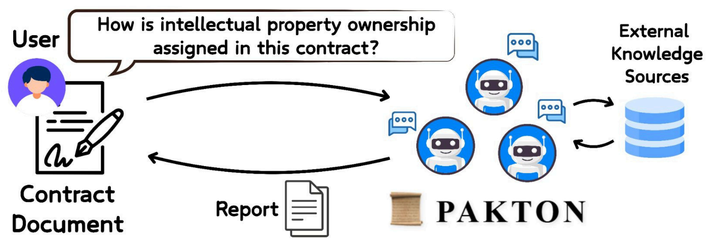
PAKTON: A MULTI-AGENT FRAMEWORK FOR QUESTION ANSWERING IN LONG LEGAL AGREEMENTS

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

THE TASK



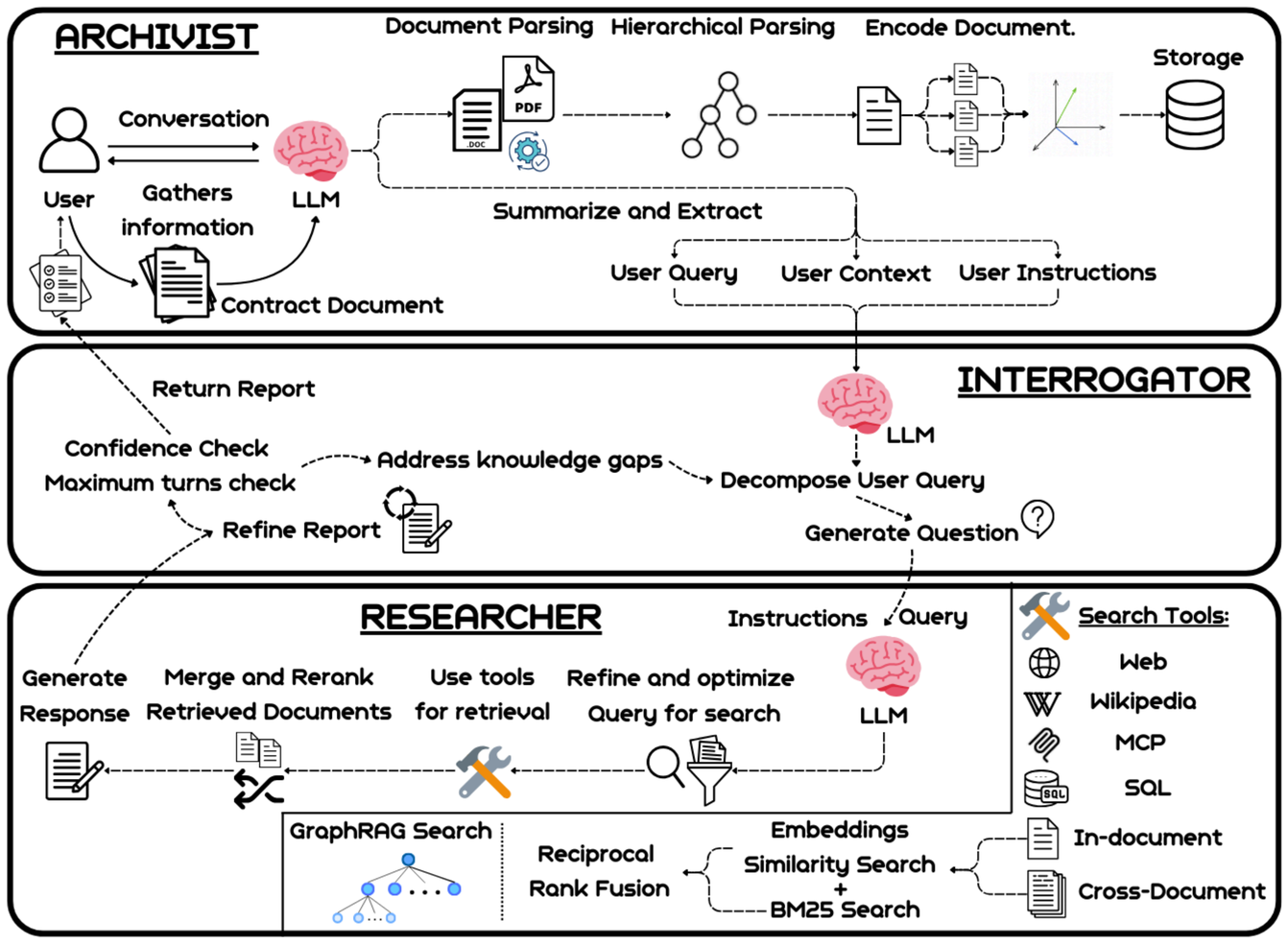
- GENERATE ANSWERS TO THE USER'S QUESTIONS USING INFORMATION FROM THE GIVEN CONTRACT AND SUPPLEMENTED BY EXTERNAL KNOWLEDGE SOURCES.
- 🎯 GOALS: ENSURE TRANSPARENT AND TRACEABLE REASONING, WHILE EXPLICITLY REFERENCING THE RELEVANT EVIDENCE SPANS FROM THE CONTRACT.

SO, WHY PAKTON?

- 💪 SUPERIOR GENERATION PERFORMANCE MEASURED, COMPARED TO BASELINES
- 💪 STATE-OF-THE-ART RETRIEVAL PERFORMANCE ACHIEVED
- 💪 PAKTON WAS PREFERRED OVER CHATGPT BY HUMANS FOR CONTRACT ANALYSIS
- 💪 LLM EVALUATION SHOWED PREFERENCE FOR PAKTON VALIDATING HUMANS

- ✅ BETTER EXPLAINABILITY THAN BLACK BOX MODELS
- ✅ OPEN-SOURCE PIPELINE
- ✅ PAKTON BRIDGES THE PERFORMANCE GAPS BETWEEN SMALLER AND LARGER MODELS
- ✅ EASY ON-PREMISE DEPLOYMENT ALLEVIATING PRIVACY CONCERNS FROM SENDING SENSITIVE CONTRACTUAL INFORMATION TO PROPRIETARY SYSTEMS

PAKTON FRAMEWORK



EXPERIMENTS AND RESULTS

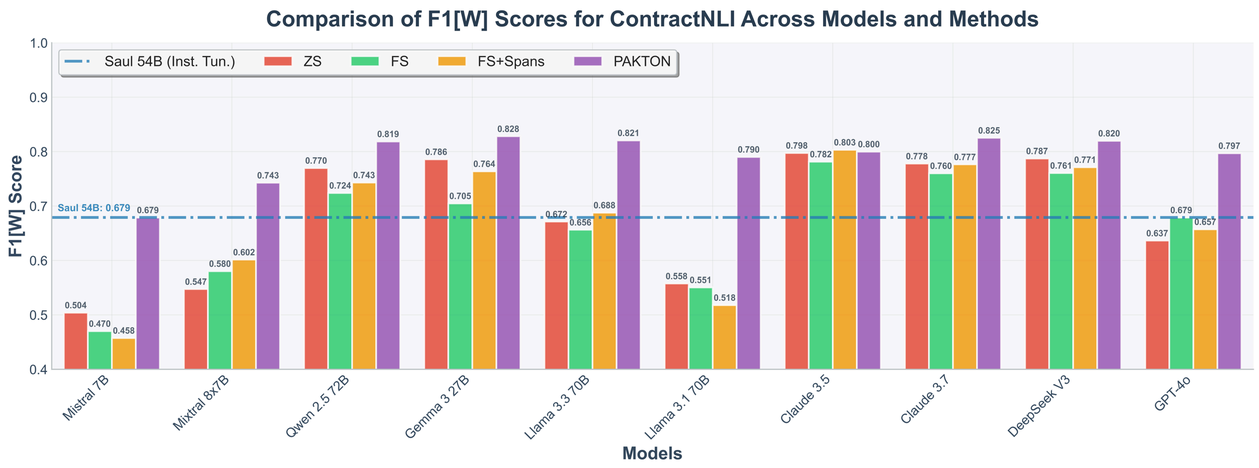
QUANTITAVE RESULTS

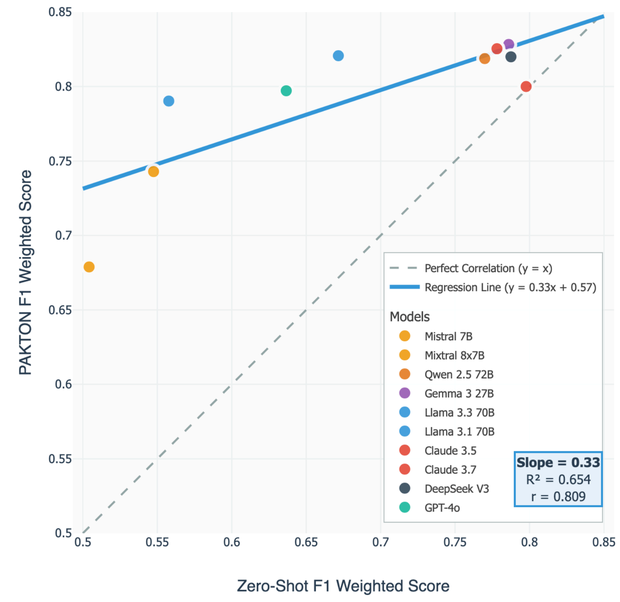
END-TO-END GENERATION PERFORMANCE

- CONTRACTNLI DATASET

TASK: "GIVEN A CONTRACT AND A HYPOTHESIS, IS THE HYPOTHESIS ENTAILED, CONTRADICTED, OR NOT MENTIONED IN THE CONTRACT?"

- PAKTON OUTPERFORMS THE BASELINE METHODS ACROSS ALL MODELS
- ROBUSTNESS ANALYSIS REVEALS THAT PAKTON COMPRESSES THE INHERENT PERFORMANCE GAPS AMONG LLMS CAUSED BY THEIR SIZE



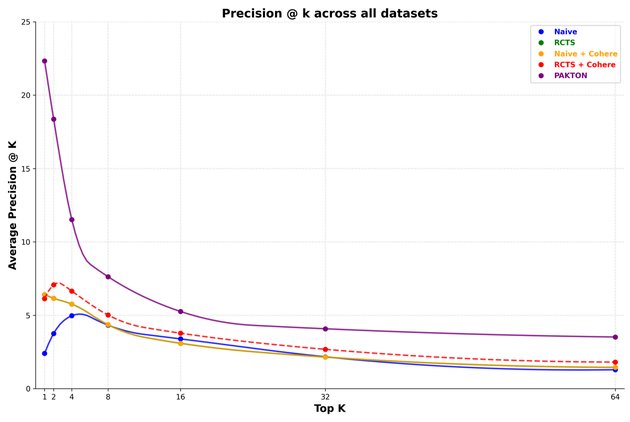


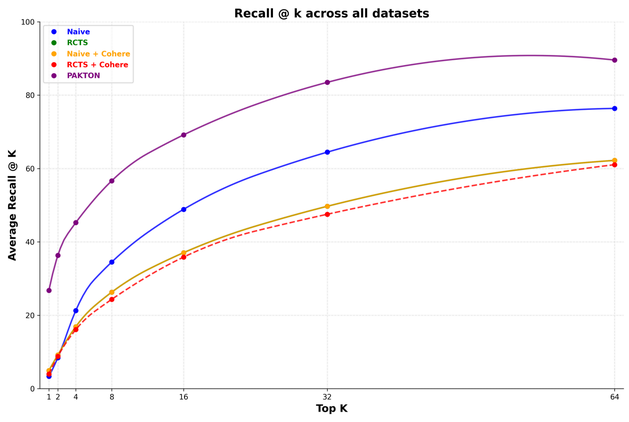
RETRIEVAL (RAG) PERFORMANCE

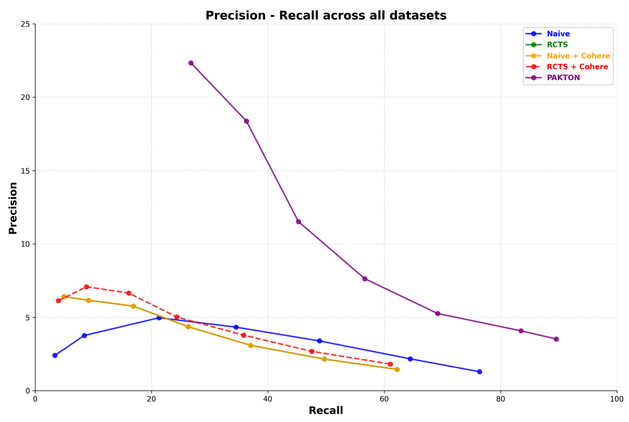
- LEGALBENCH-RAG BENCHMARK

TASK: "GIVEN A CONTRACT AND A QUESTION, DID THE SYSTEM SUCCESSFULLY RETRIEVE THE RELEVANT 'GOLDEN TRUTH' SPANS? HOW PRECISE AND COMPLETE WAS THE RETRIEVAL?"

- PAKTON OUTPERFORMS THE BASELINES AT EVERY K FOR BOTH PRECISION AND RECALL

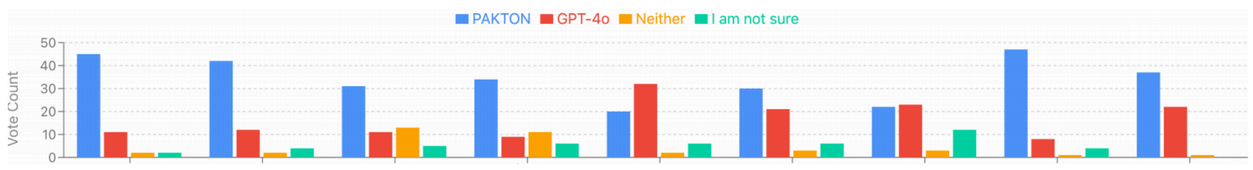






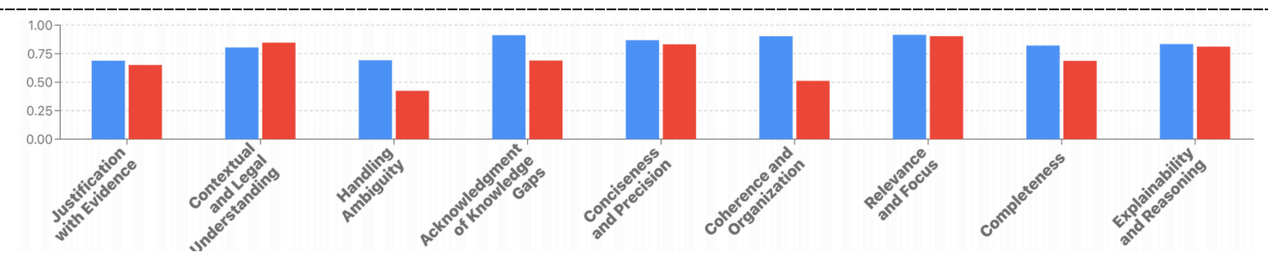
QUALITATIVE RESULTS

HUMAN EVALUATION



- HEAD-TO-HEAD COMPARISON PAKTON VS GPT-4O (RAG)
- NINE CRITERIA, 15 LEGAL QUESTIONS, 540 JUDGMENTS
- PAKTON IS FAVORED OVER CHATGPT ON THE MAJORITY OF EVALUATION DIMENSIONS.

LLM-AS-A-JUDGE



- G-EVAL SCORES FOR PAKTON AND GPT-4O (RAG)
- SAME NINE CRITERIA, 102 CONTRACT SAMPLES

- ✅ THE RESULTS FOR BOTH HUMAN EVALUATION AND LLM-AS-A-JUDGE CONSISTENTLY FAVOR PAKTON FOR MOST CRITERIA

PAKTON

EMNLP 2025 Suzhou, China 中国苏州