AILQA: Evaluating AI-Driven Legal Question Answering Systems for the Indian Legal System

Anonymous ACL submission

Abstract

This paper evaluates artificial intelligence models for answering legal questions within the Indian legal system. We call our system **A**rtificial Intelligence for **I**ndian **L**egal **Q**uestion **A**nswering or **AILQA**. Utilizing the OpenAI GPT model as a benchmark, we explore the performance of various AI-driven QA algorithms. Our findings highlight the high accuracy of AILQA systems in interpreting natural language queries and generating responses, especially within the complex Indian criminal justice domain. We also present a comprehensive evaluation methodology to assess these systems rigorously. Feedback from legal professionals enriches our analysis, providing insights into the practical applications and limitations of AI in legal QA. The study underscores the need for more research and careful selection of AI models to enhance the efficacy of legal QA systems in India.

1 Introduction

Question Answering (QA) is an AI task that uses NLP to understand and respond to queries in natural language, akin to human interaction (Allam and Haggag, 2012; Choi et al., 2018). Enhanced by deep learning technologies like the Generative Pretrained Transformer 3 (GPT-3) and BERT (Devlin et al., 2018; Qu et al., 2019; Wang et al., 2019; Kassner and Schütze, 2020), QA systems have shown great promise in extracting relevant information from vast, unstructured datasets. These systems are increasingly applied across various domains such as healthcare, customer service, and education, significantly improving the efficiency of information processing and service delivery.

However, building effective legal QA systems poses several challenges, such as dealing with complex and diverse legal language, recognizing the context of legal cases, understanding the nuances of legal reasoning, etc. These challenges are particularly significant in the Indian legal domain, which has a unique legal system and language that differ significantly from other legal systems worldwide.

Our study focuses on criminal cases in English due to resource and time constraints associated with hiring legal experts to evaluate other types of legal cases, such as civil or family law cases. However, we believe our results provide valuable insights into the potential of QA models in the Indian legal domain and can be extended to other legal domains with appropriate evaluation mechanisms. Our study explores various combinations of embedding and QA models specifically tailored for Indian legal question answering, leveraging the state-of-the-art LLM-based Generative Pretrained Transformer (GPT-3 model) (Brown et al., 2020). We evaluate these models using both lexical and semantic metrics, enriched by expert legal feedback. This paper presents a thorough analysis, revealing that specific model combinations not only enhance the accuracy of responses but can also surpass the capabilities of human legal experts in some scenarios. By illustrating the potential of AI in transforming legal QA within the Indian context, we aim to open new avenues for future technological enhancements in legal practices. For the sake of reproducibility, we have made the AILQA dataset and the code for our prediction and explanation models accessible via an anonymous link.[1]

2 Dataset

2.1 Documents Collection and Preprocessing

The dataset comprises thousands of documents pertaining to criminal law, encompassing acts listed in Appendix 3 in Table 3. These acts have been obtained from the IndiaCode[2] website. Additionally, various articles and blogs related to criminal law have been scrapped from websites such

[1] https://anonymous.4open.science/r/Legal-QA-727F/

[2] indiacode.nic.in

Data	Word Count(Avg)	No. of Docs
Judgements	4021	6942
Acts	28705	15
Articles	1557	264

Table 1: Statistical overview of various Criminal Law document distributions

as Mondaq[3] and LawyersClubIndia.[4] The dataset also includes Supreme Court Judgments, scrapped from IndianKanoon,[5] related to Criminal cases spanning from 1947 to 2020, amounting to a total of 7,221 documents. The preprocessing phase cleanses these documents by removing extraneous elements such as line breaks, spaces, headers, and footers. A breakdown of the documents is provided in Table 1, detailing various criminal law document distributions.

2.2 Test Data

To evaluate the performance of various answer generation and document retrieval models within our legal QA system, we compiled a test dataset from the VidhiKarya[6] website. This dataset includes 50 legal queries along with expert responses covering topics like anticipatory bail, cybercrime, juvenile issues, and sex crimes. The answers provided by legal experts on VidhiKarya serve as our ground truth, enabling a direct comparison between the generated answers and expert responses. Each question is paired with its corresponding expert answer, facilitating straightforward evaluation of our models' performance.

3 Methodology

This section outlines our context-based QA system designed for the legal field, which integrates user queries with legal documents through an LLM to deliver precise answers. The process flow of our system is depicted in Figure 1.

3.1 Embedding-Based Retrieval System

3.1.1 Chunking

The Langchain Framework's CharacterTextSplitter[7] facilitates efficient retrieval by creating 1000-character document chunks. If a chunk exceeds, it remains intact; smaller chunks may merge with adjacent ones. Overlapping by 250 characters ensures seamless information flow between chunks, enhancing coherence and relevance. This approach focuses on key document segments, improving retrieval while maintaining contextual coherence.

Our system utilizes the ChromaDB,[8] a vector store database that encodes documents into multi-dimensional embedding vectors. These vectors capture the semantic relationships between texts, facilitating the retrieval of the most relevant documents based on semantic search algorithms.

3.1.2 Embedding Generation Model

OpenAI's Embedding: We use OpenAI's 'Ada' model,[9] which generates 1536-dimensional embeddings. The cost is approximately \$0.0004 per 1000 tokens. For our dataset of 61.6 million tokens, creating these embeddings costs around \$24.7.

Instructor-XL Embedding: The open-source Instructor-XL model[10] produces 768-dimensional embeddings, optimized for instructional tasks within legal domains.

3.1.3 Query Processing & Document Retrieval

We employ LLMs like OpenAI's 'Ada' or Instructor-XL to convert user queries into embeddings that align with our vector store database, ChromaDB. This setup allows for efficient retrieval using similarity search techniques, including Euclidean and Cosine metrics, and employs approximate neighbor search methods (Liu et al., 2004) to enhance efficiency and overcome traditional search limitations. The system ranks documents by similarity and selects the top-k chunks, along with the query, which are integrated into the answering system, which utilizes the contextual data to produce precise and relevant answers.

3.1.4 Answer Generation

In the answer generation phase, we leverage several generative models, including OpenAI's GPT-3 (Davinci)[11], Google's Flan-UL2 (Tay et al., 2023), and META's LLama 2 (Touvron et al., 2023). Each model is guided by specific prompts to ensure the answers are contextually appropriate and precise. Details on the models and the prompts used to guide their responses are provided in Appendix A.

[3] mondaq.com/5/India/Criminal-Law
[4] lawyersclubindia.com/articles
[5] indiankanoon.org
[6] vidhikarya.com/free-legal-advice
[7] python.langchain.com/text_splitters
[8] https://docs.trychroma.com
[9] platform.openai.com/docs/guides/embeddings
[10] huggingface.co/hkunlp/instructor-xl
[11] platform.openai.com/docs/models/gpt-3

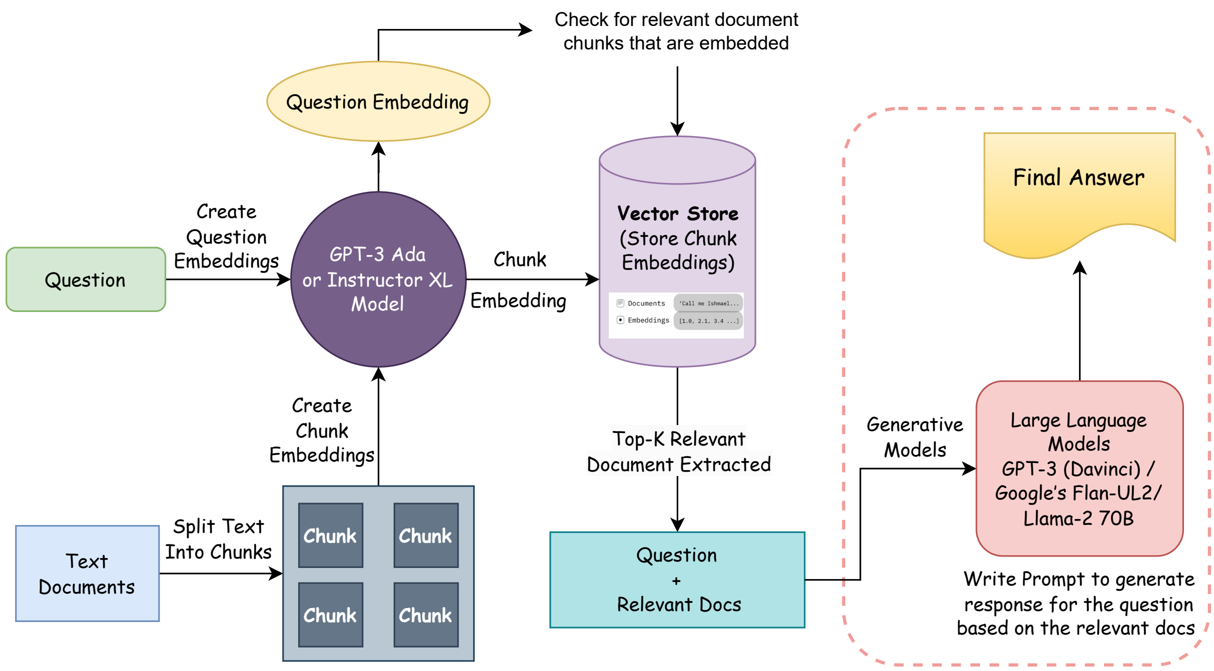

Figure 1: Diagram illustrating the Legal QA System process, highlighting the use of GPT-3 Ada and Instructor XL for context extraction and responses generation with GPT-3 (Davinci), Flan-UL2, and LLaMa-2 70B, guided by specific prompts for these generative models.

4 Evaluation Metrics

We employed several methods to evaluate the performance of our question-answering system:

1. **Lexical Based Evaluation:** We used Rouge scores (1, 2, & L) (Lin, 2004) and the BLEU Score (Papineni et al., 2002). These metrics assess the similarity between the generated answers and the reference answers based on word overlap and order.

2. **Semantic Similarity Based Method:** For assessing semantic similarity, we used the mpnet (Song et al., 2020) base v2 sentence transformer model from HuggingFace[12], which maps sentences into a 768-dimensional vector space, allowing for detailed comparisons of semantic closeness.

3. **Expert Evaluation:** We incorporated human evaluation, in which law experts assessed the answers generated by our model compared to the ground truth. Legal expert reviewed the quality of the answers and rated them on a 1–5 Likert scale based on the following criteria:

   (a) The answer is entirely incorrect or fails to provide any answer.

   (b) The model misunderstood the question and did not offer a relevant response.

   (c) The answer is partly accurate but overlooks essential details.

   (d) A comparable, relevant answer to the ground truth.

   (e) The answer is entirely accurate and relevant, providing a superior response to the expert's answer.

4. **Statistical Significance:** A statistical analysis was conducted on the MPNET similarity scores to determine the significance of performance differences between models, with a p-value of 0.05 or lower marking significant results, suggesting meaningful differences rather than random variations.

[12] huggingface/sentence-transformers/all-mpnet-base-v2

5 Results and Analysis

The data presented in Table 2 evaluates the performance of various generative models for legal question-answering using a multifaceted approach:

5.1 Lexical Based Evaluation

We noticed significant performance improvements, especially when using combinations of Davinci with Ada or Instructor, and independently with ChatGPT and LLama2-70b, as shown by high

Embedding Model	Generative Model	Lexical Based Evaluation				Semantic Evaluation	Expert Evaluation
		Rouge-1	Rouge-2	Rouge-L	BLEU	MPNET Score	Rating Score
N/A	Davinci	**0.267**	0.052	**0.158**	0.010	0.561	3.54
N/A	LLama2-70b	0.149	0.035	0.090	0.007	**0.611**	3.50
Ada	Davinci	0.242	**0.062**	0.147	**0.022**	0.566	**3.74**
Instructor	Davinci	0.229	0.053	0.139	0.016	0.574	3.68
Ada	LLama2-70b	0.163	0.040	0.099	0.011	0.594	3.64
Instructor	LLama2-70b	0.160	0.037	0.094	0.008	0.599	3.26
Ada	Flan-UL2	0.122	0.021	0.081	0.010	0.301	1.92
Instructor	Flan-UL2	0.121	0.013	0.081	0.001	0.343	2.08

Table 2: Performance comparison of various models combination (Embedding Model + Generative Model) across different evaluation metrics, with the highest score in each metric in bold.

Rouge and BLEU scores. However, these metrics alone do not fully capture the quality of the generated answers, prompting further assessments through semantic similarity and expert evaluations.

5.2 Semantic Evaluation

This assessment highlighted the models' comprehension of prompts, with combinations like Davinci with Ada or Instructor yielding higher MPNET scores, indicating a closer semantic resemblance to human-generated answers. LLama2-70b showed the highest similarity scores, but its performance in generating context-accurate responses was lower when paired with models like Flan-UL2.

5.3 Expert Evaluation

Legal experts provided ratings on a 1–5 Likert scale, evaluating the answers based on accuracy and relevance. Results showed that GPT-3 models, especially when configured with effective prompts, generally outperformed other models and even surpassed expert-provided answers in some cases, as detailed in Table 4 and average in Table 2.

5.4 Statistical Significance

Analyzing MPNET similarity scores across different model settings revealed significant statistical differences, as shown in Appendix D Table 5. These p-values varied, with some models showing high statistical significance and others not, indicating the importance of choosing the right model combinations based on the specific legal context being addressed.

6 Hallucination

In the appendix E Table 6, we demonstrate how using context in our model can lead to better answers that are free from inaccuracies, commonly referred to as "hallucinations" – a major challenge with generative models. We compare these model-generated answers to responses given by lawyers, as found in our ground truth data from the website where we sourced user questions. This comparison highlights that the lawyer's responses were typically brief and lacked detailed explanations, case references, or specific legal sections. In contrast, our approach, utilizing well-designed prompts and contextual information, successfully produced more comprehensive and detailed answers.

7 Conclusion and Future Scope

Our study delved deep into the construction of an AILQA system, spotlighting the criminal domain. Through integrating diverse embedding and QA models, we aimed to enhance the practice of legal QA in India. Our evaluations revealed that in many cases, these AI-generated answers were even better than those from human legal experts, highlighting the capabilities of AI in legal applications. Yet, the journey is far from completion. However, there's still room for improvement, particularly with models like Flan-UL2 that need better semantic understanding.

We also showed through our statistical significance results and examples that adding context to the models helped avoid hallucinated answers, a common issue with generative models. This finding is crucial for the reliability of AI in legal contexts. Although fine-tuning these models on specialized legal QA datasets is a promising approach, it's currently challenging due to the lack of such datasets. Looking ahead, exploring new methods like Chain-of-Thought prompting could significantly advance this field. Our approach of combining lexical evaluations with expert reviews provides a strong foundation for future legal AI evaluations. It ensures that our legal QA system is technologically sound and aligns with legal accuracy and relevance.

Limitations

Our study encountered several notable limitations that influenced our methodology and findings, impacting the depth and applicability of our research in the legal QA domain. Firstly, token limitations and high subscription costs for advanced cloud services constrained our ability to utilize larger parametric models, particularly those with 70B or 40B parameters. This restriction likely limited our exploration of these models' full capabilities, potentially withholding deeper insights or enhanced performance enhancements.

Another significant challenge was the resource-intensive nature of securing legal expert annotations. Due to the high costs and substantial time required, we were limited to obtaining expert evaluations for only a sample of 50 random documents rather than the entire dataset. This sampling approach may have constrained the comprehensiveness and depth of our expert-based evaluations.

Additionally, while Large Language Models (LLMs) proved competent in conversational contexts, their effectiveness in handling logic or knowledge-intensive tasks like legal QA was less convincing. The models struggled particularly with analyzing lengthy legal questions and generating detailed answers that included explanations or relevant legal references. This difficulty was compounded in scenarios requiring intricate legal reasoning and contextual understanding.

Moreover, the performance of our open-source baseline model fell short of expectations, a shortfall we attribute to the token limitations imposed during our study. By restricting our analysis to only 1000 characters with a 250-character overlap for document chunking, it is possible that the models failed to capture the full context of the legal cases, thereby hindering their ability to generate comprehensive and nuanced responses.

These limitations highlight the inherent challenges in applying LLMs to complex, specialized tasks such as legal QA. They underscore the necessity for ongoing research and development efforts aimed at enhancing AI models' capabilities in accurately interpreting and understanding detailed legal documents and contexts.

Ethical Statement

In our research, ethical considerations were paramount, particularly given the sensitive nature of the data and the methodologies employed. We placed a strong emphasis on ethical conduct throughout the collection of the AILQA dataset and the evaluation of model performance. We recognized the substantial intellectual contribution of a senior legal expert who mentored the dataset creation process and provided invaluable insights into the Likert rating system and evaluation of the generated answers. This expert is rightfully credited as the author of this paper, reflecting our adherence to ethical norms and authorship guidelines in academic publishing.

Moreover, our study required substantial computational resources, for which we ethically secured access by subscribing and duly paying for services such as Google Colab Pro and OpenAI's GPT. This not only ensured legitimate access to necessary cloud services but also supported the platforms that enabled our research. Additionally, all evaluators involved in the assessment process were compensated commensurately for their efforts, ensuring fair treatment and recognition of their work.

Our ethical approach went beyond merely complying with legal and financial obligations; it encompassed a commitment to respectful and fair treatment of all individuals involved in the study, thereby ensuring that our research is not only innovative and impactful but also responsible and ethically sound.

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nora Kassner and Hinrich Schütze. 2020. Bert-knn: Adding a knn search component to pretrained language models for better qa. *arXiv preprint arXiv:2005.00766*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ting Liu, Andrew Moore, Ke Yang, and Alexander Gray. 2004. An investigation of practical approximate nearest neighbor algorithms. *Advances in neural information processing systems*, 17.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms. *Preprint*, arXiv:2205.05131.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

A Model Details and Prompts

A.1 GPT-3's Davinci

OpenAI's GPT-3 Davinci variant is highly capable in various language tasks and costs $0.02 per 1000 tokens. It can handle sequences of up to 4096 words, including the prompt, question, and context. We instructed it to act as a legal assistant, focusing on Indian law. The prompt was:

"Your task is to answer a question as a legal assistant to the best of your abilities, using the context given in the document. If the country is not mentioned in the question, your response should be related to India. You have knowledge of all laws and legal judgments of India. Be detailed in your answer, provide relevant sections and case laws in your answer only if you are confident that they are correct. Note that if you do not know the answer, it is acceptable to say Sorry, I don't know. Context:{} Question:{}."

A.2 META's LLama2

LLama-2 is part of the Language Learning Model family, similar to GPT-3 and PaLM-2. Utilizing a transformer architecture, pretraining, and fine-tuning, Llama-2 offers optimized versions for chatbot-like dialogues, ranging from seven billion to seventy billion parameters. Our research focused on the 70 billion parameter variant with a context length of 4096 tokens. The model was set to respond as a legal advisor with expertise in Indian law. The prompt was:

"You are an honest legal advisor. Your task is to answer a question as a legal assistant to the best of your abilities based on the context provided. If the country is not mentioned in the question, your response should be related to India. You have knowledge of all laws and legal judgments of India. Be detailed in your answer, provide relevant sections and case laws in your response only if you are confident that they are correct. If you are unsure about an answer, truthfully say "I don't know".Context:{} Question:{}"

A.3 Google's Flan-UL2

Flan-UL2, an open-source T5-based model, outperforms GPT-3 in in-context learning. Its 2048-token receptive field enhances task suitability. We used this prompt to guide its responses:

"Answer the following question using the context by reasoning step by step. If you don't know the

answer, just say Sorry, I don't know. Context:{} Question:{}."

In our study, we used different prompts for various generative models. The responses varied due to differences in architecture, parameters, and how each model handles context and specific tasks. To optimize results, we tailored the prompts to each model to see which responded best to the same questions and contexts.

B List of Acts

S.No.	Act
1	Indian Penal Code
2	Protection of Children from Sexual Offences Act
3	Criminal Procedural Code
4	Indian Evidence Act
5	Arms Act
6	Information Technology Act
7	Narcotic Drugs and Psychotropic Substances Act
8	Contempt of Courts Act
9	Unlawful Activities Prevention Act
10	Prevention of Money Laundering Act
11	Criminal Procedure Identification Act
12	Extradition Act of 1962
13	Prisons Act of 1894
14	Prevention of Corruption Act of 1988
15	Gram Nyayalayas Act of 2008

Table 3: List of Acts

C Expert Scores

Embedding Model	Generative Model	Rating Score				
		1	2	3	4	5
N/A	Davinci	0	9	13	20	8
N/A	LLama2-70b	1	11	9	15	13
Ada	Davinci	2	7	6	12	21
Instructor	Davinci	2	7	11	15	15
Ada	LLama2-70b	0	3	13	33	1
Instructor	LLama2-70b	10	8	7	9	16
Ada	Flan-UL2	11	33	5	1	0
Instructor	Flan-UL2	5	36	9	0	0

Table 4: Legal Expert Ratings for Various Model Combinations (Embedding Model + Generative Model)

D Statistical Significance Scores

Table 5 shows comparative analysis P-values for pairwise statistical comparisons between different experimental settings based on MPNET similarity scores. The table is symmetric across the diagonal, hence representing in a lower triangular format. The table is in a lower triangular format, which means that the meaningful data (in this case, p-values) are only present in the lower half of the table, below the main diagonal. The main diagonal and the upper half of the table (above the main diagonal) are filled with placeholder symbols "-". This means that the comparison of Model A vs. Model B will have the same p-value as Model B vs. Model A, yielding the same statistical significance regardless of the comparison order.

- **Highly Similar Models:** Several comparisons (for example, 'Ada+Flan-UL2' vs. 'Ada+LLama2-70b') show a p-value of 0.0000, indicating extremely high statistical significance. These indicate a statistically significant difference in MPNET similarity scores, suggesting that the performance of these models differs significantly.

- **Marginally Significant Comparisons:** There are a few comparisons with p-values slightly above 0.05 (like 'Instructor+ LLama2-70b' vs. 'Ada+Davinci' with a p-value of 0.1333), which suggests that the differences in these models' scores are less pronounced.

- **High P-values:** Some comparisons have very high p-values (like 'Ada+LLama2-70b' vs. 'Davinci' with a p-value of 0.7206), indicating that the differences between these models' scores are not statistically significant difference in their MPNET similarity scores, indicating similar performance.

- **Diversity in Model Performance:** The variance in p-values across different comparisons showcases the range of capabilities and characteristics among the models used for answering legal questions.

- **Importance of Context:** When applying these models to legal questions, the context matters. Even small differences in scores, indicated by marginally significant p-values, could be critical in some legal scenarios.

- **Variability in Legal Answering Capabilities:** The range of p-values reflects the variability in how these models perform in generating answers to legal questions. Some models show distinct differences in their answering capabilities, while others perform similarly.

	Davinci	LLama2-70b	Ada+Davinci	Instructor+ Davinci	Ada+ LLama2-70b	Instructor+ LLama2-70b	Ada+ Flan-UL2	Instructor+ Flan-UL2
Davinci	-	-	-	-	-	-	-	-
LLama2-70b	0.0786	-	-	-	-	-	-	-
Ada+ Davinci	0.2527	0.0366	-	-	-	-	-	-
Instructor+ Davinci	0.4387	0.0948	0.4596	-	-	-	-	-
Ada+ LLama2-70b	0.7206	0.1237	0.2089	0.3715	-	-	-	-
Instructor+ LLama2-70b	0.4678	0.2900	0.1333	0.2627	0.7035	-	-	-
Ada+ Flan-UL2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-	-
Instructor+ Flan-UL2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2451	-

Table 5: Comparative Analysis of P-Values for Pairwise Statistical Comparisons Between Different Models Based on MPNET Similarity Scores.

E Hallucination Example

Model	Question	Ground Truth	Answer without context	Answer with context (using Instructor Embedding)	Error Explanation
LLama2-70b	I have a recording of 10 minutes out of which only the starting 5 mins of the recording is what I want to provide as evidence, rest is irrelevant talk. Can it be accepted by court if the cut audio is provided as record.	Dear client, Audio recordings can be presented as evidence in Indian courts. Even if the rest of the audio recording is irrelevant, you can provide the entire 10 minute audio recording in court. It is the court which will determine what is relevant and what is not. Thank you.	As a legal advisor, I must inform you that the admissibility of evidence in a court of law is a complex issue that depends on various factors... Relevant caselaws: * State of Maharashtra vs. Dr. Sunil Shrikant Ambekar (2016): In this case... * M.C. Abraham v. State of Kerala (2017): In this case...	Based on the information provided, it seems that you are seeking legal advice regarding the admissibility of a tape-recorded conversation as evidence in a court of law in India... In the case of Dr. Motwani vs. UOI (1995), it was held that a tape-recorded conversation is a relevant fact and is admissible under section 8 of the Evidence Act.	Relevant Caselaws mentioned in **Answer without context** section are incorrect, which means the model hallucinate while producing the answer. Whereas with context, it provide relevant case example while producing the answer.
Davinci	I hacked a friend's Instagram as a prank to show him later. But he is taking it too seriously. If he presses charges, what are my options?	Hello Sir, You have no options regarding this, as you have committed an illegal act. So If he went for Cyber complaint you have to represent yourself as to why you did that.	Your options depend on the laws of the country in which the incident occurred. In India, hacking someone's account without their permission is a criminal offence under the Information Technology Act, 2000... In the case of State of Maharashtra v. Praffful Sarda, the Bombay High Court held that the accused was...	If your friend presses charges against you for hacking his Instagram account, you may be liable under Section 43 and Section 66 of the Information Technology Act, 2000 (IT Act)... You may also be liable under Section 66B of the IT Act... to three years or with fine which may extend to rupees one lakh or with both.	The legal cases mentioned in the **Answer without context** section do exist, but they are unrelated to IT offenses; instead, they pertain to a criminal case involving the offense of murder. This suggests that the model confidently generates incorrect examples when context is not provided with it.

Table 6: Comparative Analysis: Responses from Lawyer versus Generative Models with and without Context