If You Give an LLM a Legal Practice Guide

Aaron D. Tucker [*1] **Colin Doyle** [*2 3]

Abstract

Large language models struggle to answer legal questions that require applying detailed, jurisdiction-specific legal rules. Lawyers also find these kinds of questions difficult to answer. For help, lawyers turn to legal practice guides: expert-written how-to manuals for practicing a particular type of law in a particular jurisdiction. Might large language models also benefit from consulting these practice guides? This article examines how providing LLMs with information from legal practice guides can affect their performance with answering legal questions and predicting case outcomes. Initial findings suggest that injecting relevant excerpts from practice guides into prompts for LLMs tends to improve performance. If a practice guide is used to structure a series of LLM queries that each analyze discrete issues which are then combined to answer a broader legal question, LLM performance can sometimes be substantially improved but can sometimes become worse than just using a practice guide. Results vary considerably across models and legal subject areas. These findings have implications for the potential for generative A.I. to automate legal tasks, particularly through agentic systems and retrieval augmented generation (RAG).

Introduction

Despite being trained on vast corpuses of data, LLMs often struggle to correctly answer questions that depend upon knowledge of domain-specific information. Retrieval-augmented generation (RAG) has emerged as a method for improving LLM performance by grounding LLM responses to a set of information. (Lewis et al., 2020). With RAG, information is retrieved from a knowledge database and then injected into part of the prompt given to an LLM. RAG pipelines function like an open-book exam, giving the model a chance to answer a question using retrieved information.

RAG *ought* to help LLMs answer legal questions (Ajmi, 2024). One noted shortcoming of LLM performance with legal reasoning tasks is a lack of knowledge about jurisdiction-specific rules and precedent (Magesh et al., 2024). With American law, all fifty states have their own, independent state constitutions, statutes, rules of procedure, and governing caselaw. Although the current generation of LLMs have been trained on enormous datasets, LLMs still struggle to properly answer legal questions that depend upon jurisdiction-specific rules (Dahl et al., 2024). RAG pipelines should be able to provide LLMs with this granular, jurisdiction-specific legal information.

Whether LLMs can effectively use that information is another question (Magesh et al., 2024). RAG pipelines are most successful when retrieving factual information that an LLM can copy into a response. But in a legal context, RAG pipelines don't retrieve facts for an LLM to parrot so much as they retrieve legal rules for an LLM to extract and apply within a response. Legal questions are often complex with multiple related parts and conditional logic. Even if an application retrieves the correct legal rules, the scope or complexity of those rules may overwhelm an LLM's capacity to competently apply those rules to a fact pattern (Chen et al., 2024).

This article examines how providing LLMs with information from legal practice guides can affect their performance answering related legal questions. Legal practice guides are a type of legal reference that helps attorneys become acclimated to a legal practice area without having to build that understanding from scratch by reading troves of statutes, regulations, and legal opinions (Davis). Practice guides are an ideal use case for evaluating RAG's potential for improving LLM performance at answering legal questions. Compared to other legal documents, practice guides are clear and succinct, and they are already structured as instructions for attorneys to follow. With other kinds of legal documents, the verbosity and complexity of the retrieved material might cause an LLM to provide erroneous responses. Experiments designed to test LLMs' capacity to apply legal principles

[*]Equal contribution [1]Department of Computer Science, Cornell University, Ithaca NY, USA [2]Loyola Law School, Loyola Marymount University, Los Angeles CA, USA [3]Berkman Klein Center, Harvard University, Cambridge MA, USA. Correspondence to: Aaron Tucker <aarondtucker@cs.cornell.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

based on retrieved information may instead become experiments testing LLMs' capacity to parse lengthy, confusing documents. Practice guides provide as clean as possible of an opportunity to observe how LLMs can apply a retrieved legal rule or principle in practical scenarios.

If existing practice guides can improve LLM performance at legal tasks, this has significant implications for the potential for LLMs to perform legal work in the near future. Thousands of practice guides have already been written on virtually every area of law for every jurisdiction in the United States. Practice guides may be a bountiful resource for increasing LLM performance on technical, domain-specific legal tasks as an alternative or supplement to more expensive, time-consuming processes like finetuning models on jurisdiction-specific caselaw.

Background

Legal Background

We chose to test LLM performance answering legal questions in three different legal practice areas in three different U.S. jurisdictions: California law governing the tort law doctrine of res ipsa loquitur, Minnesota law governing the state's power of eminent domain, and New Jersey criminal law concerning pretrial incarceration. Each of these practices areas includes state-specific rules that make legal questions on these topics difficult for current LLMs to answer. These practice areas were selected to test LLM performance in a variety of jurisdictions and legal subject matter areas. Each practice area belongs to a different state in a different region of the country, and each concerns a different kind of litigation: private civil litigation that depends on common law rules; public law litigation rooted in statutory law, constitutional law, and state constitutional law; and procedural rules in criminal law that depend upon recently enacted state statutes and a state constitutional amendment.

California Res Ipsa Loquitur. California's tort law doctrine of res ipsa loquitur is a set of common law rules concerning the plaintiff's evidentiary burden in a negligence cause of action. (Thomas et al.). In an ordinary negligence lawsuit, a plaintiff must prove that the defendant breached a duty of reasonable care to the plaintiff. The plaintiff must establish what would have constituted reasonable care under the circumstances and how the defendant failed to exercise reasonable care. For cases in which the defendant's negligence can be readily inferred based on the plain facts of the case, the doctrine of res ipsa loquitur makes the plaintiff's job easier. Latin for "the thing speaks for itself," res ipsa loquitur creates a presumption of negligence that the defendant has the opportunity to rebut. Under California law, for res ipsa loquitur to apply: (1) the accident must be of a kind which ordinarily does not occur in the absence of someone's negligence; (2) it must have been caused by an agency or instrumentality within the exclusive control of the defendant; and (3) the accident must not have been due to any voluntary action or contribution on the part of the plaintiff. (California, 1993) Res ipsa loquitur is a state law doctrine, but it's a foundational legal concept taught to every first-year law student, and the doctrine exists in every state in the United States with slight variation from jurisdiction to jurisdiction.

Minnesota Eminent Domain. Minnesota state law governing eminent domain is a combination of state statutory law, state constitutional law, and federal constitutional law. (Ramsey). Each of these sources of law restrict the government's power to seize private property. The U.S. Constitution sets the floor for individual rights protections against state use of eminent domain power across the United States. The government can seize private property only if the state takes that property for a public purpose and provides the original property owner with just compensation for the taking. LLMs are likely to be able to answer questions about the broader legal backdrop and basic principles of restrictions on government use of its power of eminent domain. But many states offer property owners greater protection from government takings than the federal constitution provides. Minnesota is one of those states. Compared to federal takings law, Minnesota law has a narrower interpretation of what kinds of takings can constitute a "public purpose," and Minnesota has an additional "necessity" requirement that is absent at the federal level. Minnesota also has very specific statutory rules for when a property can be classified as blighted, abandoned, or environmentally contaminated — thereby justifying a government taking. These laws regarding eminent domain are unlikely to be well represented within LLM training data because the legal rules apply only to the state of Minnesota and differ from the rules of other states.

New Jersey Pretrial Detention. New Jersey pretrial incarceration laws are set of a procedural rules that must be followed for the state to legally incarcerate a person after arrest but before trial in a criminal case. (Dorsey et al.). These rules are specific, detailed statutory and state constitutional provisions that were enacted in New Jersey in 2017 as comprehensive criminal justice reform legislation. In New Jersey, criminal defendants cannot be detained pretrial on unaffordable money bond. Rather, a prosecutor must motion for a pretrial detention hearing. At the hearing, the defense is allowed to cross-examine government witnesses and present its own evidence. To incarcerate someone pretrial, a court must consider a specific set of factors and make written findings concluding that no set of conditions of release would ensure the safety of the community, guarantee that the defendant would return to court, or prevent the de-

fendant from obstructing the judicial process. These rules are specific to the state of New Jersey and are not applicable to other states, most of which lack detailed procedural rules governing pretrial incarceration.

Computer Science Background

Retrieval Augmented Generation. Retrieval Augmented Generation was originally introduced with the motivation of being able to improve performance and factuality in knowledge-intensive tasks by allowing a model to retrieve information for its context window (Lewis et al., 2020). Most of the benchmarks in the original paper were different variants of question answering where the correct response was directly contained in or a summary of part of the text, though the FEVER task is based on deciding whether or not a given fact is entailed by different pieces of text (e.g. "ICML 2024 took place in the capital of Austria" is supported by "ICML 2024 was in Vienna" and "Vienna is the capital of Austria") (Thorne et al., 2018).

In contrast, our setting is more complicated, since legal reasoning often depends on precise technical wording, explicit legal definitions of terms, layers of logical reasoning, and practical knowledge about the world. In a legal context, rather than using RAG to augment the facts that the the LLM can access, RAG is used to augment the principles that the LLM can apply. For example, in Minnesota Eminent Domain, "public purpose" has a much more specific meaning than both the common usage of the phrase *and* the general legal definition of a phrase. Successfully using a provided practice guide then requires the LLM to be able to re-interpret the meaning of "public purpose" for Minnesota Eminent Domain specifically without interference from its usage in other contexts.

Propositional Logic. Propositional Logic is a classic topic in Philosophy and Computer Science, which analyzes the truth value of different combinations of logical statements. For this paper, we only need "not", "or", and "and". The statement not A (denoted $\bar{A}$) is true if A is false, and false if A is true. The statement A and B (denoted $A \wedge B$) is true if A and B are both true, and false otherwise. The statement A or B (denotes $A \vee B$) is false if both A and B are false, and true otherwise.

A	B	$\bar{A}$	$A \wedge B$	$A \vee B$
False	False	True	False	False
False	True	True	False	True
True	False	False	False	True
True	True	False	True	True

Both $\wedge$ and $\vee$ are associative, meaning that

$$(A \wedge B) \wedge C = A \wedge (B \wedge C) = A \wedge B \wedge C,$$
$$(A \vee B) \vee C = A \vee (B \vee C) = A \vee B \vee C.$$

These operations can also be extended into probabilistic settings. If we take True to be the value 1 and False to be the value 0, then a proposition which is true with probability p is has the value p. Not can then be implemented as $\bar{p} = (1-p)$, and is $p \wedge q = p * q$. Or is more complicated, since if A is true with probability p and B is true with probability q, then if A and B are independent events then the probability of p or q is $p + q - p * q$ to avoid double-counting the possibility that they are both true. For example, the probability that at least one of two fair coin flips is heads is 75%. Using the fact that $A \vee B = \overline{\bar{A} \wedge \bar{B}}$, we have

$$p \vee q = \overline{\bar{p} * \bar{q}} = \overline{(1-p) * (1-q)} = 1 - (1-p) * (1-q).$$

Methods

Datasets. For datasets, we chose three practice guides covering different areas of law from different U.S. jurisdictions: a California civil practice guide for tort law, a Minnesota practice guide for real estate law, and a New Jersey practice guide on criminal procedure. (Thomas et al.; Ramsey; Dorsey et al.). Within each guide, we selected a particular legal topic to test: for California, the tort law doctrine of res ipsa loquitur; for Minnesota, state statutory and constitutional law governing eminent domain; and for New Jersey, state statutes concerning pretrial detention procedures.

For each topic, we extracted from the practice guides any relevant instructions for answering legal questions on that topic. These excerpts simulate a best case scenario for information retrieval, allowing us to measure the model's performance given that the correct part of a practice guide has been included within the prompt.

For our experiments on *real cases*, we manually extracted the facts and holding of each relevant case referenced within that part of the practice guide. For each case, the facts provide background information on the parties and the legal dispute, and the holding provides the correct legal conclusion along with the reasoning behind that conclusion. This resulted in 12 California res ipsa cases and 5 Minnesota takings cases. Since the New Jersey pretrial detention reforms went into effect in 2017 and have not yet produced a sizeable body of caselaw, we did not include real cases.

For our experiments on *hypothetical cases*, a legal expert wrote hypothetical examples to cover the different elements of the legal principles contained in the practice guide, and annotated each example with the correct overall legal conclusion, along with the correct conclusion for each legal subissue. We had 13 hypotheticals for California Res Ipsa Loquitur, 20 hypotheticals for Minnesota Eminent Domain, and 14 hypotheticals for New Jersey pretrial detention.

Prompting. We used four different prompt templates to evaluate LLM performance at answering legal questions.

The first two prompt templates (*Name*) and (*Fact*) served as controls to establish the LLM's baseline performance absent any help from the practice guide. (*Name*) provided the LLM with the name of the legal case, without any facts of the case or information from the practice guide. (*Fact*) provided the LLM with only the facts of the case, without information from the practice guide. The third prompt template (+*Guide*) provided the LLM with the facts of the case and the excerpt from the practice guide. For these prompt templates we took advantage of probabilistic LLM outputs by requesting 10 responses from the LLM for each query for each proposition/question, and then averaging over the results.

The final prompt template (*Prop.*) broke the excerpt of the practice guide down into distinct components based on different parts of the relevant legal rule. A separate LLM query was made for each part of the legal rule, and the LLM was asked to evaluate whether that part of the legal rule was met. For this prompt template we requested 10 responses for each query, then if $> 50\%$ of the responses were true then we defined the proposition as true. We then combined the different propositions into an overall score for the hypothetical using the following definitions.

Res Ipsa:

the accident was...

- of a kind which ordinarily does not occur in the absence of someone's negligence
- $\wedge$ caused by an agency or instrumentality within the exclusive control of the defendant;
- $\wedge$ not due to any voluntary action or contribution on the part of the plaintiff

Minnesota Eminent Domain:

The taking was necessary

- $\wedge$ the government paid just compensation for the taking
- $\wedge$ the taking was for a public purpose: the taking was for... (
  - the possession, occupation, ownership, and enjoyment of the land by the general public or by public agencies
  - $\vee$ the creation or functioning of a public service corporation
  - $\vee$ the mitigation of a blighted area
  - $\vee$ the remediation of an environmentally contaminated area
  - $\vee$ the reduction of abandoned property
  - $\vee$ the removal of a public nuisance
  - $\vee$ the mitigation of a blighted area

)

New Jersey Pretrial Detention:

There was a lawful, valid justification for the judge to order the defendant to be detained pretrial

- $\wedge$ The court followed the correct procedures (
  - The defendant who was detained pretrial was eligible for pretrial detention
  - $\wedge$ The defendant was granted a pretrial detention hearing within the timeframe required by law
  - $\wedge$ The pretrial detention of the defendant was the result of legally required motion and hearing
  - $\wedge$ The court considered the correct factors when it decided to detain the defendant pretrial
  - $\wedge$ The pretrial detention hearing adhered to due process requirements

)

We had two variants of the propositional strategy. *Prop. 2* broke down the prompt into only the first-level propositions and combined all of the LLM subqueries for the proposition into a single LLM query. *Prop. 3* used the full propositional structure outlined for the prompt.

Results on Real Cases

LLM	Name	Facts	+Guide	Prop.
GPT-3.5	0.3	0.53	0.53	<u>0.63</u>
GPT-4	0.44	0.75	0.74	**0.83**
Claude Haiku	0.36	0.67	0.71	**0.83**
Claude Sonnet	0.34	0.61	0.64	<u>0.75</u>
Claude Opus	0.35	0.71	<u>0.77</u>	0.66

Figure 1. Accuracy on Real California Res Ipsa Loquitur Cases

LLM	Name	Facts	+Guide	Prop.
GPT-3.5	0.45	0.9	0.72	**1.0**
GPT-4	0.64	**1.0**	0.98	0.8
Claude Haiku	0.18	0.78	0.58	<u>0.8</u>
Claude Sonnet	0.2	0.76	0.53	<u>0.8</u>
Claude Opus	0.2	<u>0.98</u>	0.74	0.8

Figure 2. Accuracy on Real Minnesota Eminent Domain Cases

Our initial findings show variability between the different settings. The Claude series performance on Res Ipsa was similar to what we expected – the model has some ability to predict case outcomes based on the facts of the case, and has

improved performance when additionally given the practice guide. But the GPT series is different, and has similar performance using *Facts* and *+Guide*. More counter-intuitively, on the Minnesota Eminent Domain cases providing the practice guide consistently *harms* performance, but the more involved *Prop.* method of the practice guide where each element is given a separate was helpful for the weaker models. This is surprising, both because we expect the practice guide to lead to improvement, and because we expect the practice guide to be most helpful in domains with state-specific variation.

Another surprising finding was that increasing model capability within the Claude series on the California res ipsa cases had mixed results. For the *Facts* and *+Guide* methods, the Sonnet model does worse than the Haiku model despite its increased capability. The *Prop.* method even became *worse* with increasing model quality, exhibiting *inverse scaling* (McKenzie et al., 2024). This may be specific to res ipsa, because as the model capability increases the LLMs become increasingly creative in explaining why an accident might occur without negligence. Since res ipsa loquitur only applies if the "incident was of a type that does not generally happen without negligence" (Legal Information Institute, n.d.), more capable models may have been misled by their creative capacity to conjure up scenarios in which the incident could have happened in the absence of negligence. On the other hand, we only used 12 cases, so this result might not be statistically significant.

Limitations of our real cases. In the course of our investigation, we found that predicting the outcomes of real cases was less straightforward of an experiment than we had initially anticipated.

First, legal opinions often do not have a clean separation between the facts of the case and the legal conclusions drawn from applying legal rules to those facts. Legal opinions are persuasive texts. This raises two issues. First, the facts are often intermixed with legal reasoning about those facts. Second, the facts are often characterized in a way that supports the legal conclusions to be drawn from those facts. Although we tried to extract "clean" versions of the facts of each case without rewriting parts of the original text, the source material makes it more difficult to discern how much an LLM's legal conclusions are the product of the LLM's legal reasoning skills as opposed to the LLM's ability to extract legal conclusions from contextual clues about how facts have been characterized.

Second, given that the facts come from appellate cases, these cases tend to concern difficult, thorny legal questions concerning gray areas of the law. In contrast, practice guides tend to be concerned with the everyday routine application of the law — not the less common cases in which the legal outcome is uncertain. This makes the legal outcomes of these cases less clearly useful as a "ground truth" label of whether or not a legal practice guide is helpful for an LLM applying law to facts.

Third, the legal opinions were from appellate courts that sometimes defer to prior lower court findings or would remand the case to a lower court to reach a final conclusion on a legal issue. For example, many of the res ipsa cases cited by our practice guide were appellate cases in which the appellant claimed that a jury should have been instructed on res ipsa. Even when the appellant won the appeal, the result was rarely that the appellate court rules that res ipsa did apply. Rather, the appellate courts would rule that res ipsa could have applied and therefore the case would be remanded to the trial court for a new trial for a jury to make the final determination of whether res ipsa did apply. This makes it harder to evaluate LLM legal reasoning using real case outcomes because those case outcomes did not offer clear yes-or-no legal conclusions based on a set of facts.

Results on Hypothetical Cases

LLM	Facts	+Guide	Prop.
GPT-3.5	0.53	0.68	0.77
GPT-4	0.58	0.72	0.77
Claude Haiku	0.55	0.55	0.62
Claude Sonnet	0.49	0.71	0.77
Claude Opus	0.82	**0.88**	0.85

Figure 3. Accuracy on California Res Ipsa Loquitur Hypotheticals

	Facts	+Guide	Prop. 2	Prop. 3
GPT-3.5	0.45	0.85	0.70	0.70
GPT-4	0.62	0.91	0.90	0.80
Claude Haiku	0.71	0.87	0.70	0.75
Claude Sonnet	0.65	0.92	0.75	0.75
Claude Opus	0.76	0.93	**0.95**	0.90

Figure 4. Accuracy on Minnesota Eminent Domain Hypotheticals

	Facts	+Guide	Prop. 2	Prop. 3
GPT-3.5	0.67	0.68	0.71	0.71
GPT-4	0.61	0.74	0.79	0.71
Claude Haiku	0.61	0.62	0.71	0.79
Claude Sonnet	0.61	0.74	0.86	**0.93**
Claude Opus	0.74	0.81	**0.93**	0.71

Figure 5. Accuracy on New Jersey Pretrial Detention Hypotheticals

We also investigated the model performance on a newly-written set hypothetical cases which were designed by a legal expert to test how well LLMs could follow each individual element of each practice guide. In these experiments,

	Facts	+Guide	Prop. 2	Prop. 3
GPT-3.5	0.12	0.96	**1.00**	**1.00**
GPT-4	0.18	0.82	0.89	0.67
Claude Haiku	0.48	0.79	**1.00**	**1.00**
Claude Sonnet	0.52	0.92	**1.00**	**1.00**
Claude Opus	0.52	0.86	**1.00**	0.89

Figure 6. Accuracy on Minnesota-Specific Eminent Domain Hypotheticals

	Facts	+Guide	Prop. 2	Prop. 3
GPT-3.5	0.73	0.76	0.45	0.45
GPT-4	**0.99**	0.97	0.91	0.91
Claude Haiku	0.90	0.95	0.45	0.55
Claude Sonnet	0.75	0.92	0.55	0.55
Claude Opus	0.96	0.98	0.91	0.91

Figure 7. Accuracy on Non-Minnesota-Specific Eminent Domain Hypotheticals

we additionally broke down the *Prop.* method into two levels. *Prop. 2* breaks the practice guide down into multiple questions, and *Prop. 3* breaks the practice guide down into different questions which can themselves have subquestions.

Figures 6 and 7 offer another view on the data captured in Figure 4 on Minnesota eminent-domain hypotheticals. Figure 6 concerns a subset of hypotheticals in which knowledge of Minnesota-specific law is necessary for arriving at the correct legal conclusion. Figure 7 concerns a subset of hypotheticals in which general knowledge of eminent domain law across the country would be sufficient for arriving at the correct legal conclusion. Surprisingly, while *Prop* methods obtain the best performance on the Minnesota-specific questions, *+Guide* still gets the best performance on non-Minnesota-specific questions for most models despite the fact that there is no need for Minnesota-specific information.

These experiments have produced surprisingly variable results. Across different subject areas, +Guide, Prop. 2, and Prop. 3 each achieved the highest accuracy for a particular model. None of the methods for answering legal questions consistently produced stronger results across all of the models. Using RAG to inject relevant information from legal practice guides tended to improve performance across all models but occasionally made no difference. From model to model and subject area to subject area, the Prop. 2 and Prop. 3 methods sometimes improved and sometimes hurt LLM accuracy compared to the practice guide alone. Consider Figure 4, Accuracy on Minnesota Eminent Domain Hypotheticals. For every model except Claude Opus, the Prop. 2 method was less accurate than giving the model the guide alone. But with Claude Opus, the Prop. 2 method not only produced the most accurate results for the model, it produced the most accurate results overall across all models and methods for that subject area.

Real v. Hypothetical cases. The results are fairly different between real and hypothetical cases, even holding the legal domain fixed. In our hypothetical cases, adding the practice guide always helped. In our real cases *+Guide* was always worse than *Facts* on Minnesota Eminent domain, and *+Guide* only beat *Facts* on Res Ipsa when using the Claude series of models. Why is the practice guide more useful on the hypotheticals than on the real cases? There are many different explanations that future work could explore.

One theory is simply length: naively applying the practice guide does relatively poorly on the real Minnesota Eminent Domain cases, which has relatively long real case facts (as compared to hypotheticals) as well as a relatively long practice guide (as compared to Res Ipsa). Only GPT-4s seem to do well at using the information in this setting. This theory could possibly be tested by seeing how performance varies while adding more and more irrelevant text.

A second theory is that LLMs are trained on text found across the internet, and so they do better on evaluations that look like academic tests. Note that in our experiments the hypotheticals often have worse overall performance within a domain, however the distribution of the hypotheticals and real cases are different. A useful experiment might be to compare performance on hypothetical versions of cases to performance on the real cases.

A third theory is that the cases themselves are from quite different distributions. The hypotheticals were written to test how well the LLMs can handle each specific requirement of the practice guide, and so it makes sense that the practice guide would be more helpful.

Conclusion

Initial findings suggest that injecting relevant excerpts from practice guides into prompts for LLMs tends to improve LLM performance at answering legal questions. If a practice guide is used to structure a series of LLM queries that each analyze discrete issues which are then combined to answer a broader legal question, LLM performance can sometimes be substantially improved but can sometimes become worse than just using a practice guide alone. Results vary considerably across models and legal subject areas. These findings have implications for the potential for generative A.I. to automate legal tasks, particularly through agentic systems and retrieval augmented generation (Choi & Schwarcz, 2023).

References

Ajmi, A. Revolutionizing Access to Justice: The Role of AI-Powered Chatbots and Retrieval-Augmented Generation in Legal Self-Help. 2024.

California, S. C. Brown v. Poway Unified School Dist., 1993.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i16.29728.

Choi, J. H. and Schwarcz, D. B. AI Assistance in Legal Analysis: An Empirical Study. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4539836.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, January 2024.

Davis, J. LibGuides: Tort Law Research Guide: Practice Guides. https://lawlibguides.usc.edu/c.php?g=687841&p=4879061.

Dorsey, J. R., Gunn, B. J., Simpson, M. D., Mikhail, P. G., and Bjerkness, G. L. Chapter 10. Eminent Domain. In *25 Minn. Prac., Real Estate Law*, volume 25 of *Minnesota Practice Series*. Thomson West.

Legal Information Institute. res ipsa loquitur. https://www.law.cornell.edu/wex/res_ipsa_loquitur, n.d. Accessed: July 10th, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, May 2024.

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., Tseng, T., Korbak, T., Shen, X., Zhang, Y., Zhou, Z., Kim, N., Bowman, S. R., and Perez, E. Inverse scaling: When bigger isn't better, 2024.

Ramsey, R. Chapter 19. Criminal Justice Reform. In *Criminal Practice and Procedure*, volume 31 of *New Jersey Practice Series*. Thomson West.

Thomas, M. P., McGhee, Z. A., Kahn, B. D., and La Scala, S. L. 1:29. Presumption of breach arising from type of accident ("res ipsa loquitur"). In *Torts (California Civil Practice)*. Bancroft Whitney.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.