RAYDORF BENCHMARKS · RDF-BMK-2026-1

Turkish Legal AI *Research Platforms*

Pilot Evaluation Report

A Performance Evaluation of Leagle, Apilex and Dejure.ai

§ 0 · FOREWORD

Foreword

In Turkish legal practice, AI-assisted research platforms are no longer a curiosity; they have become part of everyday work. In-house counsel and attorneys now entrust these platforms with part of the work that runs from case-law searches to the drafting of legal opinions. What is entrusted is not small: the accuracy of a citation, whether a precedent is still current, whether an exception has been overlooked. These are questions at the very centre of professional responsibility.

Yet we have very little reliable information on how to choose these tools. The platforms' own promotional materials, by their nature, offer no common ground for comparison. What about individual trial and error? That is not enough either. It is practically impossible for a single practitioner to test a set of questions spanning different areas of law across several platforms under the same conditions, verify the results document by document, and repeat the exercise each time the platforms are updated. The time this would demand from a daily workload is precisely the time these tools are expected to save.

At the root of this gap lies a single missing piece: in Türkiye there is no agreed common standard for measuring the accuracy of legal AI tools. A lawyer's trust in a tool should rest not on impressions, habit or marketing language, but on verifiable findings produced by a disclosed method. The tradition of conformity assessment has done exactly this in other sectors for more than a century: announce the criterion in advance, apply the measurement to everyone in the same way, and publish the result together with its limitations.

The report in your hands is a first step on that path: a pilot evaluation examining three platforms against 14 common questions prepared by the Raydorf Institute. We evaluated the platforms with the same questions, under the same conditions, using pre-fixed metrics and three separate judge models; we have disclosed the method, the data collection dates and the limitations of the study. As a natural consequence of its pilot scale, the report makes no claim to a ranking. Our aim is to test the method in the field and to offer practitioners initial findings they can read in light of their own use cases. For the same reason, the selection guide does not declare a single winner.

In the next phase we are broadening the scope. We are moving to an evaluation process that will bring the market's leading legal AI companies together within the same framework. In this process, the performance criteria will be set jointly with participating companies and practising lawyers, and finalised by the Raydorf Institute; measurement and analysis will be carried out by Raydorf alone. Taking part in setting the criteria will not mean taking part in setting the result.

The Raydorf Institute's mission is to make the use of AI in professional services measurable and auditable. We regard the impartial evaluation of the tools lawyers rely on every day as a natural part of that mission, and we commit to doing so on a sustainable, periodic basis. Every critique of and contribution to our evaluation method will make future rounds better. Our door is open to them.

Dr. Taylan Yıldız

President, Raydorf Institute

§ I · SUMMARY

1. Summary

Türkiye has no common standard for measuring the accuracy of legal AI tools. This report presents the results of a pilot evaluation conducted as the first step in a process towards such a standard.

Leagle, Apilex and Dejure.ai were examined against 14 common research questions prepared by the Raydorf Institute. The questions span associations, environmental, intellectual property, tax, corporate, lease, employment, banking and enforcement, criminal and family law, as well as fundamental rights. Responses were collected on 4 August 2026 for Leagle and Apilex and on 24 June 2026 for Dejure.ai, from the live versions of the platforms on those dates. Measurement has two layers: the RAGAS framework examines the relationship between a response and the retrieved sources, while the legal-specific rubric measures citation accuracy, recency of case law, comprehensiveness and caution. Both layers were scored by three separate judge models.

Three limitations should be kept in mind when reading the results. In a 14-question sample, a difference on a single question can move a platform's average by up to 0.07 points. The findings apply only to the platform versions available on the collection dates. And scoring was performed by large language model judges rather than human experts. Small differences between platforms should therefore not be read as a ranking.

The pilot's clearest findings lie less in the differences between platforms than in patterns shared by all three tools: citation accuracy is among the two weakest rubric dimensions for all three; full grounding of response statements in the sources remains limited across all three (Response Groundedness between 0.60 and 0.64); and on some metrics the choice of judge model affects the score more than the choice of platform. Details in Section 7.

§ II · SYSTEMS

2. Systems Evaluated

- *Leagle:* Turkish legal research platform; responses collected on 4 August 2026 from the platform's live version on that date.
- *Apilex:* Turkish legal research platform; responses collected on 4 August 2026 from the platform's live version on that date.
- *Dejure.ai:* Turkish legal research platform; responses collected on 24 June 2026 from the platform's live version on that date.

The Hammurabi and Onedocs platforms, which operate in the market, are not included in any of the comparison tables below, as technical issues during response collection prevented comparable and scorable results.

§ III · DATASET

3. Dataset

The benchmark is built on a corpus of Turkish legal material and research-grade questions prepared specifically for this evaluation.

- *Question pool:* Consists of 14 research questions prepared by the Raydorf Institute. The questions span associations, environmental, intellectual property, tax, corporate, lease, employment, banking and enforcement, criminal and family law, as well as fundamental rights. The same question texts were put to all three platforms; each platform's response, the references it presented to the user and the text content of those references were recorded.
- *Legal domain coverage:* The questions cover several areas of Turkish law: Civil and Criminal Jurisdiction (Legislation, Court of Cassation, Regional Courts of Appeal), Administrative Jurisdiction (Council of State) and regulatory authorities (Capital Markets Board, Banking Regulation and Supervision Agency, Personal Data Protection Authority, etc.).

§ IV · METHODOLOGY

4. Methodology

4.1 Response collection

Each platform received the same question set through its own standard interface. For each response, the full response text, the retrieved source documents (contexts) and per-document metadata (court, chamber, docket/decision numbers, legislation identifiers) were recorded.

4.2 RAGAS metrics

Four RAGAS metrics were computed. Each metric was assessed separately by GPT-5.6 Luna, Gemini 3.5 Flash and DeepSeek V4 Flash; the main values in the report are the equally weighted average of these three judges. Answer Relevancy is embedding-based and uses the text-embedding-3-small and text-multilingual-embedding-002 models; the other metrics assess the relationship between the retrieved context and the response. Each metric ranges from 0 to 1; higher is better.

METRIC	WHAT IT MEASURES	COMPUTATION
Faithfulness	The share of statements in the response supported by the retrieved context; penalises unsupported (hallucinated) claims.	supported / total statements
Answer Relevancy	How directly the response addresses the question; penalises off-topic or evasive responses (not incorrect ones).	cosine similarity of regenerated questions × evasiveness gate
Response Groundedness	Finer-grained grounding: the degree to which each statement stays tied to the context; partial credit is given.	mean of 2 judges (0/1/2) ÷ 2
Context Relevance	Quality of the retrieval step: relevance of the retrieved documents to the question.	mean of 2 judges (0/1/2) ÷ 2

Configuration: RAGAS 0.3.1 · three independent judge models · embedding models: text-embedding-3-small and text-multilingual-embedding-002 · equally weighted judge average in main tables · 0–1 normalised score.

4.3 Legal-specific evaluation rubric

The evaluation used a legal-specific rubric adapted from PlawBench (arXiv:2601.16669) and LexRubric (arXiv:2606.09389). Citation accuracy, recency of case law, legal comprehensiveness and response caution were scored against the retrieved source material by three judge models on a 1–5 scale, then normalised to the 0–1 range.

METRIC	WHAT IT MEASURES	SCALE
Citation Accuracy	Whether cited decisions are real, correctly attributed to the right court/chamber, and traceable to the retrieved sources.	from 1 (fabricated/misattributed) to 5 (fully traceable, correct)
Case-Law Recency	Whether the response relies on current decisions rather than outdated or superseded case law.	from 1 (pre-2015 only) to 5 (current first, dates stated)
Legal Comprehensiveness	Whether the response covers the relevant legislation, core principles, exceptions, conditions and procedure.	from 1 (misses the core issue) to 5 (comprehensive)
Response Caution	Whether contested law is not presented as settled and uncertainty is expressed appropriately.	from 1 (overconfident) to 5 (calibrated, flags evolving areas)

Configuration: GPT-5.6 Luna, Gemini 3.5 Flash and DeepSeek V4 Flash · 1–5 per dimension · normalised score = raw ÷ 5 · main tables = equally weighted average of three judges.

§ V · LIMITATIONS

5. Limitations

Before turning to the findings, the limitations that determine the conditions under which the results should be read are set out below.

- *Sample size:* The common set consists of 14 questions. At this scale, a performance difference on a single question can affect a platform's average by up to about 0.07 points; small differences between platforms (±0.02) should not be interpreted as a meaningful ranking difference.
- *Versions and data collection dates:* The results reflect the live platform versions available to users on the dates responses were collected; as this product class is updated rapidly, the findings should not be generalised to later versions. Leagle and Apilex responses were collected on 4 August 2026, Dejure.ai responses on 24 June 2026. The roughly six-week gap limits the simultaneity of the comparison; Dejure.ai results should be assessed with this date difference in mind.
- *Automated judge evaluation:* Scoring was performed by large language model judges rather than human expert review. The supplementary tables in Sections 6.2 and 6.3 show considerable variance between judges on a per-metric basis. The equally weighted average of three judges reduces this variance but does not eliminate it; the results do not replace professional judgement.
- *Scope:* The Hammurabi and Onedocs platforms were left out of the comparison for the technical reasons explained in Section 2; this report does not represent the entire Turkish legal AI market.

6. Findings

6.1 Head-to-head comparison: 14-question common set

Normalised scores range from 0.00 to 1.00; higher is better. RAGAS cells are the equally weighted average of three judge models. Composite indicator = (RAGAS overall + rubric overall) ÷ 2.

METRIC	LEAGLE	APILEX	DEJURE.AI
RAGAS metrics			
Faithfulness	0.90	0.54	0.92
Answer Relevancy	0.76	0.78	0.73
Response Groundedness	0.62	0.64	0.60
Context Relevance	0.75	0.61	0.82
RAGAS overall (mean of 4 metrics)	**0.76**	**0.64**	**0.77**
Legal-specific rubric (normalised)			
Citation Accuracy	0.84	0.86	0.78
Case-Law Recency	0.90	0.87	0.84
Legal Comprehensiveness	0.91	0.87	0.84
Response Caution	0.88	0.92	0.75
Rubric overall (mean of 4 dimensions)	**0.88**	**0.88**	**0.80**
Composite overall score	**0.82**	**0.76**	**0.78**

Note: Judge-level differences are shown separately in Supplementary Tables 6.2 and 6.3. Averages in the main table were computed from unrounded values and then rounded to two decimal places; differences of ±0.01 may therefore appear against the average of the rounded rows in the supplementary tables. Differences between composite scores should not be read as a ranking, owing to the sample limitation noted in Section 5.

6.2 Judge-model sensitivity: supplementary RAGAS table

The supplementary table below shows each judge model's RAGAS scores separately on the same 14-question set. The scores in the main comparison are the equally weighted average of these rows.

JUDGE	PLATFORM	FAITHFULNESS	RELEVANCY	GROUNDEDNESS	CONTEXT REL.
GPT-5.6 Luna	Leagle	0.94	0.76	0.48	0.70
GPT-5.6 Luna	Apilex	0.40	0.78	0.55	0.57
GPT-5.6 Luna	Dejure.ai	0.93	0.74	0.48	0.82
Gemini 3.5 Flash	Leagle	0.79	0.87	0.44	0.61
Gemini 3.5 Flash	Apilex	0.39	0.87	0.46	0.55
Gemini 3.5 Flash	Dejure.ai	0.88	0.85	0.39	0.64
DeepSeek V4 Flash	Leagle	0.97	0.66	0.93	0.93
DeepSeek V4 Flash	Apilex	0.83	0.69	0.91	0.71
DeepSeek V4 Flash	Dejure.ai	0.94	0.60	0.93	0.98

6.3 Judge-model sensitivity: legal-specific rubric

The supplementary table below shows the same three judges' results on the rubric dimensions. The main rubric values are the equally weighted average of these rows.

JUDGE	PLATFORM	CITATION	RECENCY	COMPREHENSIVENESS	CAUTION
GPT-5.6 Luna	Leagle	0.77	0.81	0.80	0.76
GPT-5.6 Luna	Apilex	0.83	0.85	0.79	0.88
GPT-5.6 Luna	Dejure.ai	0.77	0.75	0.64	0.56
Gemini 3.5 Flash	Leagle	0.90	0.97	1.00	1.00
Gemini 3.5 Flash	Apilex	1.00	0.93	1.00	1.00
Gemini 3.5 Flash	Dejure.ai	0.89	0.89	1.00	0.87
DeepSeek V4 Flash	Leagle	0.84	0.91	0.91	0.87
DeepSeek V4 Flash	Apilex	0.76	0.84	0.83	0.87
DeepSeek V4 Flash	Dejure.ai	0.70	0.89	0.87	0.81

6.4 Operational characteristics

CHARACTERISTIC	LEAGLE	APILEX	DEJURE.AI
Number of common questions	14	14	14
Avg. retrieved contexts	35.64	34.93	31.07
Avg. response length (characters)	20,512	28,240	7,904
In-text citations	Yes	Yes	Yes
Access to source details	Yes	Yes	Yes
Response report export	Yes	Yes	Yes
Source content record	Full	Full	Full

§ VII · PATTERNS

7. Observed Patterns

This section first addresses patterns common to all three tools, then platform profiles. At pilot scale, shared patterns offer more robust information than small differences between platforms.

- *Citation accuracy is a shared weak point:* Citation accuracy is the lowest rubric dimension for Leagle (0.84) and Apilex (0.86), and the second lowest for Dejure.ai (0.78). The existence of a source is not enough; whether it correctly matches the legal claim in the response must be checked separately. Primary-document verification is essential, especially for restricted-access sources.
- *Grounding in sources is limited across all three tools:* Response Groundedness stays between 0.60 and 0.64 on all three platforms; for Leagle and Dejure.ai it is the lowest metric in the RAGAS layer. The value diverges markedly by judge: GPT-5.6 Luna and Gemini 3.5 Flash give scores between 0.39 and 0.55, while DeepSeek V4 Flash gives scores between 0.91 and 0.93. The finding rests on the trend two judges show consistently across all three platforms.
- *On some metrics, judge choice matters more than platform choice:* Leagle's Response Groundedness score ranges from 0.44 to 0.93 depending on the judge, while under the same judge the difference between platforms on this metric does not exceed 0.07. Apilex's Faithfulness score likewise swings between 0.39 and 0.83 depending on the judge. This finding, which shows that single-judge evaluations can be misleading, makes the joint setting of criteria and judge calibration the priority for the next round.
- *Ceiling effect:* Gemini 3.5 Flash gave a full score (1.00) to all three platforms on legal comprehensiveness and to two of them on response caution. This clustering of full scores indicates that this judge could not differentiate the platforms on these dimensions.
- *Response length and source fidelity:* Apilex, which produced the longest responses (average 28,240 characters), received the lowest Faithfulness value (0.54), while Dejure.ai, which produced the shortest (7,904 characters), received the highest (0.92); Leagle (20,512 characters, 0.90) sits between the two. Resting on three data points, this relationship should be read as a hypothesis and will be tested in the expanded round. Length alone is not a measure of quality; it should be assessed together with coverage and contextual support.
- *Platform profiles:* On the legal-specific rubric Leagle and Apilex are neck and neck (0.88); Leagle leads on comprehensiveness (0.91) and recency (0.90), Apilex on citation accuracy (0.86) and caution (0.92). On RAGAS overall, Dejure.ai (0.77) and Leagle (0.76) are close, with Apilex (0.64) behind; Dejure.ai stands out on Context Relevance (0.82), Apilex on Answer Relevancy (0.78) and Response Groundedness (0.64). Given the sample size, the differences in the composite indicator (between 0.04 and 0.06) do not constitute a definitive ranking.

8. Usage and Selection Guide for Lawyers

This section turns the findings into a practical selection framework from the perspective of lawyers, the platforms' target users. The equally weighted composite indicator provides a quick summary, but is not sufficient on its own for choosing a platform. The three platforms show different strengths in different use cases. The pairings below are based on pilot findings and serve as initial guidance; for a definitive choice, the pilot method in Section 8.3 is recommended.

8.1 Platform fit by use case

- *If you conduct comprehensive case-law research and prepare legal opinions:* taking legal comprehensiveness (0.91), case-law recency (0.90) and RAGAS overall (0.76) together, Leagle is a strong option. Every critical claim should still be verified against its source document.
- *If you prioritise responses that engage directly with the question and a cautious legal draft:* Apilex stands out on the legal-specific rubric overall (0.88), Answer Relevancy (0.78) and Response Groundedness (0.64). Given its Faithfulness result, long responses should be cross-checked against the source text.
- *If you are looking for a shorter first response closely tied to the context:* Dejure.ai is strong on Faithfulness (0.92) and Context Relevance (0.82). Before critical use, access to document details and primary-source verification capabilities should be tested separately.

8.2 Platform-independent rules

- *Verify every citation against the primary source.* The rubric results show that citation accuracy requires active checking regardless of platform choice. Every docket/decision number that goes into a pleading or legal opinion should be confirmed at its source.
- *Treat platform output as a research draft, not a final opinion.* Even the strongest comprehensiveness scores do not guarantee that exceptions and recent developments have been fully captured; responsibility for professional judgement remains with the practitioner.
- *Check the level of caution in contested and evolving areas.* Response Caution scores are 0.88 for Leagle, 0.92 for Apilex and 0.75 for Dejure.ai; particularly in lower-scoring profiles, check whether uncertainty and conflicting case law are addressed explicitly.
- *Clarify source access terms before procurement.* In this study the references presented to users and their text content are on record; in live use, confirm that this access is sustainable for citation verification and internal audit.

8.3 Where to start

Before a procurement decision, the recommended approach is a limited pilot using five to ten questions from the practitioner's own field to which they know the answers: the same questions should be put to the candidate platforms, and the accuracy of citations, the exceptions covered and the recency of case law compared with the practitioner's own knowledge. This method reveals domain-specific performance differences that general score tables cannot show, and ties the choice to the organisation's actual way of working. Choosing a platform is not a one-off decision; as platforms are updated rapidly, it is advisable to repeat the evaluation at reasonable intervals.

§ IX · INDEPENDENCE

9. Statement of Independence

No fee was received from any platform in return for inclusion in or ranking within this evaluation. The Raydorf Institute holds no commercial stake in any of the platforms evaluated. The methodology, metrics and judge instructions were fixed before responses were collected; no platform had editorial influence over the content of the results. The question set was prepared by the Raydorf Institute.

§ X · APPENDIX

10. Appendix: Metric Definitions and Judge Instructions

A. RAGAS metrics

Faithfulness: Decomposes the response into atomic statements and asks the judge whether each statement is supported by the retrieved context. Score = supported statements / total statements. Judge instruction: "Judge the faithfulness of a series of statements based on a given context..."

Answer Relevancy: Seeing only the response, the judge regenerates the question the response most likely answers (×3); each is embedded and compared with the original question by cosine similarity; if the response is flagged as evasive in all three generations, the score is set to zero. Score = mean(cosine similarity) × evasiveness gate.

Response Groundedness (NVIDIA NV): Each statement is scored against the context as 0 (not supported) / 1 (partially) / 2 (fully supported) by two independent judges at temperature 0.1; Score = mean(judge1, judge2) ÷ 2.

Context Relevance (NVIDIA NV): Each retrieved chunk is scored 0/1/2 for relevance to the question by two independent judges; Score = mean(judge1, judge2) ÷ 2. This metric evaluates retrieval, not generation.

B. Legal-specific rubric (anchor definitions)

Citation Accuracy: 1: cites decisions not found in the retrieved references or misattributes them; contains fabricated docket/decision numbers. 5: every cited decision is fully traceable to the retrieved references, attributed to the correct court and chamber, with correct docket/decision numbers; no fabricated citations.

Case-Law Recency: 1: relies only on pre-2015 decisions or presents an outdated position as current law when more recent case law exists. 5: consistently prioritises the most recent decisions, states key decision years and flags older case law that may no longer reflect current practice.

Legal Comprehensiveness: 1: misses the core legal issue or addresses only a peripheral aspect. 5: comprehensively addresses the relevant legislation, leading decisions, exceptions, conditions and procedural requirements; notes divergences between lines of case law.

Response Caution: 1: presents contested or evolving positions as settled law without any caveat. 5: calibrates confidence precisely to the strength of the authority relied on, clearly flags evolving areas and recommends professional advice where appropriate.

COLOPHON

Document code: RDF-BMK-2026-1 · Version 1.2 · Published: September 2026 · Publisher: Raydorf Institute, Istanbul · raydorf.com

Suggested citation: Raydorf Institute (2026). *Pilot Evaluation Report: Turkish Legal AI Research Platforms* (RDF-BMK-2026-1, Version 1.2). Istanbul: Raydorf Institute.

Following this pilot evaluation, an expanded evaluation round will begin with the participation of leading legal AI companies and jointly set criteria.

The Raydorf Institute is an independent certification body working in the tradition of conformity assessment. We offer four certification services for organisations and individuals in the legal sector.

LAW FIRMS

Institutional Certification and Rating

Independent assessment against the seven-dimension Raydorf Standard, with a five-tier rating. The tier is set by the weakest dimension; no total score or average is used. Certification runs on a three-year cycle with annual interim verifications.

AI-Aware · AI-Enabled · AI-Integrated · AI-First · AI-Native

LEGAL COUNSEL

Individual Competency Certificate

An individual certificate documenting a practitioner's AI competency. Assessment is by standardised psychometric examination or expert interview; the result is reported in one of four proficiency bands.

Developing · Proficient · Advanced · Distinguished

LEGAL DEPARTMENTS

AI Use Proficiency Certificate

Team-level proficiency assessment for in-house legal departments. The department's AI use practice is reviewed together with its governance framework and documented with a proficiency certificate.

Team assessment · Governance review · Department certificate

RECRUITMENT

AI Competency Tests

A standardised competency test for lawyers applying for positions. Candidates are assessed on the same scale; the employer receives a comparable score report.

Standard scale · Comparable score report

CONTACT

Write to us for information on certification processes.

contact@raydorf.com Raydorf Institute · Istanbul, Türkiye raydorf.com