Determining Legal Relevance with LLMs using Relevance Chain Prompting

Onur Bilgin and **John Licato**
Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida
{onurbilgin, licato}@usf.edu

Abstract

In legal reasoning, part of determining whether evidence should be admissible in court requires assessing its relevance to the case, often formalized as its probative value—the degree to which its being true or false proves a fact in issue. However, determining probative value is an imprecise process and must often rely on consideration of arguments for and against the probative value of a fact. Can generative language models be of use in generating or assessing such arguments? In this work, we introduce relevance chain prompting, a new prompting method that enables large language models to reason about the relevance of evidence to a given fact and uses measures of chain strength. We explore different methods for scoring a relevance chain grounded in the idea of probative value. Additionally, we evaluate the outputs of large language models with ROSCOE metrics and compare the results to chain-of-thought prompting. We test the prompting methods on a dataset created from the Legal Evidence Retrieval dataset. After postprocessing with the ROSCOE metrics, our method outperforms chain-of-thought prompting.

Introduction

In the field of information retrieval, it is commonplace to say we seek to retrieve information that is relevant to some given problem or query. However, in legal reasoning, the concept of *relevance* has a related but differing meaning: whether a piece of evidence is relevant to a given case, fact, or query has more to do with the proof-related affordance that evidence provides. Saracevic (2007) defined relevance as a “property along which parts are related and may also be considered as a measure of the strength of the related connection”. From the legal perspective, relevance is used for reasoning and admissibility of the evidence. The evidence-exclusion decisions lead to two distinct notions of relevance. First is probabilistic or logical relevance, which addresses the relevance of the evidence to increase or decrease the likelihood of the existence of the fact in a case. Second is practical relevance, which addresses in courts whether a piece of evidence is “worth hearing”, which is evaluated based on different aspects such as the processing time of the evidence, possible reactions to the evidence, or whether it will cause prejudice and wrongful conviction (Woods 2010).

Determining the relevance from the legal perspective aims to identify a chain of statements from the evidence to the fact. This process assesses a change in likelihood ratios of the fact based on the evidence that represents the probative value of the evidence. For example, the fact that someone posted on social media that they were desperate for money might have probative value for proving they robbed a bank if it can be shown that there is a link between them that increases the value of the latter (e.g.: they posted they were desperate for money, which is normally something they wouldn’t have done, which means they were amenable to being recruited by the team of bank robbers and willing to go along with them, etc.). In this work, we focused on determining the relevance between the fact and the evidence by utilizing large language models (LLMs) and evaluating the generated text with automated metrics to improve prediction accuracy on legal relevance tasks. Thus, we introduce a new prompting technique, **relevance-chain prompting**, and a new assessment method to evaluate the legal relevance between the fact and the evidence. Additionally, we introduce a new scoring method, chain score, to evaluate the generated chains. With these methods, we were able to outperform the chain-of-thought prompting. With our methodology and experiments, we aim to answer the following research questions:

- Are LLMs able to recognize the probative value of evidence; i.e., its tendency to increase or decrease the perceived likelihood of a given fact?
- How well do existing measures of chain strength reflect probative value?

Contributions: In this work, we contribute to the existing literature by:

- Proposing a new prompting technique, *relevance-chain prompting*, utilizing the relevance and a new scoring methodology for the generated text.
- Demonstrating that relevance-chain prompting outperforms chain-of-thought prompting on a legal evidence retrieval dataset.
- Evaluating different chain scoring strategies for relevance chains for a fast and standardized chain assessment by external LLMs.

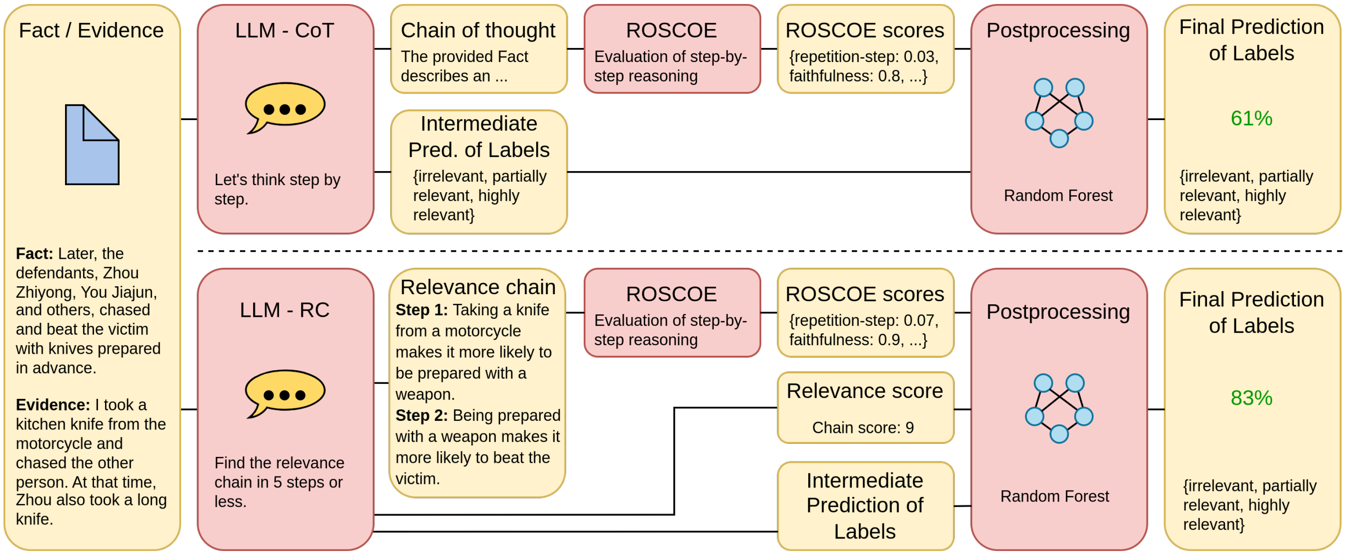

Fact / Evidence
Fact: Later, the defendants, Zhou Zhiyong, You Jiajun, and others, chased and beat the victim with knives prepared in advance.
Evidence: I took a kitchen knife from the motorcycle and chased the other person. At that time, Zhou also took a long knife.
LLM - CoT
Let's think step by step.
Chain of thought
The provided Fact describes an ...
Intermediate Pred. of Labels
{irrelevant, partially relevant, highly relevant}
ROSCOE
Evaluation of step-by-step reasoning
ROSCOE scores
{repetition-step: 0.03, faithfulness: 0.8, ...}
Postprocessing
Random Forest
Final Prediction of Labels
61%
{irrelevant, partially relevant, highly relevant}
LLM - RC
Find the relevance chain in 5 steps or less.
Relevance chain
Step 1: Taking a knife from a motorcycle makes it more likely to be prepared with a weapon.
Step 2: Being prepared with a weapon makes it more likely to beat the victim.
ROSCOE
Evaluation of step-by-step reasoning
ROSCOE scores
{repetition-step: 0.07, faithfulness: 0.9, ...}
Relevance score
Chain score: 9
Intermediate Prediction of Labels
Postprocessing
Random Forest
Final Prediction of Labels
83%
{irrelevant, partially relevant, highly relevant}

Figure 1: Our methodologies for comparing the two approaches are presented in an overview flowing from left to right. The red shapes illustrate the processing steps, and the yellow shapes illustrate the inputs and outputs of those processing steps. LLMs generate an intermediate prediction of labels, while the final label is predicted in the postprocessing step.

Related Work

Defining Relevance. Relevance is described mainly via system- or user-oriented perspectives (Saracevic 2007). In the system-oriented perspective, relevance is defined as topicality or matching queries with documents, often called topical relevance. In the user-oriented perspective, relevance is defined as the usefulness of an answer to the user or as a form of user satisfaction (Schamber and Eisenberg 1988; Saracevic 2007). Cooper (1971) used topicality as logical relevance and defined it as an important factor, but not the only factor in determining usefulness or user satisfaction.

Besides the aforementioned perspectives, relevance is studied from the communication and cognition perspective. In Sperber and Wilson's (1986) relevance theory, it is seen as an interference model where the audience has to infer the meaning from the provided utterance. If the input leads with the context or background information available to the audience to a positive cognitive effect, such as a proper conclusion, then the input is considered relevant. Any strengthening, changing, or discarding an assumption toward a proper conclusion can be regarded as a positive cognitive effect (Wilson and Sperber 2006). However, with the increasing effort required by the audience, the relevance of the input decreases. The latter is related to the vagueness as lower relevance requires more effort from the audience to clarify the meaning. Additionally, there may be many relevant inputs to choose from. Relevance theory provides a methodology for comparing and choosing the most relevant input. The optimal reference is when input is worth enough processing or input is given the speaker the most relevant utterance (Wilson and Sperber 2002). Harter (1992) studied relevance theory from an information retrieval perspective and suggested that relevance should cause a cognitive change and strengthen or weaken the assumption.

For the current paper, we use the definition that a proposition is relevant if its being true makes it more or less convincing that some target proposition is true. This is inspired by the definition in Rule 401 of Federal Rules of Evidence in the United States, which is “relevant evidence means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable that it would be without the evidence” (Wellborn III 1976).

Determining Relevance in Natural Language Processing (NLP). In Bayesian logic, the probative value of evidence is represented by a quantitative measure of support that evidence provides to combine it with one's prior beliefs to form the posterior beliefs (Koehler 1996). It is portrayed by probabilities or likelihood ratios to identify the acceptance of a proposition (Wells 2014). For the probabilistic relevance, Bayesian networks have been studied for the admissibility and probative value of the evidence (Fenton, Neil, and Lagnado 2013; Biedermann and Taroni 2012; Vlek et al. 2016; Fenton et al. 2014). In cases where probabilistic methods cannot be applied due to reasons such as missing data or type of evidence, arguments, and scenarios are evaluated for the legal relevance of the evidence (Vlek 2016; Liu, Islam, and Governatori 2021; Prakken and Kaptein 2016). Thus, deep learning methods are increasingly used in tasks such as Legal Evidence Retrieval (LER) (Yao et al. 2023). But how is the admissibility of evidence to a fact determined by LLMs? How can the reasoning steps be evaluated? Despite the increasing use of deep learning methods on legal relevance tasks, existing work still does not answer the questions. Our work sheds light on a new prompting approach and the evaluation and refinement of the LLMs' predictions about the admissibility of the evidence.

Methodology

Approach

"Chain-of-thought" (CoT) prompts LLMs to address cognitive tasks through multiple reasoning steps. It improves LLMs' reasoning performance by mimicking humans' ability to decompose a complex task into smaller steps and solve each step before giving a final answer to the task (Wei et al. 2022). In our experiments, we compare the CoT-prompting approach by Kojima et al. (2022) and our approach, "relevance-chain" (RC) prompting. We define a relevance chain as a set of connected arguments that bridges a premise and a hypothesis in predefined multiple steps. These steps increase or decrease the likelihood of the hypothesis as a whole, which can be evaluated with a chain score.

In Figure 1, the yellow shapes illustrate the inputs and outputs, and the red shapes the processing steps of our approach. First, we apply both prompting approaches to get the reasoning steps and class predictions for CoT-prompting and the relevance chain, relevance score, and class predictions for RC-prompting. The reasoning steps of both prompting approaches are evaluated with ROSCOE metrics (Golovneva et al. 2022). Additionally, we evaluate the relevance chains with external LLMs to output alternative chain scores. In the postprocessing step, the final label is predicted by random forest or support vector machines (SVMs). We compare RC-prompting results against CoT-prompting.

Dataset

We use a dataset created from the supervised LER dataset (LERD) (Yao et al. 2023). It is originally in Chinese and is comprised of facts and evidence collected from criminal cases in judgment documents (Wen 2023). Evidences are the statements that are made by involved parties in the case, provide perspectives to the facts, and are less informative than the facts. The task requires assessing the degree to which the evidence is relevant to the facts. Two lawyers annotated each sample in the supervised dataset; a third annotated the sample in case of disagreement. Evidence is annotated as either irrelevant, partially relevant, or highly relevant. The proportion of disagreements is not reported in the original work, which is a common problem with NLP datasets (Nighojkar, Laverghetta Jr., and Licato 2023).

We build a validation and a test set from LERD to apply the prompting approaches and evaluate the outputs with ROSCOE metrics to utilize those in postprocessing. For our validation set, we randomly selected 50 irrelevant, 50 partially relevant, and 50 highly relevant samples from different criminal cases, a total of 150 samples. For our test set, we randomly selected 50 irrelevant, 25 partially relevant, and 50 highly relevant samples, a total of 125. We use only a small sample of the full dataset due to the computation requirements and high costs of API usage associated with LLMs. All our test samples were from the same criminal case "intentional injury". The reason for fewer samples for partially relevant labels in the test set is the low number of samples from that criminal case in the original dataset. We translated the selected samples from Chinese to English with Google Translate (Goo 2023).

Prompting

Brown et al. (2020) showed that LLMs can solve new tasks even if only a few examples are provided. This approach, called few-shot learning, has been successfully applied in legal reasoning tasks (Blair-Stanek, Holzenberger, and Van Durme 2023; Yu, Quartey, and Schilder 2022). Relevance is essential for determining the admissibility of the evidence. If it is impossible to prove the evidence through probabilistic methods, other methods, such as argumentation and scenarios, are used to explain the legal relevance of the evidence. We introduce a new prompting technique, *relevance-chain prompting*, to better explain the relevance of possible evidence with arguments.

Relevance chains (Eq. 1) consist of n number of steps, and in each step except the last, two premises p_i are connected with "makes it more likely" or "makes it less likely" relation. All steps start where the previous step left off, with the premise connected to the hypothesis h in the last step. For example, consider the fact that a driver can not avoid high repair costs after a car crash and the evidence that the driver was driving the car with alcohol. Figure 2 shows the relevance chain in three steps. Our starting premise from the evidence is s = "driving a car with alcohol", whose relevance to hypothesis h = "avoiding high repair costs" we are trying to assess. The following premises are p_1 = "losing control of the car" and p_2 = "being guilty of a car crash". With this prompting technique, the LLM outputs whether evidence increases or decreases the likelihood of a fact.

$$RC = \{(s, p_1), (p_1, p_2), (p_2, p_3), ..., (p_{n-1}, h)\} \quad (1)$$

Step 1: Driving a car with alcohol makes it more likely to lose control of the car.

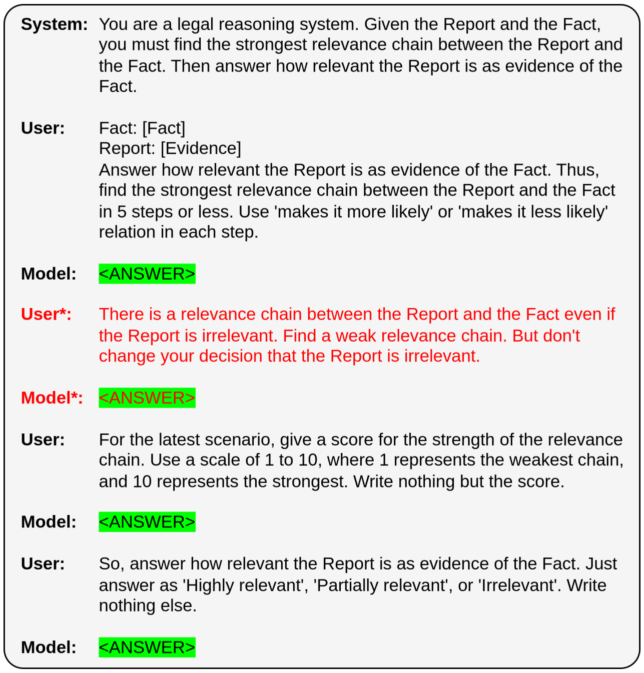
Step 2: Losing control of the car makes it more likely to be guilty of a car crash.

Step 3: Being guilty of a car crash makes it less likely to avoid high repair costs.

Figure 2: An example relevance chain. In orange is the initial evidence whose relevance to the hypothesis (green) we are trying to assess. In blue are the first premises, and in red are the second premises of each step. Each second premise is the first premise of the next step, and each step's premises are connected with a "makes it more/less likely" relation.

The relevance chain suggests a possible path to increase or decrease the likelihood of the hypothesis. In this manner, the premise affects the posterior probability of the hypothesis and thus gives a probative value to the premise. We define the maximum number of steps in the chain n_{max} as 5 because of the close context between the Fact and Evidence, leading to repeating steps after that number. As LLM, we used GPT-3.5 (`gpt-3.5-turbo`) and GPT-4 (`gpt-4`) of OpenAI and Flan-T5 (`google/flan-t5-xxl`) due to their strong chain-of-thought reasoning performance (Hagendorff, Fabi, and Kosinski 2023; Chung et al. 2022).

Figure 3 illustrates our prompting approach for RC-prompting. We use the shots to form the output in the desired format with "makes it more/less likely" relations. We used



System: You are a legal reasoning system. Given the Report and the Fact, you must find the strongest relevance chain between the Report and the Fact. Then answer how relevant the Report is as evidence of the Fact.

User: Fact: [Fact]
Report: [Evidence]
Answer how relevant the Report is as evidence of the Fact. Thus, find the strongest relevance chain between the Report and the Fact in 5 steps or less. Use 'makes it more likely' or 'makes it less likely' relation in each step.

Model: <ANSWER>

User*: There is a relevance chain between the Report and the Fact even if the Report is irrelevant. Find a weak relevance chain. But don't change your decision that the Report is irrelevant.

Model*: <ANSWER>

User: For the latest scenario, give a score for the strength of the relevance chain. Use a scale of 1 to 10, where 1 represents the weakest chain, and 10 represents the strongest. Write nothing but the score.

Model: <ANSWER>

User: So, answer how relevant the Report is as evidence of the Fact. Just answer as 'Highly relevant', 'Partially relevant', or 'Irrelevant'. Write nothing else.

Model: <ANSWER>

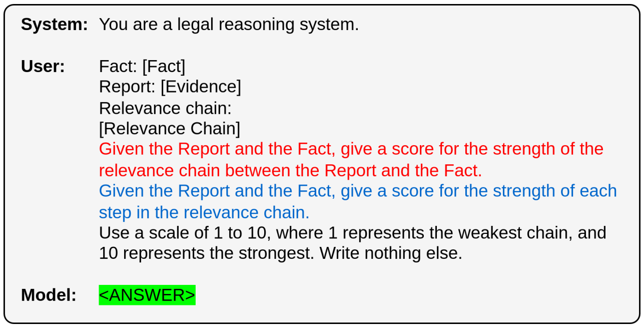
Figure 3: Our approach for RC-prompting. The model is asked a second time if the LLM does not output a relevance chain the first time (marked as User* and Model* in red).

a two-shot strategy with one relevance chain labeled "irrelevant" and the other "highly relevant" to limit the prompt size and enable the LLMs to learn the output format. We used the same shots for the validation and test, and those are from the criminal case "intentional injury". They include the fact, the evidence, the question for the relevance chain and the relevance chain itself, but no information to the label of that sample in the context except the relevance chain. Therefore, the shots of CoT-prompting give more hints about the labels because the labels were easier to understand from the context of step-by-step reasoning of CoT-prompting. In the prompts, we replaced the name of the "Evidence" feature in the data with "Report" because we don't want to mislead LLMs in case of "irrelevant" labels. If LLM does not determine relevance and cannot create a chain, we asked to build the chain a second time. Due to their probabilistic generation, we observed that the GPT models infrequently changed the prediction when the question was asked recursively in the following prompt. To counter this, we added to the prompt not to change the decision as in the User* prompt. For RC-prompting, after creating the relevance chain, we prompted the LLMs to give a chain score and the label, while for CoT-prompting, after step-by-step reasoning, we prompted the LLMs to output the label.

Evaluation

Nye et al. (2021) utilized language models in a "step-by-step" generation of outputs to show the intermediate steps of a multi-step computational operation. Wei et al. (2022) and Wang et al. (2022) showed that this new thought process, step-by-step reasoning, can improve the reasoning of LLMs. Nevertheless, these methods increase the necessity of an evaluation metric that evaluates the output from different perspectives. ROSCOE is a suite of evaluation metrics for step-by-step language generation that evaluates generated steps from four perspectives: semantic alignment, logical inference, semantic similarity, and language coherence (Golovneva et al. 2022). According to those perspectives, we utilized ROSCOE metrics to evaluate the RC-prompting and CoT-prompting. We filtered out for RC-prompting the words "Step x:" in the chain where $x \in \{1, 2, 3, ...\}$ to bring it into a continuous text and not to disrupt the logical flow of the reasoning. We applied the fine-tuned `facebook/roscoe-512-roberta-base` embeddings (Golovneva et al. 2022) to calculate the metrics because the model shows consistent performance across the datasets in the original work that introduced ROSCOE.

Additionally, we developed *chain scores* to apply LLMs to evaluate relevance chains by prompting. In Figure 3, the relevance chain is scored by the same LLM in the same dialogue. Nevertheless, other LLMs can evaluate the chain in a new instance as in Figure 4. We let LLMs evaluate the chain with a score between 1 and 10, with 10 being the highest relevance. If LLM couldn't build a chain, we assigned a score of 0. Those chain scores are used together with the ROSCOE metrics and intermediate predictions of LLMs as features to train a random forest algorithm on the validation set. We first normalized the inputs to $[0, 1]$, then employed feature selection with ANOVA (Montgomery 2017) followed by random forest (Ho 1995) algorithm. We observed in different hyperparameter settings that the RC-prompting outperforms CoT-prompting. However, we selected the hyperparameters that gave RC and CoT-prompting higher accuracy across different models. The feature selection algorithm reduces the number of features to 10. For the random forest model, the maximum tree depth is 5, and the number of estimators is 1000. Except for those, we used the default parameters in scikit-learn (Sci 2024).



System: You are a legal reasoning system.

User: Fact: [Fact]
Report: [Evidence]
Relevance chain:
[Relevance Chain]
Given the Report and the Fact, give a score for the strength of the relevance chain between the Report and the Fact.
Given the Report and the Fact, give a score for the strength of each step in the relevance chain.
Use a scale of 1 to 10, where 1 represents the weakest chain, and 10 represents the strongest. Write nothing else.

Model: <ANSWER>

Figure 4: Our approaches for external scoring of relevance chains and individual steps. The red text is for scoring the chain, and the blue text is for the individual steps.

In the following scenario, we evaluated the scoring ability of LLMs for the relevance chains created by other LLMs. Because a relevance chain consists of any given number of steps, each step is not obligated to have the same argumentation strength. While the chain score is assigned as an overall score of the chain, a weak argumentation step might break the chain's argumentation strength, which might be ignored while evaluating the chain as a whole. Thus, we additionally let external LLMs score each step to calculate a chain score from the individual steps. We used the relevance chains generated by GPT-3.5 and utilized GPT-4 and Flan-T5 (`google/flan-t5-xxl`) for external scoring. Figure 4 shows the prompts for scoring the chain and individual steps. We didn't use any shots for this approach to not influence the scoring ability with shots. For the Flan-T5 model, the system prompt is concatenated to the user prompt. The predictions and chain scores of the original LLMs are not used in the final training. We applied the same hyperparameters for the random forest algorithm as in the previous case. Besides the random forest, we additionally trained our data on the SVM classifier (Cortes and Vapnik 1995).

Results

Figures 5, 6 and 7 show the test results with final predictions. Figures 5 and 6 show that for RC-prompting, we achieved a higher test accuracy than CoT-prompting with GPT models when the ROSCOE metrics were trained either with the predictions or the chain scores. For CoT-prompting, intermediate predictions didn't benefit GPT-3.5's overall test accuracy. With its concise and focused argumentation through the steps to the hypothesis, RC-prompting improves over CoT-prompting the final prediction performance on the LER dataset. Figure 7 shows the only exception in the case of two-shot CoT-prompting. It is not immediately clear to us why this opposite effect occurs, and future work will need to explore this more in-depth.

The most significant ROSCOE metrics were according to ANOVA faithfulness-step, informativeness-step, and faithfulness-token. Those semantic alignment metrics of ROSCOE calculate a normalized cosine similarity between the hypothesis and the most similar sentence in context for each value in the reasoning alignment to measure step-wise reasoning. The faithfulness-step is the mean reasoning alignment over the reasoning steps and evaluates whether the problem is misinterpreted or the chain misuses the information. The informativeness-step evaluates the use of the information from the source text in the reasoning chain and assesses how well the hypothesis covers the information in the source text. The faithfulness-token measures the similarities in token level by extending the calculation of the reasoning alignment to token embeddings (Golovneva et al. 2022).

Introducing an External Scorer. The results thus far show that ROSCOE metrics and chain scores generally improve the final prediction accuracy, at least for GPT models. But how good are LLMs at assigning chain scores for the output they don't generate? Can we find better scoring methods? Comparing LLMs' scoring abilities might give us additional insights into their legal reasoning abilities. Next, we explore those fields using external LLMs to assign a chain score. We performed no additional hyperparameter search for random forest and applied the same hyperparame-

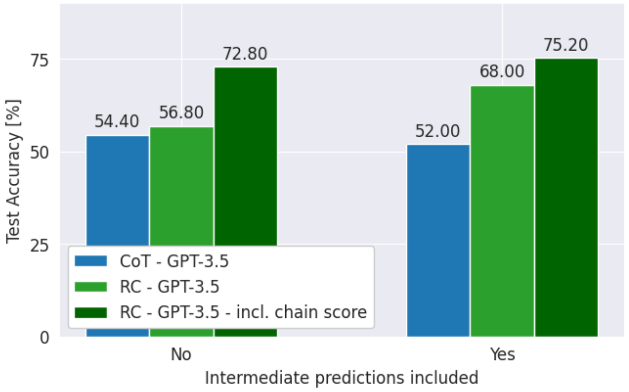


Figure 5: Test accuracy of GPT-3.5 model for RC-prompting and CoT-prompting after final training. Note that chain scores are not applicable to CoT. The x-axis shows whether LLM predictions are included in the training.

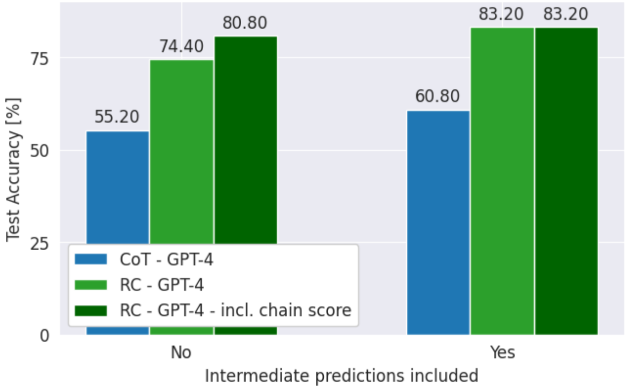


Figure 6: Test accuracy of GPT-4 model for RC-prompting and CoT-prompting after final training.

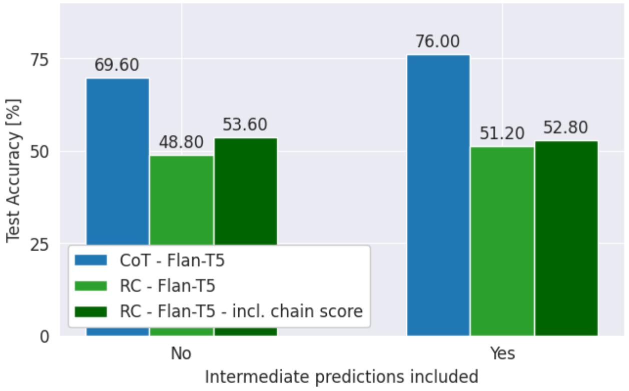


Figure 7: Test accuracy of Flan-T5 model for RC-prompting and CoT-prompting after final training. Strangely, the improvement of RC over CoT is reversed here as compared to GPT models.

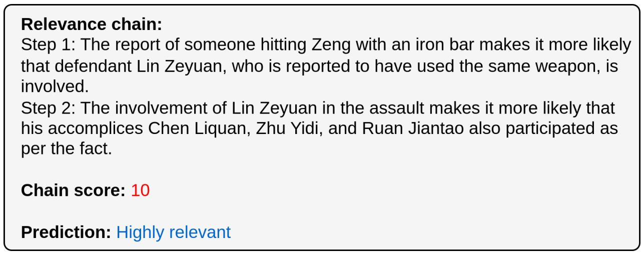
ters from the previous experiment. Table 1 shows the results of RC-prompting with an external scorer. Here, the LLMs' chain score is replaced with a score from different LLMs. We tested different methods for scoring chains by assigning scores to individual steps with LLMs and calculating the mean, minimum, or harmonic mean as the chain score. We achieved lower accuracy with both external chain and step scores. However, we find that the harmonic mean of the step scores achieves a better test accuracy than other step scoring methods and has competitive results.

Model	Train method	Score	Extern scorer	Test Acc.
GPT-3.5	SVM	Chain	GPT-4	72.80%
GPT-3.5	Random Forest	Chain	GPT-4	66.40%
GPT-3.5	SVM	Chain	Flan-T5	70.40%
GPT-3.5	Random Forest	Chain	Flan-T5	60.00%
GPT-3.5	SVM	Step-mean	GPT-4	70.40%
GPT-3.5	Random Forest	Step-mean	GPT-4	59.20%
GPT-3.5	SVM	Step-min	GPT-4	70.40%
GPT-3.5	Random Forest	Step-min	GPT-4	58.40%
GPT-3.5	SVM	Step-hmean	GPT-4	**74.40%**
GPT-3.5	Random Forest	Step-hmean	GPT-4	59.20%

Table 1: The results of the RC-prompting after training with external scorers. The predictions and chain scores of the original LLM are not used in the final training.

Discussion

Our first research question was whether the LLMs can recognize the probative value of evidence. Both prompting methods, relevance-chain and chain-of-thought, can explain the relevance of the evidence given the facts in their outputs. Figure 8 shows a relevance chain generated by GPT-4 from the test set. Both GPT models could output the relevance chain in the given format by connecting the fact and the evidence, while in the case of Flan-T5, the output was not always in the desired format. "Makes it more/less likely" relations were missing often. Flan-T5 also struggled to output a chain-of-thought prompt, as the outputs were often short explanations with one or two sentences.



Relevance chain:
Step 1: The report of someone hitting Zeng with an iron bar makes it more likely that defendant Lin Zeyuan, who is reported to have used the same weapon, is involved.
Step 2: The involvement of Lin Zeyuan in the assault makes it more likely that his accomplices Chen Liquan, Zhu Yidi, and Ruan Jiantao also participated as per the fact.

Chain score: 10

Prediction: Highly relevant

Figure 8: A relevance chain output from the test set generated by GPT-4. The chain score is in red, and the predicted class is in blue.

The text outputs of the labels were quite robust in the case of GPT models, as no text variation was observed except for the punctuation. In the outputs of Flan-T5 models, labels such as "High relevant", "High relevance", "Very relevant" or "Semi-relevant" were observed. The first three are accepted as "Highly relevant" and the last as "Partially relevant" as those preserved the meaning of the original labels. We counted the labels as wrong if the variation does not preserve the meaning of the original label. Nevertheless, the total number of variations was no more than 5% in the best results.

Our second research question was whether existing measures of chain strength reflect probative value. The training with ROSCOE metrics helped determine the admissibility of the evidence, increased prediction accuracy, and helped RC-prompting outperform CoT-prompting. We believe that one reason for RC-prompting's increased performance in post-processing is the dense information flow in the chain, which describes a clear path to the hypothesis to increase or decrease its likelihood.

Another key aspect for increased accuracy is using the chain scores in the training, improving accuracy. Thus, we experimented with external LLMs to find the best methods to assign a chain score. As shown previously, the chain score remained the best option overall, while taking the harmonic mean of the step scores has a competitive performance. One possible explanation for the performance of harmonic mean might be its ability to limit the impact of outliers, leading to a more balanced chain score. Our experiments show that LLMs successfully assign a chain score to the output. However, the restricted size of our test set might limit the robustness of the outcome.

Conclusion

We showed that LLMs can successfully utilize RC-prompting for legal tasks to assess relevant evidence. Once the outputs are postprocessed with ROSCOE metrics, it outperforms the CoT-prompting method on GPT models but has the exact opposite effect on Flan-T5. We suspect that the lower number of parameters and allowed maximum number of tokens in Flan-T5 are the reasons for the performance difference, and this suggests a deeper and qualitative difference in the kinds of reasoning both types of LLMs can produce. Future work will need to explore this more.

In our work, we applied relevance chains for the legal relevance task to determine the relevance of the evidence. We hope to explore whether it is possible to apply our method to different domains where the identification or explanation of relevance is required. For easy replicability, we provide access to our source code files.[1]

Acknowledgements

Part of this research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

[1] https://github.com/Advancing-Machine-Human-Reasoning-Lab/relevance-chain

References

Biedermann, A., and Taroni, F. 2012. Bayesian networks for evaluating forensic dna profiling evidence: a review and guide to literature. *Forensic Science International: Genetics* 6(2):147–157.

Blair-Stanek, A.; Holzenberger, N.; and Van Durme, B. 2023. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Cooper, W. S. 1971. A definition of relevance for information retrieval. *Information storage and retrieval* 7(1):19–37.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20:273–297.

Fenton, N.; Berger, D.; Lagnado, D.; Neil, M.; and Hsu, A. 2014. When 'neutral'evidence still has probative value (with implications from the barry george case). *Science & Justice* 54(4):274–287.

Fenton, N.; Neil, M.; and Lagnado, D. A. 2013. A general structure for legal arguments about evidence using bayesian networks. *Cognitive science* 37(1):61–102.

Golovneva, O.; Chen, M.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; and Celikyilmaz, A. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

2023. Google translate. `https://translate.google.com/`. Accessed: September 27, 2023.

Hagendorff, T.; Fabi, S.; and Kosinski, M. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science* 3(10):833–838.

Harter, S. P. 1992. Psychological relevance and information science. *Journal of the American Society for information Science* 43(9):602–615.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, 278–282. IEEE.

Koehler, J. J. 1996. On conveying the probative value of dna evidence: Frequencies, likelihood ratios, and error rates. *U. colo. L. rev.* 67:859.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35:22199–22213.

Liu, Q.; Islam, B.; and Governatori, G. 2021. Towards an efficient rule-based framework for legal reasoning. *Knowledge-Based Systems* 224:107082.

Montgomery, D. C. 2017. *Design and analysis of experiments*. John wiley & sons.

Nighojkar, A.; Laverghetta Jr., A.; and Licato, J. 2023. No strong feelings one way or another: Re-operationalizing neutrality in natural language inference. In *Proceedings of the 17th Linguistic Annotation Workshop*. Toronto, Canada: Association for Computational Linguistics.

Nye, M.; Andreassen, A. J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Prakken, H., and Kaptein, H. 2016. *Legal evidence and proof: statistics, stories, logic*. Routledge.

Saracevic, T. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: Nature and manifestations of relevance. *Journal of the American society for information science and technology* 58(13):1915–1933.

Schamber, L., and Eisenberg, M. 1988. Relevance: The search for a definition.

2024. Scikit-learn. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`. Accessed: March 26, 2024.

Sperber, D., and Wilson, D. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Vlek, C. S.; Prakken, H.; Renooij, S.; and Verheij, B. 2016. A method for explaining bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law* 24:285–324.

Vlek, C. S. 2016. When stories and numbers meet in court. *Constructing and Explaining Bayesian Networks for Criminal Cases with Scenarios, Diss. Groningen*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35:24824–24837.

Wellborn III, O. G. 1976. Federal rules of evidence and the application of state law in the federal courts. *Tex. L. Rev.* 55:371.

Wells, G. L. 2014. Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Current Directions in Psychological Science* 23(1):11–16.

2023. Wenshu court. `https://wenshu.court.gov.cn`. Accessed: September 23, 2023.

Wilson, D., and Sperber, D. 2002. Truthfulness and relevance. *Mind* 111(443):583–632.

Wilson, D., and Sperber, D. 2006. Relevance theory. *The handbook of pragmatics* 606–632.

Woods, J. 2010. Relevance in the law: A logical perspective. *Approaches to Legal Rationality. Springer, Dordrecht*.

Yao, F.; Zhang, J.; Zhang, Y.; Liu, X.; Sun, C.; Liu, Y.; and Shen, W. 2023. Unsupervised legal evidence retrieval via contrastive learning with approximate aggregated positive. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4783–4791.

Yu, F.; Quartey, L.; and Schilder, F. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.