GREEKBARBENCH: A Challenging Benchmark for Free-Text Legal Reasoning and Citations

Odysseas S. Chlapanis[1,2] **Dimitrios Galanis**[3] **Nikolaos Aletras**[4] **Ion Androutsopoulos**[1,2]

[1]Department of Informatics, Athens University of Economics and Business, Greece
[2]Archimedes, Athena Research Center, Greece
[3]Athena Research Center, Greece
[4]University of Sheffield, United Kingdom

Abstract

We introduce GREEKBARBENCH, a benchmark that evaluates LLMs on legal questions across five different legal areas from the Greek Bar exams, requiring citations to statutory articles and case facts. To tackle the challenges of free-text evaluation, we propose a three-dimensional scoring system combined with an LLM-as-a-judge approach. We also develop a meta-evaluation benchmark to assess the correlation between LLM-judges and human expert evaluations, revealing that simple, span-based rubrics improve their alignment. Our systematic evaluation of 13 proprietary and open-weight LLMs shows that even though the best models outperform average expert scores, they fall short of the 95th percentile of experts.

1 Introduction

As legal AI assistants become increasingly prevalent, the need for realistic legal LLM benchmarks has never been more imperative.[1] Most widely used legal Natural Language Processing (NLP) benchmarks (Chalkidis et al., 2022; Niklaus et al., 2023) focus on classification tasks, e.g., *legal judgement prediction* (Aletras et al., 2016), which have been criticized (Medvedeva and Mcbride, 2023) for being more constrained and less representative than real-world tasks. Even more recent LLM-focused legal benchmarks (Guha et al., 2023; Fei et al., 2024; Joshi et al., 2024) do not go beyond closed-form questions (e.g., multiple-choice questions), failing to capture the true complexity of legal reasoning in practice, which involves identifying, analyzing and synthesizing relevant information to reach a conclusion. Unfortunately, most existing benchmarks with challenging legal questions and free-text responses are proprietary and thus inaccessible to the research community.[2]

Facts
[1] Antonis visited his dermatologist, Ioannis, to remove facial skin tags. [2] Ioannis's assistant, Penelope, accidentally applied pure acetic acid due to a mislabeled bottle, causing burns on Antonis's face. [3] He needed plastic surgery costing €2,500 and is now seeking these costs plus €75,000 for moral damages.
Question
Which individuals are liable for the injury?
Relevant Legal Context
Civil Code 914: Anyone who unlawfully causes damage must compensate the victim. Civil Code 922: An employer is liable for unlawful damages caused by their employee during work.
Ground Truth Answer
Ioannis is responsible vicariously for Penelope's actions [2] **(Civil Code 922)** and Penelope is directly liable for her **negligence** [2] **(Civil Code 914)**. Hence, **both are liable** and must compensate Antonis.

Table 1: Cropped example (English translation) from GREEKBARBENCH. The answer requires multi-hop reasoning and citing legal articles and case facts. The spans corresponding to the scoring dimensions are highlighted in color: *Facts* (green), *Cited Articles* (blue) and *Analysis* (orange). *Important spans* are marked in bold and cited facts are denoted by square brackets. The complete example is presented in Appendix C.

Another challenge is that realistic benchmarks often require costly manual evaluation by legal experts, which limits scalability (Magesh et al., 2025; Martin et al., 2024). Automatic evaluation, using the LLM-as-a-judge framework (Zheng et al., 2023), is a promising alternative; however, its reliability has not been extensively assessed in legal reasoning (Bhambhoria et al., 2024; Li et al., 2025).

To address these issues, we present the GREEKBARBENCH, a benchmark that evaluates the rea-

[1]https://www.abajournal.com/web/article/aba-tech-report-finds-that-ai-adoption-is-growing-but-some-are-hesitant

[2]https://www.vals.ai/benchmarks

soning capabilities of LLMs on challenging legal questions across five legal areas. The questions are taken from the Greek Bar exams and require open-ended answers with citations to statutory articles and case facts. In addition, we introduce an accompanying benchmark for LLM-judges, designed to measure how well their scores correlate with those of human experts. GREEKBARBENCH is the only Greek dataset for legal reasoning. Our main contributions are the following:

- **GREEKBARBENCH**: a challenging legal reasoning benchmark that requires free-text answers citing case facts and statutory articles.
- **GBB-JME**: an accompanying dataset with human-evaluated answers from five different LLMs, to assess the quality of candidate LLM-judges in GREEKBARBENCH.
- A three-dimensional **scoring system** and an **LLM-judge framework** based on **span-rubrics** per dimension (*Facts*, *Cited Articles*, *Analysis*), which aligns well with human expert evaluation.
- A systematic **evaluation of 13 frontier and open-weight LLMs** on GREEKBARBENCH, using the best LLM-judge at GBB-JME. **Top models surpass average expert performance**, but not the 95th percentile of experts.

All resources including the two benchmarks (except for a small semi-private test set) and the prompts are publicly available.[3]

2 GREEKBARBENCH (GBB)

2.1 Greek Bar Exams

Law graduates in Greece must pass the Greek Bar exam to become licensed attorneys. The exam evaluates candidates through practical legal questions across five key areas of law: Civil Law, Criminal Law, Commercial Law, Public Law, Lawyers' Code. Greece's legal system is statutory, meaning that laws are derived from legal code documents (statutes), rather than from judicial precedents (case law). The exams are open-book; candidates have access to legal code documents and are expected to cite statutory articles from them in their answers.

The available documents include: the Civil Code and Civil Procedure Code, the Criminal Code and Criminal Procedure Code, eight Commercial Law codes, eleven Public Law codes, as well as the Lawyers' Code and the Code of Ethics for Legal Practice (see Table 3). Candidate lawyers typically approach the exam by first studying the case facts to identify the relevant legal issues. They then navigate the legal code documents to find the relevant chapter and pinpoint the exact statutory article within it to cite in support of their arguments.

[3]The URL of the GitHub repository will be provided in the camera-ready version.

Benchmark	Lang	Citations	Multi-Hop	Free-Text	Judge Eval
LegalBench	en	✗	✓	✗	✗
LexEval (5.4)	zh	✗	✗	✓	✗
CaseGen	zh	✗	✓	✓	✗
OAB-Bench	por	✗	✓	✓	✓
LLeQA	fr	✓	✓	✓	✗
GBB (Ours)	el	✓	✓	✓	✓

Table 2: Comparison of legal benchmarks. GREEKBARBENCH uniquely encompasses all challenging features essential for evaluation in realistic and practical scenarios. 'Lang': language of dataset. 'Citations': legal articles. 'Multi-hop': reasoning using multiple sources. 'Free-text': open-ended responses. 'Judge Eval': manual evaluations to compare LLM-judges.

2.2 Task

Each instance in GREEKBARBENCH is taken from Greek Bar exam papers. The input consists of (1) the case facts, (2) the legal question, and (3) a collection of potentially relevant chapters of statutory articles. The desired output is the correct free-text answer to a legal question, providing an *analysis* with citations to the case *facts* and the applicable statutory *articles*. The primary challenges are to discern significant facts, to identify the applicable articles and, finally, to analyze the outcomes of the application of the articles to answer the question (see the example in Table 1). These attributes make GREEKBARBENCH unique compared to other legal benchmarks (Table 2).

2.3 Dataset Statistics

We collect a total of 65 exam papers; 13 exam papers from each of the five aforementioned areas. The papers and suggested solutions are publicly available in a booklet in PDF format, spanning from 2015 to 2024.[4] The booklet is converted to text format and further processed (§ 2.5) to prepare the dataset. Each exam paper includes 4.7 questions on average, resulting in a total of 310 samples.

[4]The booklet is available at https://www.lawspot.gr/nomika-nea/panellinios-diagonismos-ypopsifion-dikigoron-themata-exetaseon-kai-endeiktikes-3. The authors permit its distribution for academic research only.

Law Areas	Samples	Legal Codes	Total Articles	Cited Articles	Context (tokens)
Civil	71	2	3,264	286	87k
Criminal	53	2	1,253	186	58k
Commercial	58	8	4,177	159	29k
Public	71	11	2,912	118	67k
Lawyers	57	2	4,476	182	66k
Total	310	25	16,082	931	62k

Table 3: Summary of dataset statistics. 'Legal Codes' indicates the number of distinct legal code documents in each area. 'Cited Articles' is the total number of citations to legal code articles. 'Context' denotes the average token count of the relevant legal context (chapters of legal code) provided in the input of candidate LLMs.

We keep the questions from 2024 (22 in total) as a semi-private test set, to avoid data contamination.[5] The remaining 288 samples comprise the public test set. The semi-private set will be updated each year with two more exam papers from each legal area (there are two examinations per year), and made publicly available.

Answering exam questions requires citing articles from 25 legal code documents, which we collect from the same source website as the exams.[6] Detailed statistics for these documents are presented in Table 3 per legal area. Articles are cited 931 times in total, across all exam questions. The articles within each legal code document are grouped thematically into chapters. The total number of citable articles is more than 16 thousand.

2.4 Relevant Legal Context

As mentioned in Section 2.1, the Greek Bar exams are open-book, allowing candidate lawyers to navigate legal code documents to identify relevant statutory articles for the presented case. Simulating this setup presents several challenges. One approach would be implementing a standard Retrieval-Augmented Generation (RAG) pipeline, using sparse (e.g., BM25) or dense retrievers (Karpukhin et al., 2020) to select the k most 'relevant' articles for inclusion in the LLM's input. However, this approach suffers from three significant limitations: a) candidate lawyers taking the exams do not have access to such retrieval tools, making direct comparisons with human performance problematic; b) retrievers are prone to errors, creating a substantial risk that even with large values of k, the ground truth articles might not appear among the top retrieved articles; and c) as demonstrated by Krishna et al. (2025), benchmarking RAG systems requires testing multiple configurations with varying values of k and, ideally, different retriever models, complicating fast integration of new LLMs.

Instead, we adopt a simplified yet challenging approach. For each legal case, we automatically collect all the ground truth articles cited across all questions pertaining to that case (using regular expressions), and identify the chapters containing the articles. We hypothesize that all the articles of the identified chapters have substantial relevance for all questions of the particular case. Therefore, we provide the entire text (all articles) of the identified chapters as the relevant legal context in the LLM input for *every* question of the *particular* case. Hence, for each question, the input context contains all the ground truth articles of the question, along with all the other articles from their chapters, as particularly close distractors. It also includes additional distractor articles from other chapters related to other questions of the same case, which increase the difficulty of citing the correct articles.

The total length of the legal context fed to the LLM per question is 62k tokens on average (Table 3). This makes the task manageable for recent LLMs with context windows exceeding 100k tokens, while still presenting a significant challenge, as LLMs often struggle to extract crucial information from extended contexts (Liu et al., 2024).

2.5 Fact Segmentation

To enhance the evaluation process (both manual and automatic), we require citations to facts in candidate answers, though this is not mandatory in official exams. Fact-citations help legal professionals in practical applications by enabling more efficient verification of answer accuracy. To facilitate fact-citation, we present the case facts as a numbered list of sentences (as shown in Table 1). For segmentation, we employ the Segment-Any-Text neural model (Frohmann et al., 2024).

2.6 Three-Dimensional Scoring System

The official evaluation committee of the Greek Bar Exams grades candidate answers on a scale of 1 to 10. Without explicit guidelines, they grade by comparing answers to the ground truth based on their discretion. Drawing inspiration from established legal research and evaluation practices (Clark and DeSanctis, 2013), and guided by our legal expert annotators (§ 5.2), we develop a novel three-dimensional

[5] With 'semi-private' we mean that the test set is not public, but the raw data sources are available.

[6] www.lawspot.gr

scoring system to improve the evaluation process for the benchmark. The proposed approach assesses legal reasoning across three dimensions: the *Facts*, the *Cited Articles*, and the *Analysis*. Each dimension is rated on a scale of 1 to 10, and the final score is the average of these three. This system allows explainability through detection of specific shortcomings in the reasoning abilities of LLMs. The *Facts* score measures understanding of case facts; the *Cited Articles* evaluates the accuracy and interpretation of cited legal articles; and the *Analysis* evaluates the ability to apply legal articles to the facts and reach conclusions. For instance, a low *Facts* score indicates hallucinations, a low *Cited Articles* score shows difficulty in identifying applicable articles, and a low *Analysis* score reveals weakness in legal reasoning.

3 Automatic Evaluation

To address the evaluation of free-text answers without the prohibitive cost of manual annotations, we use the LLM-as-a-judge framework (Zheng et al., 2023). LLM-judges can be categorized into two primary types: (a) *pairwise* LLM-judges, which evaluate two candidate answers and determine which is preferred (or declare a tie), and (b) *grading* LLM-judges, which assign an integer score to each individual candidate answer (Zheng et al., 2023). In our work, we focus on *grading* LLM-judges to allow cost-effective integration of new participant LLMs without the overhead of quadratically increasing pairwise comparisons.

To improve the alignment of LLM-judges with human expert annotators, we propose novel span-based rubrics; i.e., evaluator instructions in the form of annotated spans per question. These spans will guide the LLM-Judge in what to assess in the candidate answers. However, even with these question-specific rubrics, replicating the nuanced evaluation of human experts, especially in complex tasks like legal writing, cannot be guaranteed. For this reason, we also include a framework to meta-evaluate whether LLM-judges are suitable proxies for human evaluation on GreekBarBench.

3.1 Simple LLM-Judge

As an initial approach, we designed a straightforward prompt for a simple LLM-judge. The prompt outlines the evaluation task and explicitly defines the criteria for the *Facts*, *Cited Articles*, and *Analysis* scores. All necessary contextual information is provided; the facts of the case, the specific legal question, the ground truth answer with the cited articles and the candidate answer to be evaluated. This context mirrors the information provided to the human annotators for the manual evaluations (§ 5.2). The required output format is clearly specified: the model must provide an explanation for each score, followed by the integer score. The complete prompt is presented in Appendix B (Fig. 5).

3.2 Span LLM-Judge

According to Clark and DeSanctis (2013), rubrics, i.e., instructions that break down an assignment in identifiable components, can significantly improve the consistency of legal writing evaluation. To construct rubrics for our benchmark, our legal expert annotators marked reference spans (colored text in Figure 1) in the ground-truth answer for each score (*Facts*, *Cited Articles*, *Analysis*). These spans contain the information pertinent to their respective dimension. Each span is then annotated with *important* span subsets (usually a few words within the span) that are crucial for an answer to be considered correct (bold text in Table 1). Missing this crucial information should result in a lower score. We opted not to assign specific point values or 'costs' to each important subset, as previous work with rubrics (Starace et al., 2025; Pires et al., 2025), to minimize the annotation burden. The process involves simply highlighting the three scoring dimensions using different colors and then marking the important subsets within those highlighted sections. The LLM-judge is instructed to determine whether the candidate answer covers the information in the spans and then use this assessment to evaluate each scoring dimension. The complete prompt is presented in Appendix B (Fig. 6).

4 Meta Evaluation

Meta-evaluation of *grading* LLM-judges aims to quantify the alignment between LLM-generated scores and human expert annotations. Previous research has predominantly relied on Pearson's or Spearman's correlation coefficients as primary meta-metrics (Bavaresco et al., 2024; Niklaus et al., 2025), often without substantial justification. However, advancements in meta-evaluation have emerged from the machine translation domain, particularly through the WMT Metrics Shared Task (Freitag et al., 2024, 2023), where automatic evaluation frameworks have been systematically compared and refined. The task aims to identify optimal metrics for translation quality assessment by comparing system outputs against references. Recent findings demonstrate that state-of-the-art metrics are increasingly LLM-based. The task has revealed that Pearson's correlation coefficient exhibits vulnerability to outliers, while Spearman's ρ disregards the magnitude of ranking errors, applying uniform penalties. To address these limitations, WMT has adopted Soft Pairwise Accuracy (SPA) (Thompson et al., 2024), a metric that assigns partial credit for nearly correct rankings, thereby providing an evaluation framework that better reflects the alignment of metrics with human experts.

4.1 Soft Pairwise Accuracy (SPA)

SPA measures the degree of alignment in evaluation *confidence* between human experts and LLM-judges (or any other automatic metric). For example, if a human expert is *confident* that one system (e.g., a candidate LLM from GREEKBARBENCH) outperforms another, but the LLM-judge is *uncertain*, SPA penalizes the judge—even if the ranking happened to be correct. To do this, SPA approximates the *confidence* of each judge (human or LLM) on each pairwise comparison between systems using p-values of appropriate permutation tests (Fisher, 1935), as detailed below. We use the original implementation.[7] Formally, SPA between a metric m and human experts h is defined as:

$$SPA(m,h) = \binom{N}{2}^{-1} \sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} \left(1 - \left|p_{ij}^{h} - p_{ij}^{m}\right|\right)$$

where N is the number of systems being evaluated, p_{ij}^{h} is the p-value for the hypothesis that system i is better than system j according to human scores, and p_{ij}^{m} is the corresponding p-value according to the metric under evaluation. The term $\binom{N}{2}^{-1}$ normalizes the summation by the total number of systems under comparison.

SPA permutation tests: To estimate *confidence* of an evaluator (either human or automatic) in a pairwise system comparison, SPA uses permutation tests to calculate the expected mean difference under the null hypothesis that the systems are of equal quality. Specifically, a number of mock systems (1,000 in our experiments, following the original paper) are constructed as follows: for each question in the benchmark, the mock system is assigned either the score of system i or system j at random. The p-value is then computed as the proportion of mock systems for which the differences are greater than or equal to the mean difference between systems i and j, as scored by the evaluator.

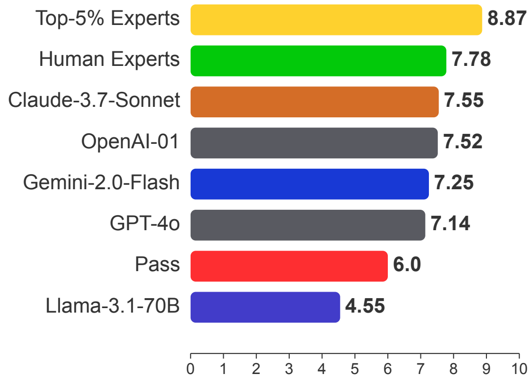

Figure 1: Manual evaluation by legal expert annotators on the semi-private test set of the 2024 exams.

5 Experiments

5.1 Models

Our experiments evaluate a diverse range of LLMs, comprising proprietary models (from OpenAI, Google, Anthropic) and open-weight models; Deepseek-R1 (DeepSeek-AI et al., 2025), Gemma-3 (Team et al., 2025), and Llama-Krikri-8B[8], a model specifically pretrained for the Greek language. We accessed proprietary models and the large open-weight Deepseek-R1 through Application Programming Interfaces (APIs) provided by OpenAI, Google, and AWS. The remaining open-weight models were deployed on a cluster of eight A100 GPUs using the vLLM framework (Kwon et al., 2023). Due to limited resources, we only evaluate a single run for each model. We used the default parameter configurations as specified by each model's provider.

Generation prompt: To generate responses from LLM candidates, we designed a system and user prompt for the questions of the benchmark. The system prompt instructs the LLM to answer with citations to Greek statutory articles. The user prompt is structured to first describe the overall task, including clear instructions on the expected output format. Then it provides the numbered *facts* of the

[7] https://github.com/google-research/mt-metrics-eval

[8] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct

Model	Simple-J (SPA)	Span-J (SPA)	Cost
GPT-4.1-mini	0.723	**0.856**	$$
GPT-4.1	0.807	0.855	$$$
Gemini-2.0-F	0.751	0.794	$
L-Krikri-8B	0.747	0.751	-
Gemma-3-27B	**0.819**	0.749	-
Gemini-2.0-L-F	0.695	0.708	$
GPT-4.1-nano	0.542	0.372	$

Table 4: Comparison of LLMs-judges on GBB-JME, using Simple-Judge and Span-Judge. Cost for input tokens per 1M is indicated as follows: $ (less than $0.3), $$ (less than $1), and $$$ (less than $3).

case, the *question* and the *relevant legal context*. The original prompts are available in Appendix B.

5.2 Manual Evaluation by Legal Experts

In this section we present the manual evaluations that we collected for GBB-JME (§ 5.3), our Judge Meta-Evaluation benchmark for assessing LLM-judges on GREEKBARBENCH. We obtain ground truth evaluations (*Facts*, *Cited Articles*, *Analysis* scores on a scale of 1 to 10) from two expert legal annotators—licensed Greek lawyers with law degrees and practical experience. The annotators were compensated for their time and expertise. They evaluated five LLMs on 87 questions drawn from three exam sessions (2024-A, 2023-A, and 2023-B), resulting in a total of 1,305 annotated samples. The models evaluated on all three exams were Claude-3.7-Sonnet, OpenAI-o1, GPT-4o, and Gemini-2.0-Flash. For the 2024 exam, we included the open-source Llama-3.1-70B; however, due to its poor performance and generation failures on several questions, we replaced it with Deepseek-R1 for the 2023-A and 2023-B exams. Annotations were managed with the open-source platform *doccano*.[9]

The average Krippendorff's α (Krippendorff, 2011) between the two annotators on the three-dimensional scores was 0.74, and the SPA was 0.85, both indicating a substantial level of inter-annotator agreement (Artstein and Poesio, 2008). For the SPA calculation, we treated one annotator's scores as ground truth and compared the other annotator's scores against them. This differs from Section 5.3, where SPA measures the correlation between LLM-generated scores and the aggregated scores of human annotators.

The results for the 2024 exam are shown in Figure 1. All LLMs except Llama-3.1-70B pass the exam. However, all LLMs lag behind the average human expert performance and the 95th percentile ('top-5%'). The best-performing models are Claude-3.7-Sonnet (7.55) and OpenAI-o1 (7.52).

[9] https://doccano.prio.org

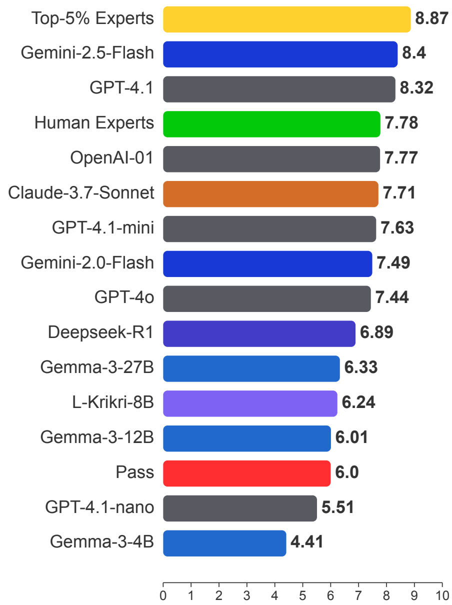

Figure 2: Comparison of closed and open-weight LLMs on GREEKBARBENCH with GPT-4.1-mini Span-Judge.

5.3 Judge Meta-Evaluation (GBB-JME)

We evaluate seven LLMs as judges on our benchmark for meta-evaluation, GBB-JME, using both the *Simple-Judge* prompt (§3.1) and the *Span-Judge* prompt with span-based rubrics (§3.2). The results are presented in Table 4. Generally, leading models (GPT-4.1, Gemini-2.0-Flash and GPT-4.1-mini) showed significant improvement when utilizing the *Span-Judge* prompt, while weaker models (Llama-Krikri-8B, Gemma-3-27B, Gemini-2.0-Lite-Flash, GPT-4.1-nano) struggled with its complexity. GPT-4.1-mini and GPT-4.1 achieved the best performance at 0.856 SPA, and 0.855 SPA, respectively. The competitive performance of the smaller model, aligns with observations reported elsewhere (Niklaus et al., 2025).

Due to its strong performance and lower cost, we adopted GPT-4.1-mini as the judge for all subsequent evaluations. The total cost for all evaluations using GPT-4.1-mini was under $60. The open-weight Gemma-3-27B model paired with the *Simple-Judge* prompt serves as a cost-effective alternative, but we also encourage researchers to evaluate new LLMs as judges on the publicly available GBB-JME benchmark.

Models	Civil Law				Criminal Law				Commercial Law				Public Law				Lawyers' Code				Overall			
	f	c	a	avg	f	c	a	avg	f	c	a	avg	f	c	a	avg	f	c	a	avg	f	c	a	avg
Top-5%	-	-	-	9.00	-	-	-	10.00	-	-	-	10.00	-	-	-	9.20	-	-	-	9.18	-	-	-	8.87
Experts	-	-	-	6.80	-	-	-	8.29	-	-	-	8.70	-	-	-	7.70	-	-	-	7.39	-	-	-	7.78
Gemini-2.5-F	**8.8**	**8.4**	**8.4**	**8.53**	8.4	7.9	**8.5**	**8.28**	**8.6**	8.0	**8.3**	**8.27**	**8.7**	**8.0**	**8.1**	**8.28**	**8.9**	**8.5**	**8.5**	**8.62**	**8.7**	**8.2**	**8.4**	**8.40**
GPT-4.1	**8.8**	8.1	**8.4**	8.44	**8.5**	**8.2**	8.2	**8.28**	8.4	**8.1**	**8.3**	**8.27**	8.6	7.7	**8.1**	8.14	8.8	8.2	8.4	8.48	8.6	8.0	8.3	8.32
Claude-3-7	8.5	7.2	7.4	7.72	8.2	6.9	7.0	7.37	7.6	6.9	7.4	7.31	8.6	7.2	7.6	7.79	8.5	8.2	8.2	8.29	8.3	7.3	7.5	7.71
GPT-4.1-mini	8.3	7.1	7.3	7.57	7.7	6.4	6.9	7.01	**8.4**	7.4	7.4	7.76	**8.4**	7.3	7.6	7.75	**8.5**	7.6	7.9	7.98	**8.3**	7.2	7.4	7.63
Gemma-3-27B	7.8	5.5	5.9	6.39	6.7	4.9	4.9	5.51	6.8	5.2	5.6	5.88	**8.1**	5.9	6.1	6.68	7.8	**6.6**	**6.6**	7.01	7.5	5.7	5.8	6.33
L-Krikri-8B	7.0	5.3	5.6	5.95	7.1	5.3	5.1	5.84	6.7	5.3	5.4	5.79	7.9	**6.3**	**6.2**	6.78	7.4	6.4	6.4	6.74	7.2	5.7	5.8	6.24

Table 5: Fine-grained comparison of proprietary and small open-weight LLMs on different legal areas: 'Civil', 'Criminal', 'Commercial', 'Public', 'Lawyers'; and for different scoring dimensions: 'Facts' (f), 'Cited Articles' (c), 'Analysis' (a). 'Experts'/'Top-5%' is the average/95th percentile score of candidate lawyers. Best LLM scores are in bold, failed scores are highlighted in red, and scores outperforming experts are highlighted in green.

5.4 Results on GreekBarBench

We conduct an extensive automatic evaluation of 13 LLMs on GreekBarBench (Figure 2). We use GPT-4.1-mini as the LLM-Judge, employing the 'Span LLM-Judge' prompt (§ 3.2). The evaluation includes proprietary models such as GPT-4o, the GPT-4.1 family (GPT-4.1-mini, GPT-4.1-nano), Gemini-2.0-flash, and Claude-3.7-Sonnet (with reasoning disabled), along with the reasoning models OpenAI-o1 and Gemini-2.5-Flash. The open-weight models include the Gemma-3 family(Gemma-3-27B, Gemma-3-12B, and Gemma-3-4B), the specialized Greek model Llama-Krikri-8B-Instruct (Krikri-8B), and the reasoning model DeepSeek-R1.

The experimental results (Figure 2) reveal that Gemini-2.5-Flash (8.4) and GPT-4.1 (8.32) demonstrate the strongest performance on GreekBarBench. They surpass the average legal expert score (7.78), though they still fall short of the 95th percentile (top-5%) of experts (8.87). OpenAI-o1 (7.77) and Claude-3.7-Sonnet (7.71), perform comparably to the average human expert (7.78). The fact that Gemini-2.5-Flash and OpenAI-o1, two reasoning models, are among the top performers, shows that leveraging inference-time reasoning is a key factor for this benchmark. The smallest models, GPT-4.1-nano (5.51) and Gemma-3-4B (4.41) are the only models that fail the exams (passing score: 6.00). The 8B Krikri model surpasses Gemma-12B and achieves performance comparable to the significantly larger Gemma-27B, highlighting the benefit of language-specific pretraining.

Table 5 presents a fine-grained comparison for four of the best proprietary LLMs (Gemini-2.5-Flash, GPT-4.1, Claude-3.7-Sonnet , GPT-4.1-mini) and two small open-weight LLMs (Gemma-3-27B, L-Krikri-8B), reporting their scores across five different legal areas (*Civil*, *Criminal*, *Commercial*, *Public*, *Lawyers*) and three scoring dimensions (*Facts*, *Cited Articles*, *Analysis*).

The fine-grained comparison shows that LLMs exhibit consistent performance across all legal areas and scoring dimensions. Notably, in the areas of 'Civil Law', 'Public Law', and 'Lawyers' Code', four of the LLMs achieve scores surpassing the average human expert performance highlighted in green in Table 5. In contrast, the smaller open models, Gemma-3-27B and L-Krikri-8B (indicated by red), struggle in certain areas, failing to meet the passing grade threshold of 6.00. The second-best model, GPT-4.1, matches the top performer, Gemini-2.5-Flash, in 'Criminal' and 'Commercial Law', but Gemini-2.5-Flash achieves slightly higher scores in the remaining three areas. Interestingly, even though 'Civil Law' seems to present the greatest challenge for human experts (average score of 6.80), this is not the case for LLMs.

Analyzing performance by dimension provides valuable insights into model capabilities. A key finding is that all models struggle most with the 'Cited Articles' dimension. The higher overall scores of Gemini-2.5-Flash and GPT-4.1 are largely attributable to their stronger performance in the *Cited Articles* and *Analysis* dimensions, when compared to Claude-3.7-Sonnet and GPT-4.1-mini. Furthermore, Gemma-3-27B's stronger performance compared to L-Krikri-8B is primarily attributable to its higher scores in the *Facts* dimension, while they get identical scores in the other dimensions.

6 Related Work

Legal domain: In the legal domain, LexGLUE (Chalkidis et al., 2022) and LEXTREME (Niklaus et al., 2023) are established benchmarks for legal classification tasks. LegalBench (Guha et al., 2023) is the standard for evaluating LLMs on legal reasoning via multiple-choice questions. More closely related to our work, task 5.4 of LexEval (Li et al., 2024) uses a similar legal examination for Chinese instead of Greek, but, unlike our approach, LexEval does not provide citations or use LLM-as-a-judge, instead evaluating with the less reliable, overlap-based ROUGE metric (Cohan and Goharian, 2016). LLeQA (Louis et al., 2024) collects pairs of everyday legal questions and answers, including citations to French statutory law articles, but they evaluate using the METEOR metric without measuring its correlation with human experts. CaseGen (Li et al., 2025) on the other hand, assesses document drafting and legal judgment generation in Chinese using the LLM-as-a-judge approach. While they do measure agreement between human and LLM evaluations, they do not compare different prompts or models. Concurrent work, OAB-Bench (Pires et al., 2025), uses data from the Brazilian Bar Examination and also provides the official guidelines as rubrics for the LLM-judges. They do evaluate different LLM-judges, but they only do it for three samples and they do not provide citations to statutory articles. Notably, the complexity of their rubrics necessitates the use of the expensive OpenAI-o1 model for evaluation, thus significantly increasing the overall cost, amounting to approximately $50 for each LLM evaluated.

LLM-as-a-judge: LLM-as-a-judge was introduced by Zheng et al. (2023), who meta-evaluated its performance against human preferences for multi-turn chat assistant dialogues. A comprehensive overview of LLM-as-a-judge and meta-evaluation resources can be found in the survey by Gu et al. (2024). Taking this concept further, JudgeBench (Bavaresco et al., 2024) introduced a general-purpose benchmark specifically for the meta-evaluation of LLM-judges. In line with our approach, other studies similarly develop separate benchmarks to meta-evaluate judges on specific tasks (Starace et al., 2025; Niklaus et al., 2025).

Evaluation Rubrics: Legal research has for long focused on creating rubrics for consistent (human) evaluation of legal writing (Clark and DeSanctis, 2013). The Brazilian Bar exams have made their rubrics for human evaluation available, so the aforementioned OAB-Bench (Pires et al., 2025) provides them to their LLM-judges. Their rubrics consist of a manually annotated ground truth answer with comments and a table with score distributions for each element of the answer. A proprietary benchmark, BigLawBench[10], describes a scoring system that uses two dimensions: the 'source' and 'answer' scores, which are analogous to our *Cited Articles* and *Analysis*. They rely on detailed instructions per question that specify explicitly the attributes that would contribute positively and negatively to the final score of candidate answers. Constructing from scratch either of these approaches is prohibitively expensive, in contrast to our simple, span-based rubrics that only require minimal annotation effort.

Greek NLP: Important Natural Language Processing resources for the Greek language include classification models (Koutsikakis et al., 2020; Saketos et al., 2024), alongside more recent LLMs pretrained on Greek like Meltemi[11] (Voukoutis et al., 2024) and Llama-Krikri[12], which we tested in our experiments (§ 5.4). Existing Greek legal datasets cover only classification and summarization tasks (Angelidis et al., 2018; Papaloukas et al., 2021; Koniaris et al., 2023). Although Greek LLM benchmarks exist for other domains, such as finance (Peng et al., 2025) and medicine (Voukoutis et al., 2024), the legal domain currently lacks one.

7 Conclusions

In this work, we introduced GREEKBARBENCH, a benchmark evaluating LLMs on legal questions requiring citations to statutory articles and case facts. We use a comprehensive scoring system and an LLM-judge for automatic evaluation. To ensure judge alignment with human experts, we developed an accompanying meta-evaluation benchmark (GBB-JME) using Soft-Pairwise Accuracy as the meta-metric. The results show that our span-based rubrics specifically designed for this benchmark improve the LLM-judges. The extensive evaluation of 13 LLMs on GREEKBARBENCH revealed that Gemini-2.5-Flash and GPT-4.1 achieved the best performance, surpassing the typical human expert, but also highlighted areas for future improvement.

Limitations

Our benchmark, GREEKBARBENCH, assumes the availability of the relevant legal code chapters for

[10] https://www.harvey.ai/blog/introducing-biglaw-bench

[11] Meltemi was excluded from our experiments, because of its relatively small context length of 8 billion tokens.

[12] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct

the *Relevant Legal Context* component (§ 2.4). We did not evaluate the performance of retrieval models on this task, which is a critical step in real-world legal applications and could pose a significant challenge not addressed by our current setup.

A notable limitation is the cost associated with evaluating models using our framework due to the primary LLM-judge being a proprietary model (GPT-4.1-mini). To mitigate this cost, we suggest utilizing Simple-Judge with the open-weight model Gemma-3-27B. While no currently available open-weight model achieves meta-evaluation performance (SPA scores on GBB-JME) on par with GPT-4.1-mini, our public release of the benchmark and meta-evaluation dataset will allow future research to test and use more accurate and cost-effective LLM-judges.

Finally, the reported legal expert performance figures (average and 95th percentile) in our comparisons (§ 1, § 2) should be interpreted as illustrative baselines rather than rigorous head-to-head comparisons under identical evaluation conditions. This is due to inherent limitations in the available human data: detailed participant statistics are not available across all exam papers, and critically, the human scores were determined by the official Greek Bar Examination grading committee, not by our developed LLM-judge framework. Nevertheless, we believe these figures provide valuable intuition regarding the current performance gap between state-of-the-art LLMs and candidate lawyers.

Ethical Considerations

The development and application of legal NLP benchmarks carry significant ethical implications and potential societal impact, particularly concerning fairness, access to justice, and responsible automation (Tsarapatsanis and Aletras, 2021). Therefore, careful consideration of their design and potential uses is essential.

Our research contributes to the development of tools that could potentially assist various types of users, including legal professionals (such as judges and lawyers), students, and individuals seeking to understand legal concepts. It is crucial to emphasize that performance on this benchmark, or any similar research benchmark, should never be considered sufficient justification for deploying automated systems that substitute human experts. We strongly caution against the uncritical reliance on models evaluated solely on benchmark performance for automating legal tasks, making legal decisions, or providing legal advice.

Despite our efforts to make GreekBarBench realistic, as a research benchmark, it overlooks two critical aspects for the safe and reliable deployment of legal AI applications in practice:

- **Data Realism**: Real-world legal problems are far more complex and nuanced than the structured, often simplified scenarios found in exam questions (Medvedeva and Mcbride, 2023). They often demand significant legal interpretation, ethical judgment and persuasion, particularly when the law does not provide an explicit answer for a given situation.

- **Safety:** Real-world applications must ensure that the AI system handles adversarial attacks effectively. Issues like guiding the decisions of the LLMs with malicious prompting (e.g., jailbreaking), and providing confident, incorrect information when asked legally unanswerable queries are unacceptable (see discussions on AI safety principles [13]).

Furthermore, the primary ethical purpose of this work is not to provide a system ready for deployment, but to advance the state of legal NLP evaluation itself. By developing a benchmark that requires free-text generation, incorporates a multi-dimensional scoring system, and uses LLM-judges with explicit evaluation criteria, we aim to encourage the development of more transparent and explainable legal AI models. These features provide greater insight into how models arrive at their answers, moving beyond simple classification or multiple-choice and offering components of explainability which are crucial for gaining trust in AI applications (Medvedeva and Mcbride, 2023).

As already mentioned (§ 2.1), the authors of the solutions of the exam papers have given approval for the public reproduction of this work, with respect to the original and strictly for academic research use. Our ground truth answers are based on the year that each exam was published. This means that if the relevant laws changed in the meantime, the solutions are no longer valid. All cases in the Greek Bar exams are fictional, created solely for educational purposes, and bear no relation to real individuals or actual legal cases.

[13]https://www.anthropic.com/news/core-views-on-ai-safety

Acknowledgments

We are grateful to our legal expert annotators, Nasia Makridou and Irene Vlachou, for their diligent work, expertise, and insightful discussions, which were invaluable to this project.

This work was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. AWS resources were provided by the National Infrastructures for Research and Technology (GRNET), with support from the EU Recovery and Resilience Facility.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named entity recognition, linking and generation for greek legislation. In *Legal Knowledge and Information Systems - JURIX 2018: The Thirty-first Annual Conference, Groningen, The Netherlands, 12-14 December 2018*, volume 313 of *Frontiers in Artificial Intelligence and Applications*, pages 1–10. IOS Press.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *CoRR*, abs/2406.18403.

Rohan Bhambhoria, Samuel Dahan, Jonathan Li, and Xiaodan Zhu. 2024. Evaluating ai for law: Bridging the gap with open-source solutions. *Preprint*, arXiv:2404.12349.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Jessica Clark and Christy DeSanctis. 2013. Toward a unified grading vocabulary: Using rubrics in legal writing courses. *J. Legal Educ.*, 63.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, and Jin Chen et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Ronald Aylmer Fisher. 1935. *The Design of Experiments*. Oliver & Boyd.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv: 2411.15594*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Marios Koniaris, Dimitris Galanis, Eugenia Giannini, and Panayiotis Tsanakas. 2023. Evaluation of automatic legal text summarization techniques for greek case law. *Information*, 14(4).

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Departmental Papers (ASC) 43, Annenberg School for Communication, University of Pennsylvania.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, Albuquerque, New Mexico. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 25061–25094. Curran Associates, Inc.

Haitao Li, Jiaying Ye, Yiran Hu, Jia Chen, Qingyao Ai, Yueyue Wu, Junjie Chen, Yifan Chen, Cheng Luo, Quan Zhou, and Yiqun Liu. 2025. Casegen: A benchmark for multi-stage legal case documents generation. *Preprint*, arXiv:2502.17943.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22266–22275.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. 2024. Better call gpt, comparing large language models against lawyers. *Preprint*, arXiv:2401.16212.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 3016–3054, Singapore. Association for Computational Linguistics.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. Swiltra-bench: The swiss legal translation benchmark. *Preprint*, arXiv:2503.01372.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on Greek legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 63–75, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. Plutus: Benchmarking large language models in low-resource greek finance. *Preprint*, arXiv:2502.18772.

Ramon Pires, Roseval Malaquias Junior, and Rodrigo Nogueira. 2025. Automatic legal writing evaluation of llms. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*.

Vasileios Saketos, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2024. The large language model greeklegalroberta. In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, SETN '24, New York, NY, USA. Association for Computing Machinery.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *Preprint*, arXiv:2504.01848.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, and Alaa Saade et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The first open large language model for greek. *Preprint*, arXiv:2407.20743.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Annotator Instructions

In this section we present the instructions given to the two legal expert annotators. The annotators possessed prior experience with the evaluation task, having previously taken the exams themselves. This existing expertise allowed for concise instructions. For the general evaluation of LLM-generated answers for the manual evaluation (§ 5.2), the instruction (translated to English) was simply to:

> "Evaluate the candidate answers on each scoring dimension (*Facts*, *Cited Articles*, and *Analysis*)."

For the creation of text spans for Span-Judge (§ 3.2), annotators were instructed to:

> "Highlight the text-spans that correspond to each scoring dimension (*Facts*, *Cited Articles*, and *Analysis*). Highlight the most important subsets of these spans with the label *important*."

B Complete Prompts

In this section we present the complete system (Fig.3) and user (Fig.4) prompts given to candidate LLMs for generation of answers, as well as the complete system prompts given to LLM-judges for the Simple- (Fig.5) and Span-Judge (Fig.6).

You are a legal assistant who answers questions in Greek, focusing on the legal system and the laws of Greece. You analyze your reasoning and respond with well-supported answers and correct references. You only respond in txt format and with only one short paragraph without headings.

Figure 3: System prompt for generation given to candidate LLMs.

You are given the numbered facts of a legal case, the current relevant legislation of Greece, and a question regarding this case. After carefully reading the entire text, you are to provide a comprehensive answer to the question, analyzing your reasoning. You should answer with references to the relevant legislation using the appropriate abbreviations for the laws (for example, you can say: "according to article X CC" to refer to article "X" of the Civil Code), where necessary. Additionally, you must provide references to the facts of the case (for example, you can say: "according to statement Y of the case data"), where necessary.

Figure 4: User prompt for generation given to candidate LLMs.

You are a legal exam evaluator. You will be given the following:

1. The facts of a case, 2. The relevant legislation, 3. A question, 4. An ideal reference answer, 5. An answer to be evaluated

You must evaluate the answer with three scores and an explanation for each. Each score consists of an integer from 1 to 10, with 10 being excellent. The reference ideal answer is considered excellent (10 in all categories). The Facts Score concerns the facts of the case. If the ideal reference answer mentions certain specific facts from the case, but the answer to be evaluated does not mention them, points should be deducted. Similarly, if the answer to be evaluated mentions facts that are not useful for the answer, points should also be deducted. The Legislation Score concerns references in the answer to the relevant articles of the laws. It is essential to refer to specific articles of laws. If such references are missing or if incorrect articles are cited, points should be deducted from the Legislation Score. Also, points should be deducted if the interpretation of the law is incorrect. The Analysis Score concerns a more general evaluation as to whether the answer to be evaluated has covered the original question, with correct and valid legal reasoning. Points already given or deducted for the above categories should not be counted here. At this point, the final conclusion of the answer is also evaluated. If the answer to be evaluated reaches the wrong conclusion or omits some critical argument, points should be deducted. Use plain text, without markdown. Your answer should follow the template shown below, where X, Y, Z are integers (1-10):

Explanation of the Facts Score: <your explanation for the score ...>, Facts Score: X

Explanation of the Legislation Score: <...>, Legislation Score: Y

Explanation of the Analysis Score: <...>, Analysis Score: Z

Figure 5: System prompt given to Simple-Judge LLM-judges.

You are a legal exam evaluator. You will be given the following:

1. The facts of a case, 2. The relevant legislation, 3. A question, 4. An ideal reference answer, 5. An answer to be evaluated, 6. The evaluation spans (json file)

The evaluation spans are verbatim spans from the text of the ideal reference answer with tags referring to each of the three scores (facts, rules, analysis). That is, for evaluating each score, emphasis should be placed on whether the information from the corresponding excerpt of the ideal reference answer is present in the answer to be evaluated, thus allowing for an appropriate score. For example, for the Facts score, the spans should be present in the answer to be evaluated. If no spans are present, it means that no very important facts (or laws or analysis) are absolutely necessary to be mentioned. However, points can still be deducted if the answer to be evaluated adds facts (or laws or analysis) that are incorrect. There are also important spans, which indicate which parts of the answer are crucial for the evaluation. You must evaluate the answer with three scores and an explanation for each. Each score consists of an integer from 1 to 10, with 10 being excellent. The reference answer is considered excellent (10 in all). The Facts score concerns the facts of the case. If the ideal reference answer mentions specific facts from the case, but the answer to be evaluated does not mention them, points should be deducted. Similarly, if the answer to be evaluated mentions facts that are not useful for the answer, points should also be deducted. The Legislation score concerns the references in the answer to the relevant articles of laws. It is essential to refer to specific articles of laws. If such references are missing or if incorrect articles are cited, points should be deducted from the Legislation score. Also, points should be deducted if the interpretation of the law is incorrect. The Analysis score concerns a more general evaluation as to whether the answer to be evaluated has covered the original question, with correct and valid legal reasoning. Points already given or deducted for the above categories are not scored here. At this point, the final conclusion of the answer is also evaluated. If the answer to be evaluated reaches the wrong conclusion or omits some critical argument, points should be deducted. Use plain text, without markdown. Your answer should follow the template shown below, where X, Y, Z are integers (1-10):

Explanation of the Facts Score: <your explanation for the score ...>, Facts Score: X

Explanation of the Legislation Score: <...>, Legislation Score: Y

Explanation of the Analysis Score: <...>, Analysis Score: Z

Figure 6: System prompt given to Span-Judge LLM-judges.

C Complete Dataset Example

In this section we present the complete version of the example that we presented in Table 1. We show the complete *Facts* and *Question* (Fig. 7), the *Relevant Legal Context* (Figures 8 and 9), the complete *Ground Truth Answer* (Fig. 10), the candidate answer by Gemini-2.5-Flash (Fig. 11) and *evaluations* of the candidate answer by the legal experts and the LLM-judge (Fig.12).

Facts:

[1] Μετά από προεξέταση και προσυνεννόηση με τον δερματολόγο του, κ. Ιωάννη (Ι), ο ασθενής Αντώνης (Α), 20χρονος φοιτητής, μετέβη στις 30-10-2011 στο ιατρείο του Ι προς αφαίρεση δερματικών θηλωμάτων στην περιοχή του προσώπου έναντι συμφωνημένης αμοιβής.

Following a preliminary examination and prior consultation with his dermatologist, Mr. Ioannis (I), the patient Antonis (A), a 20-year-old student, went on 30-10-2011 to I's clinic to remove skin papillomas on the face for an agreed fee.

[2] Πριν από την αφαίρεση ο Ι συνέστησε, όπως συνηθίζεται σε παρόμοιες περιπτώσεις, την πλύση του σημείου με διάλυμα οξικού οξέως προς εντοπισμό αόρατων θηλωμάτων.

Before removal, I recommended, as is customary in similar cases, washing the area with acetic acid solution to detect invisible papillomas.

[3] Η κυρία Πηνελόπη (Π), επί σειρά ετών βοηθός του Ι, πήρε από το ράφι ένα μπουκάλι με το υγρό και άρχισε να το επαλείφει σε επαρκή ποσότητα στο δέρμα του Α.

Mrs. Pinelopi (P), I's longtime assistant, took a bottle with the liquid from the shelf and began applying it in sufficient quantity on A's skin.

[4] Αμέσως μετά την πρώτη επάλειψη ο Α διαμαρτυρήθηκε για πόνο και η Π σταμάτησε αμέσως τη θεραπεία.

Immediately after the first application, A complained of pain and P immediately stopped the treatment.

[5] Ο Α είχε υποστεί τοπικά εγκαύματα τρίτου βαθμού.

A suffered third-degree local burns.

[6] Όπως αποδείχθηκε εκ των υστέρων, το μπουκάλι περιείχε αυτούσιο οξικό οξύ και όχι διάλυμα, όπως προδιαγράφεται από την θεραπευτική διαδικασία.

As later proven, the bottle contained pure acetic acid and not a solution, as prescribed by the treatment process.

[7] Μετά από θεραπευτική αγωγή αρκετών εβδομάδων από ειδικό εγκαυματολόγο ιατρό θεραπεύτηκαν τα εγκαύματα του Α και στη συνέχεια, χρειάσθηκε να γίνει και πλαστική εγχείρηση στο πρόσωπο, η οποία ήταν επιτυχής και οδήγησε στην πλήρη αποκατάστασή του.

After several weeks of therapeutic treatment by a specialist burn physician, A's burns healed and subsequently a plastic surgery on the face was necessary, which was successful and led to his full recovery.

[8] Ο Α είχε συνολικές ιατρικές δαπάνες 2.500 ευρώ για την αποκατάστασή του και θεωρεί ότι πρέπει να πάρει και 75.000 ευρώ ως χρηματική ικανοποίηση λόγω ηθικής βλάβης.

A had total medical expenses of 2,500 euros for his recovery and considers he should also receive 75,000 euros as compensation for moral damage.

Question:

Ποια πρόσωπα και με βάση ποιες διατάξεις ευθύνονται για τον τραυματισμό του Α;

Which persons and based on which legal provisions are responsible for the injury of A?

Figure 7: Complete *Facts* and *Question* (original and below translated in English), as given in to the candidate LLMs, for the example in Table 1.

Relevant Legislation:

Civil Code (AK)

CHAPTER FOUR - LIABILITY FROM CONTRACTS IN GENERAL

Articles 361 – 373

CHAPTER EIGHTEENTH - EMPLOYMENT CONTRACT

Articles 648 – 680

CHAPTER THIRTY-NINTH – TORTS

Articles 914 – 938

Code of Civil Procedure (ΚΠολΔ)

CHAPTER C (III) - Jurisdiction by subject matter

Articles 12 - 21

CHAPTER IA (XI) - Participation of third parties in the trial

Articles 79 - 93

Greek Constitution

BASIC PROVISIONS

Articles 1 - 2

INDIVIDUAL AND SOCIAL RIGHTS

Articles 4 – 25

Figure 8: The *Chapters* of the *Relevant Legislation* context given to candidate LLMs, for the example in Table 1. The content of the articles is not shown for brevity.

CHAPTER THIRTY-NINTH – TORTS

Article 914

Whoever unlawfully and culpably damages another person is obligated to compensate him.

Article 915

A person is not liable for damages caused without awareness of their actions or while in a mental or intellectual disorder that decisively limited the functioning of their judgment and will.

Whoever, at the time of causing the damage, brought themselves into such a state by consuming alcoholic beverages or other similar means, is liable for the damage, unless they entered that state without fault.

Article 916

A person under ten years of age is not liable for the damage caused.

Article 922

The master or the one who places another in a service (employment) is liable for damage caused unlawfully to a third party by the servant or the placed person during their service.

Article 926

If damage results from a joint act of several persons or if several are jointly liable for the same damage, all are liable severally (jointly and severally). The same applies if several acted simultaneously or successively and it cannot be determined whose act caused the damage.

Article 929

In case of harm to a person's body or health, compensation includes, besides medical expenses and damage already incurred, everything the injured party will lose in the future or spend additionally due to increased expenses. There is also an obligation to compensate a third party who legally had the right to demand services from the injured party and is deprived of them.

Article 932

In the case of a tort, regardless of compensation for property damage, the court may award monetary satisfaction at its discretion for moral harm. This especially applies to one who suffered an injury to their health, honor, or chastity, or was deprived of their freedom. In the event of a person's death, this monetary satisfaction may be awarded to the victim's family due to emotional distress.

Figure 9: Chapter Thirty-Ninth ('TORTS') from the Civil Code, which is part of the *Relevant Legislation* context given to candidate LLMs, for the example in Table 1. The *gold* cited articles are marked in bold and the articles cited by Gemini-2.5-Flash(Figure 11) are underlined.

Ground Truth Answer:

Η αμελής παράλειψη του Ι να μεριμνήσει προκειμένου να μην υπάρχει το μπουκάλι με το επικίνδυνο υγρό στο ιατρείο του ή αυτό να φέρει ακριβή και σαφώς διακριτή ένδειξη για το περιεχόμενο του ή έστω να επιστήσει την προσοχή της Π στο επικίνδυνο υγρό συνιστά αφενός, πλημμελή εκτέλεση υποχρεώσεων από τη σύμβαση και αφετέρου, **αδικοπραξία κατά την 914 ΑΚ** της οποίας το παράνομο στηρίζεται στην παράβαση της γενικής υποχρέωσης πρόνοιας, ασφάλειας και προστασίας που καθιερώνει η έννομη τάξη. Η αδικοπραξία του προστηθέντος κατά την 914 ΑΚ προϋποθέτει ανθρώπινη πράξη, υπαιτιότητα, επέλευση ζημίας, αιτιώδη σύνδεσμο μεταξύ πράξης και ζημίας και τον παράνομο χαρακτήρα της πράξης. Η παρανομία δεν περιορίζεται στην παράβαση ορισμένου κανόνα δικαίου, αλλά εκτείνεται και σε κάθε παράβαση της γενικής υποχρέωσης πρόνοιας, ασφάλειας και προστασίας που απορρέει ως ύψιστη αρχή από την έννομη τάξη μας. Όπως προκύπτει από το πραγματικό, **η Π δεν έλεγξε το περιεχόμενο της φιάλης πριν το επαλείψει στο δέρμα του Α** (παράνομη και υπαίτια πράξη). **Συνεπώς, η Π ευθύνεται κατά τη διάταξη του άρθρου 914 ΑΚ.** Σύμφωνα με τη διάταξη του άρθρου 922 ΑΚ, ο κύριος ή ο προστήσας κάποιον άλλον σε μία υπηρεσία ευθύνεται για τη ζημία που ο υπηρέτης ή ο προστηθείς προξένησε σε τρίτον παράνομα κατά την υπηρεσία του. Θεσπίζεται δηλαδή, αντικειμενική ευθύνη ενός προσώπου για άδικη πράξη άλλου υπό την προϋπόθεση ύπαρξης σχέσης πρόστησης με την ανάθεση από κάποιον σε τρίτο ορισμένης υπηρεσίας που αποβλέπει στην εξυπηρέτηση συμφερόντων του πρώτου και στοιχείου εξάρτησης στην σχέση πρόστησης υπό την έννοια της εξουσίας του προστήσαντος να παρέχει σχετικές οδηγίες και διαταγές στον προστηθέντα. Ως προς την αδικοπραξία του προστηθέντος, η παρανομία δεν περιορίζεται στην παράβαση ορισμένου κανόνα δικαίου, αλλά εκτείνεται και σε κάθε παράβαση της γενικής υποχρέωσης πρόνοιας, ασφάλειας και προστασίας που απορρέει ως ύψιστη αρχή από την έννομη τάξη μας. Εφόσον συντρέχουν οι παραπάνω προϋποθέσεις των 922 και 914 ΑΚ, ο προστήσας ευθύνεται σε αποζημίωση του ζημιωθέντος και αποκατάσταση της ηθικής βλάβης του. Εν προκειμένω **η ζημία του Α προκλήθηκε εντός του ιατρείου του Ι** από παράνομη και υπαίτια πράξη της βοηθού του Π, η οποία είναι προστηθείσα. **Νομικό έρεισμα της ευθύνης του Ι είναι οι 922 και 914 ΑΚ των οποίων οι προϋποθέσεις, όπως προκύπτει από το ιστορικό, συντρέχουν στην υπό εξέταση περίπτωση.**

*The negligent omission of I to ensure that the bottle with the hazardous liquid does not exist at his clinic, or that it has a precise and clearly distinct indication of its contents, or at least draws P's attention to the hazardous liquid, constitutes on one hand poor performance of contractual obligations and on the other **a tort under article 914 of the Civil Code**, whose illegality is based on violation of the general duty of care, safety, and protection established by the legal order. The tort of the agent under 914 requires human act, fault, damage occurrence, causal link between act and damage, and the unlawful character of the act. Illegality is not limited to breach of a specific rule of law but extends to any breach of the general duty of care, safety, and protection which arises as a supreme principle from our legal order. As it emerges from the facts, P did not check the contents of the bottle before applying it to A's skin (an unlawful and culpable act). Therefore, **P is liable according to the provision of Article 914 AK**. According to the **provision of Article 922 AK**, the principal or the person who supervises another in the performance of a service is liable for the damage that the servant or agent unlawfully caused to a third party during the course of their service. This establishes an objective liability of a person for the wrongful act of another provided that there is a supervisory relationship with the delegation by someone to a third party of a specific service aimed at serving the interests of the first and an element of dependence in the supervisory relationship, in the sense of the authority of the principal to give relevant instructions and orders to the agent. As regard the tort of the agent, the illegality is not limited to the violation of a specific legal rule but also extends to any breach of the general duty of care, safety, and protection deriving as a supreme principle of our legal order. Given that the above conditions of Articles 922 and 914 AK are met, the principal is liable to compensate the injured party and restore moral damages. In the present case, **A's damage occurred within I's clinic** due to an unlawful and culpable act of his assistant P, who is the agent. **The legal basis of I's liability is Articles 922 and 914 AK, the conditions of which, as it appears from the facts, are fulfilled in the case at hand.***

Figure 10: Ground truth answer by the legal expert. Text spans are highlighted in colors (green for *Facts*, blue for *Cited Articles* and orange for *Analysis*).

Gemini-2.5-Flash Answer:

Ο δερματολόγος κ. Ιωάννης (Ι) και η βοηθός του κυρία Πηνελόπη (Π) ευθύνονται για τον τραυματισμό του Αντώνη (Α). Ο Ι ευθύνεται πρωτίστως λόγω της συμβατικής του σχέσης με τον Α (δεδομένο 1) για την πλημμελή παροχή των ιατρικών υπηρεσιών, η οποία εκδηλώθηκε μέσω της ενέργειας της προστηθείσας βοηθού του (δεδομένα 3, 6). Παράλληλα, τόσο η Π όσο και ο Ι ευθύνονται εις ολόκληρον κατά τις διατάξεις περί αδικοπραξιών, καθώς η Π προκάλεσε παράνομα και υπαίτια σωματική βλάβη στον Α (δεδομένο 5) εφαρμόζοντας αυτούσιο οξικό οξύ (δεδομένο 6), σύμφωνα με το άρθρο 914 ΑΚ. Ο Ι ευθύνεται και ως προστήσας για την παράνομη πράξη της προστηθείσας του Π, η οποία τελέστηκε κατά την υπηρεσία της (δεδομένο 3), δυνάμει του άρθρου 922 ΑΚ. Καθώς η ζημία προήλθε από κοινή πράξη ή για την ίδια ζημία ευθύνονται παράλληλα περισσότεροι, ενέχονται όλοι εις ολόκληρον σύμφωνα με το άρθρο 926 ΑΚ. Η αποζημίωση περιλαμβάνει τα νοσήλια (δεδομένο 8) κατά το άρθρο 929 ΑΚ και χρηματική ικανοποίηση για την ηθική βλάβη (δεδομένο 8) κατά το άρθρο 932 ΑΚ, λόγω της προσβολής της υγείας του Α.

Dermatologist Mr. Ioannis (I) and his assistant Mrs. Pinelopi (P) are liable for the injury of Antonis (A). I is primarily liable due to his contractual relationship with A (fact 1) for the defective provision of medical services, which manifested through the action of his agent assistant (facts 3, 6). At the same time, both P and I are jointly liable under tort law, since P caused unlawful and culpable bodily harm to A (fact 5) by applying pure acetic acid (fact 6), according to Article 914 of the Civil Code (AK). I is also liable as principal for the unlawful act of his agent P, which was committed in the course of her service (fact 3), pursuant to Article 922 AK. As the damage resulted from a joint act or, for the same damage, several persons are liable simultaneously, all are liable jointly and severally according to Article 926 AK. Compensation includes medical expenses (fact 8) pursuant to Article 929 AK and monetary satisfaction for moral damages (fact 8) according to Article 932 AK, due to the harm to A's health.

Figure 11: The answer of Gemini-2.5-Flash for the example in Table 1. Citations are highlighted in color (green for *Facts* and blue for *Cited Articles*).

Legal Expert Evaluation:

Total score: 10/10

Facts score: 10/10, Cited Articles score: 10/10, Analysis score: 10/10

Span-Judge (GPT-4.1-mini) Evaluation:

Total score: 8/10

Facts score: 7/10

Explanation: The response under evaluation mentions the key facts regarding the role of I and P in causing the damage, as well as the structure of the contract and the service. However, it does not clearly present the crucial point that P did not check the contents of the bottle before application, which is a central element of the case and of her responsibility. Additionally, there is no mention of the fact that the damage occurred within I's clinic or that there was negligence in supervising the hazardous liquid. There are no superfluous or incorrect elements. Overall, the absence of important specific details leads to a moderate grade.

Cited Articles score: 8/10

Explanation: The response under evaluation correctly identifies the relevant provisions of the Civil Code, namely Articles 914 and 922 regarding the liability of the agent and the principal, as well as Articles 926, 929, and 932 concerning joint and several liability and compensation. Although the latter are not mentioned in the reference text, they are accurate and enrich the answer. There are no errors in the interpretation of the law nor omissions of significant provisions. The inclusion of the relevant articles on compensation and moral damages strengthens the legal argument. Therefore, the grade is high.

Analysis score: 9/10

Explanation: The response basically addresses the question with correct legal reasoning and reaches the correct conclusion of the joint liability of I and P. However, it does not provide an extensive analysis of I's negligence and deficient supervision (the factor considered crucial in the reference text). The argumentation remains general without thorough legal grounding according to the legal order and the general duties of care as presented in the reference text. Despite its completeness, the analysis is not entirely sufficient and can be improved in the presentation of the unlawfulness and the causal link. The final conclusion, however, is correct.

Figure 12: Evaluation results for Gemini's answer by Legal Experts and the LLM-Judge (GPT-4.1-mini Span-Judge). The response is perfect according to the legal experts. The LLM-judge is more strict and gives an 8/10 total score.