Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling

Hang Jiang[†∅], **Xiajie Zhang**[†], **Robert Mahari**[†◇], **Daniel Kessler**[†∅], **Eric Ma**[□],
Tal August[∞], **Irene Li**[¬⊣], **Alex 'Sandy' Pentland**[†], **Yoon Kim**[†], **Deb Roy**[†∅], **Jad Kabbara**[†∅]
[†]Massachusetts Institute of Technology, [∅]MIT Center for Constructive Communication
[◇]Harvard Law School, [□]University of Virginia School of Law
[∞]Allen Institute for AI, [¬]University of Tokyo, [⊣]Smartor.me

Abstract

Making legal knowledge accessible to non-experts is crucial for enhancing general legal literacy and encouraging civic participation in democracy. However, legal documents are often challenging to understand for people without legal backgrounds. In this paper, we present a novel application of large language models (LLMs) in legal education to help non-experts learn intricate legal concepts through storytelling, an effective pedagogical tool in conveying complex and abstract concepts. We also introduce a new dataset LEGALSTORIES, which consists of 294 complex legal doctrines, each accompanied by a story and a set of multiple-choice questions generated by LLMs. To construct the dataset, we experiment with various LLMs to generate legal stories explaining these concepts. Furthermore, we use an expert-in-the-loop approach to iteratively design multiple-choice questions. Then, we evaluate the effectiveness of storytelling with LLMs through randomized controlled trials (RCTs) with legal novices on 10 samples from the dataset. We find that LLM-generated stories enhance comprehension of legal concepts and interest in law among non-native speakers compared to only definitions. Moreover, stories consistently help participants relate legal concepts to their lives. Finally, we find that learning with stories shows a higher retention rate for non-native speakers in the follow-up assessment. Our work has strong implications for using LLMs in promoting teaching and learning in the legal field and beyond.

1 Introduction

Often individuals find themselves in certain high-stakes situations where they have to educate themselves on novel concepts such as new policies before voting, mortgage terms when buying a house or legal principles relevant to an ongoing lawsuit. Unfamiliar terms and nuanced use of language in these contexts can make it challenging for non-experts to make informed decisions, to have equal access to justice, or to participate in civic discourse and democracy. We present this work as a step towards enhancing general legal literacy, bridging the gap between non-experts and experts and promoting constructive and civic discourse.

Storytelling is an important medium to communicate science to non-experts (Dahlstrom, 2014; Martinez-Conde and Macknik, 2017) and teach professional knowledge to beginners (Abrahamson, 1998; Davidhizar and Lonser, 2003; Gallagher, 2011). In legal contexts, storytelling has been used extensively to teach abstract legal concepts such as ethics (Menkel-Meadow, 1999), and has proven effective at explaining complex legal concepts such as legal mediation to the general public (Capuano et al., 2014). However, the scalable implementation of legal storytelling education is severely limited by the high costs associated with legal experts.

Large language models (LLMs) and their impressive text generation abilities have facilitated high-quality automated explanations and stories. Recent efforts (Huang et al., 2021; Murthy et al., 2021, 2022; August et al., 2022) have leveraged LLMs to generate accessible explanations of scientific or medical concepts for diverse audiences. Savelka et al. (2023) used GPT-4 to generate explanations for legal concepts from statutory provisions. However, to the best of our knowledge, previous work has not: (1) used LLM-generated stories as a medium to explain complex concepts, especially in the under-explored legal domain, (2) generated and refined (via expert feedback) questions for the assessment of concept comprehension, nor (3) validated the effectiveness of LLM-generated stories in enhancing comprehension among non-experts.

In this work, we explore a novel application of LLMs that focuses on the use of generated stories and questions to facilitate the learning and assessment of legal concept understanding. We use a human-in-the-loop pipeline that combines LLM

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7194–7219
August 11-16, 2024 ©2024 Association for Computational Linguistics

and expert input to generate stories and multiple-choice questions. We loop in both Prolific workers and legal experts[1] to ensure that the LLM-generated content is of high-quality. Our pipeline presents a holistic approach to LLMs' application in the legal education domain, where both the learning intervention (stories) and assessment (reading comprehension questions) are generated and evaluated. By providing a reusable dataset and promising experiment results, our work has strong implications for the broader use of LLMs to enhance teaching and learning and to improve general legal literacy. Our contributions are as follows:

- We create a novel legal education dataset, LEGALSTORIES, which presents legal concepts with their definitions, LLM-generated stories and questions, and human annotations for future NLP and legal education research[2].
- We provide extensive comparisons of three LLMs, namely, LLaMA 2, GPT-3.5, and GPT-4, to generate legal stories and questions with both automatic and human evaluations.
- We conduct RCTs with both native and non-native English speakers to learn legal concepts, demonstrating that LLM-generated stories improve concept comprehension and interest in law among non-native speakers compared to Wikipedia definitions. We also find that LLM-generated stories consistently help both native and non-native participants in relating legal concepts to their personal lives.

2 Related Work

Legal NLP & Accessible Language Legal language is complex and nuanced, and this creates a challenge for non-experts navigating legal processes (Benson, 1984). This challenge has prompted research into using computational tools to improve legal reading comprehension (Curtotti and McCreath, 2013). Making legal jargon more accessible represents an impactful application of legal NLP that promises to broaden access to justice (Mahari et al., 2023). Previous work has focused on legal text simplification (Collantes et al., 2015; Garimella et al., 2022; Cemri et al., 2022), legal summarization (Farzindar and Lapalme, 2004; Manor and Li, 2019), and question answering (Khazaeli et al., 2021; Zhong et al., 2020; Martinez-Gil, 2023) to make legal language more accessible. LLMs (Zhang et al., 2023) have also been used to improve legal access such as ChatLaw (Cui et al., 2023). Recently, Savelka et al. (2023) used GPT-4 to explain legal concepts in statutory provisions. However, none of the previous legal NLP work has combined LLMs and storytelling–a widely-used technique in education and communication–as a device to bridge legal experts and non-experts.

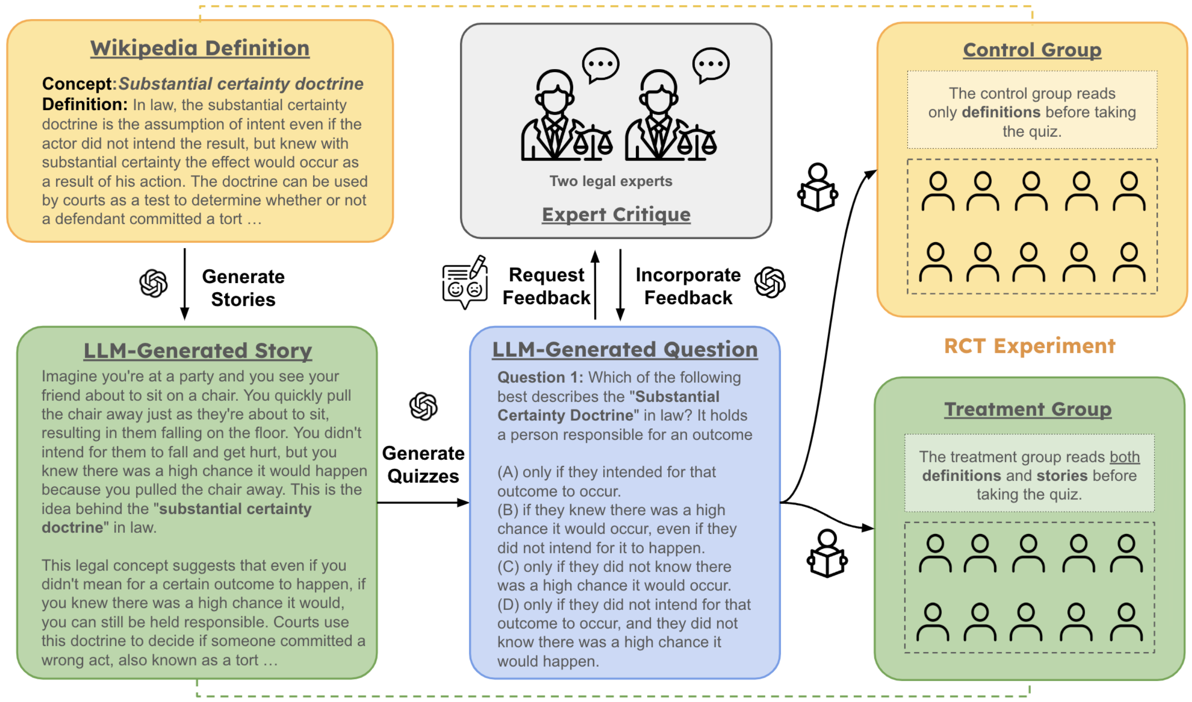
Storytelling in Education and NLP *Stories*, which we use interchangeably with the term *narratives*, are sequential depictions of actions and events (Abbott, 2020). They are an effective technique for pedagogy (Busselle and Bilandzic, 2009, 2008; Rapaport et al., 1989). Stories effectively illustrate complex concepts in various fields like math, science, and law, with law education heavily relying on "fact-patterns" (Papadimitriou, 2003). Specifically, second-person narratives are known to elicit particularly strong emotional and aesthetic-reflexive involvement from readers (Mildorf, 2016), especially when combined with more advanced narrative techniques and emotive verbs (Rembowska-Płuciennik, 2022). Automatic story generation is a long-standing task in NLP. Before LLMs, the best story generation models, even those using transformer architectures, struggled to create coherent stories with well-defined characters and story-lines (Alabdulkarim et al., 2021). Recent progress in LLMs such as ChatGPT have opened up new possibilities with storytelling, showing significantly higher quality compared to other story generation models (Xie et al., 2023; Zimmerman et al., 2022; Xu et al., 2020). However, little work has explored the use of LLM-generated stories in education. Valentini et al. (2023) experimented with LLMs to generate age-appropriate stories for children. Our work is the first to use LLM-generated stories to help non-experts in learning legal knowledge.

Educational Question Generation Generating educational questions or quizzes is important for educators to increase engagement, test reading comprehension, and improve learners' knowledge retention (Al Faraby et al., 2023). In the era of modern NLP, question generation (QG) was first tackled with seq2seq (Yuan et al., 2017; Zhou et al., 2018) and Transformer-based models (Narayan et al., 2020; Bao et al., 2020). Previous QG work tackled selecting question-worthy content (Liu et al.,

---

[1] This paper uses the term "legal experts" to refer to two people who have graduated with JD degrees or have made substantial progress towards earning JD degrees.

[2] Both the code and data are available at this repository: `https://github.com/hjian42/LegalStories`.



Concept: *Substantial certainty doctrine*
Definition: In law, the substantial certainty doctrine is the assumption of intent even if the actor did not intend the result, but knew with substantial certainty the effect would occur as a result of his action. The doctrine can be used by courts as a test to determine whether or not a defendant committed a tort …

Imagine you're at a party and you see your friend about to sit on a chair. You quickly pull the chair away just as they're about to sit, resulting in them falling on the floor. You didn't intend for them to fall and get hurt, but you knew there was a high chance it would happen because you pulled the chair away. This is the idea behind the "**substantial certainty doctrine**" in law.

This legal concept suggests that even if you didn't mean for a certain outcome to happen, if you knew there was a high chance it would, you can still be held responsible. Courts use this doctrine to decide if someone committed a wrong act, also known as a tort …

Question 1: Which of the following best describes the "**Substantial Certainty Doctrine**" in law? It holds a person responsible for an outcome

(A) only if they intended for that outcome to occur.
(B) if they knew there was a high chance it would occur, even if they did not intend for it to happen.
(C) only if they did not know there was a high chance it would occur.
(D) only if they did not intend for that outcome to occur, and they did not know there was a high chance it would happen.

The control group reads only **definitions** before taking the quiz.

The treatment group reads both **definitions** and **stories** before taking the quiz.

Figure 1: Illustration of the expert-in-the-loop pipeline. The left section demonstrates the procedure to produce an LLM-generated story from the concept. The lower section in the center shows how we use both the definition and story as input to produce LLM-generated reading comprehension (RC) questions. The center upper section shows that we first collect expert feedback on questions and regenerate questions with expert advice. The right section outlines the RCT experiment to see if LLM-generated stories improve comprehension in legal concepts.

2020; Steuer et al., 2021), modeling question types (Duan et al., 2017; Sun et al., 2018; Sugawara et al., 2018; Kang et al., 2019; Zhou et al., 2019) , or focused on specific interrogative words such as what, how, etc. (Ghanem et al., 2022). A subset of work focuses on the needs of students and teachers to tackle question generation to make them more accessible, adaptive, and educational (Le et al., 2014; Wang et al., 2018; Chen et al., 2018; Kurdi et al., 2020; Srivastava and Goodman, 2021; Wang et al., 2022; Leite and Cardoso, 2023; Laban et al., 2022). Recently, LLMs have been widely used to generate questions (Wang et al., 2022; Gabajiwala et al., 2022; Muse et al., 2023; Kasneci et al., 2023; Liang et al., 2023; Tran et al., 2023; Lu et al., 2023) in different domains such as language learning (Xiao et al., 2023), commonsense knowledge (Rathod et al., 2022), coding (Sarsa et al., 2022; MacNeil et al., 2023), and science (Bulathwela et al., 2023). However, prior research primarily evaluated generated questions via automatic metrics or simple human ratings. Differently from previous work, Steuer et al. (2022) evaluated the usability of generated questions in a practical learning setting and showed that asking non-native learners to finish questions after reading helps them understand science texts. Our work instead uses LLM-generated questions as a tool to assess the effectiveness of LLM-generated stories in legal concept learning.

3 LEGALSTORIES Dataset

In Figure 1, we present a pipeline and demonstrate how we apply it to curate a new dataset. It consists of three components: Story Generation, Question Generation, and Expert Critique. The following paragraphs illustrate how we generate stories and questions from legal doctrine definitions and refine questions with expert feedback.

3.1 Doctrine Definitions from Wikipedia

We collect 294 legal doctrines in English from the "legal doctrines and principles" page on Wikipedia[3]. A legal doctrine is a systematic framework or set of rules and procedures that evolves through legal precedent, particularly in common law systems. It serves as a guide for determining judgments in legal cases. The development of a legal doctrine often begins when a judge makes a ruling that involves a specific process, which is then outlined

---

[3] `https://en.wikipedia.org/wiki/Category:Legal_doctrines_and_principles`

and applied to the case. If this process is consistently used by other judges in similar situations and gains acceptance, it may become an established and widely accepted method for deciding similar cases in the future. A legal doctrine involves complex legal concepts whose definitions usually contain one or multiple legal terms, therefore difficult to comprehend to legal novices. We use the introduction paragraph as the definition for each legal doctrine. In our work, we focus on evaluating 101 out of 294 concepts whose definition length is between 100 and 200 words to ensure that they could be mapped to a medium-length story with around 500 words. We did not choose concepts whose definitions are too short or too long, because short definitions are not fair for participants in the control group to understand the concept and long definitions tend to have details missed by LLMs due to story length constraint. The mean and median lengths of the 101 concept definitions are 140.1 and 136.0 words.

3.2 Story and Question Generation

In this section, we describe the procedure to generate explanatory stories of legal concepts and three types of reading comprehension questions. We experiment with LLaMA 2 (Touvron et al., 2023), GPT-3.5[4] and GPT-4[5] (Achiam et al., 2023) given that they are among the state-of-the-art chat-based LLMs. See Appendix A.1 for details regarding the models and prompts.

3.2.1 Story Generation

As illustrated in the leftmost yellow and green boxes of Figure 1, we generate legal stories based on corresponding concepts and definitions. We find that a simple prompt (see Appendix A.2) is good enough to generate legal stories. We limit the story length to 500 words because the definitions are between 100 and 200 words and lengthy content tends to overwhelm the readers. For the twenty sampled concepts, the mean and standard deviation of the definition lengths are 152.0 ± 31.0 words. For the corresponding twenty stories used for human evaluation, the mean and standard deviation of their lengths are 316.8 ± 51.6 for GPT-4, 327.0 ± 50.7 for GPT-3.5, and 250.5 ± 89.9 for LLaMA 2.

3.2.2 Question Generation

Prior pedagogical research has highlighted different aspects of cognitive learning: remembering, understanding, applying, evaluating, analyzing, and creating (Adams, 2015). Inspired by this framework, we create three question types which are suitable for assessing learners' understanding of concepts. In these three cases, the model is asked to generate a multiple-choice question with a suggested answer and explanation, with each type assessing a certain kind of understanding as follows:

- **Concept question** for definition *understanding*: Here, the task requires the reader to pick the most precise description of a concept.
- **Prediction question** for *applying* the concept to scenarios: Here, the task involves asking the reader to forecast the outcome of a hypothetical situation that is related to the concept.
- **Limitation question** for *evaluating* and *analyzing* the concept's shortcomings: Here, the task requires the reader to identify a limitation or exception to the corresponding concept.

As depicted in the blue box on the lower center of Figure 1, we condition the question generation on the corresponding concept, definition, and story. The exact prompts are presented in Appendix A.3.

Question Refinement with Expertise As outlined in the central gray box at the top of Figure 1, we recruit two legal experts to read the concepts and stories, answer the questions, and provide critiques. This step aims to check (1) whether the quality of the generated questions is good, (2) whether the answers along with the explanation suggested by LLMs are correct and (3) whether they have suggestions to improve these questions or explanations. After completion, we simply ask them to provide suggestions and use these suggestions to prompt LLMs to improve the content. To implement this, we use a simple prompt which asks the model to generate new questions (see Appendix A.5.2).

4 Evaluation

A two-fold evaluation is carried out as follows: (a) an evaluation to determine the quality of the generated stories relating to doctrines, and (b) an evaluation of the generated questions and answers, as well as their efficacy in assessing comprehension.

4.1 Story Evaluation

4.1.1 Human Evaluation

For each legal concept, we generate one story based on its definition with each LLM. We recruit human

---

[4] `https://openai.com/index/chatgpt`

[5] `https://openai.com/index/gpt-4-research`

Model	RoD	RoS	Relevant	Redundant	Cohesive	Complete	Factual	Likeable	Believable
101 Concepts (Mean$_{STD}$)									
GPT-4	$3.95_{1.04}$	$4.66_{0.60}$	$4.56_{0.71}$	$4.00_{1.26}$	$4.63_{0.62}$	$4.57_{0.67}$	$4.56_{0.69}$	$4.36_{0.81}$	$4.54_{0.74}$
Sampled 20 Concepts (Mean$_{STD}$)									
GPT-4	$\mathbf{3.98}_{1.07}$	$\mathbf{4.70}_{0.46}$	$\mathbf{4.52}_{0.68}$	$3.78_{1.29}$	$\mathbf{4.57}_{0.56}$	$\mathbf{4.58}_{0.53}$	$\mathbf{4.52}_{0.62}$	$\mathbf{4.42}_{0.79}$	$\mathbf{4.48}_{0.70}$
GPT-3.5	$3.30_{1.01}$	$4.35_{0.68}$	$4.20_{0.78}$	$3.72_{0.80}$	$4.30_{0.74}$	$4.03_{0.78}$	$4.12_{0.69}$	$4.10_{0.95}$	$4.13_{0.65}$
LLaMA 2	$3.72_{1.15}$	$4.35_{0.86}$	$4.40_{0.85}$	$3.92_{1.33}$	$4.38_{0.83}$	$4.15_{1.12}$	$4.10_{1.17}$	$4.20_{1.04}$	$4.35_{0.94}$

Table 1: Human evaluation results of LLM-generated legal stories. The upper section contains scores for GPT-4 on the complete 101 legal concepts. The lower section contains scores for GPT-4, GPT-3.5, and LLaMA 2 on a subset of 20 legal concepts. RoD and RoS indicates the readability of the definition and the story respectively.

subjects with law education backgrounds on Prolific to evaluate the legal concepts and their corresponding stories generated by LLMs. Due to budget constraints, we randomly sampled 20 out of 101 concepts to compare among LLaMA 2, GPT-3.5, and GPT-4. Subsequently, we compare their performance and evaluate the full set of 101 concepts on the best model. For human evaluation, we recruit three raters to judge the **Readability of Definition (RoD)** and the following metrics for the generated stories: (1) **Readability of Story (RoS)**, (2) **Relevance**, (3) **Redundancy**, (4) **Cohesiveness**, (5) **Completeness**, (6) **Factuality**, (7) **Likeability**, (8) **Believability**. We use a 5-item Likert scale where 1 means very bad and 5 means very good. Details about the metrics and human evaluation are discussed in Appendix C.

Results In Table 1, we have several interesting observations. First, GPT-4 outperforms GPT-3.5 and LLaMA 2 in almost all the metrics except for redundancy. LLaMA 2 performs slightly better than GPT-3.5 in most metrics. By examining stories generated by LLaMA 2, we find that 8 (out of 20) generated stories are not in a story style but simplified definitions in plain language. Therefore, "stories" generated by LLaMA 2 seem shorter and more concise with high redundancy scores. In contrast, all the "stories" generated by GPT-3.5 and GPT-4 are indeed stories. Second, we observe consistently higher readability scores in stories (RoS) than the definitions (RoD), indicating that people find these stories easier to read than legal language. Additionally, GPT-4 stories receive high scores (≥ 4.5) in both readability and cohesiveness, showing they are easy to read in story format. Third, GPT-4 stories also achieve high scores (≥ 4.5) in relevance, completeness, and factuality, meaning that these stories are relevant to the definitions, and have good coverage and faithfulness reflection in the definition. Finally, human annotators find these stories decently likable ($\mu = 4.42$) and believable ($\mu = 4.48$). In practice, the stories can be further refined through expert feedback if their quality is not good enough. However, given the high ratings these generated stories have received, we have decided to provide them to participants along with their respective definitions, without the need for additional expert reviews.

4.1.2 Complexity Evaluation

Legalese usually contains long, wordy, complicated sentence structures, making it difficult for the public to understand. The readability metric in the previous section is one way to assess this. We also use multiple automatic measures of language complexity to compare concept definitions from Wikipedia and stories generated by different LLMs. These measures are not meant to be exhaustive but to provide more nuanced insights into language complexity, which is important for reader comprehension. We report the following common complexity metrics for comparison: (1) **Legal Vocabulary List (LVL) occurrences**, (2) **Top 1000 most common words (Top1K)**, (3) **Function words**, (4) **Sentence length**, (5) **Language model perplexity**, (6) **Flesch-Kincaid grade level**. More metric details can be found in Appendix B.

Results In Table 2, we compare the linguistic complexity between Wikipedia definitions and LLM-generated stories. We observe that stories from GPT-4 contain the lowest LVL proportion, GPT perplexity, sentence length, and Flesch-Kincaid score. Both GPT-4 and GPT-3.5 tend to use more function words and the top 1000 words from Thing Explainer in the stories. Across different measures, the definitions have the most linguistic complexity. Stories from GPT-3.5 and GPT-4 use a language with similar complexity but simpler than those from LLaMA 2. These observations are consistent with the RoD and RoS scores from the

Metrics	Wiki	LLaMA 2	GPT-3.5	GPT-4
LVL	0.27	0.26	0.22	**0.21**
Top1K	0.51	0.61	**0.64**	0.63
Func. Words	0.37	0.40	**0.42**	0.41
GPT PPL.	65.82	30.51	25.20	**24.96**
Sent. Length	30.11	23.41	20.26	**19.61**
FK Scores	14.79	11.35	8.61	**8.23**

Table 2: Results in complexity metrics in Wikipedia definitions and LLM-generated stories on the same subset of 20 concepts. Wiki represents the definition from Wikipedia. LLaMA 2, GPT-3.5 and GPT-4 represent stories generated by these models. The numbers in bold suggest the highest readability in that particular metric.

Model	ConceptQ	PredictionQ	LimitationQ
	101 Concepts (Mean$_{STD}$)		
GPT-4	$4.46_{0.77}$	$4.35_{0.78}$	$4.14_{0.98}$
	Sampled 20 Concepts (Mean$_{STD}$)		
GPT-4	$\mathbf{4.47}_{0.70}$	$\mathbf{4.35}_{0.71}$	$\mathbf{4.27}_{0.84}$
GPT-3.5	$4.12_{0.64}$	$3.95_{0.85}$	$3.48_{0.91}$
LLaMA 2	$4.23_{0.96}$	$4.10_{1.22}$	$4.12_{1.12}$

Table 3: Human evaluation on LLM-generated educational questions. Three questions are generated per concept, including a concept question, a prediction question, and a limitation question. The upper section contains scores for GPT-4 on the complete 101 legal concepts. The lower section contains scores for GPT-4, GPT-3.5, and LLaMA 2 on a subset of 20 legal concepts.

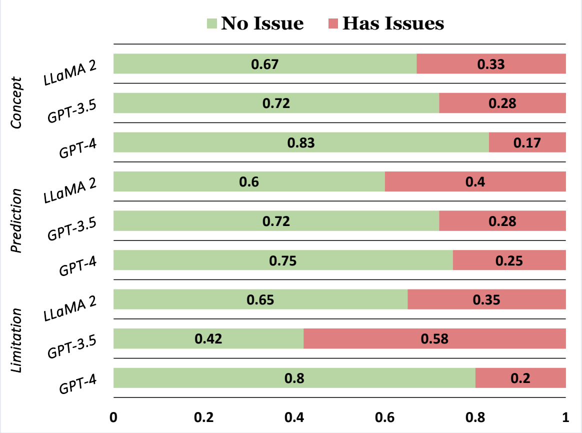


Figure 2: Distribution of questions with or without issues generated by LLaMA 2, GPT-3.5, and GPT-4.

human evaluation in Table 1.

4.2 Question Evaluation

It is challenging to evaluate generated questions since there are no gold standard questions. Therefore, we use human evaluation to assess the quality of generated questions and rely on the critique of two legal experts for improvement. We first begin with three authors of this work examining a subset of 10 concepts (30 questions), documenting emerging noticeable errors, and, following rounds of discussion, summarizing these errors as a set of six common error categories. These error categories and an “other” option for non-categorized errors are used to facilitate the evaluation process with Prolific workers and two legal experts.

4.2.1 Human Evaluation

Specifically, we recruit three Prolific human evaluators with law knowledge to judge whether the question contains the following shortcomings: (1) the question is too easy and simple, (2) the answer cannot be derived from the definition or story above, (3) the question is confusing, (4) there are more than one right answer in the 4 options, (5) there is no right answer among the 4 options, (6) the reasoning given in the suggested answer is wrong or flawed, (7) other issues not covered above. The annotators can select more than one option if multiple issues are identified. If there is no error, they can choose (8) There is no issue. The annotators also rate each question from 1 (bad) to 5 (good).

Human Ratings In Table 3, we find GPT-4 generation outperforms the other LLMs across all three types of questions. LLaMA 2 performs slightly better than GPT-3.5. In addition, the results show that concept questions and prediction questions have much higher scores than the limitation questions. We hypothesize that limitation questions are challenging for LLMs to generate because not all concepts are apparent to discuss their limitations or exceptions. The disparities in the question quality and error rate necessitate human-in-the-loop methods to improve question generation and quality control.

Error Analysis In Figure 2, with the human evaluation of LLMs-generated questions for 20 sampled concepts, we compute the percentage of questions with issues versus those without any. GPT-4 outperforms other LLMs in generation questions with no issues at 83%, 75%, and 80% for concept, prediction, and limitation questions, respectively.

Furthermore, we break down the distribution of errors found by the human annotators in the multiple-choice questions in Figure 3. We account for all the issue labels as we have allowed the annotators to select more than one issue. We observe that the six labels provided to the annotators cover most of the errors in the questions (the op-

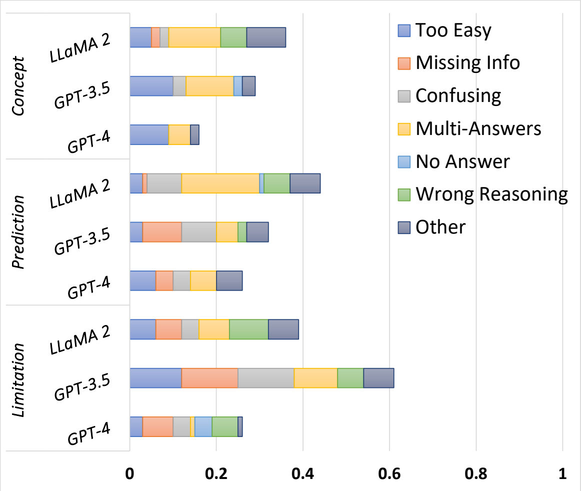


Figure 3: Distribution of different issues among the questions generated by LLaMA 2, GPT-3.5, and GPT-4.

tion "Other" is chosen when the labels do not cover certain errors). As shown in Figure 3, each LLM makes a combination of different errors; for instance, GPT-4 created fewer confusing questions and less wrong/flawed reasoning for the correct answer than other LLMs.

4.2.2 Expert Critiques

Finally, we recruit two legal experts to evaluate the efficacy of the assessment with critiques by completing the assessment on the 20 sampled concepts. The purpose of this step is to make sure the questions are answerable, given either definitions alone or definitions and stories. We randomly split 20 sampled concepts into two equal batches to avoid exposure bias. Each person completes one batch with the definition only and the other batch with the story and definition. However, the batches for two people are in reversed conditions. With this design, we can compare their agreeability without the exposure bias from the stories. At last, they are asked to categorize the difficulty level of each legal concept into easy, medium, and hard.

Inter-rater Agreeability We observe that answers from the two legal experts have overall high agreement scores in Cohen's Kappa with GPT-4 answers (ranging from 0.77 to 1.00) and each other (ranging from 0.68 to 0.86) across different question types. We include agreement scores in Table 6. The concept question has the highest agreement scores. This confirms our finding from the previous human ratings from Prolific annotators in Table 3, which shows that the concept question has the highest average human ratings. Expert 1 gave improvement advice to 6 questions and Expert 2 to 2 questions. After we prompt GPT-4 with their advice to re-generate the corresponding questions, we show them to the experts, and they are 100% approved by the experts after the first round of regeneration. This shows the effectiveness of instructing GPT-4 with expert advice to revise the questions (see one example in Appendix A.5).

5 RCT Experiment

5.1 Experiment Design

Using the stories and questions from the pipeline, we evaluate the efficacy of LLM-generated stories for helping human comprehension on a sample of 10 concepts. We design a randomized controlled trial (RCT) with two study conditions: (1) a control group that is given the legal concept definition and (2) a treatment group that is given both the definition and a story that illustrates a hypothetical situation in which the given concept applies. Both condition groups read the content and are asked to complete three multiple-choice questions to evaluate comprehension. We keep both definition and story in the treatment group (instead of just the story) because we hope to have the definition as a good reference if the story is confusing to the reader by any chance. Due to the importance of legal education for both native and non-native English speakers (such as immigrants), we recruit participants from both native-speaker and non-native speaker populations. The participants were recruited through Prolific; details of the criteria are in Appendix D. We recruit 15 to 20 people in each condition to complete batches of 5 concepts. In total, we have 65 respondents from native speakers (33 in control and 32 in treatment) and 71 respondents from non-native speakers (37 in control and 34 in treatment conditions) for a total of 136 participants. With these participants, we collect the following information through 5-point Likert scales: (1) the relevance of the legal concept, (2) their interest in engaging and learning more about the concept, (3) their familiarity with the concept, (4) their familiarity with the scenario setting in the story (treatment group only). Prior studies (Glonek and King, 2014; Kromka and Goodboy, 2019) show that stories facilitate learner recall about content. The participant comprehension is evaluated by their accuracy in RC questions and learning retention in the three-day post-study assessment, in which we present the participants with the same RC questions

Condition	ConceptQ	PredictionQ	LimitationQ
Native Speakers (Accuracy)			
Definition	**93.33**	**78.79**	77.58
Def. + Story	90.62	74.38	**84.38**
Non-native Speakers (Accuracy)			
Definition	89.19	71.89	68.65
Def. + Story	**91.18**	**81.76**	**84.71**

Table 4: Comprehension accuracy of native and non-native speakers of English. We report their performance in concept, prediction, and limitation questions.

but with the answers reordered, without providing them with the definition or story again.

5.2 Results & Discussions

Reading Comprehension We compare the answer accuracy in Table 4 and find that **legal stories improve the comprehension accuracy for the non-native English speaker group for all question types while only improving the limitation question accuracy for the native speaker group.** Specifically, the concept and prediction accuracy for native speakers decrease in the treatment group. However, the Chi-Squared test for statistical significance fails to show significant differences between two conditions in native English speakers (concept question: $X^2(1, N = 325) = 0.48, p = .487$; prediction question: $X^2(1, N = 325) = 0.65, p = .419$; limitation question: $X^2(1, N = 325) = 2.01, p = .156$). For the non-native speaker group, statistical significance is achieved for both prediction and limitation questions (concept question: $X^2(1, N = 355) = 0.20, p = .653$; prediction question: $X^2(1, N = 355) = 4.29,$ **p = .038**; limitation question: $X^2(1, N = 355) = 11.77,$ **p < .001**). We have summarized individual-level mean and standard deviation in accuracy in Appendix D.3.

The difference in accuracy for the native speakers between the treatment and control conditions is not statistically significant; thus, it could be due to the sampling error. We also notice that the native speakers' accuracy for the concept and prediction questions in the control group is the highest or close to the highest accuracy among other groups, which might imply a ceiling effect of the learning interventions. The stories may have enhanced the participant's accuracy in questions with originally lower accuracy in the control setting. For instance, both native and non-native speakers had more room for improvement in the control condition, which the stories could have potentially addressed.

Condition	ConceptQ	PredictionQ	LimitationQ
Native Speakers (Retention Rate)			
Definition	**92.55**	**88.89**	**91.03**
Def. + Story	91.58	86.84	91.01
Non-native Speakers (Retention Rate)			
Definition	86.32	82.80	91.01
Def. + Story	**98.56**	**89.60**	**92.31**

Table 5: Retention rate in delayed post-study assessment of native and non-native speakers of English. The highest possible percentage is 100%, representing a perfect retention of knowledge (in theory).

Relevance The relevance score, which rates the degree to which the content displayed to participants is relevant to their own lives and situations, shows that **participants who read the stories with the concept definition consistently felt it more relatable than the control condition for both native speaker and non-native speaker groups**. Among non-native English speakers, we observe that participants who read both definitions and stories find these legal concepts more relevant ($3.19_{1.17}$ vs. $2.47_{1.17}$) to their lives than those who read only definitions. Similarly, we observe higher relevance scores ($3.21_{1.32}$ vs. $2.63_{1.30}$) for the treatment group compared with the control group among native speakers. Mann-Whitney U tests find statistical significance in relevance scores for both native English speakers ($U = 16421.0,$ **p < .001**) and non-native English speakers ($U = 21077.5,$ **p < .001**).

Interest in Law The interest score, which rates the degree to which participants feel interested in engaging with the content displayed, shows that **the treatment condition has statistically higher interests than the control condition for the non-native speaker group but not for the native-speaker group.** Specifically, among non-native English speakers, we observe that participants who read both definitions and stories are more interested in delving into laws and legal knowledge ($4.03_{1.20}$ vs. $3.84_{0.92}$) than those who read only definitions. Among native English speakers, the treatment group shows slightly higher interest than the control group ($3.78_{0.99}$ vs. $3.67_{1.12}$). Mann-Whitney U tests reveal that statistical differences are found for the participant interest between the treatment group and control group for non-native speakers ($U = 18662.5,$ **p = .001**); however, not for the native English speaker ($U = 13800.0, p = .459$).

Knowledge Retention To investigate participants' comprehension retention, we send the follow-up assessment three days after the original study. In the end, 71% of the respondents filled out the delayed post-study test. Given participants' original and delayed assessment, the retention rate is calculated as the percentage of continued correct answers for each question. The result is summarized in Table 5. We observe that after three days, participants show different degrees of forgetting. However, **non-native speakers who read stories and definitions have a higher retention rate after three days while no such effect is found for the native speaker group**. A Chi-square test confirms significant differences for non-native speakers between the retention in the treatment and control group for the concept question $X^2(1, N = 256) = 12.74$, **p < .001**; however, no such differences are found for the other questions.

6 Conclusion

In this work, we explore a novel application of LLMs in legal concept learning through storytelling. We use an expert-in-the-loop pipeline to create the LegalStories dataset that contains educational stories and comprehension questions. Moreover, we compare the performance of several benchmark LLMs including GPT-4, GPT-3.5, and LLaMA 2 with automatic and human evaluations. While GPT-4 outperforms the others at generating legal stories and creating questions, it still exhibits certain reasoning errors, highlighting the need for human supervision when using LLMs for educational content development. Finally, through RCTs, we show that, among non-native speakers, learning with stories not only improves comprehension of legal concepts and interest but also leads to a higher retention rate in the follow-up assessment compared to learning with definitions alone. Our study suggests considerable potential for using LLMs in advancing legal education and beyond.

Limitations

Sample Size Given the limited financial budget available to conduct our research, we chose to conduct our study in a smaller data setting to obtain high-quality human feedback. Similarly, our participant pool was limited by extensive and costly surveying approaches to 65 native respondents and 71 non-native respondents, which may have negatively impacted the statistical power of our group comparison results. We want to emphasize that even at this scale the cost is nontrivial. For example, it took around 800 dollars to evaluate 101 concepts generated by GPT-4 and 320 dollars to evaluate 20 concepts generated by GPT-3.5 and LLaMA-2 in total. For human experiments, it took around 2400 dollars to run RCTs including the follow-up study with 10 sampled concepts. In an attempt to alleviate this, we have carefully chosen our current sampling strategy to ensure that the resulting samples are representative. For instance, we compare the human evaluation between 101 and 20 concepts in Table 1 and Table 3 and show that they are similar for GPT-4. Although similar studies exist with similar group sizes per condition (Lu et al., 2021; Steuer et al., 2022; August et al., 2022), and although we find our results compelling, it is possible that due to sample size limitations, we were not able to capture small effects; however, our study provides strong statistical power for observing large effect sizes, revealing several significant effects, even with limited statistical power.

Data Quality & Practicality To mitigate biases and hallucinations of LLM-generated content, we loop in crowdworkers and experts to audit and improve the generated content. Legal experts estimate that it takes at least 30 minutes to create one story and three questions from scratch for each concept. By contrast, it takes each expert around 6.5 minutes to evaluate and write feedback to each legal concept, story, and questions. Although it is a labor-saver compared to having a completely human-written dataset, it still requires human experts, thus, might suffer from scalability. In practice, we believe that having human experts (such as teachers and lawyers) in the loop is a reliable and necessary manner to create useful and less biased educational content by mitigating model errors while minimizing human effort.

RCT Design In our case, due to limited resources, we were unable to run multiple human evaluations to optimize our prompts. We also chose the most intuitive control group (definition) and treatment group (definition+story) to answer the main research question whether LLM-generated stories improve comprehension in legal concepts. The control group represents current practice for legal communication (e.g., a glossary of terms or a legal dictionary). We believe that it would be interesting for future studies to extend the work by comparing this with various prompting strategies to generate LLM-based concept explanations or elaborations such as “Explain Like I'm 5 (ELI5)”.

Ethical Considerations

Legal Experts The term "legal expert" does not constitute any suggestion or indication that the participants enlisted to provide evaluation and critique of the LLM-generated materials are admitted to practice law in any jurisdictions or are holding themselves out as attorneys qualified to provide legal advice. Their participation in this research does not involve legal representation, legal advice, or drafting of legal documents for any entity or person to any extent. The legal experts enlisted either have graduated with law degrees or have made substantial progress toward earning law degrees. As such, the participants can provide valuable feedback for the research because their legal training makes them better-suited to assess LLM-generated content than people without such backgrounds.

Code of Conduct This research follows the ACL Code of Ethics, has IRB Exempt status, and respects participants' anonymity. We used the Prolific platform for human annotation and experiments with their consent, compensated online annotators $15 per hour according to Massachusetts state law, and ensured LLM-generated content is safe and non-offensive. Exact experiment details are included in the appendix for reproducibility.

LLM-related Risks We are aware of the potential for bias that LLMs present, both in educational and in generalized contexts. It is dangerous and inappropriate to provide LLM contents to students without human supervision, because these contents might contain misleading, biased, harmful, or wrong information and education is a high-stake domain. With respect to risks such as this, our work takes a human-centered approach to loop in qualified crowdworkers and experts to audit the LLM outputs. In practice, we believe that having human experts such as teachers and lawyers in the loop is a reliable and effective manner to create useful and less biased high-quality educational content.

Information Loss Law is, by nature, a sensitive domain, and computational tools must be designed responsibly. In the context of legal education, it is critical to design comprehension tools in ways that do not over-simplify or over-generalize the nuances of legal jargon. To address these issues, we chose to collaborate with legal experts to audit the content and draw on domain-specific data, we hope to provide an approach that balances access to justice needs with responsible AI approaches.

Acknoledgement

We would like to thank Maria Antoniak, Bailin Wang, Belén Saldías, and Mingye Gao for their helpful discussions. We also thank the reviewers from ACL Rolling Review (ARR) for their constructive feedback.

References

H Porter Abbott. 2020. *The Cambridge introduction to narrative*. Cambridge University Press.

Craig Eilert Abrahamson. 1998. Storytelling as a pedagogical tool in higher education. *Education*, 118(3):440–452.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nancy E Adams. 2015. Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, 103(3):152.

Said Al Faraby, Adiwijaya Adiwijaya, and Ade Romadhony. 2023. Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*, pages 1–38.

Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: Challenges and attempts. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Robert W Benson. 1984. The end of legalese: The game is over. *New York University Review of Law & Social Change*, 13:519.

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education*, pages 327–339. Springer.

Rick Busselle and Helena Bilandzic. 2008. Fictionality and perceived realism in experiencing stories: A model of narrative comprehension and engagement. *Communication theory*, 18(2):255–280.

Rick Busselle and Helena Bilandzic. 2009. Measuring narrative engagement. *Media psychology*, 12(4):321–347.

Nicola Capuano, Carmen De Maio, Angelo Gaeta, Giuseppina Rita Mangione, Saverio Salerno, and Eleonora Fratesi. 2014. A storytelling learning model for legal education. *International Association for Development of the Information Society*, 2014:29–36.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022. Unsupervised simplification of legal texts. *arXiv preprint arXiv:2209.00557*.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Miguel Collantes, Maureen Hipe, Juan Lorenzo Sorilla, Laurenz Tolentino, and Briane Samson. 2015. Simpatico: A text simplification system for senate and house bills. In *Proceedings of the 11th National Natural Language Processing Research Symposium*, pages 26–32.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Michael Curtotti and Eric McCreath. 2013. A right to access implies a right to know: An open online platform for research on the readability of law. *J. Open Access L.*, 1:1.

Michael F Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the national academy of sciences*, 111(supplement_4):13614–13620.

Ruth Davidhizar and Giny Lonser. 2003. Storytelling as a teaching technique. *Nurse educator*, 28(5):217–221.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Atefeh Farzindar and Guy Lapalme. 2004. Legal text summarization by exploration of the thematic structure and argumentative roles. In *Text Summarization Branches Out*, pages 27–34, Barcelona, Spain. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment.

Ebrahim Gabajiwala, Priyav Mehta, Ritik Singh, and Reeta Koshy. 2022. Quiz maker: Automatic quiz generation from text using nlp. In *Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021*, pages 523–533. Springer.

Kathleen Marie Gallagher. 2011. In search of a theoretical basis for storytelling in education research: Story as method. *International Journal of Research & Method in Education*, 34(1):49–61.

Dee Gardner and Mark Davies. 2014. A new academic vocabulary list. *Applied linguistics*, 35(3):305–327.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin, Ireland. Association for Computational Linguistics.

Katie L Glonek and Paul E King. 2014. Listening to narratives: An experimental examination of storytelling in the classroom. *International journal of listening*, 28(1):32–46.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junmo Kang, Haritz Puerto San Roman, and Sung-Hyon Myaeng. 2019. Let me know what to ask: Interrogative-word-aware question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 163–171, Hong Kong, China. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Soha Khazaeli, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster, and Dhruv Sakalley. 2021. A free format legal question answering system. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 107–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Stephen M Kromka and Alan K Goodboy. 2019. Classroom storytelling: Using instructor narratives to increase student recall, affect, and attention. *Communication Education*, 68(1):20–43.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 102–111, Seattle, United States. Association for Computational Linguistics.

Nguyen-Thinh Le, Tomoko Kojiri, and Niels Pinkwart. 2014. Automatic question generation for educational applications–the state of art. In *Advanced Computational Methods for Knowledge Engineering: Proceedings of the 2nd International Conference on Computer Science, Applied Mathematics and Applications (ICCSAMA 2014)*, pages 325–338. Springer.

Bernardo Leite and Henrique Lopes Cardoso. 2023. Towards enriched controllability for educational question generation. In *International Conference on Artificial Intelligence in Education*, pages 786–791. Springer.

Gondy Leroy, Stephen Helmreich, and James R Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, 79(6):438–449.

Gondy Leroy, Stephen Helmreich, James R Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *AMIA Annual Symposium Proceedings*, volume 2008, page 394. American Medical Informatics Association.

Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. 2023. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. *arXiv preprint arXiv:2310.08395*.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.

Owen HT Lu, Anna YQ Huang, Danny CL Tsai, and Stephen JH Yang. 2021. Expert-authored and machine-generated short-answer questions for assessing students learning performance. *Educational Technology & Society*, 24(3):159–173.

Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. Readingquizmaker: A human-nlp collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Stephen MacNeil, Joanne Kim, Juho Leinonen, Paul Denny, Seth Bernstein, Brett A Becker, Michel Wermelinger, Arto Hellas, Andrew Tran, Sami Sarsa, et al. 2023. The implications of large language models for cs teachers and students. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education*, volume 2.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex Pentland. 2023. The law and NLP: Bridging disciplinary disconnects. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3445–3454, Singapore. Association for Computational Linguistics.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Susana Martinez-Conde and Stephen L Macknik. 2017. Finding the plot in science storytelling in hopes of enhancing science communication. *Proceedings of the National Academy of Sciences*, 114(31):8127–8129.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Carrie J Menkel-Meadow. 1999. When winning isn't everything: The lawyer as problem solver. *Hofstra L. Rev.*, 28:905.

Jarmila Mildorf. 2016. Reconsidering second-person narration and involvement. *Language and Literature*, 25(2):145–158.

Randall Munroe. 2015. *Thing explainer: complicated stuff in simple words*. Hachette UK.

Sonia Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jonathan Borchardt, Daniel Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A multi-document approach to generating diverse descriptions of scientific concepts. In

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 200–213, Abu Dhabi, UAE. Association for Computational Linguistics.

Sonia K Murthy, Daniel King, Tom Hope, Daniel Weld, and Doug Downey. 2021. Towards personalized descriptions of scientific concepts. In *The Fifth Widening Natural Language Processing Workshop at EMNLP*.

Hamze Muse, Sahan Bulathwela, and Emine Yilmaz. 2023. Pre-training with scientific text improves educational question generation (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16288–16289.

Shashi Narayan, Gonçalo Simoes, Ji Ma, Hannah Craighead, and Ryan Mcdonald. 2020. Qurious: Question generation pretraining for text generation. *arXiv preprint arXiv:2004.11026*.

Christos H Papadimitriou. 2003. Mythematics: in praise of storytelling in the teaching of computer science and math. *ACM SIGCSE Bulletin*, 35(4):7–9.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

William J Rapaport, Erwin M Segal, Stuart C Shapiro, David A Zubin, Gail A Bruder, Judith F Duchan, Michael J Almeida, Joyce H Daniels, Mary Galbraith, Janyce M Wiebe, et al. 1989. *Deictic centers and the cognitive structure of narrative comprehension*. State University of New York (Buffalo). Department of Computer Science.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223, Seattle, Washington. Association for Computational Linguistics.

Elissa Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, Sean Kross, Michelle Mazurek, and Hal Daumé III. 2019. Comparing and developing tools to measure the readability of domain-specific texts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4831–4842, Hong Kong, China. Association for Computational Linguistics.

Magdalena Rembowska-Płuciennik. 2022. Enactive, interactive, social—new contexts for reading second-person narration. *Narrative*, 30(1):67–84.

Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Megha Srivastava and Noah Goodman. 2021. Question generation for adaptive education. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 692–701, Online. Association for Computational Linguistics.

Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. 2021. I do not understand what i cannot define: Automatic question generation with pedagogically-driven content selection. *arXiv preprint arXiv:2110.04123*.

Tim Steuer, Anna Filighera, Thomas Tregel, and André Miede. 2022. Educational automatic question generation improves reading comprehension in non-native speakers: A learner-centric case study. *Frontiers in Artificial Intelligence*, 5:900304.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? *arXiv preprint arXiv:1808.09384*.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939, Brussels, Belgium. Association for Computational Linguistics.

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

Andrew Tran, Kenneth Angelikas, Egi Rama, Chiku Okechukwu, David H Smith IV, and Stephen MacNeil. 2023. Generating multiple choice questions for computing courses using large language models.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina Kann. 2023.

On the automatic generation and simplification of children's stories. *arXiv preprint arXiv:2310.18502*.

Zichao Wang, Andrew S Lan, Weili Nie, Andrew E Waters, Phillip J Grimaldi, and Richard G Baraniuk. 2018. Qg-net: a data-driven question generation model for educational content. In *Proceedings of the fifth annual ACM conference on learning at scale*, pages 1–10.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. 2022. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, pages 153–166. Springer.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Dell Zhang, Alina Petrova, Dietrich Trautmann, and Frank Schilder. 2023. Unleashing the power of large language models for legal applications. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5257–5258.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6032–6037, Hong Kong, China. Association for Computational Linguistics.

Brian D Zimmerman, Gaurav Sahu, and Olga Vechtomova. 2022. Future sight: Dynamic story generation with large pretrained language models. *arXiv preprint arXiv:2212.09947*.

A Dataset Generation

A.1 Model Details for Story and Question Generation

LLaMA 2 (`LLaMA 2-70b-Chat`), GPT-3.5 (`GPT-3.5-turbo-0613`), and GPT-4 (`GPT-4-0613`) are used for this experiment. For LLaMA 2, we set top p to 1.0 and temperature to 0.01 and use the default settings for the other parameters. For GPT-3.5 and GPT-4, we set the temperature to 0.0 and use the default settings for the other parameters. We use the Replicate LLaMA 2 API [6] in our experiments. The exact prompt for the story generation is shared in the main paper. We also include the exact prompts for question generation in the subsection below.

A.2 Prompt for Story Generation

Here is the exact prompt used to generate stories, where {CONCEPT} stands for the concept name and {DEFINITION} for the Wikipedia definition: Tell a story within 500 words to simplify the concept explanation below for "{CONCEPT}". Start your answer with "Concept Simplified:". Concept: "{DEFINITION}".

A.3 Prompts for Question Generation

We present the prompts used to generate each multiple-choice question and answers in the question generation phase of the pipeline. {CONCEPT} stands for the concept name, {DEFINITION} for the Wikipedia definition, and {STORY} for the generated story from the corresponding LLM.

Concept Question Prompt Read the concept explanation and story below. Please generate a multiple choice question with four candidates (only one correct answer) to test if a reader understands the concept: **which of the following answers is an accurate description of the concept "{CONCEPT}"**? Start your response with "Question:". Candidates are ordered by (A), (B), (C), (D). In the end, give the right answer with its explanation starting with "The right answer is". Concept: "{DEFINITION}". Story: "{STORY}"

Prediction Question Prompt Read the concept explanation and story below. Please generate a multiple choice question with four candidates (only one correct answer) to test if a reader understands the concept: **come up with a hypothetical scenario where the concept "{CONCEPT}" is used and ask the reader to guess the ending of the story**. Please ensure the hypothetical scenario is more challenging than the story below. Start your response with "Question:". Candidates are ordered by (A), (B), (C), (D). In the end, give the right answer with its explanation starting with "The right answer is". Concept: "{DEFINITION}". Story: "{STORY}"

Limitation Question Prompt Read the concept explanation and story below. Please generate a multiple choice question with four candidates (only one correct answer) to test if a reader understands the concept: **what is a potential limitation or exception of the rule "{CONCEPT}"**? Start your response with "Question:". Candidates are ordered by (A), (B), (C), (D). In the end, give the right answer with its explanation starting with "The right answer is". Concept: "{DEFINITION}". Story: "{STORY}"

A.4 Expert-LLM Agreement

Table 6 shows the Cohen's Kappa agreement scores between two legal experts and GPT-4 in choosing the right answers in generated questions.

A.5 LEGALSTORIES Example with Expert Feedback

A.5.1 Story Example

We show one example concept "Substantial Certainty Doctrine" from the LEGALSTORIES dataset:

Doctrine: *Substantial Certainty Doctrine*

---

[6] https://replicate.com/meta/llama-2-70b-chat

Expert 1 & GPT-4	Expert 2 & GPT-4	Expert 1 & Expert 2
Concept Question (Cohen's Kappa)		
1.00	0.86	0.86
Prediction Question (Cohen's Kappa)		
0.92	0.77	0.68
Limitation Question (Cohen's Kappa)		
0.85	0.78	0.86

Table 6: We compare expert and GPT-4 answers on LLM-generated educational questions of 20 legal concepts. We show Cohen's Kappa scores among two legal experts and GPT-4.

WiKi Definition: In law, the substantial certainty doctrine is the assumption of intent even if the actor did not intend the result, but knew with substantial certainty the effect would occur as a result of his action. The doctrine can be used by courts as a test to determine whether or not a defendant committed a tort. For example, in Garratt v. Dailey (1955), the Washington Supreme Court remanded a case back to the lower courts to determine whether or not the five year-old defendant "knew with substantial certainty that the plaintiff would attempt to sit down where the chair which he moved had been."

Story: Imagine you're at a party and you see your friend about to sit on a chair. You quickly pull the chair away just as they're about to sit, resulting in them falling on the floor. You didn't intend for them to fall and get hurt, but you knew there was a high chance it would happen because you pulled the chair away. This is the idea behind the "substantial certainty doctrine" in law.

This legal concept suggests that even if you didn't mean for a certain outcome to happen, if you knew there was a high chance it would, you can still be held responsible. Courts use this doctrine to decide if someone committed a wrong act, also known as a tort.

A famous example of this is the case of Garratt v. Dailey in 1955. In this case, a five-year-old boy moved a chair just as an elderly woman was about to sit. The woman fell and got injured. The court had to decide if the boy "knew with substantial certainty" that the woman would try to sit where the chair had been and get hurt. The case was sent back to the lower courts to make this decision.

So, in simple terms, the substantial certainty doctrine is like saying, "You knew what would happen if you did that, so you're responsible for the outcome."

Concept Question: Which of the following best describes the "Substantial Certainty Doctrine" in law?

(A) It is a doctrine that holds a person responsible for an outcome only if they intended for that outcome to occur.

(B) It is a doctrine that holds a person responsible for an outcome if they knew there was a high chance it would occur, even if they did not intend for it to happen.

(C) It is a doctrine that holds a person responsible for an outcome only if they did not know there was a high chance it would occur.

(D) It is a doctrine that holds a person responsible for an outcome only if they did not intend for that outcome to occur, and they did not know there was a high chance it would happen.

The right answer is **(B)**. The "Substantial Certainty Doctrine" in law holds a person responsible for an outcome if they knew there was a high chance it would occur, even if they did not intend for it to happen. This doctrine is used by courts to determine if someone committed a tort, or a wrongful act.

Prediction Question: In a hypothetical scenario, John, a construction worker, removes a safety barrier around a deep pit at a construction site to make his work easier. He sees his co-worker, Mike, walking towards the pit while talking on the phone, unaware of the danger. John doesn't warn Mike about the pit, thinking Mike would notice it himself. However, Mike falls into the pit and gets severely injured. If this case goes to court, under what condition might the substantial certainty doctrine be applied?

(A) The court would apply the substantial certainty doctrine if it can be proven that John knew with substantial certainty that Mike would fall into the pit.

(B) The court would apply the substantial certainty doctrine if it can be proven that John intentionally removed the safety barrier.

(C) The court would not apply the substantial certainty doctrine because John did not intend for Mike to fall into the pit.

(D) The court would not apply the substantial certainty doctrine because Mike should have been more careful.

The right answer is **(A)**. The substantial certainty doctrine could be applied if it can be proven that John knew with substantial certainty that Mike would fall into the pit. Even though John did not intend for Mike to fall, if he was aware of the high risk of such an outcome when he removed the safety barrier and did not warn Mike, this aligns with the substantial certainty doctrine, which assumes intent even if the actor did not intend the result, but knew with substantial certainty the effect would occur as a result of his action.

Limitation Question: What is a potential limitation or exception of the substantial certainty doctrine?

(A) The doctrine cannot be applied if the defendant is a minor.

(B) The doctrine cannot be applied if the defendant did not physically cause the harm.

(C) The doctrine cannot be applied if the defendant did not have a direct intention to cause harm.

(D) The doctrine cannot be applied if the defendant did not know with substantial certainty that their action would cause the effect.

The right answer is **(D)**. The doctrine cannot be applied if the defendant did not know with substantial certainty that their action would cause the effect. The substantial certainty doctrine assumes intent even if the actor did not intend the result, but knew with substantial certainty the effect would occur as a result of his action. If the defendant did not have this level of certainty, the doctrine may not apply.

A.5.2 Question before and after Expert Feedback

In this section, we walk through one example question to explain how expert feedback is used for re-generating multiple-choice questions with quality control. The prompt we used is as follows: "{EXPERTISE_SUGGESTION}. Improve the question given the feedback above." The example we use is the **Prediction Question** for *Substantial Certainty Doctrine*. The original prediction question generated by LLM is as follow:

Original Generation: In a hypothetical scenario, John, a construction worker, removes a safety barrier around a deep pit at a construction site to make his work easier. He sees his co-worker, Mike, walking towards the pit while talking on the phone, unaware of the danger. John doesn't warn Mike about the pit, thinking Mike would notice it himself. However, Mike falls into the pit and gets severely injured. If this case goes to court, how might the substantial certainty doctrine be applied?

(A) The court would not apply the substantial certainty doctrine because John did not intend for Mike to fall into the pit.

(B) The court would apply the substantial certainty doctrine because John knew with substantial certainty that Mike could fall into the pit.

(C) The court would not apply the substantial certainty doctrine because Mike should have been more careful.

(D) The court would apply the substantial certainty doctrine because John intentionally removed the safety barrier.

To this question, the answer generated by LLM is:

> The right answer is (B). The substantial certainty doctrine could be applied because John knew with substantial certainty that Mike could fall into the pit. Even though John did not intend for Mike to fall, he was aware of the high risk of such an outcome when he removed the safety barrier and did not warn Mike. This aligns with the substantial certainty doctrine, which assumes intent even if the actor did not intend the result, but knew with substantial certainty the effect would occur as a result of his action.

Expert Critiques The legal expert reviewed the multi-choice question and gave the following feedback:

1. Does your answer agree with the suggested answer?

   Response: No.

2. Is the suggested answer correct or wrong? Why?

   Response: The suggested answer seems wrong.

3. What suggestions do you have to improve the question?

   Response: I think the suggested answer is wrong/the question is a bit confusing to me. When John removes the barrier, he does not know with substantial certainty that Mike would fall in.

From this expert critique, we insert the following prompt to LLM for question re-generation:

> *I think the suggested answer is wrong/the question is a bit confusing to me. When John removes the barrier, he does not know with substantial certainty that Mike would fall in.*
>
> *Improve the question given the feedback above.*

Afterwards, the LLM takes into the critiques in the prompt, fixes the confusion about whether Jon knows the substantial certainty of his action to Mike, and produces new choices without changing the question itself:

Final Prediction Question: In a hypothetical scenario, John, a construction worker, removes a safety barrier around a deep pit at a construction site to make his work easier. He sees his co-worker, Mike, walking towards the pit while talking on the phone, unaware of the danger. John doesn't warn Mike about the pit, thinking Mike would notice it himself. However, Mike falls into the pit and gets severely injured. If this case goes to court, under what condition might the substantial certainty doctrine be applied?

(A) The court would apply the substantial certainty doctrine if it can be proven that John knew with substantial certainty that Mike would fall into the pit.

(B) The court would apply the substantial certainty doctrine if it can be proven that John intentionally removed the safety barrier.

(C) The court would not apply the substantial certainty doctrine because John did not intend for Mike to fall into the pit.

(D) The court would not apply the substantial certainty doctrine because Mike should have been more careful.

The process of expert critiques is iterative if the issue in the original generation is not resolved after one round or a new issue arises. With the renewed choices, the answer generated by LLM becomes:

> The right answer is **(A)**. The substantial certainty doctrine could be applied if it can be proven that John knew with substantial certainty that Mike would fall into the pit. Even though John did not intend for Mike to fall, if he was aware of the high risk of such an outcome when he removed the safety barrier and did not warn Mike, this aligns with the substantial certainty doctrine, which assumes intent even if the actor did not intend the result, but knew with substantial certainty the effect would occur as a result of his action.

B Automatic Complexity Metrics

Here are the details about the complexity metrics. To tokenize[7] the stories and definitions, we use:

Legal Vocabulary List (LVL) occurrences Inspired by Gardner and Davies (2014)'s Academic Vocabulary List (AVL), LVL incorporates a list of legalese from the Glossary of Legal Terms from the official US Court website[8] and the Open Legal Dictionary Project[9]. To assess the level of legal rigor in Wikipedia definitions, we quantify the proportion of LVL words employed in each definition.

Top 1000 Most Common Words out-of-vocabulary (Top1K) The popular book "Thing Explainer" utilizes a vocabulary constrained to the 1,000 most frequent English words based on Wiktionary's contemporary fiction frequency list (Munroe, 2015). To assess the simplicity and accessibility of generated definitions, we calculate the proportion of words that are from the top 1,000 words employed in the book.

Function words In health communication, the use of function words such as prepositions, auxiliary verbs, and question words is positively associated with both perceived and actual readability (Leroy et al., 2008, 2010). August et al. (2022) has also applied this to science communication.

Sentence length Sentence length is a widely used metric for assessing document-level complexity and is incorporated into numerous classic readability measures (Pitler and Nenkova, 2008; Feng et al., 2010). We only pick concepts whose definitions have 100-200 words and limit the story to 500 words. Therefore, We hypothesize that generated stories will be associated with less complex language due to elaborative simplification, a technique that involves explaining complex terms to facilitate comprehension (Srikanth and Li, 2021).

Language model perplexity Language model perplexity has demonstrated a positive correlation with perceived and actual reading difficulty (Pitler and Nenkova, 2008; Collins-Thompson, 2014). To assess the complexity of our generated stories, we utilize the GPT model to calculate language model perplexity, considering its training on common English rather than scientific text.

Flesch-Kincaid grade level This readability score (FK)[10] is derived from straightforward calculations based on sentence length, word length, and syllable counts (Kincaid et al., 1975). While studies have shown varying degrees of effectiveness for the FK score in predicting readability in scientific or medical documents (Leroy et al., 2008), it remains a widely used and standardized measure of text complexity (Redmiles et al., 2019).

C Human Evaluation

C.1 Evaluation Criteria

Details about the Prolific human evaluation, including the survey questions and their demographic details. We recruit Prolific workers who are native English speakers from United States or United Kingdom who have studied law with an approval rate between 99 and 100. Here we include the screenshots for Prolific data evaluation questions. Each batch of annotation is sent to three annotators, which contains 5 legal concepts with their stories and three generated questions. We first present the annotator with a consent form before they proceed (see Figure 4). Annotators first look at a concept definition (Figure 5) and judge

---

[7] We use the "en_core_web_sm" spaCy model for tokenization. See `https://spacy.io/`.

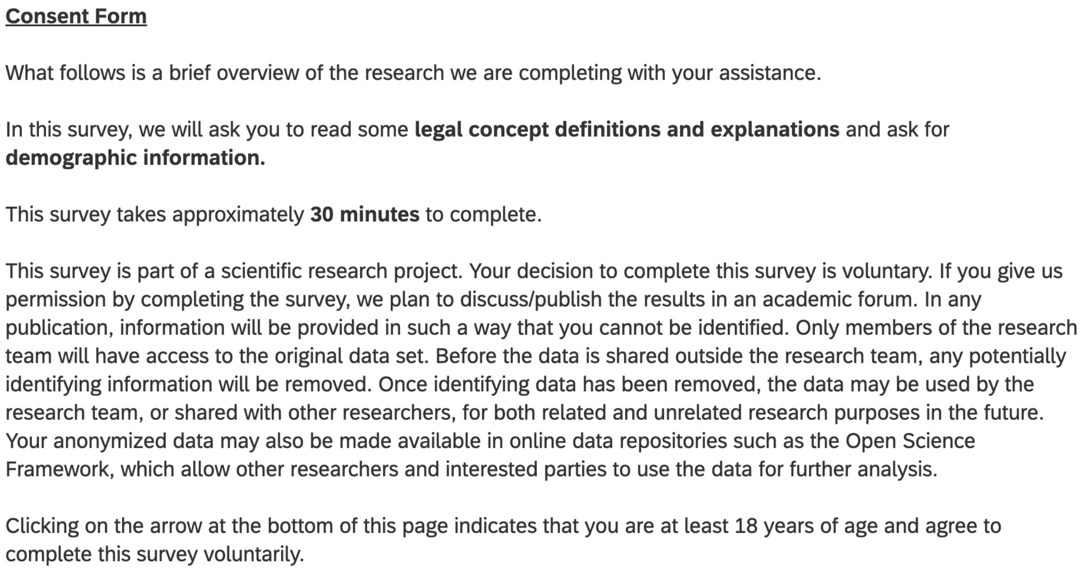
[8] `https://www.uscourts.gov/glossary`

[9] `https://openlegaldictionary.com/`

[10] We use the Readability library to compute FK scores. See `https://github.com/andreasvc/readability/`.

the **Readability of Definition (RoD)**: whether the concept definition is easy to read, well-structured, and flows naturally (Figure 6). Afterward, we ask them see the story (Figure 7) and judge the **Readability of Story (RoS)**: whether the story is easy to read, well-structured, and flows naturally (Figure 8). Later, they are asked to further evaluate the story along seven dimensions:

- **Relevance**: whether the story is highly relevant and directly addresses the given concept definition. (Figure 9)
- **Redundancy**: whether the story is concise and free from unneeded content, explaining only essential definition information. (Figure 10)
- **Cohesiveness**: whether sentences in the story fit together well. (Figure 11)
- **Completeness**: whether the story is comprehensive, accurate, and includes all relevant information. (Figure 12)
- **Factuality**: whether the story is factually accurate, supported by empirical evidence, and free from misinformation and hallucinations. (Figure 13)
- **Likeability**: whether the story is highly enjoyable or entertaining to read. (Figure 14)
- **Believability**: whether the story is plausible and internally consistent. (Figure 15)



Consent Form

What follows is a brief overview of the research we are completing with your assistance.

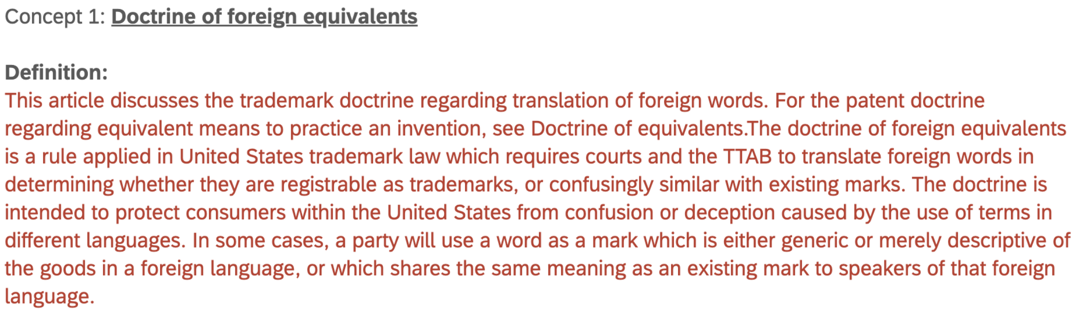
In this survey, we will ask you to read some **legal concept definitions and explanations** and ask for **demographic information.**

This survey takes approximately **30 minutes** to complete.

This survey is part of a scientific research project. Your decision to complete this survey is voluntary. If you give us permission by completing the survey, we plan to discuss/publish the results in an academic forum. In any publication, information will be provided in such a way that you cannot be identified. Only members of the research team will have access to the original data set. Before the data is shared outside the research team, any potentially identifying information will be removed. Once identifying data has been removed, the data may be used by the research team, or shared with other researchers, for both related and unrelated research purposes in the future. Your anonymized data may also be made available in online data repositories such as the Open Science Framework, which allow other researchers and interested parties to use the data for further analysis.

Clicking on the arrow at the bottom of this page indicates that you are at least 18 years of age and agree to complete this survey voluntarily.

Figure 4: Consent form on Prolific.

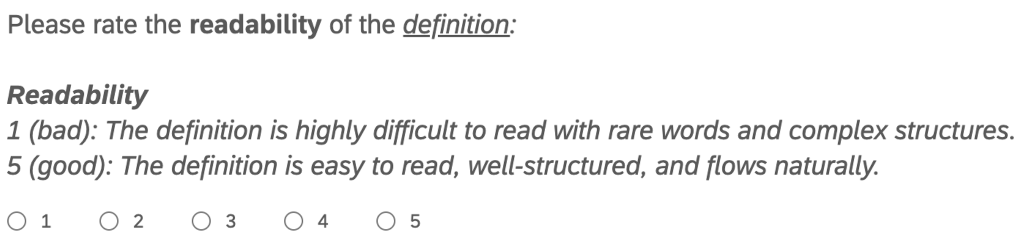


Concept 1: **Doctrine of foreign equivalents**

Definition:

This article discusses the trademark doctrine regarding translation of foreign words. For the patent doctrine regarding equivalent means to practice an invention, see Doctrine of equivalents.The doctrine of foreign equivalents is a rule applied in United States trademark law which requires courts and the TTAB to translate foreign words in determining whether they are registrable as trademarks, or confusingly similar with existing marks. The doctrine is intended to protect consumers within the United States from confusion or deception caused by the use of terms in different languages. In some cases, a party will use a word as a mark which is either generic or merely descriptive of the goods in a foreign language, or which shares the same meaning as an existing mark to speakers of that foreign language.

Figure 5: Concept definition example.

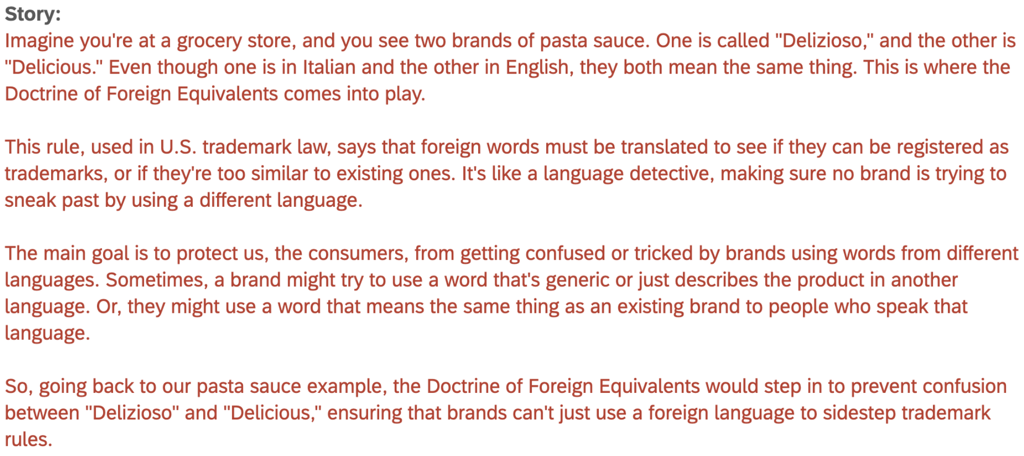


Please rate the **readability** of the *definition*:

Readability
1 (bad): The definition is highly difficult to read with rare words and complex structures.
5 (good): The definition is easy to read, well-structured, and flows naturally.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 6: Readability of definition (RoD) question.



Story:

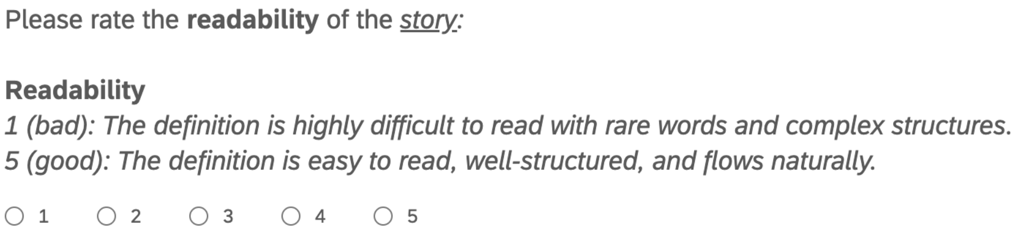
Imagine you're at a grocery store, and you see two brands of pasta sauce. One is called "Delizioso," and the other is "Delicious." Even though one is in Italian and the other in English, they both mean the same thing. This is where the Doctrine of Foreign Equivalents comes into play.

This rule, used in U.S. trademark law, says that foreign words must be translated to see if they can be registered as trademarks, or if they're too similar to existing ones. It's like a language detective, making sure no brand is trying to sneak past by using a different language.

The main goal is to protect us, the consumers, from getting confused or tricked by brands using words from different languages. Sometimes, a brand might try to use a word that's generic or just describes the product in another language. Or, they might use a word that means the same thing as an existing brand to people who speak that language.

So, going back to our pasta sauce example, the Doctrine of Foreign Equivalents would step in to prevent confusion between "Delizioso" and "Delicious," ensuring that brands can't just use a foreign language to sidestep trademark rules.

Figure 7: Story example.

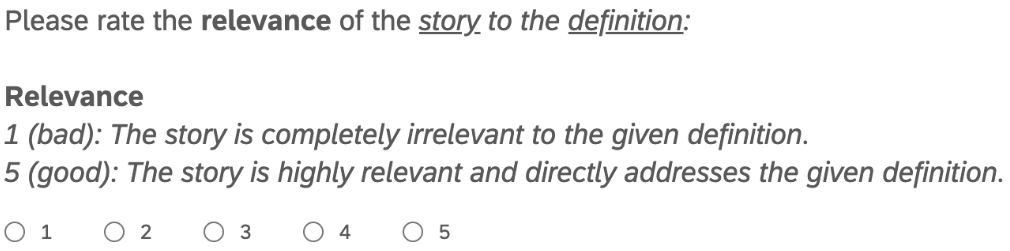


Please rate the **readability** of the *story*:

Readability
1 (bad): The definition is highly difficult to read with rare words and complex structures.
5 (good): The definition is easy to read, well-structured, and flows naturally.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 8: Readability of story (RoS) question.



Please rate the **relevance** of the *story to the definition*:

Relevance
1 (bad): The story is completely irrelevant to the given definition.
5 (good): The story is highly relevant and directly addresses the given definition.

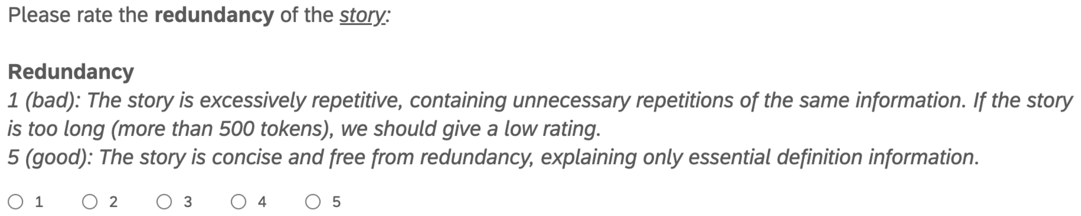
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 9: Relevance question.

After evaluating the story, these annotators are then asked to evaluate the three generated questions along with the suggested answer from LLMs. Here is one example of such question (Figire 16) and the rating about the question (Figure 17).

C.2 Annotator Demographics

We also include the demographics of 39 unique participants who contribute to evaluate the stories and questions. 26 participants are from UK and 13 from US. We show the distribution of age, sex, and



Please rate the **redundancy** of the story:

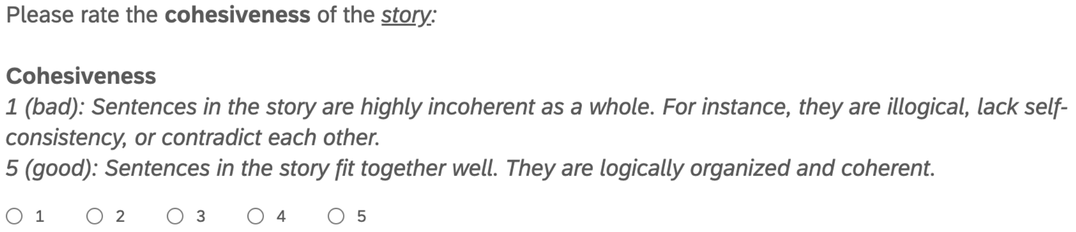
Redundancy

1 (bad): The story is excessively repetitive, containing unnecessary repetitions of the same information. If the story is too long (more than 500 tokens), we should give a low rating.

5 (good): The story is concise and free from redundancy, explaining only essential definition information.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 10: Redundancy question.



Please rate the **cohesiveness** of the story:

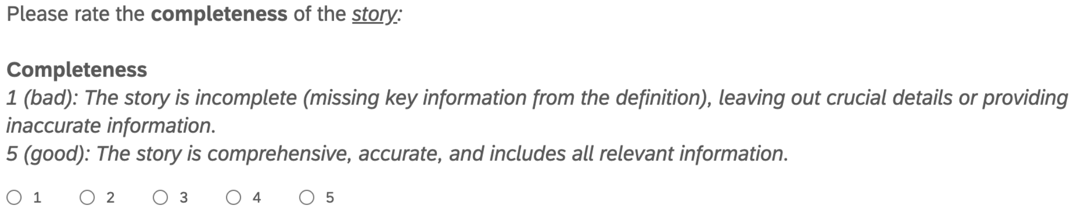
Cohesiveness

1 (bad): Sentences in the story are highly incoherent as a whole. For instance, they are illogical, lack self-consistency, or contradict each other.

5 (good): Sentences in the story fit together well. They are logically organized and coherent.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 11: Cohesiveness question.



Please rate the **completeness** of the story:

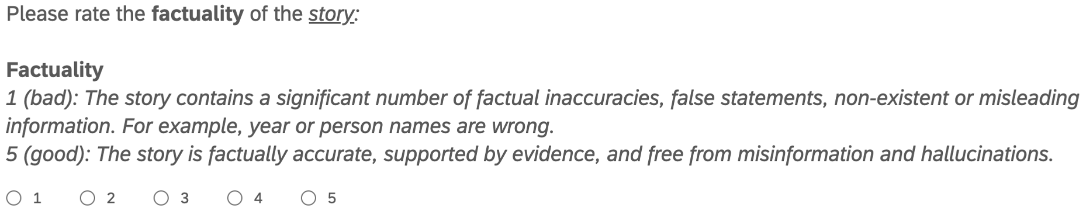
Completeness

1 (bad): The story is incomplete (missing key information from the definition), leaving out crucial details or providing inaccurate information.

5 (good): The story is comprehensive, accurate, and includes all relevant information.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 12: Completeness question.



Please rate the **factuality** of the story:

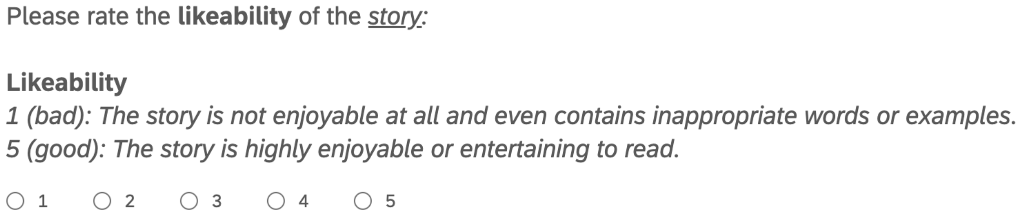
Factuality

1 (bad): The story contains a significant number of factual inaccuracies, false statements, non-existent or misleading information. For example, year or person names are wrong.

5 (good): The story is factually accurate, supported by evidence, and free from misinformation and hallucinations.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 13: Factuality question.



Please rate the **likeability** of the story:

Likeability

1 (bad): The story is not enjoyable at all and even contains inappropriate words or examples.

5 (good): The story is highly enjoyable or entertaining to read.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

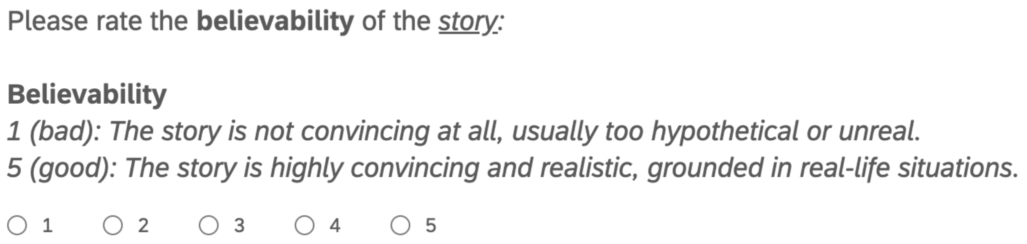
Figure 14: Likeability question.

ethnicity in the Figure 18.

D RCT Experiment Details

D.1 RCT Procedure

We recruit participants on Prolific with the following criteria: (a) have little to no law backgrounds, (b) have a bachelor's degrees as the person's highest degree, (c) lives or work in the North American region,



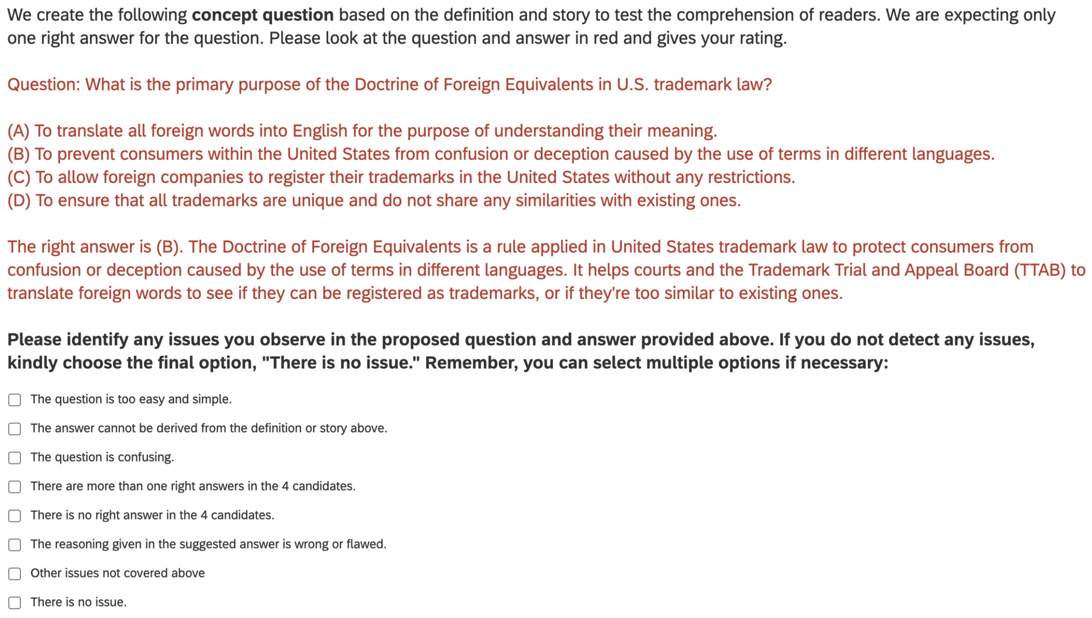
Please rate the **believability** of the *story:*

Believability

1 (bad): The story is not convincing at all, usually too hypothetical or unreal.
5 (good): The story is highly convincing and realistic, grounded in real-life situations.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 15: Believability question.



We create the following **concept question** based on the definition and story to test the comprehension of readers. We are expecting only one right answer for the question. Please look at the question and answer in red and gives your rating.

Question: What is the primary purpose of the Doctrine of Foreign Equivalents in U.S. trademark law?

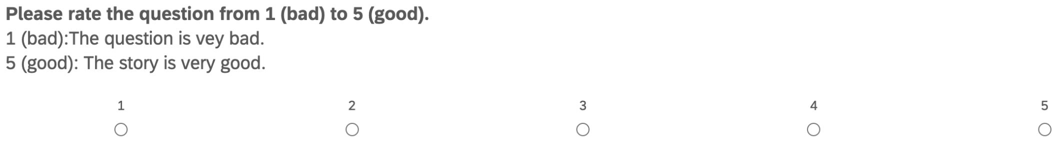
(A) To translate all foreign words into English for the purpose of understanding their meaning.
(B) To prevent consumers within the United States from confusion or deception caused by the use of terms in different languages.
(C) To allow foreign companies to register their trademarks in the United States without any restrictions.
(D) To ensure that all trademarks are unique and do not share any similarities with existing ones.

The right answer is (B). The Doctrine of Foreign Equivalents is a rule applied in United States trademark law to protect consumers from confusion or deception caused by the use of terms in different languages. It helps courts and the Trademark Trial and Appeal Board (TTAB) to translate foreign words to see if they can be registered as trademarks, or if they're too similar to existing ones.

Please identify any issues you observe in the proposed question and answer provided above. If you do not detect any issues, kindly choose the final option, "There is no issue." Remember, you can select multiple options if necessary:

☐ The question is too easy and simple.
☐ The answer cannot be derived from the definition or story above.
☐ The question is confusing.
☐ There are more than one right answers in the 4 candidates.
☐ There is no right answer in the 4 candidates.
☐ The reasoning given in the suggested answer is wrong or flawed.
☐ Other issues not covered above
☐ There is no issue.

Figure 16: Question example.



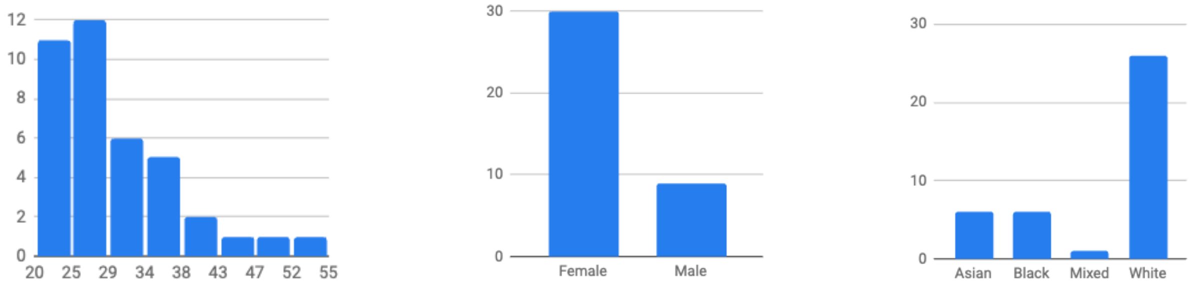
Please rate the question from 1 (bad) to 5 (good).
1 (bad):The question is very bad.
5 (good): The story is very good.

1	2	3	4	5
○	○	○	○	○

Figure 17: Rating question for the quiz.

(d) have an approval rate between 98 and 100. The criteria are set to have lower and upper bounds for language proficiency and background to limit the variance that can be sufficiently analyzed with our small sample size. After the pre-survey, the participants were randomized into two conditions based on age, gender, and educational background. We then deploy the tasks as batches of 5 concepts and recruit 15 to 20 people to complete the tasks; each person can only complete the same concept once. Due to the difference in completion levels, we have varying responses for each concept (ranging from 16 to 20).

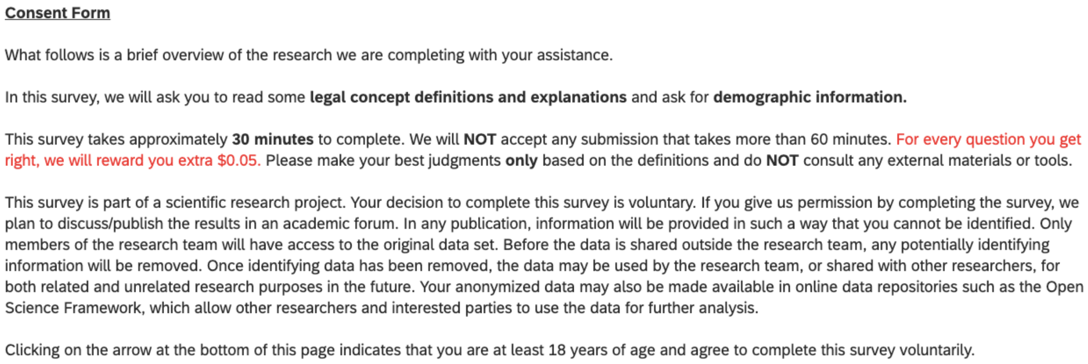
In the procedure of the RCT experiment, we begin with a consent form to ensure participants understand the task and how we want to use the data shown in Figure 19. In order to encourage participants to complete the questions accurately, we offer a reward of \$0.05 for each question that is answered correctly within 60 minutes. Afterwards, we present the participants with the concept and ask them to evaluate their familiarity of the concept (Figure 20). After reading the concept definition (Figure 21, the participant need to evaluate the perceived difficulty of the concept (Figure 22). If a participant is assigned to the treatment group, the participant will read a story (Figure 23) and evaluate the familiarity of the story setting to the person (Figure 24). Otherwise, the participant will skip this story reading part. Afterwards, the participant will answer three questions including the concept qustion (Figure 25), the prediction question (Figure 26),



(a) Age (b) Sex (c) Ethnicity

Figure 18: Distribution of age, sex, and ethnicity among the 39 Prolific annotators who evaluate the stories.

and the limitation question (Figure 27). At the end of the study, all participants will be asked if they are interested in learning more about law and legal knowledge.



Consent Form

What follows is a brief overview of the research we are completing with your assistance.

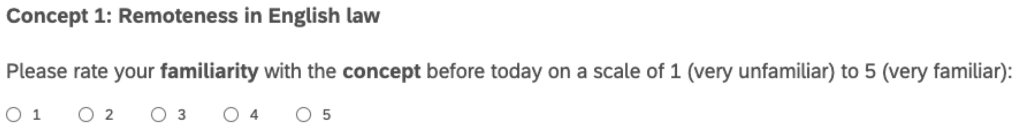
In this survey, we will ask you to read some **legal concept definitions and explanations** and ask for **demographic information.**

This survey takes approximately **30 minutes** to complete. We will **NOT** accept any submission that takes more than 60 minutes. For every question you get right, we will reward you extra $0.05. Please make your best judgments **only** based on the definitions and do **NOT** consult any external materials or tools.

This survey is part of a scientific research project. Your decision to complete this survey is voluntary. If you give us permission by completing the survey, we plan to discuss/publish the results in an academic forum. In any publication, information will be provided in such a way that you cannot be identified. Only members of the research team will have access to the original data set. Before the data is shared outside the research team, any potentially identifying information will be removed. Once identifying data has been removed, the data may be used by the research team, or shared with other researchers, for both related and unrelated research purposes in the future. Your anonymized data may also be made available in online data repositories such as the Open Science Framework, which allow other researchers and interested parties to use the data for further analysis.

Clicking on the arrow at the bottom of this page indicates that you are at least 18 years of age and agree to complete this survey voluntarily.

Figure 19: Consent form in the RCT.

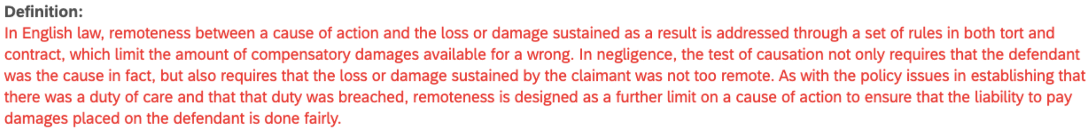


Concept 1: Remoteness in English law

Please rate your **familiarity** with the **concept** before today on a scale of 1 (very unfamiliar) to 5 (very familiar):

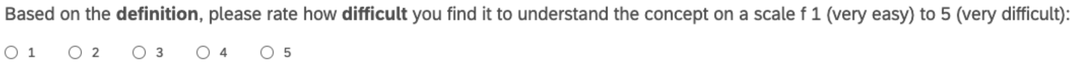
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 20: Concept familiarity in the RCT.



Definition:
In English law, remoteness between a cause of action and the loss or damage sustained as a result is addressed through a set of rules in both tort and contract, which limit the amount of compensatory damages available for a wrong. In negligence, the test of causation not only requires that the defendant was the cause in fact, but also requires that the loss or damage sustained by the claimant was not too remote. As with the policy issues in establishing that there was a duty of care and that that duty was breached, remoteness is designed as a further limit on a cause of action to ensure that the liability to pay damages placed on the defendant is done fairly.

Figure 21: Concept Definition example in the RCT.



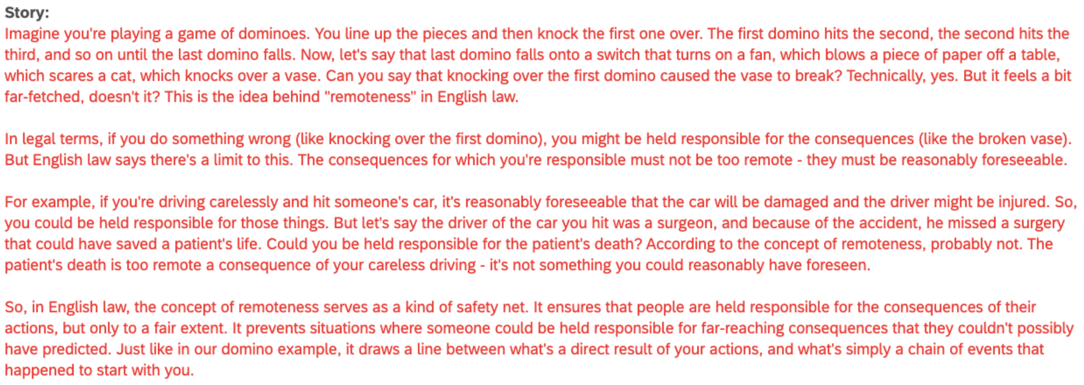
Based on the **definition**, please rate how **difficult** you find it to understand the concept on a scale f 1 (very easy) to 5 (very difficult):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 22: Perceived difficulty in the RCT.

D.2 Participant Demographics

There are 65 native respondents and 71 non-native respondents in the study. Some respondents participated in two batches and some participated in just one batch. Out of 136 respondents, there are 117 unique



Story:

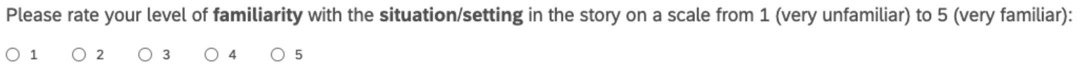
Imagine you're playing a game of dominoes. You line up the pieces and then knock the first one over. The first domino hits the second, the second hits the third, and so on until the last domino falls. Now, let's say that last domino falls onto a switch that turns on a fan, which blows a piece of paper off a table, which scares a cat, which knocks over a vase. Can you say that knocking over the first domino caused the vase to break? Technically, yes. But it feels a bit far-fetched, doesn't it? This is the idea behind "remoteness" in English law.

In legal terms, if you do something wrong (like knocking over the first domino), you might be held responsible for the consequences (like the broken vase). But English law says there's a limit to this. The consequences for which you're responsible must not be too remote - they must be reasonably foreseeable.

For example, if you're driving carelessly and hit someone's car, it's reasonably foreseeable that the car will be damaged and the driver might be injured. So, you could be held responsible for those things. But let's say the driver of the car you hit was a surgeon, and because of the accident, he missed a surgery that could have saved a patient's life. Could you be held responsible for the patient's death? According to the concept of remoteness, probably not. The patient's death is too remote a consequence of your careless driving - it's not something you could reasonably have foreseen.

So, in English law, the concept of remoteness serves as a kind of safety net. It ensures that people are held responsible for the consequences of their actions, but only to a fair extent. It prevents situations where someone could be held responsible for far-reaching consequences that they couldn't possibly have predicted. Just like in our domino example, it draws a line between what's a direct result of your actions, and what's simply a chain of events that happened to start with you.

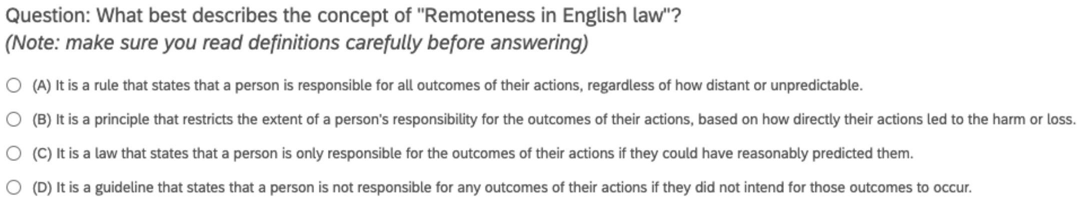
Figure 23: Story example in the RCT.



Please rate your level of **familiarity** with the **situation/setting** in the story on a scale from 1 (very unfamiliar) to 5 (very familiar):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 24: Familiarity with the story setting in the RCT.



Question: What best describes the concept of "Remoteness in English law"?
(Note: make sure you read definitions carefully before answering)

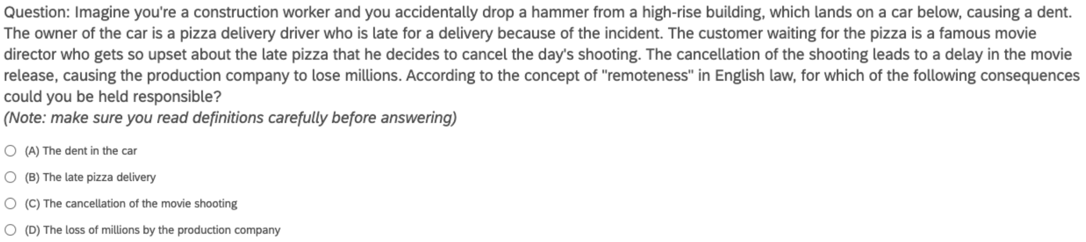
○ (A) It is a rule that states that a person is responsible for all outcomes of their actions, regardless of how distant or unpredictable.

○ (B) It is a principle that restricts the extent of a person's responsibility for the outcomes of their actions, based on how directly their actions led to the harm or loss.

○ (C) It is a law that states that a person is only responsible for the outcomes of their actions if they could have reasonably predicted them.

○ (D) It is a guideline that states that a person is not responsible for any outcomes of their actions if they did not intend for those outcomes to occur.

Figure 25: Concept question example in the RCT.



Question: Imagine you're a construction worker and you accidentally drop a hammer from a high-rise building, which lands on a car below, causing a dent. The owner of the car is a pizza delivery driver who is late for a delivery because of the incident. The customer waiting for the pizza is a famous movie director who gets so upset about the late pizza that he decides to cancel the day's shooting. The cancellation of the shooting leads to a delay in the movie release, causing the production company to lose millions. According to the concept of "remoteness" in English law, for which of the following consequences could you be held responsible?
(Note: make sure you read definitions carefully before answering)

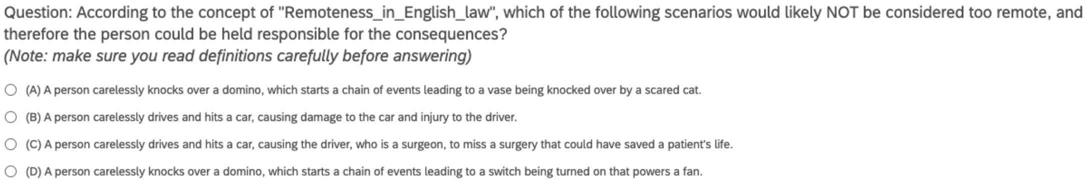
○ (A) The dent in the car

○ (B) The late pizza delivery

○ (C) The cancellation of the movie shooting

○ (D) The loss of millions by the production company

Figure 26: Prediction question example in the RCT.



Question: According to the concept of "Remoteness_in_English_law", which of the following scenarios would likely NOT be considered too remote, and therefore the person could be held responsible for the consequences?
(Note: make sure you read definitions carefully before answering)

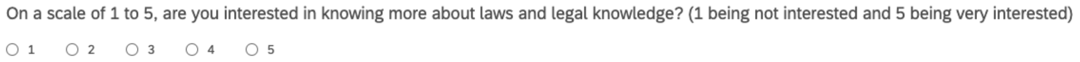
○ (A) A person carelessly knocks over a domino, which starts a chain of events leading to a vase being knocked over by a scared cat.

○ (B) A person carelessly drives and hits a car, causing damage to the car and injury to the driver.

○ (C) A person carelessly drives and hits a car, causing the driver, who is a surgeon, to miss a surgery that could have saved a patient's life.

○ (D) A person carelessly knocks over a domino, which starts a chain of events leading to a switch being turned on that powers a fan.

Figure 27: Limitation question example in the RCT.

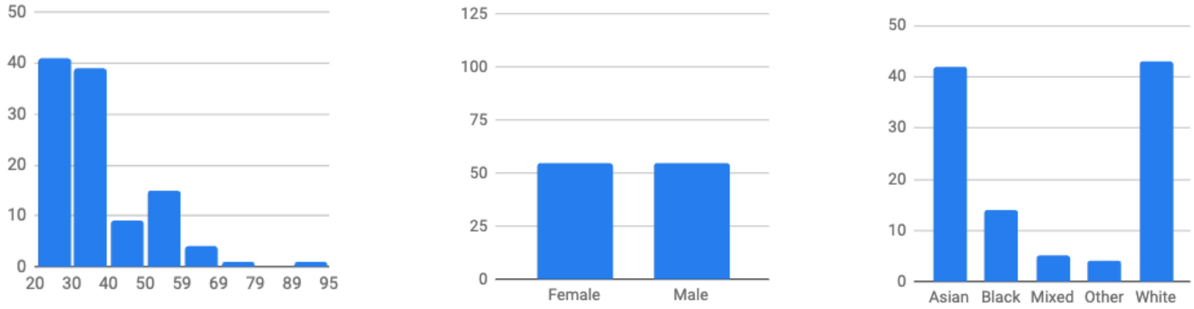


On a scale of 1 to 5, are you interested in knowing more about laws and legal knowledge? (1 being not interested and 5 being very interested)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 28: Law interst near the end of the RCT.

participants in our RCT experiments and we are able to collect the demographics of 110 participants. 35 are from Canada and 75 are from the United States. 70 report their first language as English and 40 report their first language as other languages. We include the distribution of age, sex, and ethnicity in the Figure 29.



(a) Age (b) Sex (c) Ethnicity

Figure 29: Distribution of age, sex, and ethnicity among the 110 RCT participants from Prolific.

D.3 Individual Accuracy

We provide the individual level accuracy in addition to the result presented in the main paper as a reference. In the human-subject study, we distribute the 10 concepts in 2 batches (5 concepts in each batch) on Prolific to prevent overloading the participants and losing their engagement. It is worth noting that it takes 20 to 25 minutes to complete the study with 5 concepts. In the following Table 7 and 8, we report the individual accuracy by batches.

Condition	**ConceptQ**	**PredictionQ**	**LimitationQ**
Native Speakers (Individual Accuracy - Batch 1)			
Definition	97.65 ± 9.41	78.82 ± 24.22	80.00 ± 24.73
Def. + Story	91.25 ± 19.96	77.50 ± 29.05	83.75 ± 24.71
Non-native Speakers (Individual Accuracy - Batch 1)			
Definition	89.52 ± 15.88	73.33 ± 24.94	73.33 ± 24.16
Def. + Story	92.22 ± 13.56	84.44 ± 17.07	92.22 ± 15.11

Table 7: Average and Standard Deviation of Individual Level Accuracy for Batch 1

Condition	**ConceptQ**	**PredictionQ**	**LimitationQ**
Native Speakers (Individual Accuracy - Batch 2)			
Definition	88.75 ± 12.18	78.75 ± 16.54	75.00 ± 18.03
Def. + Story	90.00 ± 17.32	71.25 ± 22.33	85.00 ± 22.91
Non-native Speakers (Individual Accuracy - Batch 2)			
Definition	88.75 ± 17.28	70.00 ± 25.50	62.50 ± 27.27
Def. + Story	90.00 ± 12.25	78.75 ± 17.98	76.25 ± 17.63

Table 8: Average and Standard Deviation of Individual Level Accuracy for Batch 2