arXiv

PAKTON: A MULTI-AGENT FRAMEWORK FOR QUESTION ANSWERING IN LONG LEGAL AGREEMENTS

Petros Raptopoulos[1], Giorgos Filandrianos[1,2], Maria Lymperaiou[1], Giorgos Stamou[1]

[1]School of Electrical and Computer Engineering,
AILS Laboratory National Technical University of Athens, Greece
[2]Instituto de Telecomunicações, Portugal
petrosrapto@gmail.com, {geofila, marialymp}@islab.ntua.gr, gstam@cs.ntua.gr

AI|LS

Contents:

Motivation

- Contract review: Complex, **time-consuming**, requiring specialized legal expertise
- **Inaccessible** to non-experts (general public), demanding even for professionals
- **High economic impact:** avg. 9.2% annual revenue loss
- **High volume:** large organizations manage 20k–40k active contracts
- **Confidentiality concerns** make proprietary models unsuitable

Contents:

Peculiarities of contracts

- Complex legal language & terminology
- Overlapping or contradictory clauses
- Frequent exceptions & cross-references
- Substantial length
- Ambiguous wording & multiple interpretations
- Jurisdictional differences

Contents:

2. PAKTON Framework - 2.1 Overview

✅ **Task:** Answer the questions of the user based on the contract provided and integrated external knowledge.

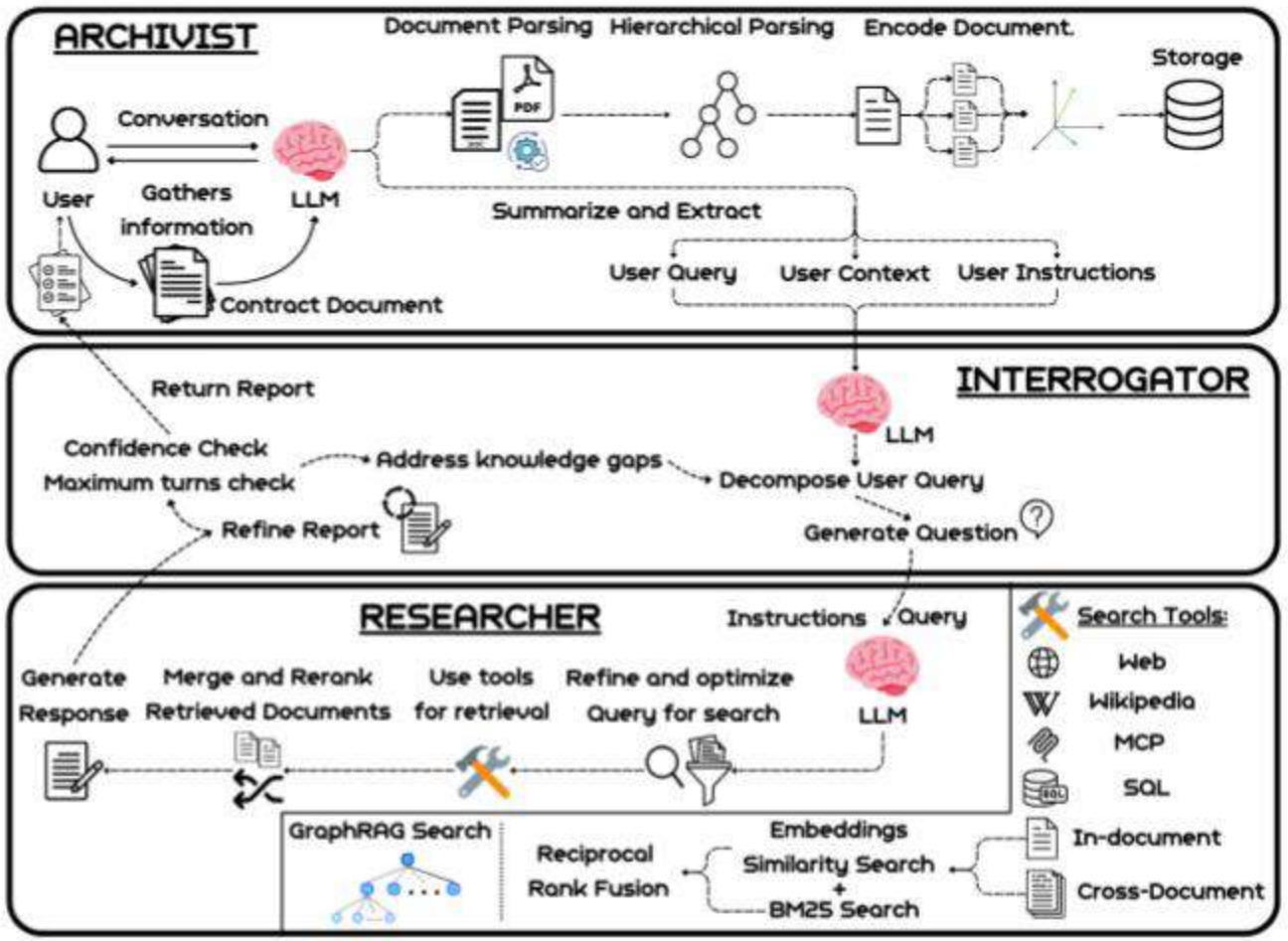
📄 **Output:** Legal Report

🎯 **Goals:**

- Transparent and traceable reasoning.
- Reference evidence spans from the contract.

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

2. PAKTON Framework - 2.1 Overview

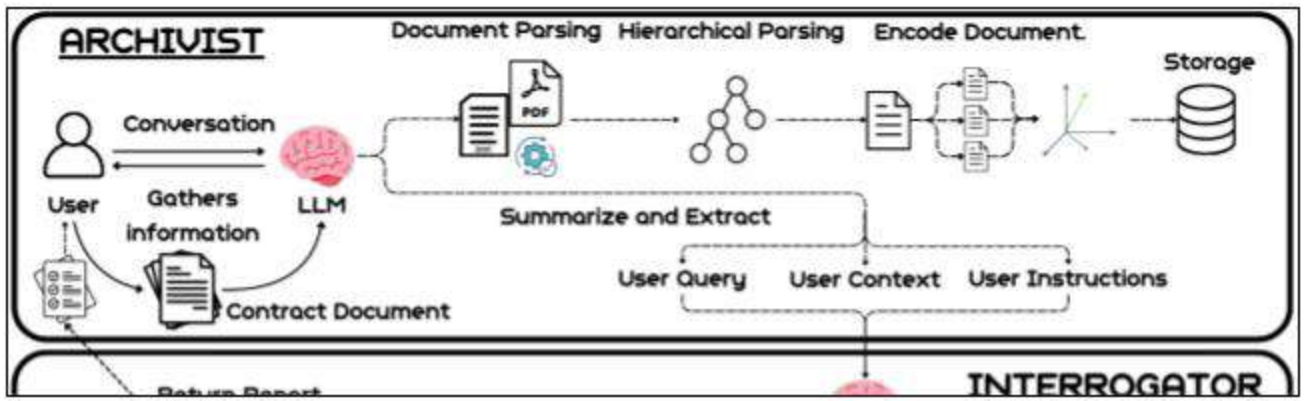


Contents:

2. PAKTON Framework - 2.2 Archivist

🤖 Role:

- Interaction with the user
- Gather/organize information
- Storage and embedding management of contracts



⚙️ Functionalities:

- Document processing
- Hierarchical Representation
- Document Segmentation and embedding

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

2. PAKTON Framework - 2.2 Archivist

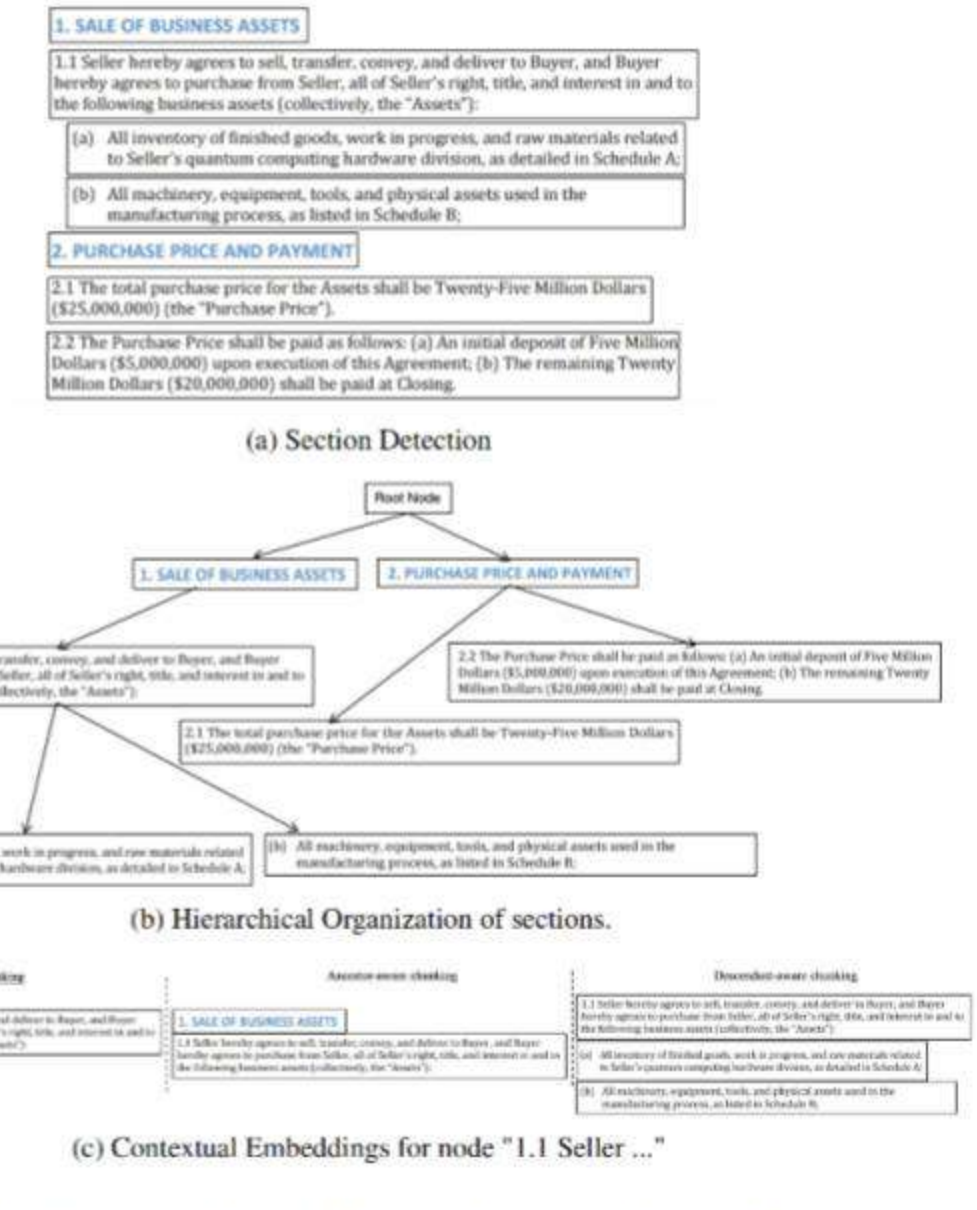
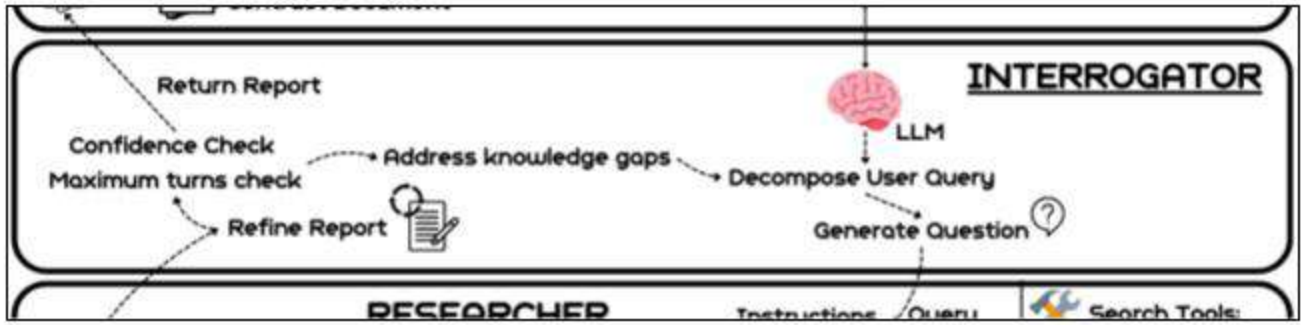


Figure 4: Tree Representation and Chunking Workflow for Contract Documents

Contents:

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

2. PAKTON Framework - 2.3 Interrogator



🤖 Role:

- Generate final report
- Orchestrate multi-step reasoning process
- Initiate interrogation of the Researcher

⚙️ Functionalities:

- Decomposition of the original query
- Generate a series of questions
- Refine report at each step

📃 Report Structure:

1) Title and Topic Summary
2) Legal Reasoning and Key findings
3) Preliminary answer and suggested research directions
4) Knowledge gaps and follow-up questions
5) Cited sources and evidentiary support

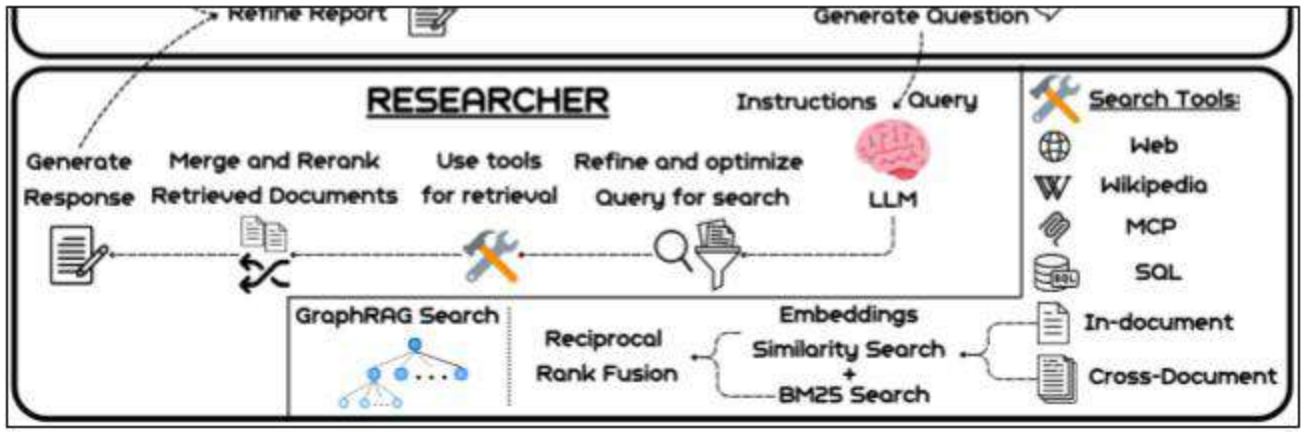
Contents:

2. PAKTON Framework - 2.4 Researcher

🤖 Role:

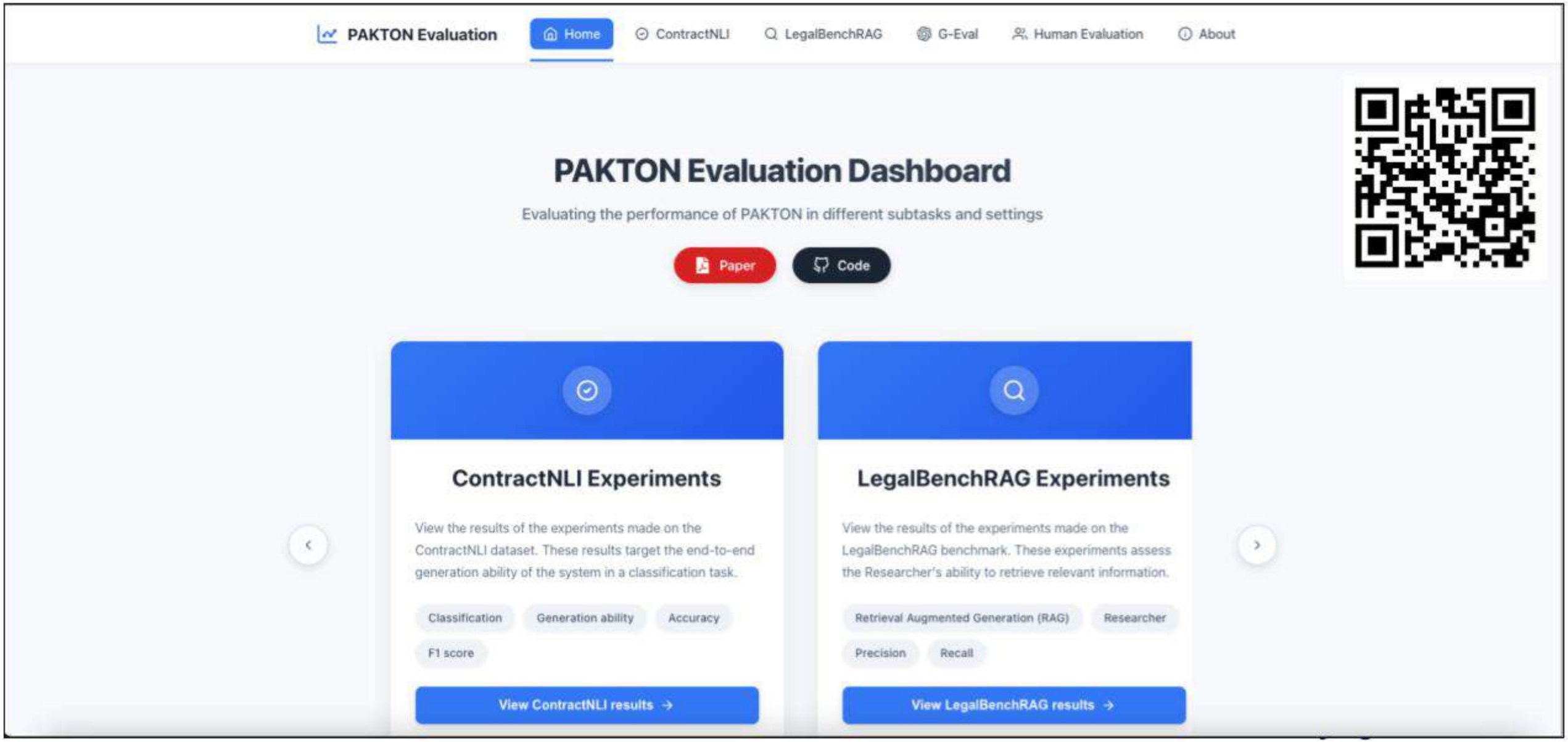
- Retrieve relevant information given a query

- Select the most suitable retrieval method or combination of methods



Contents:

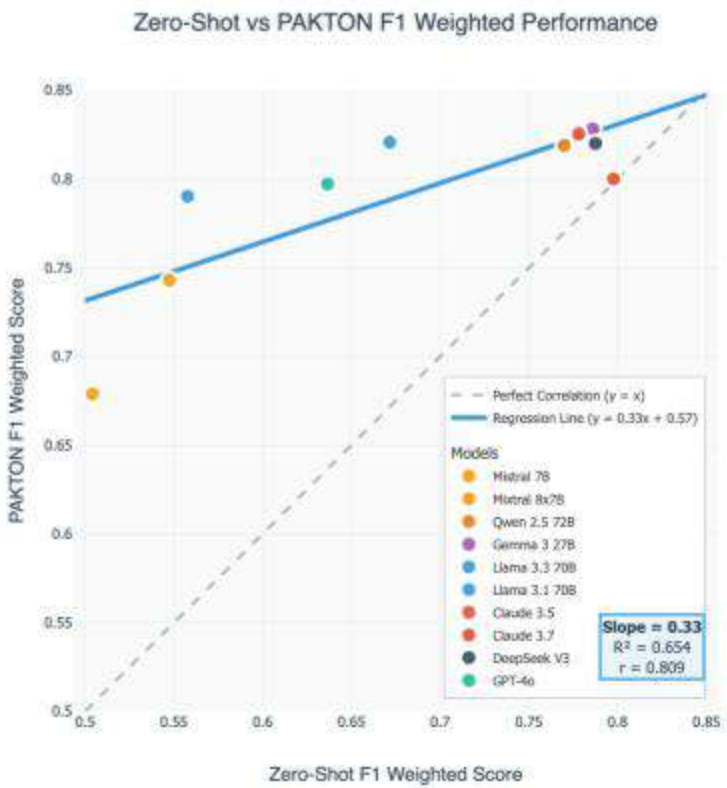
PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements



PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

3. Experiments and Results - 3.1 Quantitative Results - 3.1.1 Performance of Generation

- Assessment of the Generation ability of the end-to-end system
- ContractNLI dataset (premise, hypothesis, classification)
- PAKTON consistently outperforms the baseline methods across all models



Robustness analysis:

- Coefficient of variation (CV): 5.7% across all models for PAKTON
- ZS scores exhibit CV: 16.86% , PAKTON reduces CV by ~66%
- ANOVA: p ~= 0.075, F - statistic ~= 4.19

Model	Method	Acc.	F1 [W]	F1 [E]	F1 [C]	F1 [N]
Saul7B	Inst. Tun.	0.4196	0.2900	0.0589	0.0680	0.5920
Saul54B	Inst. Tun.	0.7020	0.6792	0.7727	0.1729	0.7024
Mistral 7B	ZS	0.5364	0.5042	0.5279	0.0248	0.5951
	FS	0.5065	0.4702	0.6053	0.0082	0.4379
	FS+Spans	0.4940	0.4576	0.6085	0.0076	0.4053
	PAKTON	0.7032	0.6789	0.7782	0.2469	0.6828
Mixtral 8x7B	ZS	0.5423	0.5475	0.6445	0.4103	0.4770
	FS	0.6002	0.5804	0.6836	0.1931	0.5642
	FS+Spans	0.6150	0.6017	0.6901	0.1951	0.6060
	PAKTON	0.7423	0.7429	0.7864	0.6655	0.7187
Qwen 2.5 72B	ZS	0.7728	0.7699	0.8248	0.5776	0.7579
	FS	0.7351	0.7241	0.8094	0.4920	0.6892
	FS+Spans	0.7484	0.7432	0.8196	0.4378	0.7357
	PAKTON	**0.8192**	**0.8188**	0.8353	0.7737	0.8132
Gemma 3 27B	ZS	0.7886	0.7860	0.8316	0.6348	0.7739
	FS	0.7191	0.7049	0.7815	0.4608	0.6891
	FS+Spans	0.7720	0.7639	0.8287	0.4728	0.7662
	PAKTON	**0.8287**	**0.8283**	0.8487	0.7546	0.8255
Llama 3.3 70B	ZS	0.6767	0.6716	0.7366	0.5378	0.6346
	FS	0.6657	0.6565	0.7326	0.4431	0.6268
	FS+Spans	0.6915	0.6879	0.7382	0.4244	0.6982
	PAKTON	**0.8217**	**0.8207**	0.8422	0.7488	0.8165
Llama 3.1 70B	ZS	0.5811	0.5577	0.5216	0.3152	0.6555
	FS	0.5729	0.5506	0.5421	0.2381	0.6358
	FS+Spans	0.5538	0.5180	0.4471	0.3014	0.6468
	PAKTON	0.7916	0.7903	0.8097	0.6846	0.7960
Claude 3.5	ZS	0.7916	0.7977	0.8757	0.5722	0.7691
	FS	0.7778	0.7816	0.8588	0.5702	0.7505
	FS+Spans	0.7999	0.8034	0.8678	0.6046	0.7826
	PAKTON	0.7990	0.8000	0.8157	0.7046	0.8072
Claude 3.7	ZS	0.7704	0.7781	0.8633	0.5602	0.7398
	FS	0.7590	0.7602	0.8463	0.5607	0.7165
	FS+Spans	0.7724	0.7766	0.8538	0.5805	0.7417
	PAKTON	0.8247	0.8254	0.8386	0.7495	0.8304
Deep-seek V3	ZS	0.7886	0.7875	0.8487	0.6117	0.7648
	FS	0.7681	0.7607	0.8346	0.6104	0.7182
	FS+Spans	0.7743	0.7714	0.8377	0.5812	0.7465
	PAKTON	0.8192	0.8200	0.8315	0.7615	0.8224
GPT-4o	ZS	0.6121	0.6366	0.7490	0.4162	0.5698
	FS	0.6640	0.6789	0.7372	0.4734	0.6666
	FS+Spans	0.6482	0.6574	0.6664	0.4636	0.6950
	PAKTON	0.7966	0.7972	0.7964	0.7592	0.8068

Table 1: Performance comparison of PAKTON and other methods across models on the ContractNLI test set. The highest accuracy and F1[w] are shown in **bold**.

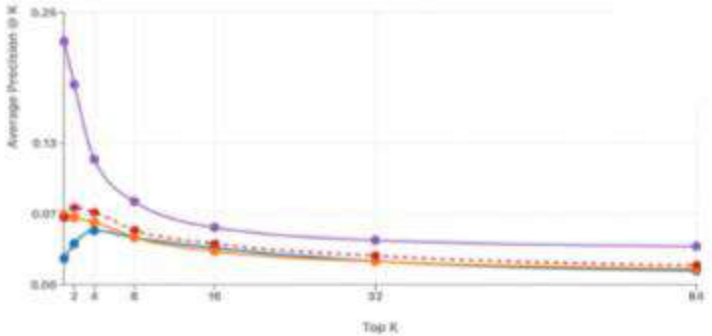
Contents:

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

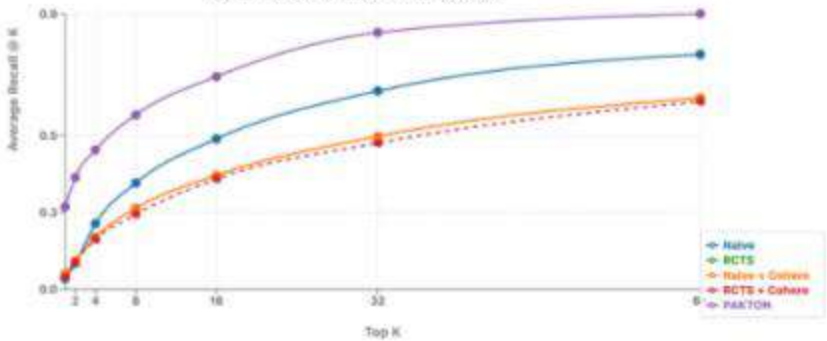
3. Experiments and Results - 3.1 Quantitative Results - 3.1.2 Performance of RAG

Dataset	Method	Precision @ k							Recall @ k						
		1	2	4	8	16	32	64	1	2	4	8	16	32	64
PrivacyQA	Naive	7.86	7.31	6.41	5.06	3.58	2.41	1.54	7.45	12.53	20.88	32.38	42.45	54.27	66.07
	RCTS	14.38	13.55	12.34	9.03	6.06	4.17	2.81	8.85	15.21	27.92	42.37	55.12	71.19	84.19
	Naive + Cohere	14.38	13.55	12.34	9.02	6.06	4.17	2.81	8.85	15.21	27.92	42.37	55.12	71.19	84.19
	RCTS + Cohere	13.94	15.91	13.32	9.57	6.88	4.68	3.28	7.32	16.12	25.65	35.60	51.87	64.98	79.61
	PAKTON	**19.94**	**16.84**	**11.44**	**8.62**	**7.38**	**6.42**	**6.08**	**13.34**	**22.43**	**32.67**	**43.39**	**61.65**	**82.30**	**89.42**
ContractNLI	Naive	16.45	14.80	12.53	9.73	6.70	4.65	3.04	11.32	19.10	29.79	45.59	56.75	69.88	86.57
	RCTS	6.63	5.29	3.89	2.81	1.98	1.29	0.90	7.63	11.33	17.34	24.99	35.80	46.57	61.72
	Naive + Cohere	6.63	5.28	3.89	2.81	1.98	1.29	0.90	7.63	11.34	17.34	24.99	35.80	46.57	61.72
	RCTS + Cohere	5.08	5.59	5.04	3.67	2.52	1.75	1.17	4.91	9.33	16.09	25.83	35.04	46.90	62.97
	PAKTON	**33.02**	**30.34**	**17.33**	**9.98**	**5.87**	**4.68**	**4.52**	**53.14**	**67.47**	**80.06**	**89.71**	**95.50**	**99.56**	**99.82**
MAUD	Naive	3.36	2.65	2.18	1.89	1.48	1.06	0.75	2.54	3.12	4.53	8.75	13.16	18.36	25.62
	RCTS	2.65	1.77	1.96	1.40	1.39	1.15	0.82	1.65	2.09	4.59	6.18	12.93	21.04	28.28
	Naive + Cohere	2.64	1.77	1.96	1.40	1.38	1.15	0.82	1.65	2.09	5.59	6.18	12.93	21.04	28.28
	RCTS + Cohere	1.94	2.63	2.05	1.77	1.79	1.55	1.12	0.52	2.48	4.39	7.24	14.03	22.60	31.46
	PAKTON	**25.47**	**17.45**	**10.51**	**7.24**	**5.08**	**3.18**	**1.85**	**23.99**	**30.09**	**34.49**	**46.42**	**59.74**	**74.96**	**82.80**
CUAD	Naive	9.27	8.05	5.98	4.33	2.77	1.77	1.09	12.60	19.47	27.92	40.70	51.02	64.38	75.71
	RCTS	1.97	4.03	4.83	4.20	2.94	1.99	1.25	1.62	8.11	17.72	31.68	44.38	60.04	74.70
	Naive + Cohere	1.97	4.03	4.83	4.20	2.94	1.99	1.25	1.62	8.11	17.72	31.68	44.38	60.04	74.70
	RCTS + Cohere	3.53	4.18	6.18	5.06	3.93	2.74	1.66	3.17	7.33	18.26	28.67	42.50	55.66	70.19
	PAKTON	**11.02**	**8.83**	**6.81**	**4.72**	**2.78**	**2.07**	**1.62**	**16.52**	**24.76**	**33.34**	**46.67**	**59.53**	**77.08**	**86.23**
ALL	Naive	2.40	3.76	4.97	4.33	3.39	2.17	1.29	3.37	8.44	21.30	34.51	48.88	64.47	76.39
	RCTS	6.41	6.16	5.76	4.36	3.09	2.15	1.45	4.94	9.19	16.90	26.30	37.06	49.71	62.22
	Naive + Cohere	6.41	6.16	5.76	4.36	3.09	2.15	1.45	4.94	9.19	16.90	26.30	37.05	49.71	62.22
	RCTS + Cohere	6.13	7.08	6.65	5.02	3.78	2.68	1.81	3.98	8.82	16.10	24.34	35.86	47.54	61.06
	PAKTON	**22.34**	**18.37**	**11.52**	**7.63**	**5.26**	**4.08**	**3.52**	**26.77**	**36.32**	**45.26**	**56.66**	**69.17**	**83.50**	**89.58**

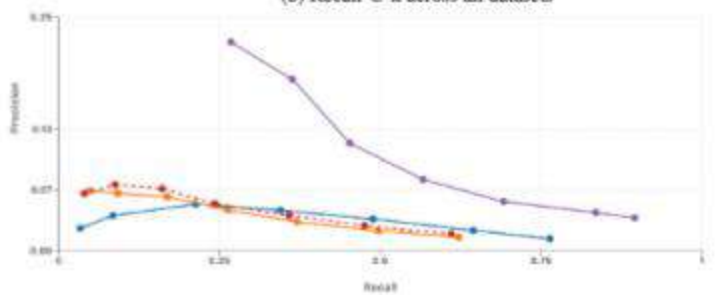
Table 2: Precision and Recall @ $k \in \{1, 2, 4, 8, 16, 32, 64\}$ for four retrieval pipelines on five legal-text datasets.



(a) Precision @ k across all datasets



(b) Recall @ k across all datasets



(c) Precision - Recall across all datasets

Figure 6: Precision and Recall values for different k across all datasets for all methods

- Assessment of the Retrieval ability
- LegalBench-RAG benchmark

Task: Given a contract and a question, did the system retrieve the "golden truth" spans? How precise and complete was the retrieval?

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

3. Experiments and Results - 3.1 Quantitative Results - 3.1.2 Performance of RAG

Dataset	LLM filtering	Precision @ k							Recall @ k						
		1	2	4	8	16	32	64	1	2	4	8	16	32	64
PrivacyQA	without	19.94	16.84	11.44	8.62	7.38	6.42	6.08	13.34	22.43	32.67	43.39	61.65	82.30	89.42
	with	29.11	30.33	25.99	23.35	22.64	22.33	22.33	10.51	18.13	23.33	26.35	27.65	28.19	28.19
ContractNLI	without	33.02	30.34	17.33	9.98	5.87	4.68	4.52	53.14	67.47	80.06	89.71	95.50	99.56	99.82
	with	59.59	51.36	46.32	45.00	45.00	44.87	44.87	38.53	45.25	51.95	54.94	58.00	58.69	58.69
MAUD	without	25.47	17.45	10.51	7.24	5.08	3.18	1.85	23.99	30.09	34.49	46.42	59.74	74.96	82.80
	with	38.87	36.99	33.54	33.12	32.77	32.33	32.29	19.06	22.60	24.06	26.52	27.51	27.64	27.64
CUAD	without	11.02	8.83	6.81	4.72	2.78	2.07	1.62	16.52	24.76	33.34	46.67	59.53	77.08	86.23
	with	29.14	29.53	29.31	28.86	28.89	28.79	28.77	25.31	30.30	34.33	37.68	38.24	38.68	38.68
ALL	without	**22.34**	**18.37**	**11.52**	**7.63**	**5.26**	**4.08**	**3.52**	**26.77**	**36.32**	**45.26**	**56.66**	**69.17**	**83.50**	**89.58**
	with	**39.17**	**37.03**	**33.78**	**32.58**	**32.26**	**32.08**	**32.05**	**23.37**	**29.07**	**33.42**	**36.36**	**37.84**	**38.29**	**38.29**

Table 3: Performance comparison on different datasets for Precision and Recall at various k values for PAKTON's Researcher and Archivist under configuration 1.

Dataset	LLM filtering	Precision @ k							Recall @ k						
		1	2	4	8	16	32	64	1	2	4	8	16	32	64
PrivacyQA	without	18.64	17.66	14.13	10.39	7.51	6.34	6.01	17.00	26.39	39.99	62.73	79.24	94.90	97.35
	with	33.02	31.85	29.33	26.66	26.11	25.70	25.68	10.35	18.20	23.12	26.95	28.20	28.58	28.76
ContractNLI	without	**42.21**	**36.86**	**19.81**	**10.58**	**6.13**	**4.63**	**4.53**	**67.67**	**83.35**	**90.87**	**95.16**	**98.19**	**99.81**	**100.00**
	with	77.51	67.15	64.02	63.02	63.22	63.23	63.23	51.78	58.93	63.10	64.52	65.07	65.24	65.24
MAUD	without	17.77	11.35	9.41	8.32	5.84	3.50	1.92	17.06	19.94	30.57	49.94	68.82	80.10	86.51
	with	37.71	33.60	31.01	32.12	32.11	31.96	31.92	17.92	21.62	24.69	28.02	29.90	30.48	30.49
CUAD	without	2.61	5.10	5.72	4.18	2.84	1.84	1.29	4.38	16.11	34.09	53.00	69.13	80.65	86.20
	with	25.51	28.62	30.28	31.13	30.27	30.21	30.22	15.54	23.74	30.06	33.11	33.60	33.60	33.65
ALL	without	20.31	17.74	12.27	8.37	5.58	4.08	3.44	26.53	36.45	48.88	65.21	78.845	88.87	92.52
	with	43.44	40.30	38.66	38.23	37.93	37.78	37.76	23.90	30.62	35.24	38.15	39.19	39.48	39.54

Table 4: Performance comparison across different datasets in terms of Precision and Recall at various k values, using PAKTON's Researcher and Archivist under Configuration 2.

Ablations

- Effect of Reranker on Performance
- Addition of LLM-filter step

Contents:

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

3. Experiments and Results - 3.2 Qualitative Results - 3.2.1 Human Evaluation

Criterion	Instructions
Explainability and Reasoning	Evaluate whether the report clearly and transparently explains not only the final conclusion, but also the reasoning process and supporting evidence in a step-by-step, understandable manner. The explanation should guide the reader through the logic in a way that supports comprehension, avoiding unexplained jumps in logic.
Justification with Evidence	Determine whether the statements and claims are explicitly justified with relevant, specific, and clearly cited evidence (e.g., direct quotations, clause references). The justification should be traceable, allowing the reader to locate the original source material.
Contextual and Legal Understanding	Assess whether the report demonstrates a deep and accurate understanding of the document, its legal terminology, and the broader context. Consider whether it correctly interprets clauses and captures implied assumptions or legal concerns behind the question.
Handling Ambiguity	Determine whether the report identifies and handles ambiguities in the source material appropriately, such as by presenting multiple interpretations or justifying a chosen one clearly.
Acknowledgment of Knowledge Gaps	Evaluate whether the report explicitly acknowledges when available information is insufficient to support a conclusion, avoiding speculation or overconfidence.
Conciseness and Precision	Assess whether the report communicates clearly and efficiently, avoiding unnecessary repetition or verbosity, while still covering all key points.
Coherence and Organization	Check whether the report is logically structured, flows smoothly, and maintains clarity across sections. Transitions between ideas should be natural and helpful.
Relevance and Focus	Evaluate whether the report stays on topic and maintains focus on answering the core question, avoiding tangents or irrelevant content.
Completeness	Assess whether the report addresses all important aspects of the question and offers a contextually broad and holistic answer. It should not omit any major points or perspectives.

Table 16: Instructions given to human annotators for each evaluation criterion used in the PAKTON vs. ChatGPT comparison. Similar instructions were given to the G-EVAL framework.

- Assessment of the quality using Human Evaluators
- Head-to-Head comparison PAKTON vs GPT-4o (RAG)
- Nine criteria, 15 legal questions, 540 individual judgments
- PAKTON is favored over ChatGPT on the majority of evaluation dimensions.

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

3. Experiments and Results - 3.2 Qualitative Results - 3.2.1 Human Evaluation

(a) Preference based on responses for a single question

(b) Overall Model Preference aggregated across all criteria and all questions

Figure 7: Comparative analysis of PAKTON vs. GPT-4o based on human evaluator judgments across different criteria

- PAKTON is favored over ChatGPT on the majority of evaluation dimensions.
- ChatGPT is preferred for conciseness.

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

3. Experiments and Results - 3.2 Qualitative Results - 3.2.1 Human Evaluation

Figure 8: The user interface (UI) of PAKTON employed during the human evaluation with study participants.

Contents:

PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements

3. Experiments and Results - 3.2 Qualitative Results - 3.2.2 LLM as a judge Evaluation

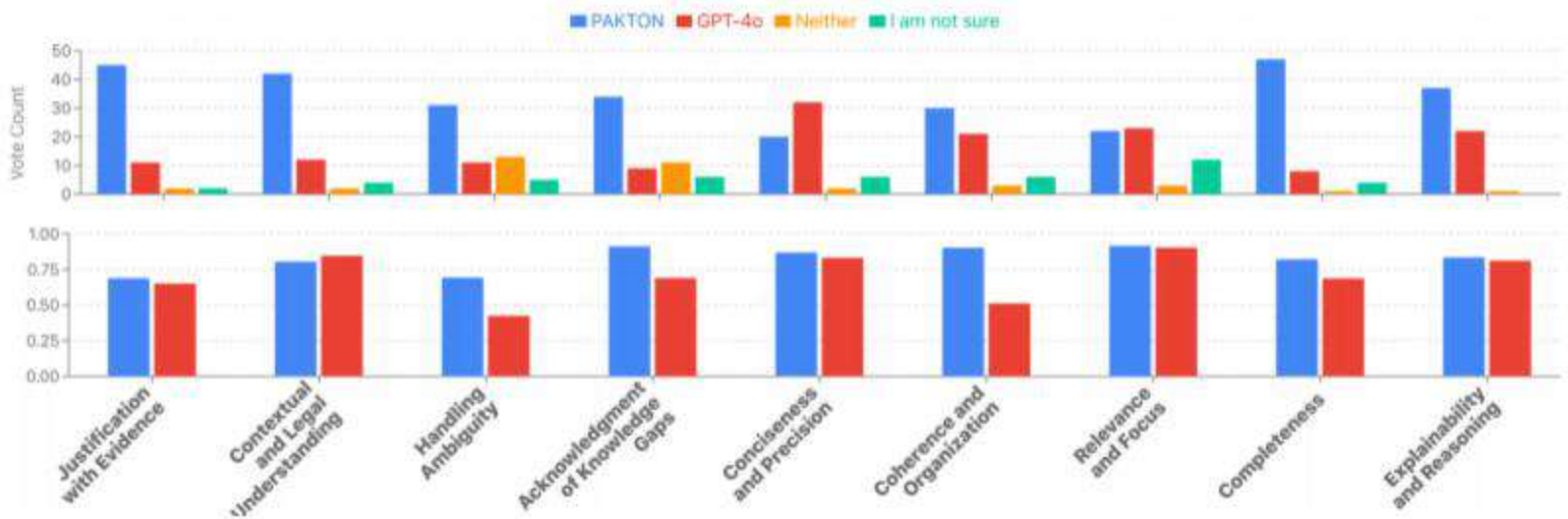


Figure 3: Comparison analysis of PAKTON and GPT-4o. Top plot presents human preferences across nine evaluation criteria aggregated for all questions. Bottom plot shows G-EVAL scores for the same criteria, aggregated across all ContractNLI outputs.

- Quality Assessment using G-Eval (LLM as a judge)
- Absolute scores across nine criteria (same as humans)
- Consistent results with humans, PAKTON is favored over ChatGPT on the majority of evaluation dimensions.

Contents:

4. Conclusion

So, why PAKTON?

💪 Superior generation performance measured, compared to baseline approaches

💪 State-of-the-art performance achieved by the Researcher component (RAG).

💪 PAKTON was preferred over ChatGPT by human evaluators for contract analysis.

💪 LLM-based evaluators demonstrated a clear preference for PAKTON over GPT-4o.

✅ Open-source pipeline

✅ Plug-and-play capabilities, end-to-end system

✅ Easy on-premise

✅ Explainability vs black box models

arXiv

Thanks for your attention!

For feedback, inquiries, or potential collaborations,
feel free to reach out at **petrosrapto@gmail.com**

AI|LS

EMNLP 2025
Suzhou, China | 中国苏州