

Objection Overruled!

Lay People can Distinguish Large Language Models from Lawyers, but still Favour Advice from an LLM

Eike Schneiders
University of Nottingham
Nottingham, UK
eike.schneiders@nottingham.ac.uk

Tina Seabrooke
University of Southampton
Southampton, UK
t.seabrooke@soton.ac.uk

Joshua Krook
University of Antwerp
Antwerp, Belgium
joshua.krook@uantwerpen.be

Richard Hyde
University of Nottingham
Nottingham, UK
richard.hyde@nottingham.ac.uk

Natalie Leesakul
University of Nottingham
Nottingham, UK
natalie.leesakul@nottingham.ac.uk

Jeremie Clos
University of Nottingham
Nottingham, UK
jeremie.clos@nottingham.ac.uk

Joel Fischer
University of Nottingham
Nottingham, UK
joel.fischer@nottingham.ac.uk

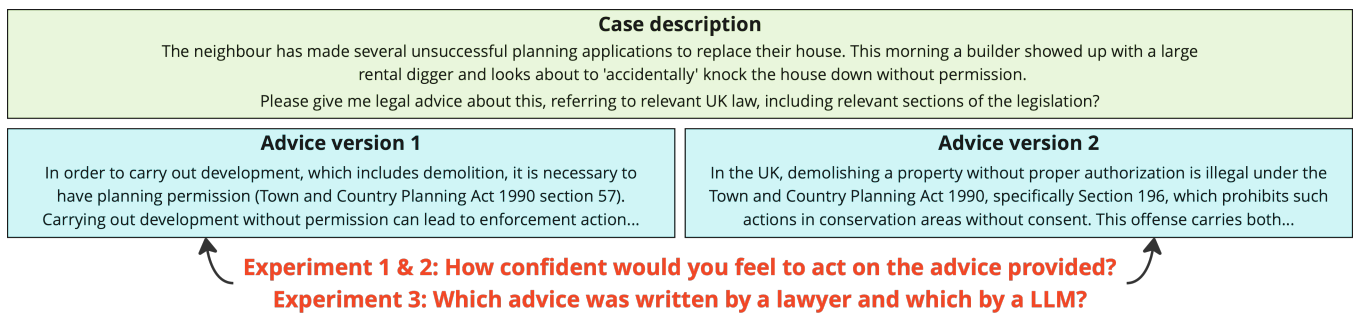


Figure 1: Legal advice provided by a lawyer (advice version 1) and a Large Language Model (LLM; advice version 2) in response to the same prompt. Experiments 1 and 2 evaluated participants' willingness to act on the legal advice, while Experiment 3 investigated participants' ability to discriminate between the advice sources.

ABSTRACT

Large Language Models (LLMs) are seemingly infiltrating every domain, and the legal context is no exception. In this paper, we present the results of three experiments (total $N = 288$) that investigated lay people's willingness to act upon, and their ability to discriminate between, LLM- and lawyer-generated legal advice. In Experiment 1, participants judged their willingness to act on legal advice when the source of the advice was either known or unknown. When the advice source was unknown, participants indicated that they were significantly more willing to act on the LLM-generated advice. This

result was replicated in Experiment 2. Intriguingly, despite participants indicating higher willingness to act on LLM-generated advice in Experiments 1 and 2, participants discriminated between the LLM- and lawyer-generated texts significantly above chance-level in Experiment 3. Lastly, we discuss potential explanations and risks of our findings, limitations and future work, and the importance of language complexity and real-world comparability.

CCS CONCEPTS

• **Human-centered computing** → *User studies; HCI theory, concepts and models*; • **Applied computing** → *Law*.

KEYWORDS

Large language model, LLM, legal advice, generative AI, ChatGPT

ACM Reference Format:

Eike Schneiders, Tina Seabrooke, Joshua Krook, Richard Hyde, Natalie Leesakul, Jeremie Clos, and Joel Fischer. 2018. Objection Overruled! Lay People can Distinguish Large Language Models from Lawyers, but still

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '25, 26 Apr.–1 May, 2025, Yokohama, Japan

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXXXX.XXXXXXX>

Favour Advice from an LLM. In *Proceedings of Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, New York, NY, USA, 13 pages. <https://doi.org/XXXXXXX.XXXXXX>

1 INTRODUCTION

The emergence of generative artificial intelligence (GenAI), particularly Large Language Models (LLMs), has shifted public attention towards the impact and tangibility of AI. OpenAI's ChatGPT, specifically, has captured the attention of the mainstream media, with many journalists focusing on its potential for misuse^{1,2}. LLMs have also generated an explosion of interest in diverse research communities, including, e.g., AI [65, 66], HCI [20, 59, 60], and CSCW [30, 58]. For instance, 21% of accepted full papers in the CHI '24 proceedings included the term 'LLM' or 'Large Language Model'³, demonstrating the clear interest in LLM research within the HCI community.

With this surge in interest in LLMs, they are expected to continue to have wide-ranging impact on society. Examples include the healthcare sector, in which medical professionals have used LLM-based platforms to obtain fast summaries of patient health information [63] and in the financial sector, where prior research has shown shortcomings in relation to the ability of LLMs to reliably retrieve basic financial information [33]. In the legal context, the use [24] and misuse of LLMs by legal experts has resulted in negative media attention^{1,2}. Similar observations have been reported with lay people, who reported high willingness to obtain legal advice from LLMs [51]. These cases highlight the need to further understand the impact that this 'new' technology, and its wide ranging accessibility, has on society.

In this paper, we present three experiments investigating lay peoples perceptions of LLMs in the context of legal advice. In doing so, we investigate two research questions:

- (1) Are lay people more willing to act on LLM- or lawyer-generated legal advice?
- (2) Are lay people, when unaware of the source, able to identify the source of legal advice presented to them?

Building on existing literature [51], Experiment 1 ($N = 100$) investigated whether lay people report higher willingness to act upon LLM- or lawyer-generated legal advice. Experiment 2 ($N = 78$) replicate the key manipulation of Experiment 1. Lastly, Experiment 3 ($N = 110$), investigated participants ability to discriminate between LLM- and lawyer-generated legal advice when the source was unknown to them. Experiment 1, replicated in Experiment 2, showed that participants reported significantly higher willingness to act on LLM-generated advice than lawyer-generated advice, when the advice source was unknown to them. In Experiment 3, participants showed above chance level when discriminating between LLM- and lawyer-generated advice, when the source was unknown.

¹ChatGPT: US lawyer admits using AI for case research: <https://www.bbc.co.uk/news/world-us-canada-65735769>

²Canada lawyer under fire for submitting fake cases created by AI chatbot: <https://www.theguardian.com/world/2024/feb/29/canada-lawyer-chatgpt-fake-cases-ai>

³Full Query Syntax for the dl.acm.org: "query": { llm OR "large language model" } "filter": { Conference Collections: CHI: Conference on Human Factors in Computing Systems, Book/Issue: CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems, E-Publication Date: (01/01/2024 TO 12/31/2024) }

We conclude the paper by discussing the importance of language complexity, the risks and strategies to mitigate risks that overtrust on LLMs can bring, as well as limitations and future work.

1.1 Openness and Transparency Statement

Experiment 1 was not pre-registered, but Experiments 2 and 3 were. For all experiments, the trial-level raw data and accompanying data analysis scripts are available on the Open Science Framework (OSF)⁴. We report the rationale for the sample sizes, all manipulations, measures, and data exclusions. All three experiments were approved by the Institutional Review Board (IRB) of the University of Nottingham (CS-2023-R22).

2 RELATED WORK

This section highlights relevant literature on the impact LLMs have had on the legal landscape to date. The section also notes additional high risk domains, beyond the legal context, that have been affected by LLMs. Lastly, we present relevant literature on trust and general perceptions towards AI.

2.1 LLMs in the Legal Landscape

LLMs are increasingly being used by legal professionals for a range of tasks, from legal research and e-discovery, to contract drafting and the filing of motions in court [2, 24]. In many cases, law firms are limiting the use of LLMs to summarising information, digesting large numbers of documents (in discovery, for example), and providing summaries for lawyers to review [29]. Litigation matters involving large companies can involve tens of thousands of documents, meaning that LLMs serve a practical and functional use-case in these scenarios.

On the other hand, some lawyers have already been caught-out relying on LLMs (such as ChatGPT and Gemini) to generate legal advice for clients, often running into the problem of hallucinated facts. In *Mata v. Avianca* [18], for example, a lawyer relied on falsely generated cases while seeking a damages claim against an airline on behalf of his injured client. The court rebuked the lawyer, stating that the provision of false or misleading cases wastes the time of the court and the opposing counsel. This misuse of the court's time interrupts the court's schedule, potentially impacting other cases, and wasting public money [56].

Courts in the United States have thus far taken a dim view of the use of LLMs for legal filings, noting the risk of errors and hallucinations and the necessary due diligence of legal professionals. In another case involving Google's Bard (now Gemini), a lawyer was again found to provide false cases to the court, which then had to try to find the non-existent cases [19]. As the problems of LLMs become more widely known, courts may start issuing harsher reprimands, or even revoking the license to practice of lawyers who provide false cases. This is in part due to the serious nature of lying to the court, something taken seriously in past legal precedent [21].

There are nevertheless circumstances where an LLM could provide high level of accuracy, while still falling below the threshold of an expert's legal advice. Nay et al. [45] investigated the use of an LLM, specifically models developed by OpenAI, in the context of legal analysis for tax law. Their findings demonstrated that "LLMs,

⁴https://osf.io/bksqa/?view_only=8c9a5893fb52478cb755870e56e686ca

particularly when combined with prompting enhancements and correct legal texts, can perform at high levels of accuracy but not yet at expert tax lawyer levels." This suggests that LLMs, even with prompt enhancements, are not yet able to reach professional performance in the domain of, e.g., tax law.

LLMs are influencing the legal landscape not only from a professional perspective, but also in ways that might have impact on non-legal experts [28, 51]. Seabrooke et al. [51] investigated lay people's willingness to obtain legal advice in a wide variety of domains (e.g., traffic, divorce, planning, property, or civil disputes). Nearly half of the participants (45%) said that they would be likely to use LLMs to inquire legal advice in the future. Furthermore, they highlight that the willingness to generate legal advice is not evenly distributed, and varies greatly depending on the specific domain (e.g., civil disputes (25%) vs. tenancy law (58%)). However, while these results suggest that lay people are willing to generate advice using LLMs, the authors did not investigate whether lay people are willing to *act* on it.

2.2 LLMs in other Domains

Large language models have been deployed and studied in numerous domains, including healthcare [32, 39, 44, 49], journalism, communication, and public messaging [3, 11, 57], and education [6, 35, 36, 53, 55]. In healthcare, Nadarzynski et al. [44] examined the acceptability of AI-led chatbot services and found that while most participants were open to using chatbots for healthcare, concerns about accuracy and lack of human touch were barriers to adoption. Miles et al. [39] further explored this issue by investigating how perceived stigma and severity of health issues influence acceptance of chatbots, finding that while chatbots might be suitable for sensitive health issues due to increased anonymity, they are less preferred for severe health conditions. Ayers et al. [4] compared chatbot responses to those of physicians on a public social media forum, demonstrating that chatbot-generated responses were often preferred for their quality and empathy. Finally, Reis et al. [49] investigated the public perception of AI-generated medical advice, highlighting an 'anti-AI bias' in which advice, even when supervised by physicians, was deemed less reliable and empathetic compared to solely human-generated advice.

Waddell [57] examined how audiences perceive news articles attributed to machine authors and found that machine-authored news is often deemed less credible due to lower perceived anthropomorphism and unmet expectations. Asscher and Glikson [3] investigated human evaluations of machine translation, particularly in ethically sensitive situations (e.g., communication between users with power imbalance), and found a negative bias against machine-translated content even among professional translators.

The effectiveness of LLMs in public messaging has also been a topic of interest. Chi [11] explored the impact of AI chatbots on individuals' attitudes towards environmental protection and their willingness to pay for conservation efforts. The study found that the problem solving capabilities of chatbots were particularly influential in encouraging financial contributions. Oviedo-Trespalcacios et al. [46] illustrated risks of using LLMs for safety-related advice, revealing that they often provide over-simplified, erroneous, or biased information.

In educational settings, LLMs have become an object of interest both for their potential as a tool for personalisation of learning and for their risk to become a tool for undetectable cheating. Liu and M'hiri [36] showed the potential of using an LLM as a virtual teaching assistant, enhancing student engagement and providing personalised feedback, while also highlighting the need for human supervision. Bernabei et al. [6] examined ChatGPT's use in engineering education, revealing its ability to improve understanding and expedite assignments, but also emphasising the importance of critical evaluation and ethical considerations. Steenstra et al. [53] demonstrated the potential of LLMs in helping the development of health education games, while underscoring the need for human oversight for the accuracy of the content. Tayan et al. [55] focused on the impact of ChatGPT on technology courses, advocating its use as a complementary tool, and providing guidelines for responsible implementation. Lim et al. [35] offered a paradoxical perspective on Generative AI in education, emphasising the need to embrace its potential as a tool to support learners while addressing academic and ethical challenges that stem from its growing capabilities as a generator of content.

2.3 Trust, Perception, and Acceptability of AI

The perception of AI influences how people feel about its outputs. Jakesch et al. [26] identified the 'Replicant Effect,' where individuals exhibit distrust toward Airbnb hosts believed to have AI-generated profiles, particularly in mixed environments with both AI and human-authored profiles. Although AI aversion is prevalent, Logg et al. [37] observed that individuals often prefer algorithmic over human judgement in specific contexts. Their research revealed that people tend to adhere more to advice when they believe it originates from an algorithm rather than a person, showcasing algorithm appreciation. This preference was consistent in various domains, including numerical estimations, forecasting tasks, and even subjective judgements such as romantic attraction.

The increasing use of LLMs for advice-seeking has prompted investigations into user characteristics and perceptions. Wester et al. [61] explored how people perceive LLM-generated advice and the influence of user characteristics on these perceptions. Their findings suggest that personality traits, particularly agreeability and technological insecurity, play a role in shaping how people evaluate and engage with LLM advice.

Acceptance of AI is not uniform and appears to be influenced by various factors. For example, Castelo and Ward [10] found that political conservatism correlated with AI aversion, particularly for consequential tasks. This aversion stems from conservatives' increased risk perception associated with AI adoption. Furthermore, Bigman and Gray [7] discovered a general aversion to machines making moral decisions, mainly due to the perception that machines lack a complete mind capable of both thinking and feeling. The perceived lack of agency and experience in machines contributes to this reluctance, suggesting that the nature of the task and the perceived capabilities of AI play a role in shaping people's acceptance of AI in decision-making roles.

Example: Traffic Law 4

Prompt:

“My father got a ticket saying he broke Section 5(1)(a) of the Road Traffic Offenses law. What happened was he rode over a speed bump and got stuck in a carriageway, preventing cars from going past him. An angry driver came over to him and started banging on his windscreen, threatening to call the police. My father ran away from this angry man before the police arrived. Can you give me some advice about my legal situation, considering relevant UK laws?”

Advice provided by LLM:

“Under Section 5(1)(a) of the Road Traffic Offenses law in the UK, it is an offense to use a vehicle in a manner that causes unnecessary obstruction of the road. In your father’s case, being stuck on a speed bump and blocking the carriageway could be seen as such an obstruction. However, it is also important to note that he fled the scene due to fear for his safety from the aggressive driver. This context could be relevant in his defense, as it demonstrates a potential threat that prompted his departure. It is advisable to gather any available evidence, such as witness statements or CCTV footage, that supports his account of the events, and to consult with a solicitor who can help argue that his actions were a response to imminent danger rather than a willful obstruction.”

Advice provided by lawyer:

“If your father was driving a car (which seems to be implied by the banging on the windscreen, but you say that he “rode” over the speed bump), I presume that your father has received a notice saying that he acted contrary to section 5 of the Road Traffic Regulation Act 1984. This prohibits acting contrary to a Traffic Regulation Order which is in force for a particular road. Road Traffic Regulation Orders can prohibit or restrict the waiting of vehicles in particular areas. If there was a Road Traffic Regulation Order in place it may be that when your father ran away from the angry driver he was seen to have breached this prohibition. Acting contrary to a Road Traffic Regulation Order is an offence which can lead to a fine of up to £1000. However, the circumstances you describe (if they can be proved, for example through dashcam footage) may amount to mitigation, and lead to a much lower penalty. Your father should consult a solicitor as soon as possible.”

Figure 2: Example of prompt for Traffic Law (#4) with corresponding LLM and Lawyer generated advice.

3 EXPERIMENT 1

In Experiment 1, we investigated participants’ willingness to act on legal advice generated by an LLM or a lawyer, both when the source of advice was known and unknown. All participants were shown a series of legal cases and were asked to rate their willingness to act on the legal advice using a scale from 1 (‘Strongly Disagree’)

to 9 (‘Strongly Agree’). Half of the participants were explicitly informed of the source of each advice provided, being either LLM- or lawyer-generated, while the remaining participants were not. The key questions were whether participants would be more willing to act on the legal advice provided by a LLM or by a lawyer, and whether this would interact with whether the source of the advice was known to participants or not.

3.1 Method

3.1.1 Participants. When considering the number of participants required for Experiment 1, we did not have a clear estimate of the effect size for the interaction between the advice source (LLM vs. lawyer) and group (source known vs. source unknown) factors. We therefore recruited 100 participants, with 50 participants in each group. A power analysis in G*Power [16] estimated that 98 participants would provide 95% power to obtain a small-to-medium effect ($n \approx 98$, $f = 0.15$, $1 - \beta = .95$, $\alpha = .05$), using the default values of 0.5 for the correlation between measures and 1 for the nonsphericity correction.

For all three experiments, participants were recruited from Prolific and we excluded any participants who failed at least two out of three attention checks. No participants were excluded in Experiment 1. As the legal prompts and responses were based on UK law, we only recruited participants who self reported fluency in English, were between 18–60 years old, and were currently living in the UK.

The final, overall sample for Experiment 1 consisted of 100 participants (60 male, 40 female, 0 prefer not to say) who were aged between 19 and 58 years ($M = 33.26$ years, $SD = 9.00$ years). The source unknown group consisted of 50 participants (29 male, 21 female, 0 prefer not to say), who were aged between 19 and 58 years ($M = 32.92$ years, $SD = 9.06$ years). The source known group consisted of 50 participants (31 male, 19 female, 0 prefer not to say), who were aged between 19 and 53 years ($M = 33.60$ years, $SD = 9.01$ years).

The median completion time was 14:53 minutes but, as progression in the experiment was self-paced, the completion time varied between participants. Participants were compensated with £9/hour.

3.1.2 Prompts. Based on prior literature [51], we selected three areas within the legal context for which participants reported a high likelihood to use LLMs for the generation of legal advice, specifically: traffic law, planning law, and property law. In consultation with legal professionals, we identified six prompts for each of these three areas, resulting in a total of 18 prompts. These prompts were inspired by real questions asked online in the subreddit [r/LegalAdviceUK](https://www.reddit.com/r/LegalAdviceUK)⁵. All prompts were based on UK law. Example prompts as well as LLM- and lawyer-generated answers for each category can be seen in Figures 2 to 4. All 18 prompts and the corresponding LLM- and lawyer-generated answers are available in the OSF (see Section 1.1). The LLM-generated prompts were generated using ChatGPT-4o, while the lawyer generated prompts were generated by UK-based lawyers with expertise within the three domains: traffic, planning, and property. The lawyers generating the legal advice did not see the LLM-generated advice, and were asked to generate succinct advice based on the provided prompts.

⁵www.reddit.com/r/LegalAdviceUK

The lawyer-generated advice was often preceded with a re-iteration of the case details, which was removed to ensure comparability in brevity between the LLM and lawyer-generated advice. We verified, with legal experts, that no text was removed which changed the meaning of the advice. No text was added to the advice provided. Each prompt was shown the same number of times in each experiment, and all three experiments used the same prompts and responses.

3.1.3 Experimental Design and Measurements. Experiment 1 followed a 2 (advice source: LLM vs. lawyer) × 2 (group: source known vs. source unknown) mixed factorial design. Advice source and group were manipulated within-subjects and between-subjects, respectively. The dependent variable was participants' ratings of their willingness to act on each legal advice provided, which were ranked on scale from 1 ('Strongly Disagree') to 9 ('Strongly Agree'). All three experiments were implemented using jsPsych⁷ and hosted using JATOS⁷.

3.1.4 Procedure. Participants were presented with the participant information sheet highlighting relevant information including details on the data being collected, participants' right to withdraw, and information on the procedure of the experiment. Once participants consented to their participation, we collected demographic data on age, gender, and language fluency. As part of the demographic data, we had a simple attention check ('Please press the highlighted key on your keyboard'). Participants were subsequently informed that some of the legal advice they would be presented with had been generated by lawyers, while other advice had been generated by an LLM. To ensure a common understanding of what an LLM was, a short, easily understandable, description was provided (see Appendix A).

Participants were randomly allocated to the source known or unknown group, ensuring that we had an equal number of participants in each group. Each participant was presented with 18 legal cases. These cases were written by either an LLM (9) or a lawyer (9) using the same 18 prompts. The LLM and the lawyer answered the same short legal prompts within each of the three categories: traffic law, planning law, and property law [51]. Two additional attention check cases were presented at random times, resulting in a total of 20 trials. In the source known condition, participants saw the statement: 'Case #: This advice has been written by [an LLM / a lawyer]' above the prompt and the advice. In the source unknown condition, by contrast, each prompt and advice was simply placed below the statement: 'Case #', with the author of the advice omitted. For each case, participants were asked to answer 'In the described situation, I would feel confident acting on the advice provided' on a 9-point scale from 1 ('Strongly Disagree') to 9 ('Strongly Agree'). Participants had to respond to each case in order to move on with the task, and progression throughout the task was self-paced.

3.2 Experiment 1: Results

3.2.1 Data Analysis. All analyses were performed in RStudio [48]. We report both frequentist analyses and Bayes factors to provide estimates of the evidence in favour of the alternative (H1) and null

⁶<https://www.jspsych.org/latest/>

⁷<https://www.jatos.org>

Example: Property Law 2

Prompt:

"My landlord has threatened to evict me from my apartment because I have a dog. They have now told me there is a "no pets" clause in the contract, but promised me it was okay to have a dog when I first moved in. Could you advise me about the legality of this in UK law, with reference to relevant legislation?"

Advice provided by LLM:

"In the UK, the legality of pet ownership in rental properties hinges on the terms of the tenancy agreement and relevant housing legislation. While landlords can include "no pets" clauses, previous verbal consent for pet ownership could constitute a waiver or variation of contractual terms, provided there is evidence to support this claim. The Housing Act 1988 allows for such clauses, but landlords must also act reasonably and fairly towards tenants, and breaches of the Consumer Rights Act 2015 may also be relevant. If negotiations fail, seeking legal advice or assistance from housing charities like Shelter is advisable, as the outcome depends on the specifics of the agreement, relevant laws, and potential legal recourse."

Advice provided by lawyer:

"On first impressions there are two legal avenues we could consider. The first is that the terms of your lease were formally varied by the conversation with your landlord. This would require the law to accept that a lease can be varied orally, in contrast to the usual practice that dealings with land need to be made in writing. This is a hard argument to make since longer residential leases tend to be varied via a Deed of Variation. What is more, if your lease agreement contains a provision that the terms can only be varied in writing, then the recent Supreme Court ruling in Rock Advertising Limited confirms the binding effect of such clauses. Nevertheless, if such a clause is absent, an argument could be made that it is possible to vary orally. For instance, the requirement for writing under the Law of Property Act 1925 relates to creating or disposing of an interest in land, but here the interest in land (the lease) has already been given to you in writing and what we would be arguing is that this is a subsequent change involving no such creation or disposal. As a result the variation can be binding even if agreed orally."

Figure 3: Example of prompt for Property Law (#2) with corresponding LLM and Lawyer generated advice.

(H0) hypotheses. The evidence categories proposed by Jeffreys [27], and the corresponding interpretations proposed by Lee and Wagenmakers [34], are shown in Table 1 and were used to interpret the Bayes factors. All Bayes factors were calculated using the default priors provided in the 'BayesFactor' [43] package (version

Example: Planning Law 6

Prompt:

“I am in the middle of building a fence with my neighbour’s property, who has objected to the proposed colour of the fence. The council has rejected my planning permission and told me that I cannot use a colour and must install a basic timber fence, and that our planned metal fence does not comply with local regulations. Please give me legal advice about this, referring to relevant sections of UK law?”

Advice provided by LLM:

“In the context of UK property law, it’s essential to adhere to local planning regulations enforced by the council, as failure to do so can result in legal consequences. The rejection of planning permission and insistence on installing a basic timber fence instead of a proposed metal one signifies non-compliance with local regulations. Sections 106 and 14 of the Town and Country Planning Act 1990 and the Planning (Listed Buildings and Conservation Areas) Act 1990 empower local authorities to enforce conditions to ensure compliance with policies, particularly in conservation areas. To avoid potential fines and enforcement actions, it’s advisable to comply with the council’s directives, seek guidance on alternative options, and consult a legal professional specializing in property law and planning regulations for further assistance.”

Advice provided by lawyer:

“You do not need to obtain planning permission to erect a fence provided that your proposed fence complies with the requirements in the Town and Country Planning (General Permitted Development) Order 2015 Schedule 2 Part 2 Paragraph A. The requirements relate to the height of the fence, and do not limit the material or colour of the fence. However, such limitations may be imposed if your property is situated in a national park or a conservation area, or if your property is a listed building. If this is the case, you need to ensure that a proposed fence meets the requirements of the conservation area or national park authority, and/or that you obtain listed building consent before erecting your fence. Failure to do this may mean that enforcement action can be taken against you, which may involve you being required to remove the fence. Failure to comply with such a requirement may amount to a criminal offence.”

Figure 4: Example of prompt for Planning Law (#6) with corresponding LLM and Lawyer generated advice.

0.9.12.4.7). Estimated Bayes factor values are reported with the error percentage of each estimate.

3.2.2 *Willingness to Act on Legal Advice.* Participants’ mean ratings for their willingness to act on the legal advice, separated by advice source and experimental group, are shown in Panel A of Figure 5.

BF_{10}	Interpretation
> 100	Extreme evidence for H1
30 - 100	Very strong evidence for H1
10 - 30	Strong evidence for H1
3 - 10	Moderate evidence for H1
1 - 3	Anecdotal evidence for H1
1	No evidence
0.33 - 1	Anecdotal evidence for H0
0.10 - 0.33	Moderate evidence for H0
0.03 - 0.10	Strong evidence for H0
0.01 - 0.03	Very strong evidence for H0
< 0.01	Extreme evidence for H0

Table 1: Bayes factor evidence categories according to Jeffreys [27] and corresponding interpretations by Lee and Wagenmakers [34].

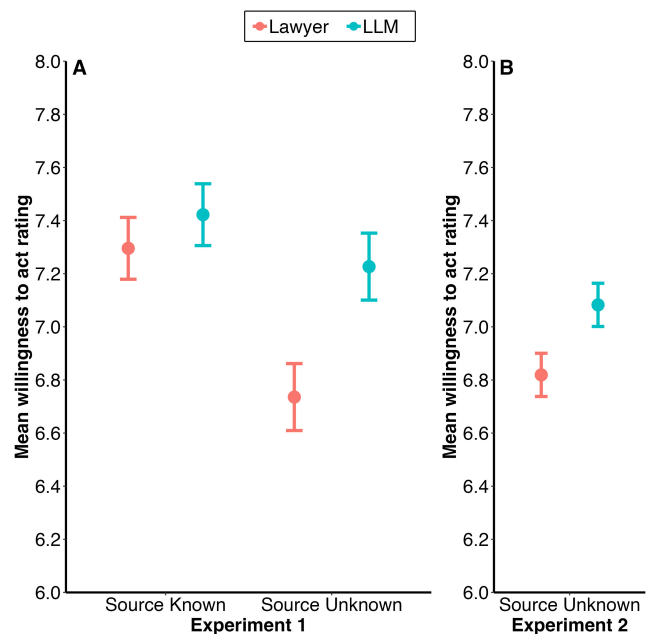


Figure 5: Mean willingness to act ratings in Experiments 1 and 2. Error bars represent difference-adjusted, within-subjects, 95% confidence intervals [5].

A 2 (advice source: LLM vs. lawyer) \times 2 (group: source known vs. source unknown) mixed analysis of variance (ANOVA) on participants’ mean ratings revealed a significant main effect of source, $F(1, 98) = 13.09, p < .001$, generalised eta squared (η_G^2) = .02. The corresponding Bayes factor provided very strong evidence for the alternative hypothesis, $BF_{10} \approx 40.27 (\pm 1.96\%)$. Participants gave significantly higher ratings to the LLM-generated advice ($M = 7.32, SD = 1.09$) than the lawyer-generated advice ($M = 7.02, SD = 1.11$). No significant main effect of group was observed, with the Bayes factor analysis providing anecdotal support for the alternative hypothesis, $F(1, 98) = 3.58, p = .06, \eta_G^2 = .03, BF_{10} \approx 1.25 (\pm 1.51\%)$.

Collapsed across LLM- and lawyer-generated advice, the source unknown group ($M = 6.98$, $SD = 1.21$) gave comparable ratings to the source known group ($M = 7.36$, $SD = 0.96$). Finally, and most interestingly, there was a significant interaction between the advice source and group factors, $F(1, 98) = 4.56$, $p = .04$, $\eta_G^2 = .01$. The Bayes factor provided anecdotal evidence for the alternative hypothesis, $BF_{10} \approx 1.58$ ($\pm 2.47\%$).

Following the significant advice source \times group interaction, paired t -tests were conducted to compare the effect of advice source in each group. In the source unknown group, participants rated their willingness to act on the LLM-generated advice ($M = 7.23$, $SD = 1.17$) significantly more highly than the lawyer-generated advice ($M = 6.74$, $SD = 1.21$), $t(49) = 3.91$, $p < .001$, Cohen's $d_z = 0.55$. Moreover, the Bayes factor provided very strong evidence in favour of the alternative hypothesis, $BF_{10} = 89.70$ ($\pm 0\%$). The source known group, by contrast, showed no significant difference in ratings for advice that was generated by the LLM ($M = 7.42$, $SD = 1.00$) and the lawyers ($M = 7.30$, $SD = 0.93$), $t(49) = 1.09$, $p = .28$, $d_z = 0.15$. The corresponding Bayes factor provided moderate support for the null hypothesis, $BF_{10} \approx 0.27$ ($\pm 0.05\%$).

4 EXPERIMENT 2

With HCI originating, in part, from psychology [9, 13], HCI research would benefit from the open science and research integrity practices that psychology has adopted in recent decades. Dating as far back as 1968, psychologist Jane Loevinger urged the American Psychological Association to place greater emphasis on replicability of experimental research [1]. Around 40 years later, psychology found itself in the midst of a replication crisis. Mirroring Loevinger's thoughts, over the last 15 years the HCI community has also seen an increasing call for replicability through open science practices and a greater emphasis on replications [12, 14, 17, 25, 62].

In Experiment 2, we therefore sought to replicate the key result from Experiment 1. Specifically, we aimed to replicate the finding that participants reported that they were more willing to act on LLM-generated legal advice than advice that had been generated by lawyers. To this end, we re-ran the source unknown condition from Experiment 1 with a new set of participants.

4.1 Method

4.1.1 Participants. In Experiment 1, the effect size for the source manipulation in the source unknown group was $d_z = 0.55$. A power analysis in G*Power [16] indicated that 45 participants would provide 95% power to replicate an effect of this size ($n \approx 45$, $d_z = 0.55$, $1 - \beta = .95$, $\alpha = .05$). We preregistered that we would therefore oversample slightly and recruit 50 participants. However, upon recruiting a random set of 50 participants from Prolific, we found that the gender distribution was highly imbalanced. We therefore recruited the minimum number of additional participants ($n = 29$) required to balance the gender distribution as closely as possible, while taking care to match the number of participants (and gender) within each counterbalancing condition. The recruitment of these additional participants was not preregistered.

After recruiting the additional 29 participants, one participant was excluded because they failed two out of three attention checks

(as per our preregistered exclusion criteria). The final sample consisted of 78 participants (39 female, 38 male, 1 prefer not to say), with 39 participants in each counterbalancing condition. The participants were aged between 18 and 60 years ($M = 37.86$ years, $SD = 11.01$ years) and all reported that they spoke fluent English. The experiment was not advertised to participants who had participated in Experiment 1.

As Experiment 1, progression in Experiment 2 was self-paced and the completion time varied between participants. The median completion time was 14:32 minutes. Participants were compensated with £9/hour.

4.1.2 Experimental Design and Measurements. The experiment followed a repeated-measures design with one independent variable: advice source (LLM vs. lawyer). As in Experiment 1, the dependent variable was participants' rating of their willingness to act on each legal advice shown.

4.1.3 Prompts and Procedure. The prompts and procedure were identical to those of Experiment 1, except that only the source unknown condition from Experiment 1 was included.

4.2 Experiment 2: Results

Participants' mean willingness to act ratings are shown in Panel B of Figure 5. A paired sample t -test demonstrated that participants gave significantly higher ratings to the texts that were generated by an LLM ($M = 7.08$, $SD = 1.17$) than those generated by lawyers ($M = 6.82$, $SD = 0.99$), $t(77) = 3.22$, $p = .002$. The effect size was small-to-medium, $d_z = 0.37$, and the Bayes factor indicated that there was strong evidence for the alternative hypothesis, $BF_{10} = 14.00$. This finding replicates the key result from Experiment 1, thereby demonstrating that it is robust and replicable.

5 EXPERIMENT 3

Previous research (see Section 2.3, e.g., [11, 37]) has demonstrated that participants, at times, tend to be more prone to adhere to algorithmic advice over human generated advice. Our findings from Experiment 1 and 2 align with these findings, showing that participants also report higher willingness to act on LLM-generated legal advice.

In Experiment 1 and 2, participants rated their willingness to act on the advice provided differently depending on whether the source of the advice was known or not. This finding raises the question of whether participants were able to discriminate between the LLM- and lawyer-generated advice when the source was not presented to them. Experiment 3 addressed this question by investigating participants' ability to identify the source of legal advice. Participants were presented with the same 18 legal cases and advice texts from Experiments 1 and 2, and were asked to rate the extent to which they thought each advice text had been generated by an LLM or lawyer.

5.1 Method

5.1.1 Participants. As per our preregistration, we recruited 110 participants (55 male, 53 female, 2 prefer not to say) between 18 and 57 years ($M = 35.20$ years, $SD = 9.59$ years) using Prolific. All participants reported that they spoke fluent English. We applied the

same inclusion/exclusion criteria as in the previous experiments, but additionally excluded all participants who had participated in either Experiment 1 or 2. As in the previous two experiments, participants had to answer two of the three attention check questions correctly in order to be included in the final dataset. We did not exclude any participants due to failed attention check questions.

Consistent with Experiments 1 and 2, progression in Experiment 3 was self-paced and the completion time varied between participants. The median completion time was 12:12 minutes and participants were compensated with £9/hour.

5.1.2 Experimental Design and Measurements. Experiment 3 followed a repeated-measures design with one independent variable: advice source (LLM vs. lawyer). The dependent variable was participants' rating of their confidence in whether the presented advice was generated by an LLM or a lawyer for each advice presented. For each advice text, participants were asked to indicate their confidence in the source by answering 'Please rate to what extent you think the text has been generated by an LLM or a lawyer' on a 6-point scale from 1 ('Definitely LLM generated') to 6 ('Definitely lawyer generated').

5.1.3 Prompts and Procedure. The prompts were identical to those of the first two experiments. However, as Experiment 3 aimed to identify participants' ability to identify the source of the advice when it was unknown, we made slight changes to three LLMs responses: 'Planning advice 2', 'Traffic advice 3', and 'Property advice 4'. Specifically, we removed the three/four first words stating 'As a lawyer,...' or 'As your legal advisor,...'. As in Experiment 2, we only included the source unknown condition. Apart from the change in question and the rating scale, see above, Experiment 3 followed the same procedure as Experiment 1 and 2.

5.2 Experiment 3: Results

As per our preregistration, we used a measure derived from signal detection theory to estimate participants' ability to discriminate between the LLM- and lawyer-generated advice [38]. Specifically, we calculated the area under the receiver operating characteristic (ROC) curve (AUC). Unlike mean ratings [40, 41], and mean rating difference scores [23], ROC analysis allows discrimination to be measured separately from response bias. In the current context, discrimination refers to participants' ability to accurately distinguish between the LLM- and lawyer-generated texts. Response bias, by contrast, refers to participants' overall tendency to rate all texts as LLM- or lawyer-generated. We focus on AUC as a bias-free measure of participants' discrimination performance, although note that the ROC curve also provides a visual representation of response bias.

ROC analysis can be employed wherever participants are tasked with discriminating between two categories using graded discrimination ratings on a multi-point scale. In our discrimination task, participants made such graded ratings by judging the extent to which the advice was generated by an LLM or a lawyer on a scale from 1 ("Definitely LLM generated") to 6 ("Definitely lawyer generated"). Therefore, given the upper boundary of the scale refers to lawyer-generated advice, it is helpful to conceptualise the task as one requiring participants to detect lawyer-generated advice. Under this definition, trials in which lawyer-generated advice is presented

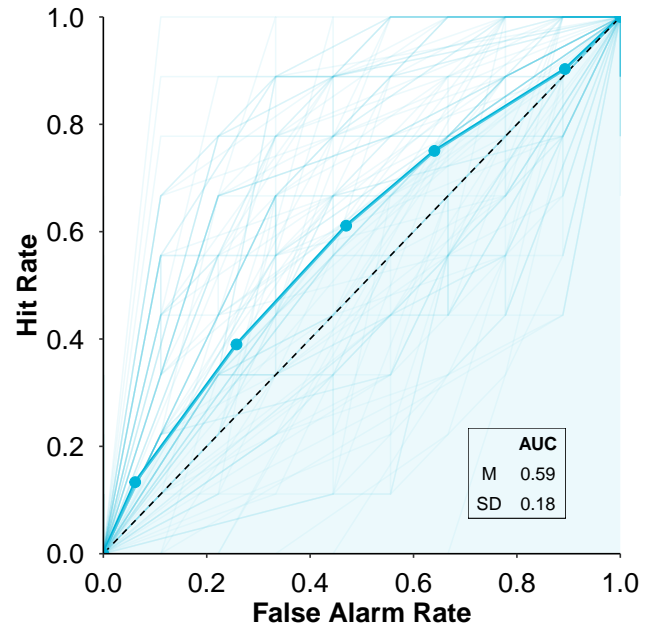


Figure 6: ROC curve for Experiment 3 indicating, through the bowing towards the top-left corner, above chance discrimination of LLM- and lawyer-generated legal advice. The area under the ROC curve (AUC) is highlighted in light blue ($M = .59$, $SD = .18$). Faint lines represent individual participant ROC curves.

can be considered 'signal' trials, whereas trials that contain LLM-generated advice can be considered 'noise' trials. When making judgements, participants are assumed to base their responses on a 'decision variable' that is determined by their subjective evaluation of the amount of signal in the text [52]. They are also assumed to adopt a decision threshold, which is known as a 'criterion'. If the decision variable is sufficiently high that it meets or exceeds the decision threshold, the participant will give a positive response, thereby classifying the trial as a signal trial. Traditionally, signal trials (i.e., lawyer advice trials in the current context) in which a participant correctly concludes that a signal is present are considered 'hits'. Conversely, noise trials (i.e., LLM advice trials) in which a participant incorrectly concludes that a signal is present are considered 'false alarms'.

To measure discrimination, each point on the rating scale is treated as a separate threshold reflecting different LLM/lawyer criteria. For each participant, the proportion of hits and false alarms (i.e., the hit and false alarm rates, HRs and FARs) is then calculated for each point on the rating scale. For example, for scale point 2, the HR and FAR would be defined as the proportion of lawyer- and LLM-generated advice that received a rating of 2–6, respectively. The HR and FAR rate for each scale point can be plotted against each other to calculate an ROC curve, which can be used to visualise both discrimination and response bias.

The ROC curve for Experiment 3, with the mean HRs and FARs plotted against each other, is presented in blue in Figure 6. The

points on the ROC curve reflects the different scale values, with scale point 1 depicted on the [1,1] coordinates. The diagonal, dashed line that runs from the [0,0] to [1,1] coordinates is a reference line that represents chance discrimination performance. The more the ROC curve bows away from the diagonal line, towards the top-left corner of the plot, the better the discrimination. The position of the scale points also provides a visual indication of response bias. Specifically, points clustering towards the bottom-left corner of the plot indicate a conservative response bias, i.e., participants tending to rate the presented advice as LLM-generated, whereas points clustering towards the top-right of the plot indicate a liberal response bias, with participants tending to rate the presented advice as lawyer-generated.

As noted above, we quantified discrimination by estimating the AUC, which we calculated using the trapezoidal rule [22, 47]. AUC values vary between 0 and 1, with .50 representing chance discrimination performance (i.e, the diagonal, dashed line in Figure 6) and 1 representing perfect discrimination. Following our preregistration, we compared participants' AUC values to a theoretical mean of .50 to establish whether participants could discriminate between the lawyer- and LLM-generated advice significantly above chance. This analysis confirmed that participants could discriminate the source of the advice presented significantly above chance ($M = .59$, $SD = .18$). The Bayes factor provided extreme evidence in favour of the alternative hypothesis, $t(109) = 5.51$, $p < .001$, $d = 0.53$, $BF_{10} = 5.10 \times 10^4$.

6 DISCUSSION

In this paper, we have presented three experiments. Experiment 1 investigated lay people's willingness to act on LLM- and lawyer-generated legal advice. The results showed that, when the source of the advice was unknown to participants, they were more willing to act on the LLM-generated advice than the lawyer-generated advice. This result was replicated in Experiment 2. In Experiment 3, we investigated lay people's ability to distinguish between the LLM- and lawyer-generated advice when the source was unknown. Participants discriminated between the advice significantly above chance, but their discrimination performance was far from perfect.

The remainder of the discussion focuses on the importance of the phrasing of the language used and the risks of overtrust in LLM-generated content and strategies to mitigate this. Lastly, we will present limitations while also integrating future research directions at the intersection of LLMs and the legal domain throughout.

6.1 Use of Language for Legal Advice

In Experiments 1 and 2, participants reported a significantly higher willingness to act on the legal advice provided by the LLM than the lawyers when the source was unknown. While the presented work here cannot answer *why?* with certainty—and this is therefore a challenge for future work—we now discuss potential explanations for our results.

6.1.1 Language Complexity and Advice Length. While the LLM- and lawyer-generated advice was broadly comparable in length, we chose not to equate the word counts or textual complexity exactly. Indeed, the average number of words used in the lawyer-generated advice was 170 (min: 107; max: 276), while the average number of

words used by the LLM-generated advice was 124 (min: 93; max: 176). We also did not restrict the complexity of the language that the lawyers or the LLM used. The LLM-generated advice appeared to be more complex, as measured by an average Lix score [8] of 72 (min: 60; max: 79) versus 57 (min: 49; max: 65) for the lawyer-generated advice⁸. These Lix scores correspond to 'Very Difficult' and 'Difficult' to comprehend for the LLM- and the lawyer-generated advice respectively [8, Table 4]. While balancing both of these metrics could have been done, and would have increased the advice similarity between sources, we were concerned that doing so would reduce realism and impose experimenter bias on the advice. Limiting the LLM or the lawyer to an arbitrary language complexity or word count would introduce constraints that would distort the advice that both an LLM and lawyer would otherwise generate. Therefore, while such experimenter-imposed constraints increase experimental control, they also reduce realism and ecological validity.

One possible explanation, is that participants were more willing to act on the LLM-generated advice as they conflated complexity with quality of advice. To test this possibility, future work could investigate people's willingness to act on legal advice, when the complexity is matched between the LLM and the lawyers.

6.1.2 Comparability to Real World Legal Advice. Our study directly compared participants' responses to LLM- and lawyer-generated advice. For practical purposes, we stipulated that the reply to our prompt had to be a short 'summary' of the legal advice to be provided to a potential client. In practice, a lawyer would rarely, if ever, provide such a short one-paragraph summary to a client, nor would an LLM without this added stipulation.

Lawyers do not typically provide one-paragraph answers to a legal problem, nor are they trained to do so in an expeditious manner. While the LLM could provide an answer all-but instantaneously, the lawyers often took days or weeks to provide legal advice following our queries. There are practical reasons for this difference. A lawyer must worry about their professional reputation, their licensing certificate, and the risk of being sued for negligence, when advising clients regarding legal matters. By nature, these precautions do not lend themselves to one-paragraph answers. Instead, lawyers are often prone to hedge their bets, providing both sides of an argument, and cautioning against certainty and finality when providing solutions to legal problems.

LLMs have the opposite problem, as they often appear overly confident when providing advice. Such overconfidence—which might be a consequence of the vocabulary, sentence structure, and syntax used—can lead to clients being exposed to decisively-voiced, hallucinated facts. In part, LLMs are trained on the rules of grammar, to present convincing, structured responses with clear introductions, arguments, conclusions, sign-posting, and other features of good writing. These structural advantages cannot easily be compared to a human writer responding off-the-cuff to a prompt. To attain the same level of writing structure and quality, a human writer may need to edit, or have their work edited, something which was not done in the context of our experiments.

⁸Lix ranking: Very Easy: 20; Easy: 30; Average: 40; Difficult: 50; Very Difficult: 60.

6.2 Causes for Higher Ratings and Overtrust

Our findings of highlight (Experiment 1 and 2) that participants, while being able to distinguish the source above chance (Experiment 3), have a higher willingness to act on the legal advice when generated by an LLM. In this section we discuss possible causes for this, risks that might result of this overtrust, as well as possible strategies addressing the overtrust in LLM-generated legal advice.

6.2.1 Ratings are Influenced by more than just the Source. When examining our findings, a contradictory observation can be made. In Experiment 1 and 2, participants in the source unknown condition, reported a higher willingness to act on LLM-generated advice than on the lawyer-generated advice. On the other hand, Experiment 3 showed that when participants were directly asked to identify the source, they were able to distinguish between LLM- and lawyer-generated advice at an above chance rate. Although these findings were observed in separate experiments, when considered together, they suggest that participants might have been aware of the source in Experiment 1 and 2 even when it was not disclosed to them. Interestingly, while participants seemed to be—at least to some extent—aware of the source, even when it was not provided to them, they behaved differently than in the source known condition of Experiment 1. Specifically, participants increased their willingness to act ratings when explicitly informed that the advice presented had been generated by a lawyer. This pattern may indicate a social acceptability bias [64] in which participants thought that they *should* be more willing to act on lawyer-generated advice. Participants may have given higher ratings to the lawyer-generated advice than the LLM-generated advice in the source known condition, as this aligns with perceived social norms that humans should be trusted over an algorithm.

6.2.2 Risks of Overtrust. Overtrust in LLM-generated advice carries an abundance of risks, some of which are being specifically addressed through policies and regulations such as the European Union Artificial Intelligence Act (EU AI Act). For instance, the EU AI Act - article 50.2⁹ emphasises that “*Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated...*” [15]. While these regulations aim to ensure the safe use of AI-based systems, including LLMs, there has been a growing focus on how to effectively implement them. One strategy directly addressing the above article is the implementation of watermarks [31, 54]. However, while watermarks may allow machines to detect AI-generated content, they may not (necessarily) improve the transparency of AI-generated text for people. Although watermarks enable the generation of “*artificially generated or manipulated text in a machine-readable format...*” [15] they are still “*invisible to humans*” [31].

While machine detectable indicators of AI-generated content are highly valuable to enable automatic detection, improving the general public’s AI literacy will become increasingly important as AI-infused systems become more sophisticated and prevalent. In Experiment 3, participants were significantly above chance when discriminating LLM- from lawyer-generated advice, but there was still clear room for improvement. Within the related field of fake news detection, current research [42] has demonstrated that interventions—in the form of short interactive training session—can significantly improve participants’ discrimination of true and fake news headlines. We suggest that future work draws inspiration from the fake news detection literature, in order to develop similar interventions that improve the general public’s AI literacy, and equips lay people with the skills to discriminate human- from AI-generated content. Such interventions may be particularly useful for user groups who are especially susceptible to trusting LLM-generated content, such as people with high ‘agreeableness’ [61].

Beyond watermarking and human interventions, the EU AI Act contains further provisions (Article 13)¹⁰ for ‘high-risk’ systems, including transparency around the disclosure of risks, limitations of the systems, accuracy level, and impacts on health, safety, and fundamental rights. These ‘high risk’ provisions may also apply to AI systems offering legal services. The obligation on providers of these systems to offer further information on risks and disclaimers could address some of the problems of watermarking. However, the extent to which these safeguards are adopted in AI-based systems varies greatly. For instance, the current disclaimer provided by ChatGPT-4o simply states that “*ChatGPT can make mistakes. Check important info.*” while Google’s Gemini states that “*Gemini may display inaccurate info, including about people, so double-check its responses*”¹¹. In addition to the brevity of the disclaimers, their usefulness depends on users reading, understanding, and acting upon the information provided. Prior work has shown that users rarely read or engage with online terms and conditions containing legal information [50]. Therefore, informing users of potential risks of the use of LLMs may not be sufficient.

6.3 Limitations

To compile the lawyer-generated advice, we recruited three specialists lawyers within the domains of traffic, planning, and property law. These lawyers likely have a personal writing style that may not transfer to other lawyers. Likewise, we used ChatGPT-4o to create the LLM-generated advice. Different results might be achieved with other LLMs. Therefore, future work should test the boundary conditions of our findings with different advice sources.

Furthermore, as our focus was on lay peoples’ perceptions of LLM- and lawyer-generated advice, we did not evaluate the accuracy of the advice. Given that we were interested in participants’ willingness to act on and discriminate the LLM- and the lawyer-generated advice, the accuracy of the advice is inconsequential for this purpose. Nevertheless, future work would benefit from establishing the validity of the advice—particularly that of the LLM—as LLMs are prone to create hallucinations and participants have reported a higher willingness to act on its advice.

⁹Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems

¹⁰Article 13: Transparency and Provision of Information to Deployers

¹¹Both disclaimers were observed on the 20.08.2024 and might change in the future.

7 CONCLUSION

Given the human-like nature of responses generated by LLMs, it becomes increasingly important to understand how lay people use this technology, especially in the context high-risk domains such as the legal context. In this paper, we have presented three experiments (total $N = 288$). Experiment 1 investigated if lay people are willing to act on legal advice—for advice on traffic, planning, and property law—when the source of the advice was either known or unknown. Experiment 2 replicated the key manipulation of Experiment 1 using only the source unknown condition. Experiment 3 investigated if participants, when the source of advice was unknown, are able to discriminate the source.

Findings of Experiment 1, successfully replicated in Experiment 2, show that participants, when the source of legal advice was unknown, report significantly higher willingness to act on the LLM-generated legal advice compared to the lawyer-generated advice. When the source of advice was known, no significant differences could be observed. Experiment 3 demonstrated, that even when participants were unaware of the source, they were able to discriminate the LLM- from the lawyer-generated advice significantly above chance. Lastly, this paper discusses the importance of language used (e.g., complexity and advice length) when providing legal advice, possible causes for the change in ratings—going beyond the source—depending on if the source was known or not, risks associated with overtrust in LLMs and strategies to mitigate it, as well as limitations and future work.

ACKNOWLEDGMENTS

We would like to thank Horia Maior for his valuable comments on the manuscript. This project was supported by the Engineering and Physical Sciences Research Council [grant number EP/V00784X/1] UKRI Trustworthy Autonomous Systems Hub and Responsible AI UK [grant number EP/Y009800/1].

REFERENCES

- [1] Andrew Ahlgren. 1969. A modest proposal for encouraging replication. *American Psychologist* 24, 4 (1969), 471a.
- [2] Allen & Overy LLP. 2023. *A&O announces exclusive launch partnership with Harvey*. <https://www.aoshearman.com/en/news/ao-announces-exclusive-launch-partnership-with-harvey> Accessed: 2024-08-16.
- [3] Omri Asscher and Ella Glikson. 2023. Human evaluations of machine translation in an ethically charged situation. *New Media & Society* 25, 5 (2023), 1087–1107.
- [4] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.
- [5] Thom Baguley. 2012. Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior research methods* 44 (2012), 158–175.
- [6] Margherita Bernabei, Silvia Colabianchi, Andrea Falegnami, and Francesco Costantino. 2023. Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Computers and Education: Artificial Intelligence* 5 (2023), 100172.
- [7] Yochanan E Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34.
- [8] C. H. Björnsson. 1983. Readability of Newspapers in 11 Languages. *Reading Research Quarterly* 18, 4 (1983), 480–497. <http://www.jstor.org/stable/747382>
- [9] JOHN M. CARROLL. 1997. Human-computer interaction: psychology as a science of design. *International Journal of Human-Computer Studies* 46, 4 (1997), 501–522. <https://doi.org/10.1006/ijhc.1996.0101>
- [10] Noah Castelo and Adrian F Ward. 2021. Conservatism predicts aversion to consequential Artificial Intelligence. *PLoS one* 16, 12 (2021), e0261467.
- [11] Nguyen Thi Khanh Chi. 2024. The effect of AI chatbots on pro-environment attitude and willingness to pay for environment protection. *SAGE Open* 14, 1 (2024), 21582440231226001.
- [12] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (jul 2020), 70–79. <https://doi.org/10.1145/3360311>
- [13] Alan Dix. 2017. Human-computer interaction, foundations and new paradigms. *Journal of Visual Languages & Computing* 42 (2017), 122–134. <https://doi.org/10.1016/j.jvlc.2016.04.001>
- [14] Florian Ehtler and Maximilian Häußler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI EA '18). Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3170427.3188395>
- [15] European Union. 2024. Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems. <https://artificialintelligenceact.eu/article/50/>. Accessed: 2024-08-20.
- [16] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
- [17] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Pawel W. Woźniak, and Albrecht Schmidt. 2019. The Role of HCI in Reproducible Science: Understanding, Supporting and Motivating Core Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI EA '19). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3312905>
- [18] United States District Court for the Southern District of New York. 2023. *Mata v. Avianca, Inc.*, No. 1:2022cv01461 - Document 54 (S.D.N.Y. 2023). <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/>. UNITED STATES DISTRICT COURTSOUTHERN DISTRICT OF NEW YORK.
- [19] United States District Court for the Southern District of New York. 2024. *United States v. Michael Cohen*. <https://www.nysd.uscourts.gov/sites/default/files/2024-03/18cr602%20Cohen%20Opinion.pdf> Accessed: 2024-08-16.
- [20] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 838, 21 pages. <https://doi.org/10.1145/3613904.3642139>
- [21] Lady Hale, Lord Kerr, Lord Wilson, Lady Black, and Lord Sales. 2023. *Villiers v. Villiers*.
- [22] Philip A Higham and D Paul Higham. 2019. New improved gamma: Enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. *Behavior Research Methods* 51 (2019), 108–125.
- [23] Philip A. Higham, Ariana Modirrousta-Galian, and Tina Seabrooke. 2024. Mean rating difference scores are poor measures of discernment: The role of response criteria. *Current Opinion in Psychology* 56 (2024), 101785. <https://doi.org/10.1016/j.copsyc.2023.101785>
- [24] Péter Homoki and Zsolt Zódi. 2024. Large language models and their possible uses in law. *Hungarian Journal of Legal Studies* (2024).
- [25] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? on the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (CHI '14). Association for Computing Machinery, New York, NY, USA, 3523–3532. <https://doi.org/10.1145/2556288.2557004>
- [26] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [27] Harold Jeffreys. 1961. *The theory of probability (3rd ed.)*. OUP Oxford.
- [28] Hang Jiang, Xijie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex 'Sandy' Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling. [arXiv:2402.17019](https://arxiv.org/abs/2402.17019) [cs.CL] <https://arxiv.org/abs/2402.17019>
- [29] Sayash Kapoor, Peter Henderson, and Arvind Narayanan. 2024. Promises and pitfalls of artificial intelligence for legal applications. [arXiv:2402.01656](https://arxiv.org/abs/2402.01656) [cs.CY] <https://arxiv.org/abs/2402.01656>
- [30] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (apr 2023), 29 pages. <https://doi.org/10.1145/3579592>
- [31] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 17061–17084. <https://proceedings.mlr.press/v202/kirchenbauer23a.html>

- [32] Joshua Krook, Jennifer Williams, Tina Seabrooke, Eike Schneiders, Jan Blockx, Stuart E Middleton, and Sarvapali Ramchurn. 2023. AI large language models inquiry: TASHub Response. <https://eprints.soton.ac.uk/481740/>
- [33] Kausik Lakkaraju, Sara E Jones, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath C Muppasani, and Biplav Srivastava. 2023. LLMs for Financial Advice: A Fairness and Efficacy Study in Personal Decision Making. In *Proceedings of the Fourth ACM International Conference on AI in Finance* (Brooklyn, NY, USA) (ICAIF '23). Association for Computing Machinery, New York, NY, USA, 100–107. <https://doi.org/10.1145/3604237.3626867>
- [34] Michael D. Lee and Eric-Jan Wagenmakers. 2014. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.
- [35] Weng Marc Lim, Asanka Gunasekara, Jessica Leigh Pallant, Jason Ian Pallant, and Ekaterina Pechenkina. 2023. Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The international journal of management education* 21, 2 (2023), 100790.
- [36] Mengqi Liu and Faten M'hiri. 2024. Beyond Traditional Teaching: Large Language Models as Simulated Teaching Assistants in Computer Science. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 743–749.
- [37] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
- [38] Neil A Macmillan. 2002. Signal detection theory. *Stevens' handbook of experimental psychology: Methodology in experimental psychology* 4 (2002), 43–90.
- [39] Oliver Miles, Robert West, and Tom Nadarzynski. 2021. Health chatbots acceptability moderated by perceived stigma and severity: a cross-sectional survey. *Digital health* 7 (2021), 20552076211063012.
- [40] Ariana Modirrousta-Galian and Philip A Higham. 2023. Gamified inoculation interventions do not improve discrimination between true and fake news: Re-analyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General* 152, 9 (2023), 2411.
- [41] Ariana Modirrousta-Galian, Philip A Higham, and Tina Seabrooke. 2023. Effects of inductive learning and gamification on news veracity discernment. *Journal of Experimental Psychology: Applied* 29, 3 (2023), 599.
- [42] A. Modirrousta-Galian, T. Seabrooke, Y. Hanooh, N. J. Kelley, and P. A. Higham. 2024. An Inductive Learning Intervention to Improve News Veracity Discernment. <https://doi.org/10.31234/osf.io/6j7fr>.
- [43] Richard D. Morey and Jeffrey N. Rouder. 2023. BayesFactor: Computation of bayes factors for common designs. R package version 0.9.12-4.7. <https://CRAN.R-project.org/package=BayesFactor>
- [44] Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health* 5 (2019), 2055207619871808.
- [45] John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence. *arXiv preprint arXiv:2306.07075* (2023).
- [46] Oscar Oviedo-Trespalacios, Amy E Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, JE Rod, Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, et al. 2023. The risks of using ChatGPT to obtain common safety-related information and advice. *Safety science* 167 (2023), 106244.
- [47] Irwin Pollack and Robert Hsieh. 1969. Sampling variability of the area under the ROC-curve and of d'e. *Psychological Bulletin* 71, 3 (1969), 161.
- [48] R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [49] Moritz Reis, Florian Reis, and Wilfried Kunde. 2024. Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine* (2024), 1–3.
- [50] Eric P. Robinson and Yicheng Zhu. 2020. Beyond "I Agree": Users' Understanding of Web Site Terms of Service. *Social Media + Society* 6, 1 (2020), 2056305119897321. <https://doi.org/10.1177/2056305119897321>
- [51] Tina Seabrooke, Eike Schneiders, Liz Dowthwaite, Joshua Krook, Natalie Leesakul, Jeremie Clos, Horia Maior, and Joel Fischer. 2024. A Survey of Lay People's Willingness to Generate Legal Advice using Large Language Models (LLMs). In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems* (Austin, TX, USA) (TAS '24). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3686038.3686043>
- [52] Harold Stanislaw and Natasha Todorov. 1999. Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers* 31, 1 (1999), 137–149.
- [53] Ian Steenstra, Prasanth Murali, Rebecca B Perkins, Natalie Joseph, Michael K Paasche-Orlow, and Timothy Bickmore. 2024. Engaging and Entertaining Adolescents in Health Education Using LLM-Generated Fantasy Narrative Games and Virtual Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [54] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The Science of Detecting LLM-Generated Text. *Commun. ACM* 67, 4 (mar 2024), 50–59. <https://doi.org/10.1145/3624725>
- [55] Omar Tayan, Ali Hassan, Khaled Khankan, and Sanaa Askool. 2023. Considerations for adapting higher education technology courses for AI large language models: A critical review of the impact of ChatGPT. *Machine Learning with Applications* (2023), 100513.
- [56] First tier Tribunal Tax Chamber. 2023. Harber v. The Commissioners for his Majesty's Revenue and Customs. <https://www.bailii.org/uk/cases/UKFTT/TC/2023/TC09010.html>. FIRST-TIER TRIBUNAL TAX CHAMBER.
- [57] T Franklin Waddell. 2018. A robot wrote this? How perceived machine authorship affects news credibility. *Digital journalism* 6, 2 (2018), 236–255.
- [58] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 84 (apr 2024), 26 pages. <https://doi.org/10.1145/3637361>
- [59] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 303, 21 pages. <https://doi.org/10.1145/3613904.3641960>
- [60] Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Soellner. 2024. LegalWriter: An Intelligent Writing Support System for Structured and Persuasive Legal Case Writing for Novice Law Students. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1052, 23 pages. <https://doi.org/10.1145/3613904.3642743>
- [61] Joel Wester, Sander De Jong, Henning Pohl, and Niels Van Berkel. 2024. Exploring People's Perceptions of LLM-generated Advice. *Computers in Human Behavior: Artificial Humans* 2 (2024), 100072. Issue 2.
- [62] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. 2012. RepliCHI SIG: from a panel to a new submission venue for replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) (CHI EA '12). Association for Computing Machinery, New York, NY, USA, 1185–1188. <https://doi.org/10.1145/2212776.2212419>
- [63] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 73 (may 2024), 35 pages. <https://doi.org/10.1145/3659625>
- [64] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 156, 26 pages. <https://doi.org/10.1145/3613904.3642385>
- [65] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19632–19642.
- [66] Feiyu Zhu and Reid Simmons. 2024. Bootstrapping cognitive agents with a large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, 1101 Pennsylvania Ave, NW, Suite 300, Washington, DC, 655–663. <https://doi.org/10.1609/aaai.v38i1.27822>

A WHAT IS A LARGE LANGUAGE MODEL (LLM)

The description presented to participants in all three experiments prior to the first case:

'Large language models (LLMs) are computer programs that can read and generate human-like text by learning from vast amounts of written language. They can answer questions, write essays, or even create poetry, mimicking the style and content of the texts they were trained on. They can produce coherent and contextually relevant content, but their understanding is derived purely from the data they are trained on.' [51]