Journal of Psychology and AI

ISSN: 2997-4100 (Online) Journal homepage: www.tandfonline.com/journals/tpai20

Evaluating firearm examiner testimony using large language models: a comparison of standard and knowledge-enhanced AI systems

Francesco Pompedda, Pekka Santtila, Eleonora Di Maso, Thomas J. Nyman, Yongjie Sun & Angelo Zappala

To cite this article: Francesco Pompedda, Pekka Santtila, Eleonora Di Maso, Thomas J. Nyman, Yongjie Sun & Angelo Zappala (2025) Evaluating firearm examiner testimony using large language models: a comparison of standard and knowledge-enhanced AI systems, Journal of Psychology and AI, 1:1, 2503343, DOI: 10.1080/29974100.2025.2503343

To link to this article: https://doi.org/10.1080/29974100.2025.2503343

JOURNAL OF PSYCHOLOGY AND AI
2025, VOL. 1, NO. 1, 2503343
https://doi.org/10.1080/29974100.2025.2503343

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Evaluating firearm examiner testimony using large language models: a comparison of standard and knowledge-enhanced AI systems

Francesco Pompedda[a*], Pekka Santtila[b,c*], Eleonora Di Maso[d], Thomas J. Nyman[e], Yongjie Sun[b,f] and Angelo Zappala[d,g]

[a]INVEST Research Flagship Center, INVESThub, University of Turku, Turku, Finland; [b]Faculty of Arts and Sciences, NYU Shanghai, Shanghai, China; [c]Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, New York University Shanghai, Shanghai, China; [d]Faculty of Arts, Psychology and Theology, Åbo Akademi University, Turku, Finland; [e]School of Psychology and Clinical Language Sciences, University of Reading, Berkshire, UK; [f]School of Psychology and Cognitive Science, East China Normal University, Shanghai, China; [g]Psychology, IUSTO - Salesian University Institute Torino Rebaudengo (CRIMELAB), Torino, Italy

ABSTRACT

This study evaluated the decision-making of Large Language Models (LLMs) in interpreting firearm examiner testimony by comparing a standard LLM to one enhanced with forensic science knowledge. The present study is a replication study. We assessed whether LLMs mirrored human decision patterns and if specialised knowledge led to more critical evaluations of forensic claims. We employed a $2 \times 2 \times 7$ between-subjects design with three independent variables: LLM configuration (standard vs. knowledge-enhanced), cross-examination presence (yes vs. no), and conclusion language (seven variations). Each model condition performed 200 repetitions per scenario. This yielded a total of 5,600 measures of binary verdicts, guilt probability ratings, and credibility assessments. LLMs showed low conviction rates (9.4%) across conditions, with logical variations as a function of the way in which the firearm expert's conclusion was formulated. Cross-examination produced lower guilt assessments and scientific credibility ratings. Importantly, knowledge-enhanced LLMs demonstrated significantly more conservative evaluations of firearm evidence across all match conditions compared to standard LLMs. LLMs, particularly when enhanced with domain-specific knowledge, showed advantages in evaluating complex scientific evidence compared to human jurors in Garrett et al. (2020), suggesting potential applications for AI systems in supporting legal decision-making.

ARTICLE HISTORY Received 24 December 2024; Accepted 1 May 2025

KEYWORDS Firearm examination evidence; Large Language Models (LLMs); knowledge-enhanced artificial intelligence (AI) systems; legal decision-making; expert testimony evaluation

CONTACT Pekka Santtila pekka.santtila@nyu.edu

*Shared first authors with equal contributions, order of the two authors determined alphabetically.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/29974100.2025.2503343

1. Introduction

1.1. Scientific validity and firearm analysis

The scientific validity of forensic evidence and how it is evaluated in legal proceedings represent a critical challenge for the criminal justice system. This is particularly true in relation to forensic feature-comparison methods, where experts compare characteristics from evidence found at a crime scene against an item from the suspect (e.g. a suspect's gun or fingerprint) to determine if they have the same source. Despite mounting scientific evidence questioning the validity of some of these methods, courts and juries continue to rely heavily on forensic testimony, creating a disconnect between scientific understanding and legal practice (Curley et al., 2022). In fact, forensic science evidence significantly influences jury decisions and can prove decisive in determining case outcomes (Lieberman et al., 2008). This influence persists even when the underlying methods lack scientific validity.

Firearm analysis is a case in point. In 2021, approximately 20,958 of the 26,031 murders (about 80%) in the U.S. involved firearms (Gramlich, 2025). In 2023, 9% of violent victimisations in the U.S. involving a firearm amounted to over half a million firearm-related cases (2024). This means that firearm evidence features prominently in criminal proceedings. The conventional practice involves experts visually comparing toolmarks on bullets or cartridge casings to determine whether they were fired from a specific weapon. For over a century, firearm experts have testified in U.S. courts that weapons leave distinctive toolmarks on ammunition, allowing them to visually examine and conclusively match bullets or cartridge casings to specific firearms (Lanigan, 2012). This practice has been formalised through professional standards such as the Association of Firearms and Tool Mark Examiners (AFTE) theory of identification. This framework instructs practitioners to declare "identifications" based on "sufficient agreement" between compared samples, relying primarily on the examiner's training and experience (AFTE Glossary, 1998). However, this approach has faced increasing scrutiny from the scientific community, particularly regarding the potentially subjective nature of the conclusions (allowing impact of confirmation bias) and lack of empirical validation.

In fact, a number of scientific bodies have identified fundamental problems with firearm analysis methodology. The National Academy of Sciences' 2008 report on ballistics concluded that definitive associations were not scientifically supported (National Research Council, 2008). A subsequent review emphasised that categorical conclusions regarding firearms or toolmarks lacked adequate validity, highlighting that examiners make "subjective decisions based on unarticulated standards and no statistical foundation for estimation of error rates" (National Research Council, 2009). Most notably, the President's Council of Advisors on Science and

Technology (the President's Council of Advisors on Science and Technology, 2016) reviewed existing studies of firearm examiner performance and concluded that firearm comparison methods are not "foundationally valid". These scientific critiques have prompted various stakeholders to modify how firearm evidence is presented in court.

1.2. Source identification language and cross-examination

Garrett et al. (2020) analysed recent legal responses to scientific criticisms of forensic firearm evidence. They explain how some judges have begun restricting expert testimony, such as requiring experts to state only that it is "More likely than not" that the ammunition came from the defendant's firearm (Center for Statistics and Applications in Forensic Evidence, 2008). Garrett et al. (2020) also identified other rulings that have mandated conclusions to be expressed within a "reasonable degree of ballistic certainty" (Center for Statistics and Applications in Forensic Evidence, 2006, 2007) or even limiting experts to saying they "cannot exclude" a firearm as the source of the bullet found at the crime scene (Center for Statistics and Applications in Forensic Evidence, 2019). Finally, the Department of Justice has responded to the situation with new guidelines prohibiting categorical source conclusions in federal cases, instead requiring so-called "source identification" language.

However, these linguistic modifications may not be effective in addressing the problems. The study by Garrett et al. (2020) examined how the wording utilised by a firearm expert in formulating their conclusion impacted mock jurors' evaluation of firearm evidence. Through two experiments, they demonstrated that any language indicating a match significantly increased guilty verdicts compared to "Inconclusive" findings, regardless of how the match was phrased. Only the most limited "Cannot be excluded" language produced significantly lower guilty verdicts. These findings suggest that current approaches to reforming forensic testimony may be insufficient to ensure appropriate evaluation of this type of evidence.

However, Gutierrez et al. (2024) found that defence expert testimony challenging prosecution conclusions and highlighting scientific limitations reduced guilty verdicts by undermining perceptions of case strength and ballistic analysis reliability in mock jurors. Other researchers have studied the effects of cross-examination on jurors' evaluation of expert evidence more generally. Lieberman et al. (2008) conducted a study where undergraduate students read a trial description that included DNA evidence from an independent laboratory with proficiency testing or from an unaccredited police department laboratory. Cross-examination focused on either the expert's credibility or potential issues with DNA evidence, showing that when the lab was deemed reliable, participants had more confidence in

guilt with evidence-focused cross-examination than with expert-focused cross-examination; the opposite occurred with unreliable labs. Similarly, Salerno and McCauley (2009) found that cross-examinations increased culpability ratings and decreased defence expert credibility, affecting jurors' judgements by helping jurors recognise a flawed expert witness. Finally, Austin and Kovera (2015) found that scientifically informed cross-examinations can promote critical evaluations of scientific evidence, particularly in recognising threats to internal validity, such as research lacking a control group.

1.3. Large language models (LLMs) in forensic science

Recent advances in artificial intelligence, particularly Large Language Models (LLMs), offer novel opportunities to examine these challenges in forensic evidence evaluations. LLMs have demonstrated sophisticated capabilities in complex reasoning tasks, natural language understanding, and systematic information processing (Chang et al., 2024; Webb et al., 2023). Unlike human decision-makers, LLMs can process large volumes of information systematically without fatigue Eigner & Händler, (2024). Their ability to integrate multiple sources of information while maintaining logical consistency makes them potentially suitable for evaluating scientific evidence in legal contexts.

However, LLMs trained on general text data may simply reproduce common assumptions about forensic science rather than applying appropriate scientific scepticism. Research has shown that these models can absorb and replicate prevalent beliefs about the reliability and validity of various methods, including scientific and technical fields. In Hofmann et al. (2024), medical doctors evaluated LLM-generated suggestions regarding interventional radiology. The results showed that LLMs did not always incorporate critical aspects such as the risks and alternatives of the procedure in their suggestion demonstrating lack of depth of knowledge. Therefore, without specialised knowledge, LLMs might default to the same expectations about firearm analysis that have been observed in human jurors – namely, high confidence in the scientific validity of the method. In addition, as a natural language processing tool, LLMs are highly sensitive to subtle changes in language. Huda et al. (2024) tested the extent to which GPT-3.5 and LLMs such as Gemini perceive subtle emotional cues (misspellings, tone of voice, word choice, sentence length), and found that even subtle and contradictory linguistic cues can significantly affect the response of LLM. Yan et al. (2025) provided LLMs with the same medical statement, varying gender, symptoms, and outcomes, and asked LLMs to make a medical diagnosis. The results showed that LLMs reacted to these subtle language changes, as reflected in the significant impact of the variations on the diagnosis results.

One promising direction for improving LLM performance involves enhancing these systems with specialised expert knowledge. Unlike traditional approaches that rely solely on general training data, knowledge-enhanced LLMs incorporate specific scientific (or other) understanding into their decision-making processes (Wang et al., 2024; Yang et al., 2024). This approach has shown promise in other domains requiring specialised expertise. Li et al. (2023) trained a medical LLM using expert knowledge. The results showed that the professionally trained LLM significantly outperformed the larger general LLM in the medical field, despite a smaller model size. In the context of forensic evidence evaluation, providing LLMs with authoritative scientific knowledge about the limitations and proper interpretation of forensic methods could also potentially lead to more informed and appropriate evidence assessment.

The relationship between specialised knowledge and evidence evaluation is particularly relevant for forensic science. While human jurors typically rely on general assumptions about scientific validity, knowledge-enhanced LLMs could potentially identify specific methodological limitations while evaluating case evidence. This might produce more accurate assessments of firearm evidence that would consider known error rates and empirical validity studies (or lack of them). However, while studies have examined how human jurors evaluate forensic evidence (McQuiston-Surrett & Saks, 2009; Thompson et al., 2013) and how LLMs perform in some legal reasoning tasks (Chalkidis et al., 2022), the potential of LLMs – particularly knowledge-enhanced versions – to evaluate forensic evidence remains unexplored.

1.4. Aims and hypotheses

We examined how Large Language Models (LLMs) evaluated firearm examiner testimony and whether their judgements were in line with human decision-making. By replicating the experimental paradigm of Garrett et al. (2020) using LLMs as participants, we sought to understand whether LLMs exhibit similar patterns of evaluation and potential biases as human jurors. Additionally, we investigated whether providing LLMs with specialised forensic science knowledge would affect their decision-making processes. We tested four main pre-registered hypotheses:

Hypothesis 1. Impact of Identification (i.e. Match) Conclusions: We hypothesised that, similar to the human participants in Garrett et al. (2020), identification (i.e. match) conclusions would lead to higher guilt estimates compared to more cautious language. Specifically, we predicted that both binary guilty verdicts and probability of guilt ratings would be higher in all other conditions compared to the "Cannot be excluded" and "Inconclusive" conditions.

Hypothesis 2. Influence of Cautious Language: We hypothesised that LLMs, different from the human participants in Garrett et al. (2020), would be sensitive to variations in the conclusion language used by the experts, with more cautious or qualified statements leading to lower guilt assessments. Specifically, we expected fewer guilty verdicts and lower probability of guilt ratings in all identification (i.e. match) conditions with cautious or qualified statements compared to the simple identification statement. This prediction contrasts with previous findings in human mock jurors, where such linguistic variations often had minimal impact.

Hypothesis 3. Effect of Cross-Examination: We also predicted that the presence of cross-examination, different from the human participants in Garrett et al. (2020), would reduce both guilty verdicts and probability of guilt ratings made by LLMs across all conditions with the exception of the "Inconclusive" condition. This hypothesis was based on the assumption that LLMs would integrate critical information about the limitations of the firearm analysis presented during cross-examination into their decision-making process.

Hypothesis 4. Role of Forensic Knowledge: We predicted that knowledge-enhanced LLMs equipped with specialised forensic science information (through exposure to the President's Council of Advisors on Science and Technology (2016) report, hereafter the PCAST (2016 report) would assign lower guilt estimates in all identification (i.e. match) conditions. This hypothesis reflected our expectation that additional domain knowledge would lead to more critical evaluation of forensic evidence claims.

For each hypothesis, we expected similar patterns to emerge across both our primary dependent measures: binary guilty/not guilty verdicts and continuous probability of guilt ratings (0–100%). We also anticipated that these effects would be reflected in the ratings of scientific credibility made by the LLMs.

2. Method

This study did not involve human subjects, so ethical review and approval were not required.

2.1. Pre-registration

The present study was pre-registered on 27 October 2024 (AsPredicted #196113). No data had been collected at the time of pre-registration. Any departures from the pre-registration have been noted below.

2.2. Participants

We employed two configurations of the GPT-4o language model: standard GPT-4o and GPT-4o with specialised expert knowledge. Each configuration provided 200 responses per experimental condition, yielding a total of 5,600 total responses. This sample size was selected to match the power specifications of the original study (Garrett et al., 2020), providing sufficient power (>.90) to detect small effects (d = 0.35) at $\alpha = .05$.

The base LLM configuration comprised GPT-4o with its standard training, accessed through the OpenAI API. This configuration received only the case materials and experimental manipulations without additional specialised knowledge about forensic science. The knowledge-enhanced configuration was identical to the standard configuration but additionally received the contents of the PCAST (2016) report on forensic science prior to case material presentation to consider during decision-making. This content specifically addressed the validity of feature-comparison methods in forensic science. The specialised expert knowledge formed part of the initial prompt for each query in the knowledge enhanced condition.

To ensure independence of responses, each query represented a distinct interaction with the model, with no information retained between queries. The implementation involved initiating each query as a new conversation, maintaining consistent prompt formatting across all conditions, and including all relevant experimental materials (case synopsis and cross-examination when applicable) within each individual query. All dependent measures (verdict, likelihood rating and scientific credibility ratings) were collected within the same query. The models provided responses in a standardised format: a binary verdict (guilty/not guilty), a likelihood rating (0–100%) and six scientific credibility ratings on 1–7 Likert scales. Responses that did not conform to the required format were discarded and the query repeated.

Given the potential for model updates and variations across implementations, we specify that this study used the gpt-4o-6 August 2024 version, which supported structured outputs for the first time, accessed between [10/27/2024–11/01/2024] through the OpenAI API.

2.3. Design

The study employed a 2 × 2 × 7 between-subjects factorial design examining how Large Language Models (LLMs) evaluate firearm examiner testimony. The independent variables were (1) LLM Configuration (standard LLM vs. knowledge enhanced LLM, where the latter received specialised forensic

science knowledge); (2) Cross-examination (present vs. absent), manipulated through the inclusion of challenging expert testimony; and (3) Conclusion Language (seven variations).

The conclusion language conditions were identical to those employed by Garrett et al. (2020). There were two *non-match* conditions:

(1) Inconclusive: Expert testified that "there was not sufficient agreement among the characteristics to determine that the toolmarks have been produced by the same tool"
(2) Cannot be excluded: Expert testified that "the defendant's gun Cannot be excluded as the gun that fired the bullet recovered at the crime scene," explaining that "we observe some agreement of a combination of individual characteristics suggesting that the toolmarks have been produced by the same tool"

There were five *match* conditions:

(3) Simple identification: Expert stated "I identified the crime scene bullet as having been fired by the defendant's gun"
(4) Ballistic certainty: Expert provided a simple identification plus the statement that this "means that the bullet recovered from the crime scene was identified, to a reasonable degree of ballistic certainty, as having come from the defendant's firearm"
(5) More likely than not: Expert provided a simple identification plus the statement that this "means that it is More likely than not that the bullet recovered from the crime scene came from the defendant's firearm"
(6) Complete agreement: Expert testified that "I concluded that the crime scene bullet has markings consistent with being fired by the defendant's gun," adding that "the class characteristics were in complete agreement"
(7) Department of Justice (DOJ) standard: Expert provided a simple identification plus the statement that this "means that the bullet recovered at the crime scene and a bullet fired through the defendant's gun originated from the same source. That all the class characteristics are in agreement and I would not expect to find that same combination of agreement in another source"

2.4. Materials

We used case materials developed by Garrett et al. (2020), which presented a criminal case involving a defendant charged with discharging a firearm during an attempted convenience store robbery. The case synopsis described

how an unidentified perpetrator wearing a mask fired a gun into the floor during an unsuccessful robbery attempt, with no injuries occurring. The case materials detailed how, 2 days later, police conducted a routine traffic stop of the defendant and confiscated a 9-mm handgun, which was subsequently compared to a 9-mm bullet recovered from the crime scene. The materials did not include other trial elements such as opening statements, closing arguments or additional witness testimony.

The firearm expert's testimony was presented in a structured question-and-answer format. The testimony began with the expert's qualifications, including 12 years at the state crime lab and prior experience at the FBI as a firearm and toolmark examiner. The expert then described their methodology using the ACE-V approach (Analysis, Comparison, Evaluation, and Verification), explaining how they examine class characteristics (such as calibre, number of lands and grooves, and twist direction) and individual characteristics using comparison microscopy. The expert described examining a 9 mm semiautomatic pistol and comparing test-fired bullets to the crime scene bullet. While the methodological description remained constant across conditions, the expert's conclusion varied according to the seven conditions described in the design section.

In conditions where cross-examination was present, we included a cross-examination transcript where the defence attorney systematically challenged the expert's testimony. The questioning highlighted: (1) the subjective nature of interpretation in firearm analysis, (2) the existence of individual differences between the crime scene bullet and test-fired bullets that the expert deemed "not meaningful," (3) the lack of formal rules for what constitutes a match, relying instead on personal judgement, (4) the expert's claims of never making errors and having a 0% error rate, and (5) the expert's lack of awareness of the PCAST (2016) report questioning the scientific validity of firearm analysis and mainstream scientists' views of the field. The cross-examination demonstrated that the expert relied heavily on personal judgement and experience rather than objective criteria.

For the knowledge enhanced LLM condition, we provided content from the PCAST (2016) report focusing on forensic feature-comparison methods, particularly the sections on firearm analysis. This specialised knowledge included : (1) the requirement for empirical testing to establish scientific validity, (2) the need for black-box studies to assess accuracy for subjective methods, (3) critiques of traditional firearm analysis as lacking scientific validation, (4) findings that only one appropriately designed black-box study had been conducted showing a false-positive rate that could be as high as 1 in 46 cases, and (5) recommendations that testimony about firearm evidence should clearly report error rates and not imply greater certainty than demonstrated by empirical evidence. This additional content was

presented to the model before the case materials in the knowledge enhanced LLM condition, while being entirely absent in the standard LLM condition.

Following Garrett et al. (2020), we employed three primary dependent measures. The first was a dichotomous verdict decision, where the LLM was required to provide either a "guilty" or "not guilty" response after receiving standard reasonable doubt instructions. The second measure was a likelihood rating, where the LLM estimated on a continuous scale from 0% to 100% the probability that the defendant fired the weapon. The third measure assessed the scientific credibility of the evidence through six items rated on 7-point Likert scales. These items evaluated the reliability of the specific firearm evidence presented, the general reliability of firearm analysis, the scientific basis of the methodology, confidence in the expert's conclusions, potential for error in the analysis, and the overall validity of the technique. The six items were averaged to create a composite scientific credibility score. Before this they were reverse coded so that higher values indicate higher reliability, and/or credibility, and/or validity. Reliability analysis showed Cronbach's $\alpha = .84$. The item *How often do firearm examiners make mistakes when determining whether bullets were fired through the same gun?* performed poorly (correlations with other items ranged between .046 and .206). Deleting this item would have improved Cronbach's α to .88. However, given that the improvement was not substantial and in order to preserve comparability with Garrett et al. (2020), we decided to compute a composite score with all six items, dividing their sum by six as in the original article.

2.5. Procedure

Each experimental trial began with a standardised query to the GPT-4o model using the OpenAI API (gpt-4o-6 August 2024) with OpenAI's default token limits: 128K tokens context window and 16K tokens output limit and temperature setting of 1. In the standard LLM condition, the model received the prompt: "You are a juror in a legal case that is attached. Please read the case and answer the questions at the end of it." In the knowledge-enhanced LLM condition, the model received both the PCAST (2016) report content and the prompt: "You are an expert juror in a legal case that is attached. Your expertise is based on the also attached report on forensic science. Please, read the case and answer the questions at the end of it."

The case materials were then presented in a fixed sequence, beginning with the case synopsis, followed by the expert testimony containing one of the seven conclusion language variations. In conditions where cross-examination was present, this was provided immediately after the expert testimony. Following the case materials, each model was asked to provide responses to the dependent variables.

To ensure response independence, each of the 200 repetitions per condition constituted a completely new query, with no information carried over between trials. Each API call was implemented using Python 3.9, with responses stored in JSON format containing timestamp, condition identifiers, and all dependent measures. The API implementation included error handling for rate limits and automatic retry logic for failed requests, with a 1-s delay between individual responses. Under LLM condition, it automatically pauses for 60 s after every 18 responses, while under knowledge enhanced LLM condition, this pause occurs after every 4 responses, to comply with API rate limitations per minute.

In cases where a model provided incomplete or improperly formatted responses, the query was discarded and repeated with the same materials until a valid response was obtained.

2.6. Statistical analyses

The analyses in this study followed the pre-registration in AsPredicted (#196113) unless otherwise specified.

First, we examined the relationship between scientific credibility ratings and verdict decisions using point-biserial correlations, and between credibility ratings and likelihood estimates using Pearson correlations.

Next, we conducted a series of analyses to examine the effects of our independent variables on verdict decisions and likelihood ratings. For the binary guilty/not guilty decisions, we performed logistic regression analyses examining the effects of LLM type (standard vs. knowledge-enhanced), conclusion language (seven levels), and cross-examination (present vs. absent). After looking at the descriptive statistics (see Tables 5 and 6) and running the logistic regressions, we noticed that in some instances we were dealing with quasi-perfect separation which resulted in unreliable regression coefficients. This was due to low or zero frequencies of guilty verdicts in some conditions. This was evident in particular when looking at "Inconclusive" and "Cannot be excluded" conclusion categories. Given that the regression models proved unstable, decreasing the validity of the inference, we decided to modify our analytical approach. We moved these above analyses to supplementary materials (see S1) and ran Fisher's Exact Tests and chi-square analyses to illustrate the associations in the main text. We also tested penalised methods (Firth); however, they showed convergence errors and we abandoned this approach.

For the same reason, for the continuous likelihood ratings (0–100%), we also opted for more simple one-way ANOVAs and *t*-tests (instead of the pre-registered 3-way ANOVAs) investigating the effects of LLM type, conclusion language, and cross-examination. We used Bonferroni-corrected post-hoc comparisons to examine specific differences between conclusion language

conditions. To analyse scientific credibility ratings, we performed identical one-way ANOVAs and *t*-tests examining the same factors' effects on perceived credibility scores. As with the likelihood ratings, we employed Bonferroni corrections for multiple comparisons.

Effect sizes were reported using partial eta-squared (η_p^2) or eta-squared (η^2) for ANOVAs. Given the multiple analyses conducted on related dependent variables, we employed a Bonferroni-corrected alpha level of .0125 (.05/4) to maintain a familywise error rate of .05 across our four primary dependent variables (verdict decisions, likelihood ratings, scientific credibility, and LLM type comparisons). For multiple comparisons, Bonferroni corrections were also employed.

When the assumption of homogeneity of variances was violated, we reported the corrected *df* values where necessary and used Welch's *F* and Games-Howell correction for the post hoc comparisons. We did not transform the data nor go for non-parametric options and kept the outliers as they were real data and because we had pre-registered the analyses without specifying such operations.

2.7. LLM use statement

Claude 3.5 Sonnet was used for language editing and for preliminary ideas for structuring the introduction and discussion sections. The final text in all sections has been extensively edited and added to by the authors.

2.8. Data availability

To facilitate replication, all materials, prompts and response formats are documented and available at [https://osf.io/byjtn/?view_only=bc8c5112895f401f87dea21eb3627ec3].

3. Results

3.1. Descriptive statistics

In total, 9.4% ($n = 575$) of LLM decisions were to convict the defendant compared to 50.4% ($n = 715$) in Garrett et al. (2020). The proportions of guilty verdicts and likelihood ratings of whether the defendant fired the gun in the different conditions in the present study and in Garrett et al. (2020) are reported in Table 1 (Figure 1 illustrates the results for the present study only). Clearly, the proportions of guilty verdicts were much lower in the present study. Interestingly, visual inspection of the frequencies suggested more variation as a function of the experimental conditions in the present study. In particular, when the report was "Inconclusive" the LLMs never chose to convict compared to 21% of

Table 1. Proportions of guilty verdicts and mean likelihood ratings that the defendant fired the gun in each experimental condition in the present study and in Garrett et al. (2020).

LLM				Humans in Garrett et al. (2020)
	Proportion Guilty Verdicts			Proportion Guilty Verdicts
	Proportion	*95% CI Upper*	*95% CI Lower*	*Proportion*
Inconclusive	.000	.000	.000	.210
Cannot be excluded	.003	.006	−.000	.390
Simple identification	.122	.145	.100	.610
Ballistic certainty	.186	.213	.159	.550
More likely than not	.096	.117	.076	.600
Complete agreement	.059	.075	.042	.550
DOJ	.190	.217	.163	.620
	Likelihood the Defendant Fired the Gun			Likelihood the Defendant Fired the Gun
	M	*95% CI Upper*	*95% CI Lower*	*M*
Inconclusive	.395	.403	.387	.377
Cannot be excluded	.512	.519	.506	.588
Simple identification	.622	.632	.611	.691
Ballistic certainty	.615	.627	.603	.679
More likely than not	.614	.624	.604	.695
Complete agreement	.587	.596	.577	.655
DOJ	.625	.637	.613	.693

In Garrett et al. (2020), *CI*s and *SD*s were not reported. Three decimal places used given close to floor effects for the proportions of guilty verdicts by LLM participants.

the human participants doing so in Garrett et al. (2020). As in Garrett et al. (2020), the highest proportion of guilty verdict was observed in the DOJ condition. In contrast, the continuous assessments of the likelihood that the defendant fired the gun were much more similar in the two studies.

3.1.1. Correlations between items measuring scientific reliability, the likelihood the defendant fired the gun, and dichotomous guilty decisions

Subsequently, we examined the relationship between scientific credibility ratings and verdict decisions using point-biserial correlations, and between credibility ratings and likelihood estimates using Pearson correlations (see Table 2). All six scientific reliability items as well as their total score were correlated with the likelihood that the defendant fired the gun and with the dichotomous guilty decision. These findings suggest that the LLMs were responding to the different questions in a coherent manner.

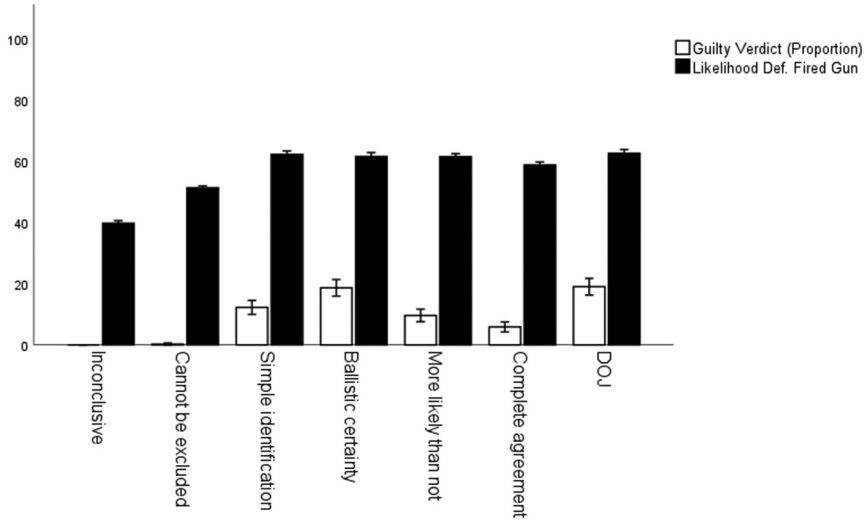

Figure 1. Proportions of guilty verdicts and mean likelihood ratings that the defendant fired the gun in each experimental condition. DOJ = Department of Justice. Whiskers represent 95% *CI* around the mean. For Inconclusive proportion and *CI*s were 0.

3.2. Hypothesis 1: impact of identification conclusions

We had anticipated that match conclusions would lead to higher guilt estimates compared to more cautious language. Specifically, Hypothesis 1 predicted that guilty verdicts would be more likely and that probability of guilt ratings would be higher in all other conditions compared to the "Inconclusive" and "Cannot be excluded" conditions. With the new analysis for categorical variables, we tested both independence within each conclusion category and if the distribution between categories will be different from the expected distribution.

3.2.1. Dichotomous guilty decisions

First, we ran an omnibus Fisher's Exact Test, comparing the combined first two categories (Inconclusive and Cannot be excluded) with the remaining five levels. Results were significant ($p < .001$, Cramer's $V = .201$) showing that overall match conclusions resulted in more guilty verdicts than non-match conclusions. (See Table 3 for the frequencies of guilty verdicts in the different conclusion categories). Subsequently, we ran a series of Fisher's Exact Tests comparing the two non-match conclusions separately with each of the match conclusions. All comparisons between Inconclusive and all of the match conclusions, $p < .001$, Cramer's Vs = .174 to .324 as well as between Cannot be excluded and all of the match conclusions, $p < .001$, Cramer's Vs = .163 to .314, were significant.

Table 2. Correlations between the items and sum score of scientific credibility, guilty verdicts and likelihood that the defendant fired the gun.

	M	SD	1	2	3	4	5	6	7	8	9
1 how reliable do you think the firearm evidence in this case is?	3.537	1.031									
2 In general, do you think the firearm evidence is reliable?	3.741	1.047	.514**								
3 how scientific do you think the firearm evidence in this case is?	3.940	1.293	.594**	.686**							
4 Did you find the testimony of Patrick West - the firearm examiner- to be credible?	3.535	1.161	.495**	.590**	.556**						
5 how often do firearm examiners make mistakes when determining whether bullets were fired through the same gun?	3.625	0.974	.046**	.194**	.081**	.139**					
6 In general, do you think the firearm evidence is scientific?	3.724	1.038	.495**	.916**	.697**	.594**	.206**				
7 Scientific Credibility	3.683	0.813	.711**	.879**	.834**	.771**	.349**	.881**			
8 Would you convict this defendant, based on the evidence that you have heard?	.094	.292	−.371**	−.502**	−.477**	−.413**	−.111**	−.512**	−.541**		
9 What would you say is the numerical probability that the defendant is the man who fired the shot in the convenience store?	56.715	16.497	−.425**	−.621**	−.611**	−.514**	−.344**	−.639**	−.712**	.588**	

Scientific Credibility refers to the mean of variables 1–6.

Table 3. Frequencies of guilty vs. Not guilty verdicts in different expert conclusion formulation categories.

	Guilty	
Expert's Conclusion Formulation	No	Yes
Non-Match conclusions		
Inconclusive	800	0
Cannot be excluded	798	2
Match Conclusions		
Simple identification	702	98
Ballistic certainty	651	149
More likely than not	723	77
Complete agreement	753	47
DOJ	648	152

DOJ = Department of Justice.

3.2.2. Likelihood the defendant fired the gun

Next, we ran a one-way ANOVA with the firearm expert's conclusion as IV and the likelihood that the defendant fired the gun as DV. Normality was violated at all levels of the IV (Kolmogorov–Smirnov test, *p*s < .001) and outliers more than 1.5 *SD* from the mean were present in the "Cannot be excluded" and "Simple identification" -conditions. The assumption of homogeneity of variances was also violated ($p < .001$).

The estimated likelihood that the defendant fired the gun was significantly different for different levels of the firearm expert conclusion (Welch's $F(6, 24567.80) = 357.19$, $p < .001$, $\eta^2 = .23$). Post hoc comparisons showed how the Inconclusive ($M = 39.50$, $SD = 11.81$) and Cannot be excluded ($M = 51.24$, $SD = 9.01$) conclusions had significantly lower likelihood estimates compared to all other levels ($p < .001$).

Detailed results of the post-hoc comparisons are reported in Table 4.

Table 4. Pairwise comparison with Games-Howell correction for the likelihood the defendant fired the gun and scientific credibility for the different levels of the expert conclusion.

	Likelihood the Defendant Fired the Gun		Scientific Credibility	
Expert's Conclusion	*M*	*SD*	*M*	*SD*
Inconclusive	39.50a	11.81	3.95 a	0.64
Cannot be excluded	51.24b	9.00	4.24b	0.59
Simple Identification	62.17c	15.55	4.48c	0.78
Ballistic certainty	61.48c	17.60	4.41c	0.92
More likely than not	61.39c	14.15	4.38c	0.78
Complete agreement	58.71d	13.73	4.22b	0.92
DOJ	62.51c	17.53	4.52c	0.85

DOJ = Department of Justice. $N = 5600$, $n = 800$ per level. Roman letters were used for column comparisons (within). Means sharing different subscripts differed significantly at .05 using Bonferroni correction. Higher values on the scientific credibility variable indicate that firearm expert statements were seen as more valid and reliable.

3.2.3. Scientific reliability

Next, we ran a one-way ANOVA with the firearm expert's conclusion as our IV and the scientific reliability variable as DV. Normality was violated at all levels of the IV (Kolmogorov-Smirnov test, ps < .001) and outliers more than 1.5 *SD* from the mean were present in all categories with the exception of the "ballistic certainty" -conclusion category. The assumption of homogeneity of variances was also violated ($p < .001$).

The scientific reliability score was significantly different for different levels of the firearm expert's conclusion (Welch's $F(6, 2477.80) = 60.44$, $p < .001$, $\eta^2 = .05$). Post hoc comparisons showed how the Inconclusive conclusion ($M = 3.95$, $SD = .64$) had significantly lower scientific reliability compared to all other levels ($p < .001$) while the Cannot be excluded conclusion ($M = 4.24$, $SD = .59$) had significantly lower scientific reliability compared to all other conclusion categories ($p < .001$) with the exception of the Complete agreement conclusion ($p = .996$).

Detailed results of the post-hoc comparisons are reported in Table 4.

Overall, the results offered strong support for Hypothesis 1. The results concerning perceived scientific reliability of the evidence mirrored the findings concerning guilty verdicts and perceived likelihood of the accused person being the shooter.

3.3. Hypothesis 2: influence of cautious language

First, we ran an omnibus Fisher's Exact Test, comparing the frequency of guilty verdicts between "Simple identification" and the remaining four match conclusions. The result was not significant ($p = .482$, Cramer's $V = .012$). Subsequently, we ran a series of Fisher's Exact Tests comparing "Simple identification" with each of the individual match conclusions. "Simple identification" was not statistically significantly different from "More likely than not" ($p = .109$), while all other comparisons were significant, ps < .001, Cramer's Vs = .088 to .111. However, the results supported Hypothesis 2 only partially, with the "Complete agreement" formulation resulting in a lower frequency of guilty verdicts while others (DOJ and "Ballistic certainty") going in the opposite direction.

Based on the same ANOVA ran for Hypothesis 1 on the likelihood of the defendant firing the gun, "Simple identification" ($M = 62.17$, $SD = 15.55$) was significantly different only from the "Complete agreement" conclusion ($p < .001$). Interestingly, there were a number of other differences between the different match formulations: "Ballistic certainty", "More likely than not" as well as "DOJ", all led to higher estimated likelihoods compared to "Complete agreement".

Additionally, in terms of scientific reliability, "Simple identification" ($M = 4.48$, $SD = 0.78$) was significantly different only from the "Complete agreement" category ($p < .001$) again mirroring the results for likelihood of the defendant firing the gun. Detailed results of the post hoc comparisons are reported in Table 4.

The results offer partial support for Hypothesis 2 with the LLM interpreting some of the more cautious or qualified statements (especially the "Complete agreement") as less indicative of guilt/the accused person having fired the gun compared to the conclusion "Simple identification". However, this was not true for all of the supposedly more cautious language categories.

3.4. Hypothesis 3: effect of cross-examination

We predicted that a cross-examination would reduce both guilty verdicts and probability of guilt ratings except in the "Inconclusive" conclusion condition.

3.4.1. Dichotomous guilty decisions

To test this hypothesis, we first ran an omnibus Fisher's Exact Test testing for differences in guilty verdicts between the cross-examination (four guilty verdicts) and no cross-examination (521 guilty verdicts) conditions which was significant $p < .001$. Later, we ran a series of Fisher's Exact Tests on the frequency of verdicts in the cross-examination vs. no cross-examination conditions for each type of Expert Conclusion (except for the "Inconclusive" conclusion which had no guilty verdict). Results showed how with the exception of "Cannot be excluded" there were statistically significant differences in all other categories with cross-examination decreasing guilty verdicts (see Table 5).

3.4.2. Likelihood the defendant fired the gun

First, we ran a series of *t*-tests separately for each firearm expert's conclusion category with likelihood the defendant fired the gun as DV and cross-

Table 5. Guilty verdicts divided by expert conclusion and No cross-examination vs. Cross-examination.

	No Cross-examination (n = 400)		Cross-examination (n = 400)		
	Count	Proportion	Count	Proportion	*p*
Cannot be excluded	1	.003	1	.003	1.00
Simple Identification	98	.245	0	.000	<.001
Ballistic Certainty	147	.368	2	.005	<.001
More likely than not	76	.190	1	.003	<.001
Complete Agreement	47	.118	0	.000	<.001
DOJ	152	.380	0	.000	<.001

DOJ = Department of Justice. n = 800 for each level of expert conclusion per experimental variable. p refers to the Fisher's Exact test result.

examination as IV. Normality was violated in both levels of the IV (Kolmogorov-Smirnov test, *p*s < .001) and for every level of firearm expert's conclusion. Outliers both over 1.5 and 3 *SD* from the mean were present with the exception of the "More likely than not" and "Complete agreement" conclusions. The assumption of homogeneity of variances was also violated with the exception of the "Cannot be excluded" conclusion.

Results showed as expected that the presence of cross-examination decreased the perceived likelihood that the defendant fired the gun with the exception of the "Inconclusive" conclusion category. Here, a reverse effect was observed. Detailed results are reported in Table S1.

3.4.3. Scientific reliability

Next, we ran a series of *t*-tests separately for each firearm expert's conclusion category with scientific reliability as DV and cross-examination as IV. Normality was violated in both levels of the IV (Kolmogorov-Smirnov test, *p*s < .001) and for every level of expert conclusion. Outliers over 1.5 *SD* from the mean were present with the exception of the "More likely than not" and "Complete agreement" -conclusion categories. The assumption of homogeneity of variances was also violated with the exception of the "Complete agreement" -conclusion category. Results consistently showed that cross-examination lowered the perceived scientific reliability. Detailed results are reported in Table S1 in supplementary materials S1 and Figure 2.

The analyses of the continuous likelihood variable clearly supported Hypothesis 3. Whenever an effect was observed, it was in the predicted direction. The results regarding scientific reliability again mirrored those on the other variables.

3.5. Hypothesis 4: role of expert forensic knowledge

Here, we predicted that LLMs with expert forensic knowledge would assign lower guilt estimates in all match conditions (that is, except for the "Inconclusive" and "Cannot be excluded"-conclusion conditions). Proportions for the dichotomous guilty verdict variable are reported in Table 6.

3.5.1. Dichotomous guilty decisions

First, we ran an omnibus Fisher's Exact Test looking for associations in verdicts between knowledge-enhanced (1 guilty verdict) and not knowledge-enhanced (524 guilty verdicts) LLMs which was significant $p < .001$. Later, we ran a series of Fisher's Exact Tests on the frequency of guilty verdicts delivered by the knowledge-enhanced vs. the not knowledge-enhanced LLMs separately for the different conclusion categories (excluding Inconclusive). Results showed that there were statistically significant associations in all other categories with expert knowledge decreasing guilty verdicts (see Table 6).

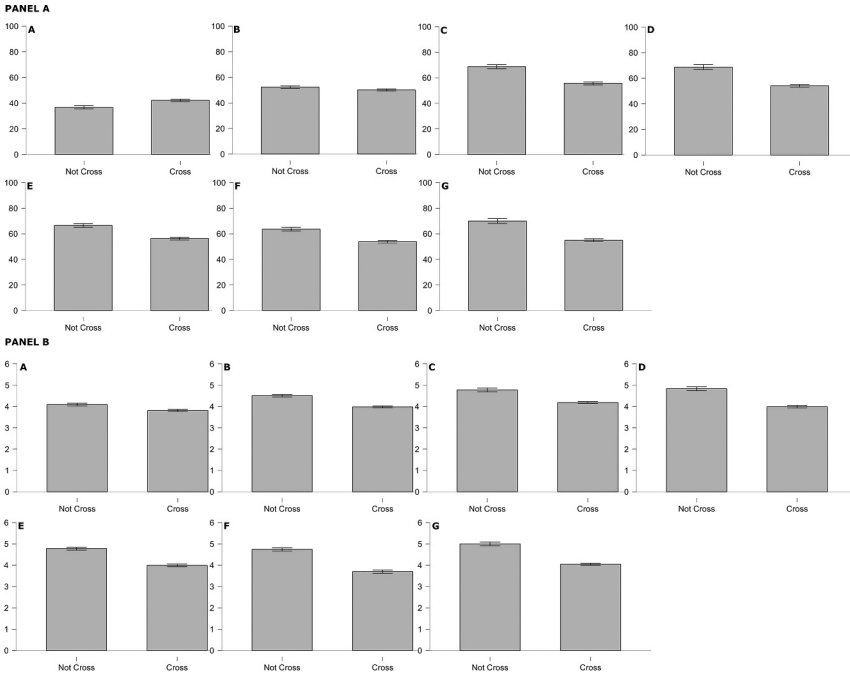

Figure 2. Means and 95% CI for the effects of expert conclusion divided by no cross-examination and cross-examination on the likelihood the defendant fired the gun and scientific reliability scores. Means comparing No Cross-Examination vs. Cross-Examination conditions within Each Firearm Expert's Conclusion Category (A: Inconclusive, B: Cannot be excluded, C: Simple Identification, D: Ballistic certainty, E: More likely than not, F: Complete agreement, and G: DOJ = Department of Justice). On the y-axis the Likelihood of the Defendant Fired the Gun (Panel A) and the Scientific Reliability scores (Panel B). Whiskers represent 95%CI. All comparisons were significant with $p < .001$. Cannot be excluded (B) in Panel A had a $p = .001$.

Table 6. Fisher's exact tests and proportions of conviction verdicts divided by expert conclusion and no specialist knowledge vs. Specialist knowledge.

	No Specialist Knowledge (n = 400)		Specialist Knowledge (n = 400)		
	Count	Proportion	Count	Proportion	p
Simple Identification	98	.245	0	.000	<.001
Ballistic Certainty	148	.370	1	.003	<.001
More likely than not	77	.193	0	.000	<.001
Complete Agreement	47	.118	0	.000	<.001
DOJ	152	.380	0	.000	<.001

DOJ = Department of Justice. n =800 for each level of expert conclusion per experimental variable. p refers to the Fisher's Exact test result.

3.5.2. Likelihood the defendant fired the gun

First, we ran a series of *t*-tests separately for each firearm expert's conclusion category with the likelihood the defendant fired the gun as DV and LLM having expert forensic knowledge as IV. Normality was violated in both levels of the IV (Kolmogorov-Smirnov test, *p*s < .001) and for every level of expert conclusion. Outliers both over 1.5 and 3 *SD* from the mean were present in all analyses. The assumption of homogeneity of variances was also violated with the exception of the "Cannot be excluded" -conclusion category. Results showed as expected that the LLM having expert knowledge decreased the estimates of the likelihood that the defendant fired the gun in all firearm expert's conclusion categories and that the effect sizes were large. Detailed results are reported in Table S2 in supplementary materials S1 and Figure 3.

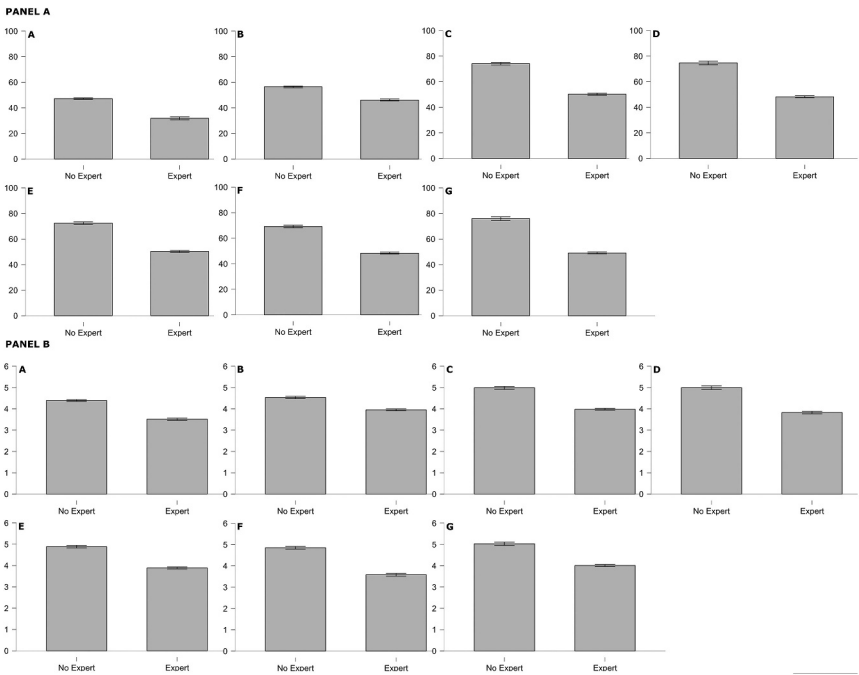

Figure 3. Means and 95% CI for the effects of expert conclusion divided by not receiving expert knowledge or receiving expert knowledge on the likelihood the defendant fired the gun and scientific reliability scores. Means comparing No Expert Knowledge vs. Expert Knowledge conditions on the y-axis within Each Firearm Expert's Conclusion Category (A: Inconclusive, B: Cannot be excluded, C: Simple Identification, D: Ballistic certainty, E: More likely than not, F: Complete agreement, and G: DOJ = Department of Justice). On the y-axis, the Likelihood the Defendant Fired the Gun (Panel A) and the Scientific Reliability scores (Panel B). Whiskers represent 95%CI. All comparisons were significant with $p < .001$.

3.5.3. Scientific reliability

Next, we ran a series of corresponding *t*-tests separately for each firearm expert's conclusion category with scientific reliability as DV and LLM having expert forensic knowledge as IV. Normality was violated in both levels of the IV (Kolmogorov-Smirnov test, *p*s < .001) and for every level of expert conclusion. Outliers over 1.5 *SD* from the mean were present with the exception of the level "Complete agreement" -conclusion category. The assumption of homogeneity of variances was also violated with the exception of "Inconclusive" and "Complete agreement" -conclusion categories. Results showed how LLM having expert forensic knowledge (vs. not) resulted in lower scientific reliability assessments in all firearm expert's conclusion conditions and that the effect sizes were large. Detailed results are reported in Table S2.

The analyses of the continuous likelihood variable clearly supported Hypothesis 4. Additionally, once again, probably due to floor effects, the results were not as clear for the dichotomous verdict variables. However, whenever an effect was observed, it was in the predicted direction. The results regarding scientific reliability again mirrored those on the other variables.

4. Discussion

4.1. Main findings

Replicating the study by Garrett et al. (2020) by replacing human mock jury participants with Large Language Model (LLM) mock jury participants, we examined how the LLM participants evaluated firearm examiner testimony, finding both similarities and differences compared with human decision-making patterns. The substantial divergence in conviction rates between LLMs and human jurors (9% vs. 50%) underscores fundamental differences in decision-making frameworks. LLMs seemed to adhere to formal standards of scientific evidence reliability, rejecting causal inferences that exceed frequency-based probabilities (e.g. refusing to extrapolate from limited firearm analysis data without explicit error rates). This mirrors a textualist interpretation of legal standards like the Daubert criteria. In contrast, human jurors may be more likely to apply "plausible reasoning presumptions" permitted in case law – resolving ambiguities through commonsense heuristics (e.g. associating spatial proximity with guilt likelihood). While the LLM's objectivity mitigates cognitive biases (e.g. anchoring on prosecutors' narratives), its fidelity to quantifiable standards would reshape trial outcomes. Historical conviction benchmarks (e.g. 30% firearm-only conviction rates in U.S. courts) suggest that human decisions balance formal rules with socially contextualised rationality; a balance the LLMs in the present study

did not replicate. It is important to recognise, however, that it is unclear which approach would be desirable. Moreover, like human mock jurors in Garrett et al. (2020), the LLMs in the present study showed sensitivity to different conclusion language, with "Inconclusive" and "Cannot be excluded" statements leading to significantly lower guilt assessments compared to more definitive conclusion formats. However, unlike humans who showed minimal response to cross-examination challenging the firearm expert (Garrett et al., 2020), the LLMs exposed to cross-examinations lowered their guilt assessments and scientific credibility ratings. Most importantly, the LLMs enhanced with forensic science knowledge (through exposure to the PCAST (2016) report) showed consistently more conservative evaluations of firearm evidence across all match conditions, suggesting that knowledge-enhanced LLMs could help address the disconnect between scientific understanding and legal practice in forensic evidence evaluation (Curley et al., 2022). This finding is particularly promising given longstanding concerns about the ability of jurors to appropriately evaluate complex scientific evidence (Hans & Saks, 2018).

Unlike human participants who showed minimal sensitivity to linguistic variations in match statements, the LLMs demonstrated differences in responses to different conclusion formulations. For instance, the "Complete agreement" language led to significantly lower guilt assessments compared to "Simple identification" statements. Particularly striking was the LLMs' complete rejection of guilty verdicts in the "Inconclusive" condition (0% guilty verdicts), compared to the 21% conviction rate among human mock jurors in Garrett et al. (2020). This difference may reflect a more rigid logical processing and application of the reasonable doubt standard in LLMs. In sum, LLMs may be better able to distinguish between different levels of certainty in expert testimony, an important finding given ongoing debates about appropriate conclusion language in firearm analysis testimony (National Research Council, 2009, the President's Council of Advisors on Science and Technology, 2016). Specifically, the LLMs showed the lowest likelihood of convicting defendants when exposed to evidence-negating expressions such as "Inconclusive" and "Cannot be excluded". Conversely, they demonstrated the highest conviction rates under positively identifying language conditions: "Simple identification," "Ballistic certainty," "More likely than not," and the DOJ-recommended phrasing. Notably, despite being the most definitive formulation, the "Complete agreement" condition resulted in significantly lower firearm discharge likelihood assessments compared to these four positive identification conditions. The difference may be explained by the research of Gu et al. (2025), which found that LLMs exhibit stronger endorsement of objective evidence (i.e. "Ballistic certainty" in the current study) compared to subjective judgement (e.g. "Complete agreement" in the current study).

Based on our results, the investigated LLMs were clearly sensitive to the information from the cross-examination. While Garrett et al. (2020) found that cross-examination challenging the scientific basis of firearm analysis had minimal impact on human juror decisions, the LLMs showed consistently lower guilt assessments and scientific credibility ratings when expert testimony was challenged through cross-examination. This finding aligns with Austin and Kovera's (2015) research suggesting that scientifically informed cross-examination can promote critical evaluation of scientific evidence. The LLMs' response to cross-examination was particularly pronounced in conditions using conclusion language, such as "Ballistic certainty" and "More likely than not" statements. This pattern suggests that LLMs may be better equipped to integrate critical information about methodological limitations into their overall assessment of forensic evidence, addressing a key concern identified by Devine and Macken (2016) regarding jurors' difficulty in evaluating scientific validity. It should be noted that the low baseline conviction rate (9% overall) necessitates cautious interpretation of shifts in outcomes as a function of the experimental manipulations. Notably, the analysis revealed a marked relative reduction in guilty verdicts following cross-examination ($\Delta n = 517$, 64.6% decrease compared to non-cross-examination condition, $p < .001$), a pattern aligning with established findings regarding human jurors' tendency towards evidentiary conservatism post-interrogation (Scanlon et al., 2023). This parallelism substantiates LLMs' potential to implement Daubert-compliant evidentiary screening mechanisms during pretrial phases. Through automated documentation of methodological deficiencies, LLMs could systematically filter unreliable testimony – resolving a critical challenge in trials involving forensic evidence (Koehler et al., 2023). While the current study used abbreviated and simplified trial materials, future investigations could enhance ecological validity through high-fidelity simulated environments or real-world adjudicative contexts. However, it is somewhat surprising that LLMs without knowledge enhancement did not do so spontaneously in the condition without a cross-examination given that they probably had access to this or other applicable information in their training data. The fact that they did not apply it, points to the importance of the instructions given to LLMs in decision-making tasks.

Our most significant finding was the impact of knowledge-enhancement on LLM performance. When equipped with forensic science knowledge from the PCAST (2016) report, the LLMs demonstrated consistently more conservative evaluations of firearm evidence across all match conditions, with significantly lower guilt assessments and scientific credibility ratings. This finding addresses a fundamental challenge in the legal system identified by Judge Learned Hand in 1901 - the difficulty of evaluating technical evidence without relevant expertise (Hans & Saks, 2018). While traditional juries must

rely on expert testimony and cross-examination alone, our knowledge-enhanced LLMs effectively operated as "special jurors" with relevant domain expertise, similar to historical special juries that were composed of members chosen for their subject matter knowledge (Oldham, 1983). The enhanced LLMs' evaluation of the evidence aligned with current understanding of the limitations of forensic feature-comparison methods (National Research Council, 2009, the President's Council of Advisors on Science and Technology, 2016). This finding is particularly important given that human jurors tend to place substantial weight on forensic testimony even when the underlying methods lack scientific validation (Lieberman et al., 2008). This finding suggests a potential pathway for improving the evaluation of complex scientific evidence in legal settings.

The distinction between an LLM's potential knowledge (information incorporated during training) and its actual ability to autonomously activate such knowledge without explicit prompting is an important one. This phenomenon is linked to prompt dependence (Lu et al., 2022; Zhao et al., 2021): LLMs often require explicit cues to determine which information to activate. Without clear guidance, they may, for example, fail to recall relevant criticisms of forensic science, even when such concepts are present in their training data. However, it is also important to acknowledge that a similar pattern is present in human jurors who do not inherently process all available information. Studies have shown that court instructions and the framing of information significantly influence their decision-making process (Hans & Saks, 2018). Thus, prompting might be viewed not as an artifice but as analogous to the legal education provided to human jurors to enhance their ability to evaluate complex evidence. However, this study showed that even when optimising prompts were used, the standard LLMs' performance was still inferior to the knowledge-enhanced LLM in critically assessing the evidence. This discrepancy may stem from the fact that the standard LLM has insufficient knowledge of forensic science, or that it actually possesses forensic science knowledge but is unable to apply it to the task at hand, whereas the Knowledge-enhancement (i.e. PCAST) provides the tool to externalise their knowledge. To test this possibility, we created 90 true/false questions (See Supplementary material S2) covering the content of the PCAST after the formal analyses without pre-registered, and asked both the standard (gpt-4o-6 August 2024, temperature = 1, without knowledge enhancement) and the knowledge-enhanced models (gpt-4o-6 August 2024, temperature = 1, with the PCAST uploaded as the knowledge-enhancement) to answer them. The results showed that the standard model had a correct response rate of 55.6%, with no difference between that and chance level (binomial test, $p = .343$). In contrast, the correct response rate of

the knowledge-enhanced model was 93.3%, significantly higher than that of the standard model (binomial test, $p < .001$), suggesting that the standard model actually either did not have the relevant knowledge and that the improvement of the LLMs in a specific domain relies on knowledge-enhancement. However, Burns et al. (2024) have demonstrated that much of the latent information in language models remains inaccessible unless triggered by very specific prompts, suggesting that an LLM may "know the answer" without producing it unless the request is formulated in the appropriate manner.

The potential advantages of LLMs extend beyond their demonstrated ability to critically evaluate scientific evidence. Unlike human jurors, LLMs are not subject to psychological stress that can impact decision-making National Center for State Courts, 1998), and they process information without emotional interference (Kerr, 2010; Nunez et al., 2016; Semmler & Brewer, 2002). This emotional detachment could be valuable when evaluating complex scientific evidence, where emotional responses could interfere with rational assessment of validity of evidence.

However, any benefits must be weighed against fundamental principles of the jury system. The jury institution serves multiple functions beyond fact-finding, including ensuring civic participation, representing community values, and providing a check on state authority (Abramson, 2000; Kalven & Zeisel, 1966; Langbein, 2005; Odgers, 1957). Most crucially, the concept of jury nullification: where juries can render verdicts contrary to evidence when strict application of law would produce unjust results, represents an important mechanism to challenge unjust laws that might be lost with the involvement of AI systems in decision-making (Conrad, 2013; Crispo et al., 1997; Leipold, 1996).

Finally, although the low conviction rate especially in the cross-examination and knowledge enhanced conditions may reflect that LLMs apply the beyond reasonable doubt standard correctly, this could also make the implementation of LLMs in real-world legal contexts problematic. Human jurors likely employ various cognitive heuristics that LLMs do not, such as considering the spatio-temporal proximity and other association of the accused to the crime scene (in this case, being stopped with a similar calibre weapon two days after the incident) as damning evidence. While potentially biased, such heuristics may be fundamentally accurate. Additionally, human jurors may weigh the social costs of false positives against false negatives differently than LLMs.

4.2. Strengths and limitations

A key strength of the present study is the application of knowledge-enhanced LLMs to legal decision-making. To the best of our knowledge, this represents

the first empirical test of whether AI systems enhanced with domain-specific expertise can improve the evaluation of forensic evidence. Also, the study's systematic replication of Garrett et al. (2020) experimental paradigm allows for direct comparisons with human decision-making patterns, while the sample size (5,600 responses) provides robust statistical power for detecting effects.

However, several limitations must be considered. The first limitation relates to statistical power. As an exact replication, we adhered to the experimental parameters outlined by Garrett et al. (2020), employing identical condition structures and accumulating 200 responses per cell for a comprehensive dataset of 5,600 observations, which provided a sufficient power (>.90) to detect small effects ($d = 0.35$) at $\alpha = .05$. The sample size also surpassed typical benchmarks in LLM research – notably exceeding Suri et al."s (2024) 60-response paradigm and Abdurahman et al. (2024), pp. 1, 000-response moral reasoning investigation. However, a critical methodological consideration is related to LLMs" attenuated variability. Empirical evidence from Abdurahman et al. (2024) demonstrates that GPT-3.5 exhibits 43–121-fold reduced variance compared to human populations in moral judgement tasks. The difference of variance was also observed between our current examination of legal decision-making variables and Garrett et al. (2020)'s study against humans. Our analysis reveals systematically diminished variance in LLM-generated responses for both guilt likelihood assessments and conviction propensity measures relative to Garrett et al. (2020)'s human data. This fundamental discrepancy in response distributions necessitates caution when looking at the results of the regressions in supplementary materials; the smaller variances introduce larger effect sizes, which may overestimate the actual difference between groups. The temperature coefficient is the control for the randomness of the output content of the LLMs, and future research should consider increasing the temperature coefficient of the model to increase the variability and better align it with the variance of the human participants. Moreover, several statistical assumptions were violated, particularly affecting the logistic regression analyses. The quasi-complete separation observed in the "Inconclusive" and "Cannot be excluded" conditions limited the interpretability of these results. The study relied on a single LLM architecture (GPT-4), and results may not generalise to other AI systems or future iterations.

Additionally, while the knowledge enhancement was based on an authoritative source (the PCAST (2016 report), future research should examine how different types and combinations of expert knowledge affect the performance of LLMs. Moreover, the use of simplified case materials may not fully capture the complexity of real-world jury deliberations, where, for example, the extent of materials to be processed could result in different outcomes. In an actual trial, jurors hear live testimony from the firearms inspector and

learn about the ammunition in question in graphic or video form. However, limited by LLM development technology, in this study, courtroom testimony was presented entirely in written form. Studies have found that the form of information influences individuals' attitude through peripheral route and central route, which in turn affects the outcome of the decision (Richardson & Nash, 2022). In addition to this, jurors in a real trial should be informed about the background and training associated with the expert. There has been a great deal of research demonstrating that the background knowledge of the information provider affects the extent to which the recipient receives the information (e.g. Raju et al., 1995). Therefore, these shortcomings lead to limited generalisability of the results of this study. Future research should explore multimodal testimonies, including videos and pictures, as well as provide more information about the background and training of the LLM. In addition, we have diverted from the pre-registration for what concerns some analyses, this due to unforeseeable circumstances and a misspecification in the pre-registration.

4.3. Future directions

Future research should explore the potential of knowledge-enhanced LLMs across different types of evidence. For eyewitness testimony, LLMs could be enhanced with established findings about memory malleability, confidence-accuracy relationship, and factors affecting identification accuracy. In the domain of confession evidence, systems could incorporate basic knowledge about false confessions and interrogation techniques that promote them. Following the present study, the scope could easily also be expanded beyond firearms to other pattern-matching domains like fingerprints, toolmarks, and shoe prints, incorporating relevant validation studies and information about error rates. Critically, transparency and explainability of LLM decision processes, and integration with existing legal procedures and safeguards need careful examination. The study showed that providing LLMs with domain-specific expertise significantly improved their judgement in legal evidence assessment, making them more inclined to make decisions based on scientific criteria. This finding suggests that future LLM training should focus more on structured knowledge integration, especially in domains involving professional judgement (e.g. justice, medicine), where structured criteria and critical analysis frameworks need to be directly embedded into the model's body of knowledge rather than relying solely on the general training corpus. In addition, this study revealed the sensitivity of LLM to evidence in cross-examination situations, which suggests that the training process needs to strengthen the adversarial learning mechanism and enhance the model's ability to recognise logical contradictions and information reliability by simulating the questioning and rebuttal scenarios in legal debates.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the New York University Shanghai [10109_Academic Research]; Spring Crocus s.r.l., managing body of CBT. ACADEMY.

ORCID

Francesco Pompedda http://orcid.org/0000-0001-9253-0049
Pekka Santtila http://orcid.org/0000-0002-0459-1309
Eleonora Di Maso http://orcid.org/0009-0008-7244-1709
Thomas J. Nyman http://orcid.org/0000-0002-6409-2528
Yongjie Sun http://orcid.org/0009-0005-9616-9875
Angelo Zappala http://orcid.org/0000-0002-3579-1987

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *Proceedings of the National Academy of Sciences Nexus*, *3*(7), 245. https://doi.org/10.1093/pnasnexus/pgae245

Abramson, J. B. (2000). *We, the jury: The jury system and the ideal of democracy : With a new preface*. Harvard University Press.

AFTE Glossary. (1998). Theory of identification as it relates to Toolmarks. *AFTE Journal*, *30*(1), 86–88.

Austin, J. L., & Kovera, M. B. (2015). Cross-examination educates jurors about missing control groups in scientific evidence. *Psychology, Public Policy, and Law*, *21*(3), 252–264. https://doi.org/10.1037/law0000049

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2024). Discovering latent knowledge in language models without supervision (arXiv:2212.03827). arXiv. https://doi.org/10.48550/arXiv.2212.03827

Center for Statistics and Applications in Forensic Evidence. (2006). *United States v. Monteiro, 407 F.Supp.2d 351*. Center for Statistics and Applications in Forensic Evidence. https://forensicstats.org/cases/united-states-v-monteiro-407-f-supp-2d-351-d-mass-2006/

Center for Statistics and Applications in Forensic Evidence. (2007). *United States v. Diaz, 2007 WL 485967*. Center for statistics and Applications in Forensic Evidence. https://forensicstats.org/cases/united-states-v-diaz-2007-wl-485967-n-d-cal-2007/

Center for Statistics and Applications in Forensic Evidence. (2008). *United States. V. Glynn, 578 F.Supp.2d 567*. Center for Statistics and Applications in Forensic Evidence. https://forensicstats.org/cases/united-states-v-glynn-578-f-supp-2d-567-s-d-n-y-2008/

Center for Statistics and Applications in Forensic Evidence. (2019). *U.S. v. Tibbs, 2019 WL 4359486*. Center for Statistics and Applications in Forensic Evidence. https://forensicstats.org/cases/u-s-v-tibbs-2019-wl-4359486-d-c-super-2019/

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English (arXiv:2110.00976). arXiv. https://doi.org/10.48550/arXiv.2110.00976

Chang, Z., Li, M., Jia, X., Wang, J., Huang, Y., Wang, Q., Huang, Y., & Liu, Y. (2024). What external knowledge is preferred by LLMs? Characterizing and exploring chain of evidence in imperfect context (arXiv:2412.12632). arXiv. https://doi.org/10.48550/arXiv.2412.12632

Conrad, C. S. (2013). *Jury nullification: The evolution of a doctrine*. Cato Institute.

Crispo, L., Slansky, J., & Yriarte, G. (1997). Jury nullification: Law versus anarchy. *Loyola of Los Angeles Law Review*, *31*(1), 1.

Curley, L. J., Munro, J., & Dror, I. E. (2022). Cognitive and human factors in legal layperson decision making: Sources of bias in juror decision making. *Medicine, Science, and the Law*, *62*(3), 206–215. https://doi.org/10.1177/00258024221080655

Devine, D. J., & Macken, S. (2016). Scientific evidence and juror decision making: Theory, empirical research, and future directions. In *Advances in psychology and law* (pp. 95–139). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-43083-6_4

Eigner, E., & Händler, T. (2024, February 27). *Determinants of LLM-Assisted decision-making*. arXiv.Org. https://arxiv.org/abs/2402.17385v1

Garrett, B. L., Scurich, N., & Crozier, W. E. (2020). Mock jurors' evaluation of firearm examiner testimony. *Law and Human Behavior*, *44*(5), 412–423. https://doi.org/10.1037/lhb0000423

Gramlich, J. (2025, March 5). *What the data says about gun deaths in the U.S.* Pew Research Center. https://www.pewresearch.org/short-reads/2025/03/05/what-the-data-says-about-gun-deaths-in-the-us/

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025). A survey on LLM-as-a-judge (arXiv:2411.15594). arXiv. https://doi.org/10.48550/arXiv.2411.15594

Gutierrez, R. E., Scurich, N., & Garrett, B. L. (2024). *The impact of defense experts on juror perceptions of firearms examination testimony*. *SSRN scholarly paper 4966326*. Social Science Research Network. https://doi.org/10.2139/ssrn.4966326

Hans, V. P., & Saks, M. J. (2018). Improving Judge & jury evaluation of scientific evidence. *Daedalus*, *147*(4), 164–180. https://doi.org/10.1162/daed_a_00527

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Huda, N. U., Sahito, S. F., Gilal, A. R., Abro, A., Alshanqiti, A., Alsughayyir, A., & Palli, A. S. (2024). Impact of contradicting subtle emotion cues on large language models with various prompting techniques. *International Journal of Advanced Computer Science and Applications*, *15*(4). https://doi.org/10.14569/IJACSA.2024.0150442

Kalven, H., & Zeisel, H. (1966). *The American jury*. Books. https://chicagounbound.uchicago.edu/books/725

Kerr, N. L. (2010). Explorations in juror emotion and juror judgment. In B. H. Bornstein & R. L. Wiener (Eds.), *Emotion and the Law: Psychological perspectives* (pp. 97–132). Springer. https://doi.org/10.1007/978-1-4419-0696-0_4

Koehler, J. J., Mnookin, J. L., & Saks, M. J. (2023). The scientific reinvention of forensic science. *Proceedings of the National Academy of Sciences*, *120*(41), e2301840120. https://doi.org/10.1073/pnas.2301840120

Langbein, J. H. (2005). *The origins of adversary criminal trial*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199287239.001.0001

Lanigan, B. (2012). Firearms identification: The need for a critical approach to, and possible guidelines for, the admissibility of ballistics evidence. *Suffolk Journal of Trial and Appellate Advocacy*, *17*(1), 54.

Leipold, A. D. (1996). Rethinking jury nullification. *Virginia Law Review*, *82*(2), 253–324. https://doi.org/10.2307/1073635

Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digital Medicine*, *6*(1), 1–14. https://doi.org/10.1038/s41746-023-00979-5

Lieberman, J. D., Carrell, C. A., Miethe, T. D., & Krauss, D. A. (2008). Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, and Law*, *14*(1), 27–62. https://doi.org/10.1037/1076-8971.14.1.27

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In S. Muresan, P. Nakov & A. Villavicencio (Eds.). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8086–8098). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.556

McQuiston-Surrett, D., & Saks, M. J. (2009). The testimony of forensic identification science: What expert witnesses say and what factfinders hear. *Law and Human Behavior*, *33*(5), 436–453. https://doi.org/10.1007/s10979-008-9169-1

National Center for State Courts. (1998). *Through the eyes of the juror: A manual for addressing juror stress*. https://cdm16501.contentdm.oclc.org/digital/collection/juries/id/310/rec/1

National Research Council. (2008). *Ballistic imaging*. National Academies Press. https://doi.org/10.17226/12162

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press. https://doi.org/10.17226/12589

Nunez, N., Estrada-Reynolds, V., Schweitzer, K., & Myers, B. (2016). The impact of emotions on juror judgments and decision-making. In B. H. Bornstein & M. K. Miller (Eds.), *Advances in psychology and law: Volume 2* (pp. 55–93). Springer International Publishing. https://doi.org/10.1007/978-3-319-43083-6_3

Odgers, F. J. (1957). Trial by jury. *The Cambridge Law Journal*, *15*(2), 242–244. https://doi.org/10.1017/S000819730008212X

Oldham, J. C. (1983). The origins of the special jury. *The University of Chicago Law Review*, *50*(1), 137–221. https://doi.org/10.2307/1599384

the President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*.

Raju, P. S., Lonial, S. C., & Glynn Mangold, W. (1995). Differential effects of subjective knowledge, objective knowledge, and usage experience on decision

making: An exploratory investigation. *Journal of Consumer Psychology*, *4*(2), 153–180. https://doi.org/10.1207/s15327663jcp0402_04

Richardson, B. H., & Nash, R. A. (2022). 'Rapport myopia' in investigative interviews: Evidence from linguistic and subjective indicators of rapport. *Legal and Criminological Psychology*, *27*(1), 32–47. https://doi.org/10.1111/lcrp.12193

Salerno, J. M., & McCauley, M. R. (2009). Mock jurors' judgments about opposing scientific experts: Do cross-examination, deliberation and need for cognition matter? *American Journal of Forensic Psychology*, *27*(3), 37–60.

Scanlon, F., Lester, M. E., Brace, T., Batastini, A. B., & Morgan, R. D. (2023). Tried and true? A psychometric evaluation of the psychological inventory of criminal thinking styles-short form. *Assessment*, *30*(6), 1985–1997. https://doi.org/10.1177/10731911221132500

Semmler, C., & Brewer, N. (2002). Effects of mood and emotion on juror processing and judgments. *Behavioral Sciences & the Law*, *20*(4), 423–436. https://doi.org/10.1002/bsl.502

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology. General*, *153*(4), 1066–1075. https://doi.org/10.1037/xge0001547

Tapp, S. N., & Coen, E. J. (2024). *Criminal victimization, 2023*. Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/criminal-victimization-2023

Thompson, W. C., Kaasa, S. O., & Peterson, T. (2013). Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies*, *10*(2), 359–397. https://doi.org/10.1111/jels.12013

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *Npj Digital Medicine*, *7*(1), 1–9. https://doi.org/10.1038/s41746-024-01029-4

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Yan, C., Fu, X., Xiong, Y., Wang, T., Hui, S. C., Wu, J., & Liu, X. (2025). LLM sensitivity evaluation framework for clinical diagnosis. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio & S. Schockaert (Eds.). *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 3083–3094). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.207/

Yang, Y., Liu, X., Jin, Q., Huang, F., & Lu, Z. (2024). Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, *4*(1), 1–6. https://doi.org/10.1038/s43856-024-00601-z

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models (arXiv:2102.09690). arXiv. https://doi.org/10.48550/arXiv.2102.09690