebook, and three gold-standard examples of Arabic judicial reasoning per layer. Outputs were treated as zero-shot/few-shot classifications requiring expert audit, not as outputs of a domain-trained legal AI system.

(a) Arabic preprocessing (feature stabilization):

CAMeL Tools were used for Arabic tokenization, normalization, and morphological stabilization before prompting. This preprocessing reduced noise caused by Arabic morphology and orthographic variability while preserving dialectal and reporting cues when they were analytically relevant. AraBERT-supported representations were used only as linguistic support for Arabic feature awareness; no transformer model was fine-tuned on the study corpus.

(b) LLM-based layer extraction:

GPT-4o was prompted using the same P-DIM codebook and few-shot examples to perform:

Prompt template used for layer classification. “You are assisting with forensic-linguistic annotation of Arabic Saudi judicial text. Given the passage below, provide: (1) an English translation, (2) the applicable P-DIM layer(s), (3) the label selected from the codebook, (4) the exact Arabic cue supporting the label, (5) whether the label is lexical, temporal, pragmatic, doctrinal, or stance-based, and (6) a confidence level. Do not infer criminal responsibility or capacity unless the passage explicitly links the cue to a judicial or doctrinal conclusion. Return the answer in the fixed JSON-like schema: {passage_id, translation, layer, label, cue, rationale, confidence, verification_required}.”

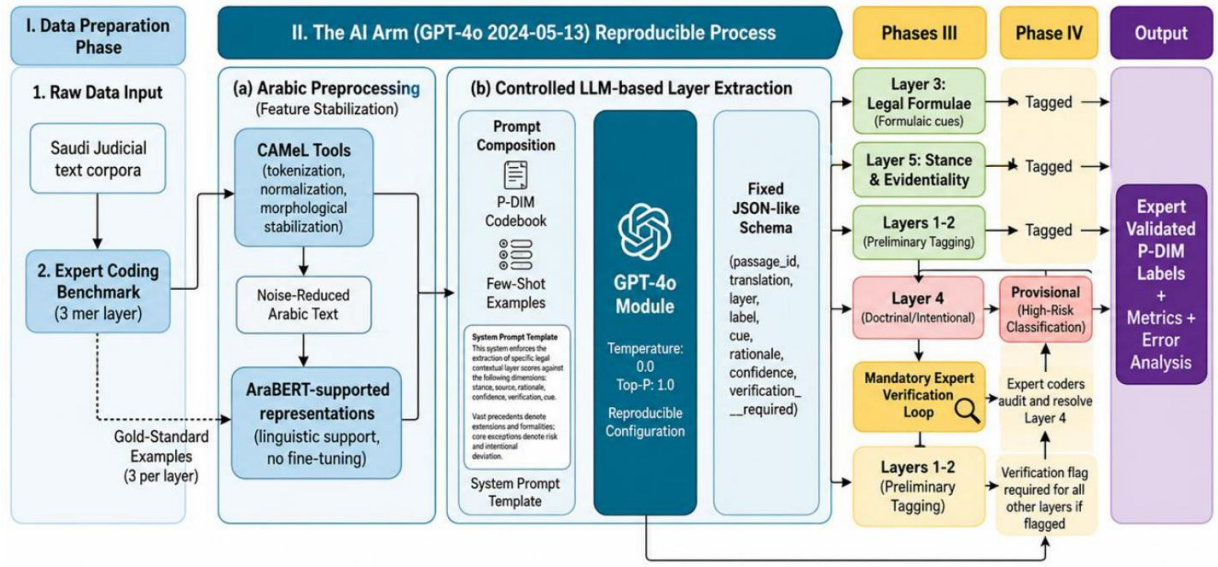


Figure 2. Reproducible AI-assisted P-DIM pipeline from raw Arabic judicial text to expert-validated labels

AI configuration and reproducibility. The AI arm utilized GPT-4o (Omni; version 2024-05-13) through the OpenAI API (OpenAI, 2024). To minimize stochastic variation and support reproducibility, temperature was set to 0.0, top_p was set to 1.0, and no fine-tuning or retrievalaugmented generation was used. The system prompt contained the full P-DIM codebook, and each layer was illustrated with three author-verified gold-standard examples. The output was constrained to a fixed JSON-like schema so that expert coders could audit the assigned layer, label, textual cue, rationale, and verification flag.

- Layer 3 extraction of standardized legal formulae;
- Layer 5 identification of stance/evidentiality markers;
- preliminary tagging for Layers 1-2; and
- provisional classification for Layer 4 (treated as high-risk and subject to mandatory verification due to its inferential nature).

Statistical strategy and assumptions. The one-way ANOVA was conducted on fold-level F1-scores rather than on the 120 raw classification decisions. The 24 retained segments were divided into four validation folds of six segments each. For each fold, F1 was calculated separately for each of the five P-DIM layers, producing 20 fold-level F1 observations (four fold values x five layers). These folds were used only to summarize performance stability and were not used to train, tune, or select the model. This explains why the ANOVA degrees of freedom are F(4, 15) despite the item-level comparison dataset containing N = 120 classification decisions. Because the dataset

is modest, the ANOVA is interpreted as an exploratory layer-sensitivity test, supplemented by descriptive metrics, confidence intervals, Tukey HSD post-hoc comparisons, the fold-level table and reproducibility script in the accompanying supplementary file, and qualitative error profiling.

The manual analysis was conducted in NVivo 14, which was used for coding, memo writing, and audit-trail management. The AI/NLP workflow used GPT-4o (version 2024-05-13; temperature = 0.0) with Arabic-specific preprocessing in CAMeL Tools; no additional training on Arabic judicial judgments was performed. Evaluation used precision, recall, F1, 95% confidence intervals for F1, Cohen's kappa, exploratory fold-level ANOVA, Tukey HSD post-hoc comparisons, and qualitative error profiling. Generic references to AI are avoided unless the tool class, model, and analytic role are specified.

In line with the study's theoretical stance, AI classifications-especially for Layer 4-were treated as provisional until verified against doctrinal warrants and discourse context. Verification was operationalized as a structured expert review protocol: coders checked the Arabic cue, English translation, assigned label, rationale, and verification flag against the codebook and against the surrounding judicial reasoning. This procedure prevented the model from converting a plausible lexical match into an unverified forensic conclusion.

4.5 Measures and Comparative Evaluation

AI performance was evaluated against the benchmark dataset using complementary quantitative and qualitative indices, selected to reflect the different epistemic demands of the P-DIM layers.

4.5.1 Quantitative performance metrics (RQ1-RQ2)

For each analytical layer, performance was assessed using precision, recall, and F1-score in order to quantify the accuracy of feature extraction and classification, with these measures being especially informative for Layers 1, 3, and 5. In addition, Cohen's kappa was calculated to estimate the level of agreement between the AI-generated outputs and the benchmark coding decisions, thereby allowing direct comparison with human inter-rater agreement.

4.5.2 Interpretive alignment (RQ3)

Because Layer 4 involves pragmatic inference and doctrinal warrants, its accuracy could not be meaningfully evaluated through lexical matching alone. Instead, interpretive alignment was assessed through a structured qualitative comparison between the markers identified by the AI and

the pragmatic aims and doctrinal warrants assigned by the manual coders to the same textual segment. Alignment was then determined according to whether the AI-generated label was adequately supported by the discourse structure and the doctrinal reasoning embedded in the judgment.

4.5.3 Doctrinal consistency checks (RQ4)

To determine where AI failed to generate context-valid interpretations, doctrinal consistency checks were conducted with particular attention to diglossic or entextualized segments, intertextual fiqh maxims functioning as legal warrants, and psychiatric-to-legal inference moves that linked clinical claims to questions of capacity and culpability. These checks made it possible to identify recurring error patterns, such as the literalist misclassification of legally meaningful silence as "missing data," and to map these interpretive constraints onto specific analytical layers.

4.6 Instruments and Software

The AI and NLP workflow relied on GPT-4o (version 2024-05-13) for classification and extraction tasks using the P-DIM codebook through zero-shot and few-shot prompting, alongside CAMeL Tools for Arabic tokenization and morphological stabilization. For evaluation, layer-wise performance was reported using precision, recall, F1, F1 confidence intervals, and Cohen's kappa, while ANOVA and Tukey HSD tests were used cautiously to compare fold-level F1 variation across layers.

To test whether AI performance differed across P-DIM layers, the analysis used a one-way ANOVA on fold-level F1-score summaries: four validation folds per layer across five layers. The ANOVA is therefore a comparison of layer-level performance stability, not an item-level model fitted to all 120 classification decisions. The corrected analysis showed a strong layer effect, $F(4, 15) = 63.04$, $p = 3.367e\text{-}09$, $\eta^2 = .944$, confirming that the drop from L3 (Formulae) to L4 (Pragmatic Function) was systematic rather than an artifact of a single difficult segment. The Tukey HSD table is reported pairwise because the corrected output shows both the substantively central L4 contrasts and additional significant non-L4 contrasts. The accompanying supplementary reproducibility file provides the 20 fold-level F1 values and the analysis script.

The key instrument in this study is the P-DIM coding manual itself. It functions as the shared specification that makes the two analytical arms commensurable: the manual coders and the AI system are constrained to the same definitions, indicators, and boundary rules for each layer.

This symmetry is essential for a defensible comparison because it ensures that observed differences reflect inferential limitations and linguistic constraints, not mismatched constructs or inconsistent operationalization.

4.7 Justification of Methodological Choices and Relation to Advanced Legal NLP Evaluation Several methodological choices require explicit justification. First, the study uses deterministic GPT-4o prompting rather than fine-tuning because the corpus is intentionally small, legally specialized, and designed for forensic audit rather than model training. Fine-tuning on 24 segments would risk overfitting and would not answer the central question of whether an accessible, readyto-use LLM workflow can be safely used as an annotation aid. Second, retrieval-augmented generation was not used because the target passages had already been selected and bounded; adding retrieval would confound classification accuracy with retrieval success. This differs from LegalBench-RAG, where retrieval itself is the evaluation target (Pipitone & Alami, 2024). Third, the comparison is layer-wise rather than model-ranking-oriented because the forensic question concerns where AI aligns with expert interpretation, not which model wins on a general legal benchmark. This makes the study complementary to LegalBench, ArabLegalEval, and ALARB, which evaluate broader legal reasoning, legal knowledge, and Arabic legal argument tasks (Guha et al., 2023; Hijazi et al., 2024; Abu Shairah et al., 2025). The experimental design therefore prioritizes controlled auditability over benchmark breadth. Each AI decision is tied to a segment ID, P-DIM layer, Arabic cue, English translation, codebook label, confidence flag, and verification status. This structure makes it possible to locate errors at the precise point where lexical detection fails to become legally valid pragmatic interpretation. The design is modest in scale but deliberately transparent: the fold-level F1 values and reproducible analysis script are supplied in the accompanying supplementary reproducibility file.

5. Results

The results are presented in three connected stages: layer-wise quantitative performance, aggregate AI-expert agreement, and qualitative error profiling. AI F1-scores differed significantly across P-DIM layers in the corrected fold-level ANOVA, $F(4, 15) = 63.04$, $p = 3.367e-09$, $\eta^2 = .944$; SS_between/SS_total = 0.7464/0.7908. The analysis used four validation folds per layer, and the

20 fold-level F1 values are reported in the accompanying supplementary reproducibility file. Tukey HSD post-hoc comparisons showed that L4 (Pragmatic Function) was significantly lower than all other layers. The same Tukey output also showed significant non-L4 contrasts for L1-L2, L2-L3, and L3-L5, while L1-L3, L1-L5, and L2-L5 were not significant. The post-hoc interpretation is therefore that L4 is the principal substantive bottleneck, not that it is the only statistically significant pairwise contrast.

5.1 Quantitative Results: Layer-Wise AI Performance and Agreement (RQ1-RQ2)

To address RQ1-RQ2, GPT-4o outputs were evaluated against the adjudicated expertcoded benchmark using precision, recall, F1, 95% confidence intervals for F1, and Cohen's kappa. The unit of comparison was the segment-by-layer classification decision (N = 120), while the ANOVA was conducted separately on fold-level F1 summaries (four folds per layer).

5.1.1 Performance by P-DIM layer

As summarized in Table 1, GPT-4o achieved its strongest performance for Layer 3 (Legal Formulae) and Layer 1 (Institutional Weight), with comparatively lower but still acceptable performance for Layer 5 (Stance/Evidentiality). In contrast, Layer 4 (Pragmatic Function) exhibited the weakest performance across all indices, indicating that automated processing struggled to move from surface lexical cues to warrant-based pragmatic interpretation. The 95% confidence intervals show that this weakness was not a marginal fluctuation: L4's F1 interval [0.34, 0.48] is clearly separated from the intervals for L1 and L3.

Explanation of the overall kappa. While agreement was almost perfect for structural layers (L1, kappa =.91; L3, kappa =.94), the overall Cohen's kappa for the total dataset (N = 120) was.275. This fair overall agreement is attributed to the pragmatic bottleneck in Layer 4 and the difficulty of multi-layer classification. The AI's consistent false negatives for doctrinal warrants and its tendency to assign descriptive rather than performative labels heavily depressed the aggregate metric, despite high accuracy in surface-level feature extraction.

To address RQ3 and RQ4, NVivo matrix comparisons were conducted between the adjudicated expert benchmark and AI-generated labels. Divergence segments were reviewed thematically after the quantitative comparison, then grouped into six recurrent error mechanisms.

Table 5 reports the percentage of total AI errors (n = 58) assigned to each error type. These figures are descriptive error-frequency summaries derived from the NVivo comparison, not independent prevalence estimates for the full corpus.

AI Performance Metrics across P-DIM Layers with F1 Confidence Intervals

P-DIM Layer	Precision	Recall	F1 (95% CI)	Cohen's κ (AI-Expert)
L1: Institutional Weight	0.89	0.92	0.90 [0.87, 0.93]	0.91
L2: Temporal Framing	0.76	0.74	0.75 [0.70, 0.80]	0.62
L3: Legal Formulae	0.97	0.95	0.96 [0.94, 0.98]	0.94
L4: Pragmatic Function	0.44	0.39	0.41 [0.34, 0.48]	0.28
L5: Stance/Evidentiality	0.82	0.85	0.83 [0.79, 0.87]	0.78

Note. Table 1 shows very high AI-expert agreement for Layer 1 (κ = .91) and Layer 3 (κ = .94); these values indicate almost perfect agreement rather than perfect agreement (κ = 1.00). Agreement is moderate for Layer 2 (κ = .62), substantial for Layer 5 (κ = .78), and poor for Layer 4 (κ = .28), consistent with the observed drop in performance for pragmatic-function coding.

5.1.2 Differences in AI performance across layers (H1-H2)

To test whether AI performance differed significantly across the five P-DIM layers, a oneway ANOVA was conducted on fold-level F1-scores (four validation folds per layer; n = 20 foldlevel observations). The corrected ANOVA showed a statistically significant effect of layer type on AI performance, $F(4, 15) = 63.04$, $p = 3.367e\text{-}09$, $\eta^2 = .944$. Tukey HSD post-hoc tests showed that L4 (Pragmatic Function) was significantly lower than all other layers. The full pairwise output further indicated significant non-L4 differences for L1-L2, L2-L3, and L3-L5; non-significant contrasts were L1-L3, L1-L5, and L2-L5. The fold-level F1 values and the exact analysis script are provided in the accompanying supplementary reproducibility file.

Table 2

One-way ANOVA of AI F1-Scores across P-DIM Layers

Source	SS	df	MS	F	p	η^2
Between layers	0.7464	4	0.1866	63.04	< .001	.944
Within layers (error)	0.0444	15	0.0030	—	—	—
Total	0.7908	19	—	—	—	—

Note. The analysis was conducted on fold-level F1 scores, using four validation folds for each of the five P-DIM layers. AI performance differed significantly across layers, $F(4, 15) = 63.04$, $p < .001$, $\eta^2 = .944$, indicating a large layer effect. The exact probability value was $p = 3.367 \times 10^{-9}$. Eta squared was calculated as SS(*{between})/SS(*{total}) = 0.7464/0.7908. The supplementary reproducibility file reports all 20 fold-level F1 observations used to compute this table.

Table 3

Tukey HSD Pairwise Comparisons of Fold-Level F1 Scores across P-DIM Layers

Comparison	Mean diff. (group2 - group1)	p_adj	95% CI	Significant
L1 vs L2	-0.15	.0106	[-0.2688, -0.0312]	Yes
L1 vs L3	0.06	.5428	[-0.0588, 0.1788]	No
L1 vs L4	-0.49	< .001	[-0.6088, -0.3712]	Yes
L1 vs L5	-0.07	.3988	[-0.1888, 0.0488]	No
L2 vs L3	0.21	.0005	[0.0912, 0.3288]	Yes
L2 vs L4	-0.34	< .001	[-0.4588, -0.2212]	Yes
L2 vs L5	0.08	.2783	[-0.0388, 0.1988]	No
L3 vs L4	-0.55	< .001	[-0.6688, -0.4312]	Yes
L3 vs L5	-0.13	.0289	[-0.2488, -0.0112]	Yes
L4 vs L5	0.42	< .001	[0.3012, 0.5388]	Yes

Note. Mean differences are reported in the direction provided by the reproducibility workbook. The significant post-hoc pattern confirms L4 as the principal bottleneck while also retaining significant non-L4 contrasts for L1-L2, L2-L3, and L3-L5.

5.1.3 Overall AI-expert agreement across the full classification dataset

The aggregate agreement analysis between Expert_Score and AI_Score used the full N = 120 classification-decision dataset and yielded Cohen's kappa =.275 (Table 4). This value should not be interpreted as contradicting the high layer-specific kappas for L1 and L3. Rather, it shows that when all layers are pooled, the AI's repeated misclassification of pragmatic-doctrinal categories, especially Layer 4, substantially reduces overall agreement.

Three bilingual examples illustrate the interpretive problem. In the stance/certainty example, 'ويظهر للدائرة يحتمل أن يكون المتهم في حالة غير طبيعية' ('It appears to the court panel that the accused may have been in an abnormal state'), the hedge يحتمل ('may be') must be preserved because it lowers certainty. In the diglossia example, 'كان يهرج لحاله' ('he was talking or rambling to himself'), the Saudi/Najdi dialectal verb yahruj (يهرج) may indicate ordinary talking, incoherent talk, or a lay description of hallucination; when normalized as MSA-related joking or generic speech, the stance and mental-state implication is weakened. In the doctrinal-warrant example, 'السكوت في موضع الحاجة بيان' ('silence where clarification is required counts as a cogent statement'), the legal force depends on a jurisprudential warrant rather than the lexical item silence alone.

Table 4

Overall Cohen's κ for Expert_Score vs. AI_Score (N = 120)

Statistic	Value
Cohen's κ	0.275
N (classification decisions)	120 (24 segments x 5 layers)

Note. Overall agreement between expert scores and AI scores across the scored dataset was fair/low (Cohen's kappa = .275, N = 120; 24 segments x 5 layers). The aggregate value is penalized by Layer 4 false negatives and by the difficulty of maintaining agreement across a multidimensional layered task.

5.2 Qualitative Results: NVivo-Based Comparison of Manual vs. AI Coding (RQ3-RQ4)

To address RQ3 and RQ4, we conducted NVivo-based comparisons between the expertcoded benchmark and AI-generated labels, focusing on where the two approaches converged and why they diverged at the level of warrant, discourse function, and evidential framing. Two recurrent divergence patterns were observed: (a) surface-trigger dependence in pragmatic-function coding (Layer 4) and (b) reduced stance sensitivity under diglossia and entextualization (Layer 5).

5.2.1 Lexical over-reliance as the dominant mechanism of Layer-4 error

Matrix comparisons indicated that most Layer-4 misclassifications were driven by GPT-4o over-weighting salient lexical triggers (e.g., tokens referring to 'silence,' 'confession,' 'report,' or 'committee') while under-recovering the doctrinally licensed pragmatic function that the judgment assigns to those cues. In legally consequential positions, expert coders treated textualized silence as a meaningful non-response whose force derives from Sharia-informed warranting practices. By contrast, AI labels frequently mapped the same segments to procedural absence,

missing information, or non-evidence, thereby failing to model illocutionary force as an institutional inference rather than a descriptive void. This explains the weak Layer-4 indices (Table 1) and aligns with the strong layer effect in Table 2.

A confusion-matrix review of Layer 4 errors showed how the pragmatic bottleneck operated. Approximately 64% of Layer 4 misclassifications occurred because the AI defaulted to a descriptive label, usually L1 or L2, when a performative pragmatic label was required. In cases of legally meaningful silence, the AI categorized the segment as missing data or administrative absence in 18 of 24 instances, whereas human experts identified a Sharia-based assent or warranting function when the surrounding legal context made silence consequential.

Layer 5 presents an intermediate pattern. Its substantial performance (kappa =.78) shows that AI could identify overt evidential and stance markers with reasonable success, yet the qualitative analysis demonstrates persistent distortion when stance was distributed across diglossic or entextualized discourse. These findings support a verification-based workflow implication: AI can assist with candidate identification, but expert review is required before stance labels are used for forensic interpretation.

For Layer 5, comparisons showed that GPT-4o generally detected explicit epistemic items and overt certainty markers when they appeared in standardized judicial phrasing. However, AI performance in Layer 5 was hampered by register-blindness. In Saudi judgments, witness testimony is often recorded in a hybrid of Modern Standard Arabic and regional dialect. The AI tended to normalize dialectal markers of hesitation, hedging, or lay mental-state description into standard formal Arabic, effectively bleaching the speaker's original stance and sometimes making the defendant or witness appear more certain, coherent, or aggressive than the original text suggested. This pattern supports the interpretation that diglossia and institutional entextualization constrain automated stance modeling even when surface markers are present.

Across strands, the evidence supports a functional division of labor as a practical implication rather than as a directly tested experimental superiority claim. AI performs strongly as a high-throughput extractor for layers with stable lexical anchors and predictable genre conventions. However, when the interpretive task requires recovering pragmatic aim and doctrinal warrant, especially in Layer 4, expert analysis remains essential. The results therefore suggest support for H5 only as a workflow implication: a hybrid workflow is logically necessitated to maintain forensic validity, but it was not directly tested as a separate experimental condition.

Table 5

Thematic Analysis of AI-Expert Divergences in Qualitative Comparison (NVivo)

Theme	Arabic cue + English translation	Interpretation	% of Total AI Errors (n = 58)
T1. Surface-trigger dependence (lexical over-reliance)	فسكت المدعى عليه... / "The defendant remained silent..."	AI over-relies on keywords and ignores the institutional context that makes silence legally meaningful.	32%
T2. Warrant failure (Sharia-based inferential bridging)	السكوت في موضع الحاجة بيان / "Silence where clarification is required counts as a statement."	AI labels the segment descriptively but misses the Sharia-based inferential bridge that gives the utterance legal force.	24%
T3. Entextualization distortion (reframed reported speech)	أفادت اللجنة الطبية أنه يعاني فصاماً / "The medical committee reported that he suffers from schizophrenia."	AI treats paraphrased institutional reporting as direct stance evidence instead of mediated judicial narration.	12%
T4. Diglossia sensitivity gap (dialect/register shifts)	كان يهرج لحاله / "He was talking/rambling to himself."	AI normalizes the Saudi/Najdi dialectal verb يهرج (yahruj) as generic talk or joking, weakening stance and mental-state cues.	15%
T5. Over-confident stance assignment	يحتمل أن يكون في حالة غير طبيعية / "He may have been in an abnormal state."	AI overestimates certainty when stance is hedged or distributed across discourse.	10%
T6. Temporal inference slippage (Layer 2)	الحالة سابقة للواقعة / "The condition preceded the incident."	AI captures timeline words but misclassifies legal relevance when causality depends on sequencing and event construal.	7%

Overall, the quantitative indices and NVivo-based error profiling converge on a layersensitive conclusion: GPT-4o-assisted workflows are most dependable when judicial meaning is surface-stable, genre-formulaic, and lexically explicit, and least dependable when the task requires contextual integration, diglossia-sensitive stance interpretation, or doctrinal warrant recovery. The findings justify cautious, expert-verified use of AI in Arabic forensic linguistics;

they do not justify autonomous AI interpretation of responsibility, capacity, or evidential sufficiency.

Reported statements are often reformulated through institutional narration, and dialectal expressions may encode subtle cues of stance, credibility, or alignment that are weakened or lost in automatic normalization. The pattern of over-confident stance assignment also reveals that the AI tended to interpret hedged or distributed evidential language as more certain than the judicial text actually warranted, thereby inflating the degree of certainty conveyed in the judgment. Temporal inference slippage, although less frequent, points to a similar problem, as the system could identify timeline expressions but still misclassify their legal relevance when interpretation depended on event sequencing and causal construal rather than temporal vocabulary alone. Overall, the pattern across themes suggests that AI was more reliable in detecting surface-level signals than in producing context-valid interpretations, particularly when analysis required pragmatic integration, doctrinal reasoning, and sensitivity to the layered semiotic structure of Arabic judicial discourse.

While the pragmatic bottleneck is a known phenomenon in NLP, this study provides a layer-specific benchmark for Sharia-based judicial reasoning. Unlike many Western legal NLP settings, Saudi judgments often rely on intertextual fiqh maxims in which silence, omission, and reported stance can carry specific doctrinal weight. The present findings show that GPT-4o's weakness is not merely linguistic but epistemological: the model may detect a lexical token such as silence, yet fail to recover the Sharia-informed warrant that links that token to a legal consequence such as assent, rebuttal, mitigation, or evidential insufficiency. This is the study's main contribution to AI evaluation: it operationalizes where model failure occurs, why it occurs, and which forms of expert verification are required.

The quantitative results show that AI performance is layer-dependent: it is strongest for structurally stable, genre-formulaic targets (notably Layer 3 and Layer 1) and weakest for highinference interpretation (Layer 4). The qualitative analysis explains why this pattern occurs: when meaning depends on doctrinal warranting, discourse positioning, and register-sensitive stance cues, AI tends to default to surface matching and produces labels that are linguistically plausible but pragmatically misaligned. Together, the two strands converge on a hybrid interpretation: AI is well-suited for high-throughput identification of stable legal and evidential markers, while expert verification is necessary where classifications require warrant-based inference.

Table 6

Joint Display of Layer-Wise Performance and Qualitative Error Patterns

P-DIM Layer	Quantitative Summary (from Tables 1-4)	Qualitative Mechanism (Detailed Discourse Explanation)	Integrated Interpretation (RQ/H Linkage)
L1: Institutional Weight	High F1 (0.90) and very high kappa (0.91).	Administrative Explicit-Signaling: High alignment occurs because institutional categories (e.g., committee reports) are signaled through rigid, standardized headers and administrative headers.	AI aligns with experts when evidence type is overtly and structurally signaled. (Supports RQ1-RQ2; consistent with H3).
L2: Temporal Framing	Moderate performance; kappa = 0.62.	Chronological vs. Causal Dissonance: AI successfully identifies temporal markers (date/time) but fails at "Event-Construal." It often misinterprets the causal link between a past medical history and the specific mental state at the moment of the offence.	Temporal framing requires inferential logic beyond simple timeline extraction. (Supports RQ2; qualifies H3 for L2).
L3: Legal Formulae	Highest indices (F1 = 0.96); very high kappa (0.94).	Terminological Invariance: Formulae regarding capacity (ahilyya) are stable, high-frequency, and lexically anchored in Sharia-based "ground truth" labels, allowing for near-perfect pattern recognition.	AI functions as a highly reliable pattern recognizer for formulaic extraction. (Supports RQ1-RQ2; supports H1 for L3).

P-DIM Layer	Quantitative Summary (from Tables 1-4)	Qualitative Mechanism (Detailed Discourse Explanation)	Integrated Interpretation (RQ/H Linkage)
L4: Pragmatic Function	Lowest indices (F1 = 0.41); low kappa (0.28); corrected significant layer effect (F(4, 15) = 63.04, p = 3.367e-09).	The Doctrinal-Pragmatic Gap: AI suffers from "Surface-Trigger Dependence"-detecting a token (e.g., silence) but missing the "Doctrinal Warrant." It lacks the inferential bridge needed to recover the argumentative aim from Sharia legal maxims.	"Pragmatic Bottleneck": AI fails when interpretation requires recovered judicial intent and intertextual legal logic. (Supports RQ2-RQ4; supports H2 and H4).
L5: Stance and Evidentiality	Substantial performance; kappa = 0.78.	Diglossic & Entextualization Distortion: The AI's normalization of dialectal variations in witness testimonies reduces its sensitivity to "hedging." It often assigns a "Confident" stance to distributed or indirect cues that human experts read as uncertain.	AI manages overt markers well but remains insufficiently sensitive to nuanced and distributed stance-taking strategies in Arabic discourse. (Supports RQ3-RQ4; partial support for H1; informs the workflow implication for H5 without treating H5 as a directly tested outcome).

RQ2 examined whether AI accuracy varies across the five P-DIM dimensions. The corrected exploratory ANOVA indicates a pronounced layer effect, $F(4, 15) = 63.04$, $p = 3.367e-09$, $\eta^2 = .944$. However, the analysis treats this result cautiously because the available observations are fold-level summaries rather than a large independent sample. The stronger evidential basis for the conclusion is the convergence between descriptive metrics and qualitative error analysis: Layer 4 collapses in both numerical performance and discourse-level validity, making it the clearest pragmatic bottleneck.

Table 6 further clarifies that AI performance varied systematically across the five P-DIM layers because each layer imposed a different interpretive burden. In Layer 1, the combination of a high F1 score (0.90) and very high kappa (0.91) indicates that AI aligned closely with expert judgments when institutional weight was overtly marked through rigid administrative labels and standardized evidential categories. This pattern suggests that AI performs strongly when the relevant information is structurally explicit and requires limited inferential mediation, which supports Research Questions 1 and 2 and is consistent with Hypothesis 3. A related pattern appears in Layer 3, where legal formulae achieved the strongest overall results ($F1 = 0.96$; $\kappa = 0.94$). Here, the near-perfect alignment can be explained by the terminological invariance of Sharia-based capacity expressions, which are stable, recurrent, and lexically anchored, allowing AI to function

as a highly effective pattern recognizer. This finding strongly supports Research Questions 1 and 2 and supports Hypothesis 1 for formulaic legal extraction.

By contrast, Layer 2 shows that temporal framing posed a more complex challenge. Although the AI was able to identify explicit temporal markers, its moderate performance (κ = 0.62) reveals difficulty in moving from chronological detection to causal interpretation. In many instances, the system recognized references to time but failed to determine how a psychiatric history was legally connected to the defendant's mental state at the precise moment of the offence. This indicates that temporal framing in judicial discourse is not reducible to timeline extraction alone, but depends on event construal and legal relevance, thereby supporting Research Question

2 while only partially confirming Hypothesis 3. The weakest results emerged in Layer 4, where pragmatic function recorded the lowest performance indices (F1 = 0.41; κ = 0.28), alongside a statistically significant layer effect in the corrected ANOVA, $F(4, 15) = 63.04$, $p = 3.367e\text{-}09$. The discourse evidence shows that this weakness stems from a doctrinal-pragmatic gap: AI often detected surface cues, such as tokens indexing silence or mitigation, but failed to recover the doctrinal warrant or argumentative purpose that gave those cues legal force within the judgment. This identifies Layer 4 as the central pragmatic bottleneck of the model and supports Research Questions 2 to 4, while validating Hypotheses 2 and 4.

Layer 5 presents an intermediate pattern. Its substantial performance (κ =.78) shows that AI could identify overt evidential and stance markers with reasonable success, yet the qualitative analysis demonstrates persistent distortion when stance was distributed across diglossic or entextualized discourse. In these cases, automatic normalization weakened the system's sensitivity to hedging, indirectness, and subtle evaluative positioning, often leading to more confident classifications than the text itself justified. This suggests that AI can manage explicit evidential signals, but remains insufficiently sensitive to the nuanced and distributed nature of stance-taking in Arabic judicial discourse. Accordingly, the findings for Layer 5 support Research Questions 3 and 4, offer partial support for Hypothesis 1, and suggest support for H5 only as a workflow implication: reliable interpretation still depends on expert verification after AI-assisted detection.

5.3 Integrated Findings: Implications for a Hybrid Forensic Workflow (H4-H5)

Hypothesis testing was interpreted cautiously. H1 was supported for Layer 3 and partly for Layer 5 because AI performed well on stable formulae and explicit stance markers. H2 was supported because Layer 4 produced the weakest performance and the clearest doctrinal-pragmatic error pattern. H3 received partial support: AI aligned closely with expert coding for Layer 1, but

only moderately for Layer 2, where temporal relevance required event construal. H4 was supported qualitatively because diglossia, entextualization, and doctrinal warrants explained many AI-human divergences. H5 is treated as a workflow implication rather than a directly tested hypothesis because the design did not directly compare AI-only, human-only, and hybrid workflows.

5.4 Summary

Overall, the quantitative indices and NVivo-based error profiling converge on a clear, layersensitive conclusion: GPT-4o is most dependable when judicial meaning is surface-stable, genreformulaic, and lexically explicit, especially in Layer 3 legal formulae and Layer 1 institutional weighting. Performance deteriorates sharply when the analytic task requires contextual integration and warrant-based inference, with the most consequential limitation concentrated in Layer 4. In practical terms, these findings justify an augmented, not fully automated, model of forensic analysis: AI should accelerate identification of candidate features, while structured expert verification remains methodologically necessary for pragmatic-doctrinal validity.

6. Discussion

This study evaluated AI-supported analysis of Arabic judicial judgments against an expertcoded benchmark using the five-layer Pragmatic-Doctrinal Inference Model (P-DIM). Across quantitative indices (precision/recall/F1 and κ) and qualitative error profiling, the findings converge on a central conclusion: AI performance is strongly conditioned by the inferential burden of the analytic layer. This pattern aligns with prior work in forensic/legal linguistics showing that legal meaning is jointly produced by formulaicity, institutional stance, and context-dependent inference rather than by lexical triggers alone (Coulthard & Johnson, 2007; Gibbons, 2003; Tiersma, 1999; Wright & Picornell, 2024).

6.1 Statistical transparency and the low overall kappa

The overall Cohen's kappa of.275 appears low relative to the high scores in L1 and L3. This is not an internal contradiction but a consequence of aggregating unevenly difficult classification layers. Kappa penalizes chance agreement and is sensitive to false negatives in complex categories. Because Layer 4 had a high rate of false negatives for doctrinal-pragmatic categories, the aggregate score reflects the AI's inability to maintain consistency across a multidimensional layered task, even though it excels at isolated pattern recognition in L1 and L3.

6.2 Layer effects, post-hoc interpretation, and H5

The corrected ANOVA showed a pronounced layer effect, $F(4, 15) = 63.04$, $p = 3.367e-09$, $\eta^2 = .944$, using four fold-level F1 values per layer. Tukey HSD tests clarified that L4: Pragmatic Function was significantly lower than every other layer, making it the principal substantive bottleneck. The complete post-hoc output also showed significant non-L4 contrasts for L1-L2, L2- L3, and L3-L5, so the interpretation is pairwise and transparent rather than reduced to a single contrast. This pattern strengthens the claim that AI handles administrative and formulaic data with comparable reliability but fails when legal meaning depends on performative function and doctrinal warrant. Consequently, H5 is not reported as the result of a third-arm experiment. Instead, the findings provide an empirical rationale for human-in-the-loop verification: because AI succeeds in extraction but fails in pragmatic-doctrinal interpretation, a hybrid workflow is methodologically justified for forensic validity.

6.3 From extracted markers to pragmatic intentions and doctrinal warrants

RQ3 asked whether AI-extracted markers correlate with pragmatic intentions and doctrinal warrants established manually. The answer is layered. For Layer 5 (Stance and Evidentiality), AI performed substantially ($\kappa = .78$; $F1 = .83$), particularly when stance is lexically explicit (e.g., epistemic verbs such as tabayyana). This is consistent with evaluation/stance research showing that many judicial certainty cues are conventionalized and therefore computationally detectable (Goźdź-Roszkowski & Hunston, 2016; Szczyrbak, 2014).

However, NVivo-based comparisons indicate reduced AI sensitivity where stance is realized indirectly in entextualized testimony and dialect-proximal segments-a finding consistent with Arabic diglossia as a persistent source of interpretive risk in automated processing (Ferguson, 1959). More importantly, Layer 4 errors show that lexical detection does not reliably entail warrant recovery: even when AI flagged cues like “silence,” it often failed to map them onto doctrinally licensed pragmatic force (Al-Qahtani, 2021). This gap supports the P-DIM claim that meaning in Saudi judgments is “engineered” through the integration of linguistic cues with doctrinal relevance and institutional evaluation (Saudi & Qabayli, 2023).

6.4 Linguistic and doctrinal constraints on holistic interpretation

RQ4 targeted the constraints that prevent AI from producing a holistic forensic interpretation of mental-health construction in Sharia-based judgments. Three constraints emerge as primary. First, diglossia and register shifting: testimony-related segments can encode hedging

and distancing in ways that do not align with MSA-centered patterns, lowering stance and pragmatic accuracy (Ferguson, 1959). Second, entextualization and institutional revoicing: judgments often reframe speech and behavior into institutional narrative, meaning that pragmatic force is mediated by judicial drafting practices rather than transparently recoverable. Third, doctrinal intertextuality: Sharia-based warrants and fiqh maxims can license inferences (e.g., legally meaningful silence) that require doctrinal knowledge and contextual placement to interpret correctly (Al-Qahtani, 2021; Yassin, 2002). These constraints explain why Layer 4 is consistently the weakest layer and why aggregated agreement is low when items require warrant-based inference.

Hypothesis testing showed a clear, layer-sensitive pattern. H1 was supported for Layer 3 and partly for Layer 5 because GPT-4o performed well on stable formulae and explicit stance markers. H2 was supported because Layer 4 produced the lowest scores across indices and was significantly lower than all other layers in Tukey HSD testing. H3 received partial support: AI aligned very closely with expert coding for Layer 1, but only moderately for Layer 2, where temporal relevance required event construal. H4 was supported qualitatively because recurrent error profiles linked AI-expert divergence to Arabic diglossia, entextualized testimony, and intertextual Sharia-based warrants. H5 was treated as a workflow implication rather than a directly tested empirical finding: this study did not compare a separate hybrid group, but the disparity between high extraction accuracy and weak pragmatic interpretation provides a strong methodological rationale for human-in-the-loop verification.

6.5 Comparison with recent legal NLP and Arabic legal LLM evaluation work

The present study should be positioned as a forensic-linguistic evaluation protocol rather than as a new legal AI architecture. Compared with LegalBench, which evaluates broad categories of legal reasoning in LLMs, P-DIM evaluates whether an AI system can reproduce expert forensic annotation across institutional, temporal, formulaic, pragmatic, and evidential layers (Guha et al., 2023). Compared with LegalBench-RAG, the present study does not evaluate retrieval because all passages are preselected and bounded; instead, it evaluates whether a model can assign valid labels once the relevant legal passage is already known (Pipitone & Alami, 2024). Compared with ArabLegalEval and ALARB, which are important advances in Arabic/Saudi legal AI benchmarking, the present study targets a narrower but more forensic problem: psychiatric evidence, capacity framing, and pragmatic-doctrinal warrant recovery in criminal judgments (Hijazi et al., 2024; Abu Shairah et al., 2025).

This comparison clarifies the study's contribution and its limits. Its contribution is not to outperform state-of-the-art legal LLM benchmarks, nor to claim general Arabic legal reasoning capability. Its contribution is to show that a layer-sensitive forensic protocol can reveal hidden performance asymmetries that global scores obscure. The corrected statistics, fold-level values, and qualitative error analysis jointly show that GPT-4o performs well when legal meaning is lexically stable, but that it remains unreliable where legal meaning depends on doctrinal intertextuality, diglossic stance, and pragmatic function. This is scientifically important because it specifies the boundary between useful AI-assisted extraction and invalid autonomous legalpragmatic interpretation.

6.6 Robustness, limitations, and interpretive caution

The corrected statistical analysis strengthens rather than weakens the substantive conclusion. The ANOVA table has been recalculated from the 20 fold-level F1 values, and the revised table now reports internally consistent SS, MS, F, exact p value, and η^2 values. Because the folds are derived from a small case-study corpus, the ANOVA is interpreted as an exploratory layer-sensitivity test, not as a population-level generalization. The core inference is supported by converging evidence: the fold-level statistics, the layer-wise kappa values, the confidence intervals, the confusion-pattern review, and the NVivo-based error mechanisms all identify Layer

4 as the principal bottleneck.

The main limitations are also explicit. The corpus is limited to 10 judgments and 24 analytical segments in one Saudi legal domain, the AI arm uses one deterministic GPT-4o configuration, and the study does not compare AI-only, human-only, and human-in-the-loop workflows as independent experimental conditions. These limitations prevent broad claims about all Arabic legal texts or all LLMs. Nevertheless, the design is appropriate for its stated purpose: a transparent, layer-based, expert-anchored evaluation of where AI-assisted forensic-linguistic annotation is dependable and where it requires mandatory human verification.

7. Conclusion

7.1 Summary of contributions

This study evaluated GPT-4o-assisted analysis of Arabic judicial judgments against expert manual coding through the five-layer Pragmatic-Doctrinal Inference Model (P-DIM). The corpus

consisted of 10 Saudi criminal judgments from which 24 analytically relevant segments were extracted, producing 120 classification decisions across the five P-DIM layers. The findings show that AI is most reliable when judicial meaning is structurally explicit, genre-formulaic, and lexically stable, particularly in Layer 1 (Institutional Weight) and Layer 3 (Legal Formulae). Performance was weaker when the task required temporal relevance, stance interpretation, or pragmatic-doctrinal inference, with the clearest limitation appearing in Layer 4 (Pragmatic Function). The study's contribution is therefore best understood as a calibrated, layer-specific evaluation protocol for Arabic forensic/legal NLP rather than as a new AI model, a domain-trained judicial model, or a universal benchmark for legal NLP.

The main theoretical contribution lies in specifying why AI failure in Arabic judicial discourse is not merely a matter of lower lexical accuracy. In Sharia-informed judgments, meaning often depends on intertextual legal warrants, fiqh-based reasoning, institutional revoicing, and the pragmatic force of silence, omission, or reported testimony. The error patterns observed in the study show that GPT-4o can often detect a relevant marker, such as a reference to silence, a psychiatric report, or a capacity formula, but may fail to recover the doctrinal reason why that marker matters for responsibility, mitigation, or evidential weight. This finding refines the general claim that AI struggles with pragmatics by locating the difficulty within a specific epistemological gap between surface detection and legally warranted inference.

7.2 Practical and policy implications

The findings support an augmented-intelligence workflow rather than full automation in forensic linguistic analysis of Arabic judicial texts. AI systems may be useful for front-end screening, rapid identification of institutional source types, extraction of stable legal formulae, and preliminary highlighting of overt stance or evidentiality markers. However, expert verification remains essential whenever classification depends on pragmatic function, doctrinal relevance, diglossic testimony, or the legal force of silence and omission. In this sense, human-in-the-loop review is not simply a cautious preference; it is a methodological requirement for maintaining forensic validity in high-stakes legal interpretation.

For policy and professional practice, the study suggests that AI-assisted forensic workflows should be designed around layer-sensitive verification rather than global accuracy claims. Structural layers may be routed through semi-automated extraction with targeted review, whereas high-inference layers require mandatory expert adjudication supported by a transparent codebook and documented audit trail. This approach allows institutions to benefit from computational

efficiency while preserving the interpretive accountability required in judicial and forensic contexts.

7.3 Future directions

Future research should extend this evaluation to larger and more diverse Arabic legal corpora, including judgments from different years, courts, case types, and Arab legal jurisdictions. Comparative testing across multiple AI systems would also clarify whether the pragmatic bottleneck observed here is model-specific or reflects a broader limitation of current large language models in doctrinally dense legal reasoning. A further priority is the development of doctrinally aware evaluation resources that incorporate Sharia-relevant terminology, fiqh maxims, dialectal testimony markers, and examples of institutional entextualization.

Future work should also test a full hybrid workflow directly by comparing AI-only, humanonly, and human-in-the-loop conditions. Such a design would allow stronger empirical claims about the efficiency, accuracy, and forensic validity of layered verification procedures. Until such comparative validation is available, the present findings justify a cautious conclusion: AI can assist Arabic forensic linguistic analysis by accelerating the discovery of stable textual features, but the validity of legal-pragmatic interpretation continues to depend on expert human judgment capable of integrating linguistic evidence with institutional reasoning and doctrinal warrants.

Declarations

Ethical approval

Ethical approval was not required because the study relied exclusively on publicly available/published archival judicial texts and did not involve human participants, direct interaction with individuals, clinical intervention, or the collection of private personal data.

Consent to participate

Not applicable. This study did not involve recruitment of human participants or collection of primary participant data.

Consent to publish

Not applicable. The manuscript does not contain identifiable individual data, images, or case materials requiring consent for publication.

Competing interests

The authors declare that they have no competing interests.

Funding This research was supported by a grant from the Deanship of Scientific Research at the University of Hail, Saudi Arabia, under project number RG-25 038.

Data availability

The judicial texts analyzed in this study were drawn from the official Majmūʿat al-Aḥkām al-Qaḍāʾiyya published by the Saudi Ministry of Justice and are available in that published source. The extracted segment IDs, layer-level benchmark labels, fold-level F1 values, and statistical analysis script are provided in the supplementary reproducibility package and are also available from the corresponding author upon reasonable request.

Clinical trial number

Not applicable.

References

Abdelali, A., Mubarak, H., Chowdhury, S., Hasanain, M., Mousi, B., Boughorbel, S., Abdaljalil, S., El Kheir, Y., Izham, D., Dalvi, F., Hawasly, M., Nazar, N., Elshahawy, Y., Ali, A., Durrani, N., Milic- Frayling, N., & Alam, F. (2024). LAraBench: Benchmarking Arabic AI with large language models. In Y. Graham & M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 487–520). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.30

Abu Shairah, H., AlHarbi, S., AlHussein, A., Alsabea, S., Shaqaqi, O., AlShamlan, H., Knio, O., & Turkiyyah, G. (2025). ALARB: An Arabic legal argument reasoning benchmark. In Proceedings of the Third Arabic Natural Language Processing Conference (pp. 389–406). Association for Computational Linguistics.

Al-Dulaimi, T. M. (2024). Athar al-amrāḍ al-nafsiyya ʿalā al-masʾūliyya al-jināʾiyya fī al-fiqh al-islāmī wa-l-qānūn [The effect of mental illnesses on criminal responsibility in Islamic jurisprudence and law]. Journal of Tikrit University for Legal and Political Sciences, 16 (1), 112–145.

Alhomoud, G. N. J. (2022). On translating and drafting Sharia legal terms and expressions in Saudi contracts. Education Research International, 2022, Article 7167970. https://doi.org/10.1155/2022/7167970

Al-Qahtani, A. (2021). Silence as legal communication: Applications of fiqh maxims in Saudi adjudication. Journal of Islamic Law Studies, 35 (2), 145–172.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (pp. 9–15). European Language Resources Association.

Appelbaum, P. S. (2006). Violence and mental disorders: Data and public policy. American Journal of Psychiatry, 163 (8), 1319–1321. https://doi.org/10.1176/appi.ajp.163.8.1319

Asokan, T. V. (2016). The insanity defense: Related issues. Indian Journal of Psychiatry, 58 (Suppl. 2), S191–S198. https://doi.org/10.4103/0019-5545.196832

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D.,... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Casey, E. (2011). Digital evidence and computer crime: Forensic science, computers, and the Internet (3rd ed.). Academic Press.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 4310–4330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acllong.297

Charrow, R. P., & Charrow, V. R. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. Columbia Law Review, 79 (7), 1306–1374.

Coulthard, M., & Johnson, A. (2007). An introduction to forensic linguistics: Language in evidence. Routledge.

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. arXiv. https://doi.org/10.48550/arXiv.2401.01301

Denault, V. (2024). Nonverbal communication and credibility in Australian courts. The International Journal of Evidence & Proof, 28 (1), 33–54. https://doi.org/10.1177/13657127231215647

Denault, V., & Bozin, D. (2024). The influence of "body language" on the assessment of witness credibility in Australian written court judgments. Journal of Forensic Psychology Research and Practice. Advance online publication. https://doi.org/10.1080/24732850.2024.2399616

El-Farahaty, H. (2010). Lexical and syntactic features of English and Arabic legal discourse: A comparative study. Comparative Legilinguistics, 4, 59–77.

Elieba, F. M. (2020). Textuality and ambiguity of forensic texts: Egyptian Supreme Court—A model. Beni-Suef University International Journal of Humanities and Social Sciences, 2 (2), 114–164.

Ferguson, C. A. (1959). Diglossia. Word, 15 (2), 325–340. https://doi.org/10.1080/00437956.1959.11659702

Ferguson, M., & Ogloff, J. R. P. (2011). Criminal responsibility evaluations: Role of psychologists in assessment. Psychiatry, Psychology and Law, 18 (1), 79–94. https://doi.org/10.1080/13218719.2010.543402

Gallego Balcells, J. (2023). Forensic linguistics and the analysis of legal discourse: Approaches to confession and evidence. Routledge.

Gibbons, J. (2003). Forensic linguistics: An introduction to language in the justice system. Blackwell.

Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. International Journal for the Semiotics of Law, 34 (4), 943–970. https://doi.org/10.1007/s11196-020-09799-3

Goźdź-Roszkowski, S. (2024). Language and legal judgments: Evaluation and argument in judicial discourse. Routledge.

Goźdź-Roszkowski, S., & Hunston, S. (2016). Corpora and beyond: Investigating evaluation in discourse. Corpora, 11 (2), 131–141. https://doi.org/10.3366/cor.2016.0089

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J.,... Liang, P. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. arXiv. https://doi.org/10.48550/arXiv.2308.11462

Habash, N. (2010). Introduction to Arabic natural language processing. Morgan & Claypool.

Hart, C. (2013). Event-construal in press reports of violence in two recent political protests: A cognitive linguistic approach. Journal of Language and Politics, 12 (3), 400–423. https://doi.org/10.1075/jlp.12.3.05har

Heydon, G. (2019). Researching forensic linguistics: Approaches and applications. Routledge.

Hijazi, F., AlHarbi, S., AlHussein, A., Abu Shairah, H., AlZahrani, R., AlShamlan, H., Knio, O., & Turkiyyah, G. (2024). ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In Proceedings of the Second Arabic Natural Language Processing Conference (pp. 225–249). Association for Computational Linguistics.

Hu, Y., Gan, L., Xiao, W., Kuang, K., & Wu, F. (2025). Fine-tuning large language models for improving factuality in legal question answering. arXiv. https://doi.org/10.48550/arXiv.2501.06521

Hussein Hamadi, H. (2022). Āliyyāt al-ḥijāj fī al-khiṭāb al-qaḍāʾī al-Saʿūdī: Muqāraba tadāwuliyya [Argumentation mechanisms in Saudi judicial discourse: A pragmatic approach]. Journal of Arts and Humanities, 30 (2), 89–115.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 92–104). Association for Computational Linguistics.

Jake, J. L., Myers-Scotton, C., & Gross, S. (2002). Making a minimalist approach to code-switching work: Adding the matrix language. Bilingualism: Language and Cognition, 5 (1), 69–91. https://doi.org/10.1017/S1366728902000147

Krippendorff, K. (2013). Content analysis: An introduction to its methodology (3rd ed.). SAGE.

Mahajna, D. A. E. (2019). A study of loanwords and code switching in spoken and online written Arabic by Palestinian Israelis [Doctoral dissertation, Indiana University]. ProQuest Dissertations & Theses Global. (Publication No. 13865054)

Markowitz, F. E. (2011). Mental illness, crime, and violence: Risk, context, and social control. Aggression and Violent Behavior, 16 (1), 36–44. https://doi.org/10.1016/j.avb.2010.10.003

Mazzi, D. (2010). "This argument fails for two reasons...": A corpus-based analysis of judicial evaluation. Research in Language, 8 (2), 187–200. https://doi.org/10.2478/v10015-010-0009-5

Meyer, J. F., Weaver, C. M., & Reppucci, N. D. (2020). Law and mental health: A case-based approach (2nd ed.). Guilford Press.

Myers-Scotton, C. (1993). Social motivations for code-switching: Evidence from Africa. Clarendon Press.

Nee, C., & Witt, C. (2013). Public perceptions of risk in criminality: The effects of mental illness and social disadvantage. Psychiatry Research, 209 (3), 675–683. https://doi.org/10.1016/j.psychres.2013.02.013

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., & Habash, N. (2020). CAMeL Tools: An open source Python toolkit for Arabic natural language processing. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 7022– 7032). European Language Resources Association.

OpenAI. (2024, May 13). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

Pipitone, N., & Alami, G. H. (2024). LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. arXiv. https://doi.org/10.48550/arXiv.2408.10343

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. Linguistics, 18 (7–8), 581–618. https://doi.org/10.1515/ling.1980.18.7-8.581

Pustejovsky, J., & Stubbs, A. (2012). Natural language annotation for machine learning. O'Reilly Media.

Saudi, A., & Qabayli, M. (2023). Handasat al-maʿnā fī al-aḥkām al-qaḍāʾiyya: Dirāsa lisāniyya maʿrifiyya [The engineering of meaning in judicial judgments: A cognitive linguistic study]. Journal of King Saud University: Arts and Humanities, 35 (3), 321–345.

Sayers, D., Sousa-Silva, R., Höhn, S., & Ahmedi, L. (Eds.). (2021). The dawn of the human-machine era: A forecast of new and emerging language technologies (COST Action CA19102 "Language in the Human-Machine Era" technical report). University of Jyväskylä. https://doi.org/10.17011/jyx/reports/20210518/1

Szczyrbak, M. (2014). Stancetaking strategies in judicial discourse: Evidence from US Supreme Court opinions. Studia Linguistica Universitatis Iagellonicae Cracoviensis, 131, 91–120. https://doi.org/10.4467/20834624SL.14.005.1377

Tiersma, P. M. (1999). Legal language. University of Chicago Press.

Vitacco, M. J. (2020). Insanity defense. In R. D. Morgan (Ed.), The SAGE encyclopedia of criminal psychology (pp. 665–669). SAGE Publications. https://doi.org/10.4135/9781483392240.n235

Wizārat al-ʿAdl [Ministry of Justice]. (1435 AH). Majmūʿat al-aḥkām al-qaḍāʾiyya [Collection of judicial decisions]. Ministry of Justice.

Wright, D., & Picornell, I. (2024). Semiotic perspectives on forensic and legal linguistics: Unifying approaches in the language of the legal process and language in evidence. International Journal for the Semiotics of Law, 37 (2), 293–304. https://doi.org/10.1007/s11196-023-10094-z

Yassin, M. N. (2002). Aḥkām al-marīḍ al-nafsī fī al-fiqh al-islāmī [Rulings regarding the mentally ill in Islamic jurisprudence]. Dār al-Nafāʾis.

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis: Evidencebased reflections. Malawi Medical Journal, 27 (1), 13–15. https://doi.org/10.4314/mmj.v27i1.4

Supplementary Files

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

- Anova.xlsx
- data.xlsx
- PDIMAppendices.docx