Assessing Verbal Eyewitness Confidence Statements Using Natural Language Processing

Rachel Leigh Greenspan
UNIVERSITY OF MISSISSIPPI - CRIMINAL JUSTICE AND LEGAL STUDIES

Alex Lyman
BRIGHAM YOUNG UNIVERSITY, PH.D.

Paul Heaton
UNIVERSITY OF PENNSYLVANIA CAREY LAW SCHOOL

This paper is not the copy of record and may not exactly replicate the authoritative document published in *Psychological Science*. The final article is available, upon publication, at: DOI: 10.1177/09567976241229028.

Assessing Verbal Eyewitness Confidence Statements Using Natural Language Processing

Rachel Leigh Greenspan[1], Alex Lyman[2], Paul Heaton[2]

[1]Department of Criminal Justice and Legal Studies, University of Mississippi

[2] University of Pennsylvania Carey Law School

Author Note:

Correspondence concerning this article should be addressed to Rachel Greenspan, Department of Criminal Justice and Legal Studies, University of Mississippi, M313 Mayes Hall, University, MS 38677, USA. Email: rlgreen1@olemiss.edu.

This paper was presented at the 2023 American Psychology-Law Society conference in Philadelphia, PA.

Abstract

After an eyewitness completes a lineup, officers are advised to ask witnesses how confident they are in their identification. While researchers in the lab typically study eyewitness confidence numerically, confidence in the field is primarily gathered verbally. In the current study, we used a natural language processing approach to develop an automated model to classify verbal eyewitness confidence statements. Across a variety of stimulus materials and witnessing conditions, our model correctly classified adult witnesses' (N = 4,541) level of confidence (i.e., high, medium, or low) 71% of the time. Confidence-accuracy calibration curves demonstrate that the model's confidence classification performs similarly in predicting eyewitness accuracy compared to witnesses' self-reported numeric confidence. Our model also furnishes a new metric, confidence entropy, that measures the vagueness of witnesses' confidence statements and provides independent information about eyewitness accuracy. These results have implications for how empirical scientists collect confidence data and how police interpret eyewitness confidence statements.

Keywords: eyewitness confidence, verbal confidence, natural language processing

Statement of Relevance: After an eyewitness completes a lineup, police officers are advised to ask witnesses how confident they are in their identification. Confidence, from an unbiased lineup, can help predict whether a witness has made an accurate identification. While researchers in the lab typically study eyewitness confidence numerically, confidence in the field is primarily gathered verbally, in the witness' own words. We developed a machine learning model to read and classify eyewitness confidence statements and made it freely available online for use by researchers and practitioners (https://huggingface.co/spaces/psheaton/eyewitness_confidence_classifier). Across a variety of lineup types, our model correctly classified witnesses' level of confidence (i.e., high, medium, or low) 71% of the time. We further demonstrate that the model's confidence classification serves as a reliable tool for identifying accurate witnesses. These results have implications for how empirical scientists collect confidence data in the lab and how police interpret eyewitness confidence statements in the field.

Assessing Verbal Eyewitness Confidence Statements Using Natural Language Processing

Of the over 3,300 exonerations recorded to date by the National Registry of Exonerations (n.d.), over 900 are due, at least in part, to eyewitness misidentifications. Nearly all these cases involved an eyewitness testifying that they were highly confident in their identification (Garrett, 2011). Highly confident witnesses are persuasive to jurors, influencing their verdict decisions, judgments of guilt, and perceptions of a witness' accuracy (Slane & Dodson, 2022). Researchers often refer to eyewitness confidence as a *reflector variable* (Wells, 2020). Reflector variables are witness behaviors that occur during or after an identification procedure that are associated with witness accuracy. Eyewitness confidence is a reflector variable as high confidence, from an unbiased lineup, indicates accuracy (Wixted & Wells, 2017). For this reason, officers are advised to document witness confidence immediately after any identification procedure (National Research Council, 2014; Wells et al., 2020).

Empirical research on the relationship between confidence and accuracy largely measures witness confidence numerically (Smalarz et al., 2021). However, officers in the field typically ask for witness confidence verbally, in the witness' own words (National Research Council, 2014). While verbal confidence is the most common method of documenting confidence, it has distinct drawbacks. Interpretations of verbal confidence are inconsistent and individuals' numeric translations of verbal confidence vary widely (Theil, 2002). Imagine a situation in which a witness makes an identification from a lineup and then states they are "fairly certain" the person they selected is the person who committed the crime. How does a police officer interpret this confidence statement? Does the officer believe that this witness has made a highly confident identification or not? If the officer misinterprets the intended meaning of the witness' confidence

statement (e.g., believes the witness is highly confident when they are not), then this impairs the ability of confidence to act as a cue to witness accuracy.

Compounding this problem, interpretations of verbal confidence are impacted by base rate (Wallsten et al., 1986) and contextual information (Brun & Teigen, 1988). For example, researchers have documented a featural justification effect (Dodson & Dobolyi, 2015), wherein witnesses justifying their reasons for their level of confidence (e.g., "I remember his bushy eyebrows") can lead to more misinterpretations of confidence than when justification information is not provided. Thus, not only are verbal confidence statements prone to misunderstandings, but additional contextual information provided by witnesses tends to worsen, rather than improve, this problem.

One way to improve comprehension and reduce the ambiguity of verbal confidence statements is to use machine-learning approaches rather than human evaluators to classify verbal confidence. In one of the only papers to date using this approach, a "bag of words" model revealed that verbal confidence was predictive of accuracy and that the content of verbal confidence statements contained additional diagnostic cues beyond the information provided by numeric confidence (Seale-Carlisle et al., 2021). However, the confidence statements used in this study included both verbal statements of confidence as well as verbal statements of justification, which are currently not typically collected in the field. Additionally, the language classifier used in this study relied on a basic linguistic model that simply counted the use of individual words.

Using machine learning approaches to evaluate verbal eyewitness confidence statements has several potential benefits. It is faster than human coders, especially with large amounts of data. It is also more replicable and can be easily implemented for both researchers and

practitioners. Machine learning approaches may also be less influenced by cognitive biases such as the influence of pre-existing information or cultural biases (Grabman & Dodson, 2019).

In the current study, we developed a Transformer-based Large Language Model (LLM) to categorize witness confidence statements. The goal of this model is to interpret the intended meaning of a witness' verbal confidence statement. To do this, we analyze a sample of witnesses who explain their confidence both in their own words and using numbers. We use these witness-provided numeric translations to identify the "ground-truth" of how confident witnesses actually are in their identification. We define confidence statements as low confidence (0-25%), medium confidence (26-74%), or high confidence (75-100%).

After developing the model, we then tested the performance of our LLM by applying it to data previously unseen by the model (hereafter called external data), specifically samples of verbal eyewitness confidence statements from studies in which researchers also collected self-reported numeric confidence, furnishing a reference measure of ground-truth confidence. Our LLM is freely available for use by other researchers and practitioners—including functionality to batch process large collections of confidence statements—at https://huggingface.co/spaces/psheaton/eyewitness_confidence_classifier.

Open Practices Statement

The data and materials for the pilot study are available on the Open Science Foundation (OSF): https://osf.io/9cbt6/. Appendix S2 details how the remaining datasets were obtained. This study was not pre-registered.

Methods

Pilot Data

We recruited 989 participants on Amazon Mechanical Turk using CloudResearch's MTurk Toolkit. Participants watched a short video of a robbery (Kenchel et al., 2021) and were randomly assigned to view a six-person either target-present or target-absent lineup. After completing the lineup, all participants were asked to explain their confidence in their own words and then to translate their confidence to a number using an 11-point scale. The data and materials for this study are available on OSF. A full description of the methods and outcomes of the pilot study is available in the supplemental materials. This study was approved by the University of Mississippi Institutional Review Board.

Datasets

To ensure that the model was as generalizable as possible, we sought to obtain all existing datasets for which participants (1) made a lineup identification, (2) expressed their confidence in their own words, and (3) translated their verbal confidence into a numeric response. To obtain these datasets, the first author reviewed the eyewitness confidence literature and contacted authors of published articles that contained data that met these conditions. Seven datasets were used in the current study: the pilot data described above as well as data from six additional papers (Bergold & Heaton, 2018; Dobolyi & Dodson, 2018; Grabman & Dodson, 2022; Grabman et al., 2019; Kenchel et al., 2021; Smalarz et al., 2021). Only participants who made a lineup identification (i.e., not a lineup rejection) from a target-present lineup were included in our analyses. A more detailed description of each of these datasets can be found in the supplemental materials. Table S1 compares the key features of these datasets. See Figure S1 for a flowchart about the distribution of data to the test and external datasets.

Modeling Approach

Our model relies on the Transformer (Vaswani et al., 2017). The Transformer is a neural-net-based architecture featuring multi-headed self-attention, where the relative importance between different words in an input is calculated. This enables Transformers to handle dependencies between words, a key feature in natural language understanding. Using this Transformer architecture, researchers have trained several LLMs to achieve unprecedented performance on language modeling tasks. The first of these LLMs was BERT (Devlin et al., 2018), which was developed at Google and introduced in 2018.

BERT is trained using masked language modeling. During training, tokens (usually words) are masked out or hidden from the model, and the transformer neural net is trained to guess the missing token based on the tokens in the surrounding context. As this process is repeated over millions of tokens, the language model effectively learns the probability distribution of the language. From there, the language model can be fine-tuned to perform a variety of downstream tasks such as text classification and extractive question answering.

In the current study, we rely on RoBERTa (Liu et al., 2019), an improved successor to BERT, which is pre-trained on massive amounts of text including all of English Wikipedia and 100GB of other web-crawled data from the Internet. We used transfer learning to adapt the RoBERTa model to our task. Transfer learning begins with a model pre-trained on a very large amount of generic data; the model is then fine-tuned by providing it with ancillary data specific to a particular downstream task—in this case classifying a particular post-lineup confidence statement as reflecting low, medium, or high confidence. This allows the fine-tuned model to leverage all the general semantic information gained during the pre-training step (e.g., that the word “very” denotes greater intensity of quality or belief), with customization towards performing a particular task (e.g., that the response “not very” maps onto low confidence).

Model Training

We used the Transformers library from HuggingFace to train our model on the classification task. The training data included three separate datasets: the pilot data as well as data from two previous studies (Dobolyi & Dodson, 2018; Grabman et al., 2019) using a similar methodology (Figure S1). Responses from the three studies were comingled and treated equally.

We randomized the training data and split it into training, test, and validation sets using roughly an 80%, 10%, 10% split. This approach, standard in machine learning, predicts an output for a given set of model parameters or weights, compares this to the known true output in the training dataset, and then updates the weights to iteratively achieve better classification accuracy. To assess how well the model is learning, we periodically tested performance using the test set, allowing us to gauge improvement and terminate the iterative learning process once we reached the model's maximum achievable classification accuracy. Finally, we used the validation set, containing the 10% of the initial training dataset the model has never seen, to measure model performance. Beyond providing a classification for each verbal confidence statement, the LLM also outputs a probability distribution over the three categories (i.e., probability that a given confidence statement is low, medium, and high confidence).

Across all the datasets, respondents often included numeric language in their verbal confidence statements (e.g., "I am 100% sure"). To help the model learn a numerical baseline, we performed data augmentation. We added 115 additional samples to only the training set that use numeric language in sample verbal confidence reports (see Table S2). This data augmentation ensured the model saw examples (both numeric and text) of different numbers and percentages to help it learn how to classify words associated with numeric confidence.

We used the Trainer API from HuggingFace to fine-tune RoBERTa, adding a randomly initialized classification head and training for 5 epochs at a learning rate of 5e-6. During the validation step, our final model correctly predicted 71% of the validation set. We define a correct or accurate classification as one in which the model's categorization matches the witness' intended level of confidence. That is, if a witness states they are "pretty certain" and defines this as 60% then the "ground truth" would be this witness is moderately confident. If the model outputs a categorization of moderate confidence, then this would be an accurate outcome.

Results

Model Performance

Table 1 reports the classification accuracy of the LLM when applied to four different external datasets not used in the model development or training process. These results illustrate the range of performance that might be expected when applying the model to new, unseen data. Tables S3-S6 provide full confusion matrices for each dataset.

Across the four external datasets, when classifying statements as low, medium, or high confidence, the LLM correctly classifies 71% of the confidence statements. When classifying confidence as either highly confident (75% or above) or not highly confident (less than 75%), the model performance improves to an average of 83% accuracy.

We believe that comparing accuracy to a perfect 100% rate is not appropriate for this model. The LLMs cannot achieve 100% accuracy because the classification process requires a *unique* mapping between a particular statement and an accuracy level. In reality, different participants may use the same words to describe different degrees of numeric confidence (e.g., participant 1 is 80% confident (high) and describes their confidence as "pretty confident" versus participant 2 who is 60% confident (medium) but also describes their confidence as "pretty

confident"). The final column of Table 1 reports the maximum possible accuracy achievable by any classification process for each dataset, taking into account the incidence of such conflicting responses. For three of the four external datasets, the LLM achieves greater than 75% of the maximum possible accuracy.

The main reason for measuring confidence is to permit inferences about the likely accuracy of a particular identification. In the laboratory, researchers often use confidence-accuracy calibration curves as a way of characterizing the relationship between confidence and accuracy for a given eyewitness task. Do the LLM-based classifications yield similar information about accuracy as would be available with direct, witness-reported numeric confidence? To examine this question, Figure 1 plots confidence-accuracy curves for each of the four external datasets based on true confidence measured numerically ("actual" dashed line with circles in blue) and imputed confidence measured by our model based on verbal confidence ("imputed" dotted line with squares in orange).

The external datasets exhibit a range of actual calibration patterns. Figures 1A and 1B show examples of tasks with weak calibration—in the true data, there is a modest (Figure 1A: 13.2 percentage points; Figure 1B: 13.1 percentage points) but statistically significant (Figure 1A: $p = .003$; Figure 1B: $p = .011$) increase in the likelihood of a correct identification for highly confident eyewitnesses as compared to those with medium or low confidence, but no measurable difference between low and medium confidence respondents (Figure 1A: $p = .144$; Figure 1B: $p = .610$). Figures 1C and 1D show examples of tasks with good calibration—the likelihood of a correct identification is monotonically increasing in the level of confidence, it is substantially higher for high-confidence respondents as compared to low-confidence respondents, and the

absolute rate of correct responses is high for highly confident respondents (Figure 1C: 86%; Figure 1D: 85%).

The model-imputed confidence classifications appear to perform well. True accuracy levels generally fall within the 95% confidence intervals of those estimated based on the LLM classification, and the LLM-based confidence-accuracy curves yield qualitatively comparable insights to the true curves, demonstrating similarly weak calibration in Figures 1A and 1B and good calibration in Figures 1C and 1D.

To more formally test for differences between the calibration curves, we estimated regression models where we predict accuracy using indicators for the LLM-based confidence levels (low, medium, and high) as primary predictors and indicators for the actual confidence levels as auxiliary predictors. We conducted an *F*-test for joint significance of the actual confidence levels, which in essence tests statistically whether the actual confidence levels provide any information about accuracy over and above what is available from the LLM model. For two of the datasets we failed to reject the null of no difference (Figure 1A, $F(2, 241) = 1.73$, $p = .180$, $\eta_p^2 = .014$; Figure 1B, $F(2, 113) = 0.38$, $p = .685$, $\eta_p^2 = .007$), while for two of the datasets, true confidence measured numerically did provide some additional explanatory power (Figure 1C, $F(2, 1673) = 54.28$, $p < .001$, $\eta_p^2 = .062$; Figure 1D, $F(2, 2028) = 39.49$, $p < .001$, $\eta_p^2 = .037$). For the two datasets where there are statistically significant differences, the qualitative differences remain modest—for example, the expected accuracy based on the LLM classification is always within 8 percentage points of the actual accuracy based on true confidence, despite the fact that accuracy varies by over 40 percentage points across confidence levels.

To examine whether our results are particular to our choice of RoBERTa as the base LLM, we also fine-tuned OPT (Zhang et al., 2022), another transformer-based language model,

using the same data. OPT, developed by Meta AI, has a similar architecture to the GPT class of models and can incorporate substantially more parameters than RoBERTa (1.3 billion versus 125 million). Rather than predicting what word should fill in a mask, as in RoBERTa's masked language modeling, OPT is simply trained to predict the next word in a given sequence. Our results using OPT are statistically and qualitatively similar to those achieved with RoBERTa (see Table S7), demonstrating the robustness of our results.

To further probe the real-world usefulness of the model, we applied it to the 35 confidence phrases extracted by Behrman and Davey (2001) and Behrman and Richards (2005) from statements from actual eyewitnesses obtained by the Sacramento Police Department during the investigation of 183 real criminal cases. Behrman and Richards (2005) employed 84 human coders to classify these statements into low, medium, and high confidence categories. Our model's categorization of these 35 statements is shown in Figure 2. Although Behrman and Richards (2005) used different confidence level cutoffs from the present study—meaning that conventional confusion matrices would likely be uninformative—we can assess statistically whether our model replicates human interpretations by conducting a MANOVA test where the outcomes are our three model-generated categories and the main predictors are the Behrman and Richards (2005) low, medium, and high groupings. This tests for whether the LLM assigns systematically different probability ratings to statements categorized differently by humans. For the overall joint test (Wilks' lambda = .35; $F(6, 60) = 6.93$; $p < .001$) and for each individual categorical comparison (low vs. medium: $F(1, 32) = 10.29$; $p = .003$; high vs. medium: $F(1, 32) = 13.80$; $p < .001$; high vs. low: $F(1, 32) = 46.72$; $p < .001$) we reject the null of no difference, demonstrating that when humans distinguish particular real-world eyewitness confidence statements, the model also distinguishes them.

Confidence Entropy

Our model also furnishes a new metric containing independent information about eyewitness confidence. Adapting a concept from information theory, we define the *confidence entropy* of a particular statement as $-\sum_{k\in\{low,medium,high\}} p(k)log_2(p(k))$ where *p(k)* represents the probability that a given statement is low, medium, or high confidence, and the relevant term evaluates to 0 when *p(k)=0*. Confidence entropy in essence measures the vagueness of the confidence statement; a statement belonging to one of the categories with probability 1 would have zero entropy, representing complete clarity [0 + 0 - 1· $log_2(1)$], whereas a fully diffuse probability distribution (1/3 for each category) has a confidence entropy of 1.58 $\left[-\frac{1}{3}log_2\left(\frac{1}{3}\right)-\frac{1}{3}log_2\left(\frac{1}{3}\right)-\frac{1}{3}log_2\left(\frac{1}{3}\right)\right]$. An example of a confidence statement with very low entropy is "I remember the guy's face. It is him for sure. I am completely, 100% confident." The model assigns a 99% probability this statement is highly confident, 1% probability of medium confidence, and <1% probability of low confidence for an entropy value of 0.10. On the other hand, an example of a confidence statement with very high entropy is "I don't recognize any of them, I got a decent look at the person but I don't see him. But it's absolutely possible he's there." The model assigns a 20% probability this statement is highly confident, a 42% probability the statement is moderately confident, and a 37% probability this statement is low confidence for an entropy value of 1.52.

As shown in Figure 3, which depicts a histogram of confidence entropy measures for statements that are low, medium, and high confidence drawn from the combined four external datasets (*N* = 4,541), confidence entropy is a distinct concept from confidence itself. It is possible, for example, for someone to express high certainty with little ambiguity in interpretation (e.g., "I'm 100% certain") or for someone to express high certainty but in a way

that leaves more room for interpretation (e.g., “I believe I saw this one.”). Whereas confidence level provides information about whether the participant believes they made an accurate identification, confidence entropy measures how well they feel they can assess their accuracy. Table S8 provides examples of actual confidence statements that exhibit differing combinations of confidence level and confidence entropy.

To assess whether entropy provides useful additional information about eyewitness accuracy over and above the confidence level itself, Figure 4 plots confidence-accuracy curves for the two well-calibrated test datasets (Grabman & Dodson, 2022; Smalarz et al., 2021) that also differentiate between responses with low, medium, and high confidence entropy (defined by terciles). For high confidence identifications, confidence entropy appears clearly related to accuracy, with higher entropy (i.e., vaguer) responses associated with lower accuracy. To more formally test the predictive power of entropy, we estimated regressions identical to those described previously (i.e., with a full set of interactions between true and predicted confidence levels) as main predictors, but with confidence entropy as an auxiliary predictor. For both datasets, after conditioning on the confidence level, entropy was negatively and statistically significantly related to accuracy (Smalarz et al. (2021), $F(1, 1668) = 5.39$, $p = .020$, $\eta_p^2 = .003$; Grabman & Dodson (2022), $F(1, 2489) = 39.49$, $p < .001$, $\eta_p^2 = .016$).

To examine whether entropy can add predictive power relative to current best practices, we estimated regression models using these two datasets where the outcome was a correct identification, and the predictors were a full set of indicators for all numeric confidence levels reported by participants. Saturating the model with predictors in this manner incorporates all possible information obtainable from self-reported numeric confidence. We then entered entropy as an additional predictor and tested its significance. For the Smalarz et al. (2021) dataset, which

allows respondents to make a nuanced reporting of numeric confidence (i.e., 0-100%), we find that entropy remains negative and statistically significantly associated with confidence even after fully controlling for numeric confidence (Smalarz et al. (2021), $F(1, 1613) = 3.86$, $p = .049$, $\eta_p^2 = .002$; Grabman & Dodson (2022), $F(1, 2492) = 0.42$, $p = .519$, $\eta_p^2 < .001$).

Together, these analyses suggest that confidence entropy—a measure unavailable previously, but now readily producible by applying natural language processing to confidence statements–merits further investigation as a potential new reflector variable that can be used to better characterize eyewitness accuracy.

Discussion

Under appropriate conditions, eyewitness confidence measured at the time of an identification procedure can be a valuable diagnostic cue for identification accuracy (Wixted & Wells, 2017). Moreover, the U.S. Supreme Court has specifically identified a witness's confidence at the time of the identification as a relevant factor for courts to consider in evaluating the admissibility of lineup evidence (Neil v Biggers, 1972). However, when lineups are administered in the field, eyewitnesses are rarely asked to provide numeric or other structured ratings of their confidence (Police Executive Research Forum, 2013). Until now there has been no efficient, systematic, reproducible method to interpret verbal or textual descriptions of confidence. Our Transformer-based LLM accomplishes this in a manner that largely reproduces the categorization one would obtain had the eyewitness been asked to rate their confidence numerically. Moreover, our LLM-based categorization provides similar information about eyewitness accuracy as would be obtainable with a numeric confidence measure as well as providing the new metric of confidence entropy.

Our work extends the existing literature about the relationship between confidence and accuracy across confidence scale types. Past work has shown similar confidence-accuracy relationships between different numeric confidence scales (Tekin & Roediger, 2017), between numeric and graded verbal scales (Weber et al., 2008), and between numeric and freely reported verbal confidence (Smalarz et al., 2021). Our model replicates this pattern of findings showing a similar confidence-accuracy relationship between the model's confidence classification and participants' self-reported numeric confidence.

Our model has several practical applications. Initial confidence recorded from an unbiased lineup is predictive of accuracy (Wixted & Wells, 2017). Most known misidentification cases had an eyewitness who testified at trial they were highly confident in their identification, but were not highly confident at the time of the lineup (Garrett, 2011). We believe our LLM is an efficient, low-cost solution to help officers better understand a witness' initial confidence statement. Outside evaluators often have differing evaluations of verbal confidence statements (Greenspan & Loftus, 2024). Our model provides a way for officers to reliably, simply, and replicably interpret the intended meaning of a witness' initial confidence statement. In situations where the identification procedure is video recorded—a recommended best practice in lineup administration (Wells et al., 2020)— or recorded verbatim in writing, the LLM could categorize the witness's description of their confidence at any point in the future—including, potentially many years after the original procedure—and then evaluate that statement free from contextual bias. The model also provides a way to adjudicate ambiguous cases where human coders may disagree as to whether a particular statement denotes high confidence ("I'm thinking I'm right") by essentially leveraging a large body of data from our training data. The model also offers a linguistically informed method to infer confidence when eyewitnesses offer unusual or

unexpected statements that might be difficult for humans to interpret (“that’s a clown question, bro”).

Confidence, and now confidence entropy, are two reflector variables that can help indicate witness accuracy. One additional reflector variable that might be at play here is decision time. In addition to high confidence, fast identification decisions are predictive of witness accuracy (Quigley-McBride & Wells, 2023). Future research using this LLM model could explore the interplay of witness confidence, confidence entropy, and decision time to further understand factors related to accurate and inaccurate identifications.

The model also has considerable potential to support the expansion of academic research on verbal confidence statements. One significant impediment to experimental research is that any quantitative analysis of verbal confidence statements has traditionally required researchers to hire and train human coders, who then review each verbal statement and classify it manually. This process is expensive, time-consuming, and non-replicable across coders. The LLM can process thousands of confidence statements almost instantaneously, and the version of the model at https://huggingface.co/spaces/psheaton/eyewitness_confidence_classifier includes functionality to accept file uploads for batch processing of statements. We anticipate that this model should substantially reduce the cost and complexity of coding verbal confidence statements, thus removing barriers for researchers to use the more ecologically valid measure of verbal confidence in their studies. Moreover, the LLM can readily improve over time both as the underlying language model is upgraded and through incorporating data from additional studies into the training process. Whereas human coding typically requires starting anew with a fresh set of coders for each study, the LLM can draw from the accumulated knowledge embedded in

thousands or even tens of thousands of identification responses generated by multiple researchers.

Acknowledgments

We acknowledge and thank Amanda Bergold, Chad Dodson, Jesse Grabman, Jillian Kenchel, and Laura Smalarz for sharing their data for this study.

References

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior*, *25*(5), 475–491. https://doi.org/10.1023/A:1012840831846

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, *29*(3), 279–301. https://doi.org/10.1007/s10979-005-3617-y

Bergold, A. N., & Heaton, P. (2018). Does filler database size influence identification accuracy? *Law and Human Behavior*, *42*(3), 227–243. https://doi.org/10.1037/lhb0000289

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404. https://doi.org/10.1016/0749-5978(88)90036-2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/ARXIV.1810.04805

Dobolyi, D. G., & Dodson, C. S. (2018). Actual vs. Perceived eyewitness accuracy and confidence and the featural justification effect. *Journal of Experimental Psychology: Applied*, *24*(4). https://doi.org/10.1037/xap0000182

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior*, *39*(3), 266–280. https://doi.org/10.1037/lhb0000120

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press. https://doi.org/10.4159/harvard.9780674060982

Grabman, J., & Dodson, C. (2022). *Unskilled, Underperforming, or Unaware? Testing Three Accounts of Individual Differences in Metacognitive Monitoring Sensitivity*. https://doi.org/10.17605/OSF.IO/4XZNC

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, *8*(2), 233–243. https://doi.org/10.1016/j.jarmac.2019.02.002

Grabman, J. H., & Dodson, C. S. (2019). Prior knowledge influences interpretations of eyewitness confidence statements: 'The witness picked the suspect, they must be 100% sure'. *Psychology, Crime & Law*, *25*(1), 50–68. https://doi.org/10.1080/1068316X.2018.1497167

Greenspan, R. L., & Loftus, E. F. (2024). Interpreting eyewitness confidence: Numeric, verbal, and graded verbal scales. *Applied Cognitive Psychology*, *38*(1), e4151. https://doi.org/10.1002/acp.4151

Kenchel, J. M., Greenspan, R. L., Reisberg, D., & Dodson, C. S. (2021). "In your own words, how certain are you?" Post-identification feedback distorts verbal and numeric expressions of eyewitness confidence. *Applied Cognitive Psychology*, *35*(6), 1405–1417. https://doi.org/10.1002/acp.3870

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://doi.org/10.48550/ARXIV.1907.11692

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. The National Academies Press.

Neil v Biggers, 409 U.S. 188 (1972).

Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. https://www.ojp.gov/pdffiles1/nij/grants/242617.pdf

Quigley-McBride, A., & Wells, G. L. (2023). Eyewitness confidence and decision time reflect identification accuracy in actual police lineups. *Law and Human Behavior*. https://doi.org/10.1037/lhb0000518

Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2021). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001152

Slane, C. R., & Dodson, C. S. (2022). Eyewitness confidence and mock juror decisions of guilt: A meta-analytic review. *Law and Human Behavior*, *46*(1), 45–66. https://doi.org/10.1037/lhb0000481

Smalarz, L., Yang, Y., & Wells, G. L. (2021). Eyewitnesses' free-report verbal confidence statements are diagnostic of accuracy. *Law and Human Behavior*, *45*(2), 138–151. https://doi.org/10.1037/lhb0000444

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*(1), 1–13. https://doi.org/10.1186/s41235-017-0086-z

The National Registry of Exonerations. (n.d.). *Interactive data dispaly*. Retrieved June 14, 2023, from https://www.law.umich.edu/special/exoneration/Pages/Exonerations-in-the-United-States-Map.aspx

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: A meta-analysis. *Journal of Risk Research, 5*(2), 177–186. https://doi.org/10.1080/13669870110038179

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. https://doi.org/10.48550/ARXIV.1706.03762

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, *25*(5), 571–587. https://doi.org/10.1016/0749-596X(86)90012-4

Weber, N., Brewer, N., & Margitich, S. (2008). The confidence-accuracy relation in eyewitness identification: Effects of verbal versus numeric confidence scales. In K. H. Kiefer (Ed.), *Applied Psychology Research Trends* (pp. 103–118). Nova Publishers.

Wells, G. L. (2020). Psychological science on eyewitness identification and its impact on police practices and policies. *American Psychologist, 75*(9), 1316–1329. https://doi.org/10.1037/amp0000749

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*(1), 3–36. https://doi.org/10.1037/lhb0000359

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). *OPT: Open Pre-trained Transformer Language Models*. https://doi.org/10.48550/ARXIV.2205.01068

Table 1

LLM Model Classification Accuracy for External Datasets

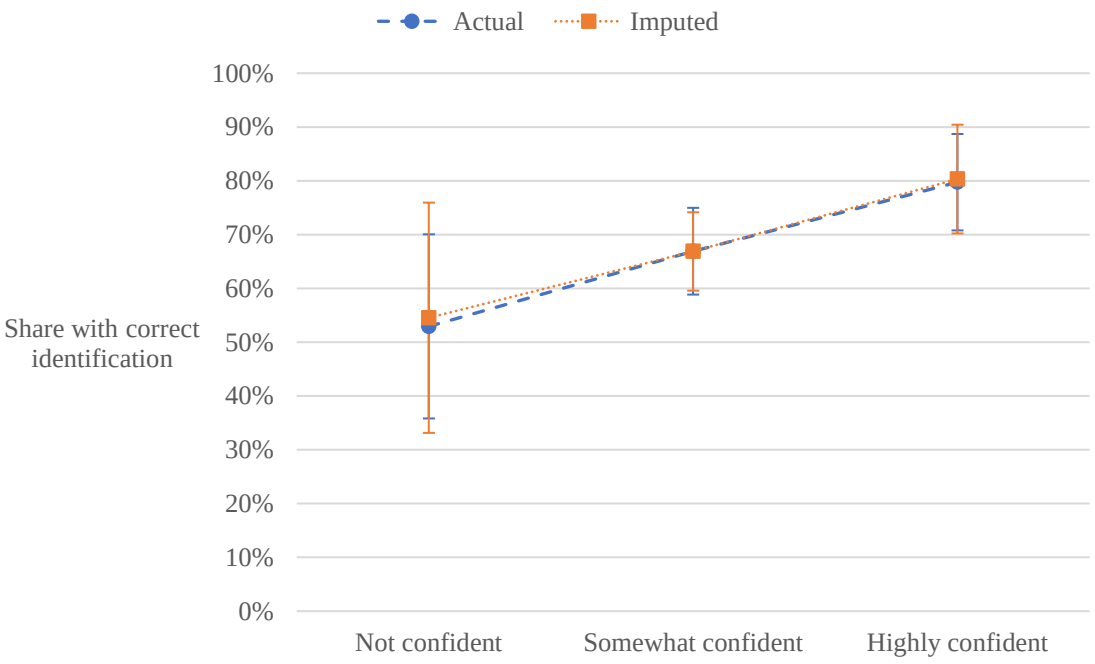
Dataset	*N*	Classification accuracy Low/Medium/High	Classification accuracy Not High/High	Practical maximum
Bergold & Heaton (2018)	246	65.9% (59.9% - 71.8%)	77.6% (72.4% - 82.9%)	99.5%
Kenchel et al. (2021)	118	72.9% (64.8% - 80.9%)	77.1% (69.5% - 84.7%)	97.5%
Smalarz et al. (2021)	1,678	68.7% (66.4% - 70.9%)	83.6% (81.8% - 85.3%)	90.4%
Grabman & Dodson (2022)	2,499	72.6% (70.9% - 74.4%)	83.7% (82.2% - 85.1%)	85.4%
Overall	4,541	70.8% (69.5% - 72.1%)	83.1% (82.0% - 84.2%)	

Note: 95% confidence intervals are reported in parentheses.

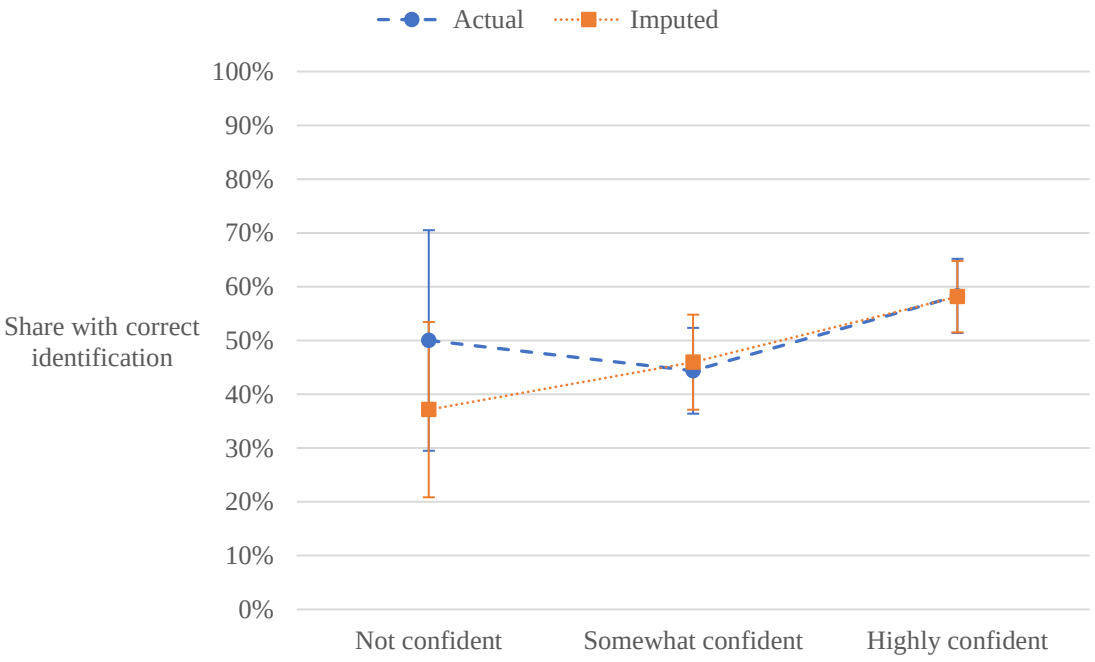
Figure 1A-D

Confidence-Accuracy Curves for Four External Datasets

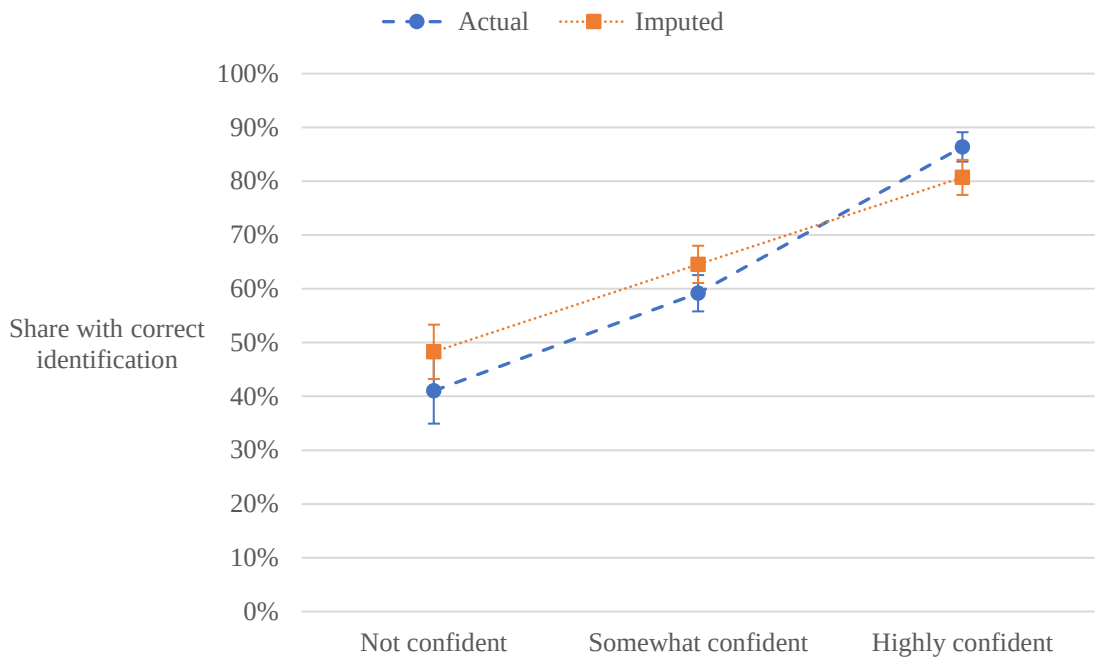
A. Bergold & Heaton (2018)



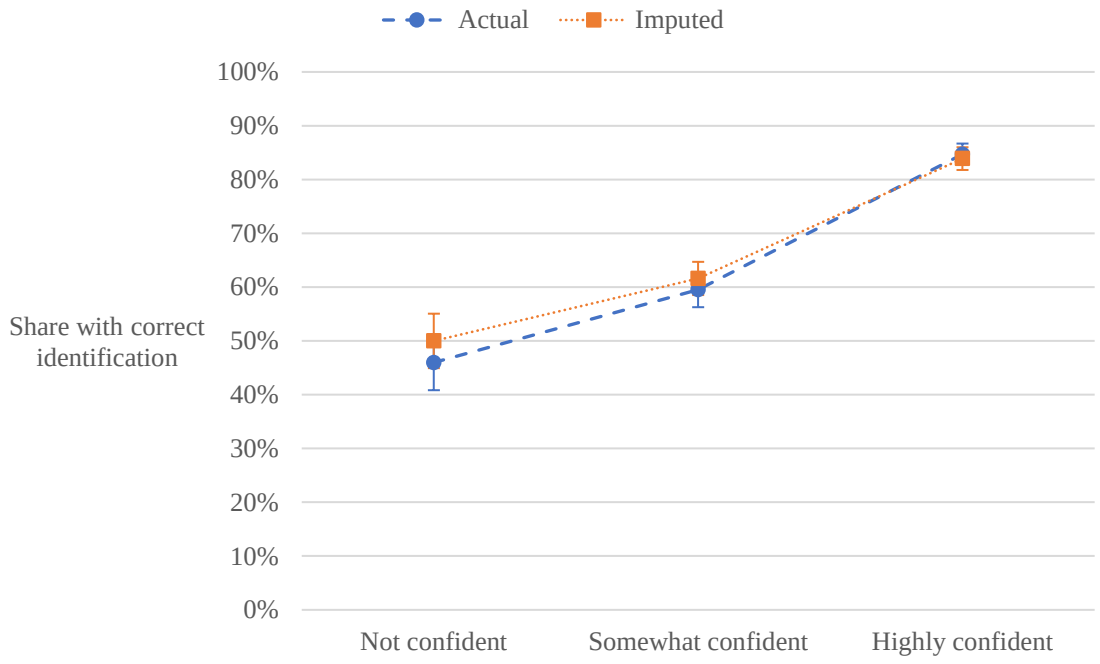
B. Kenchel et al. (2021)



C. Smalarz et al. (2021)



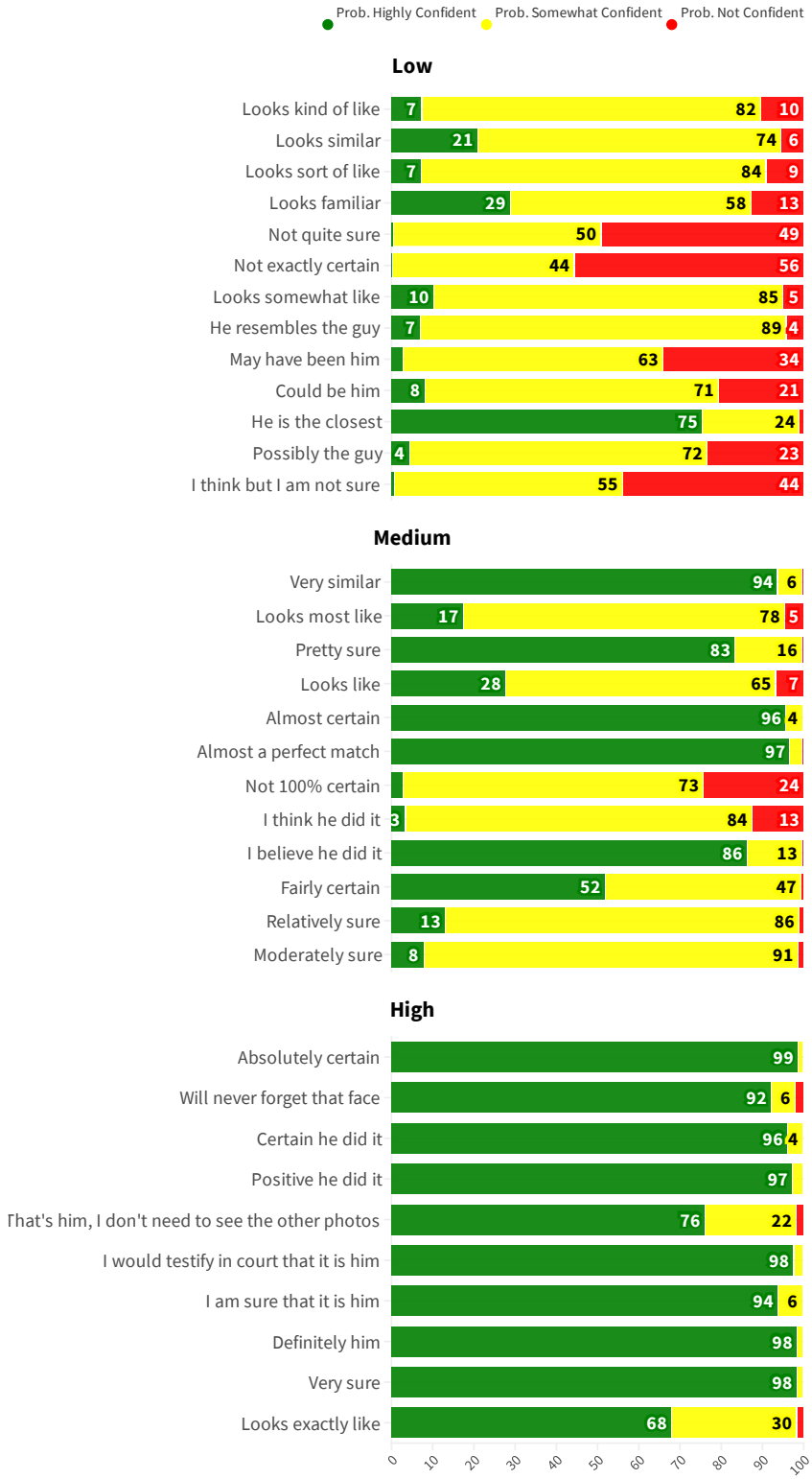
D. Grabman & Dodson (2022)



Note: Whiskers denote 95% confidence intervals for each accuracy level.

Figure 2

Model Output for 35 Confidence Statements from Real Witnesses from Behrman and Richards (2005) by Human Categorization



Note: Low, medium, and high categorizations were made by 84 human coders as described in Behrman and Richards (2005). The LLM model probabilities for low, medium, and high confidence are depicted with the green, yellow, and red bars in the figure. The LLM's final classification would be given by the longest of the three bars. Note Behrman and Richards (2005) define low confidence as 0-4, medium confidence as 5-7, and high confidence as 8-10.

Figure 3

Entropy Distribution by Self-Reported Numeric Confidence Level

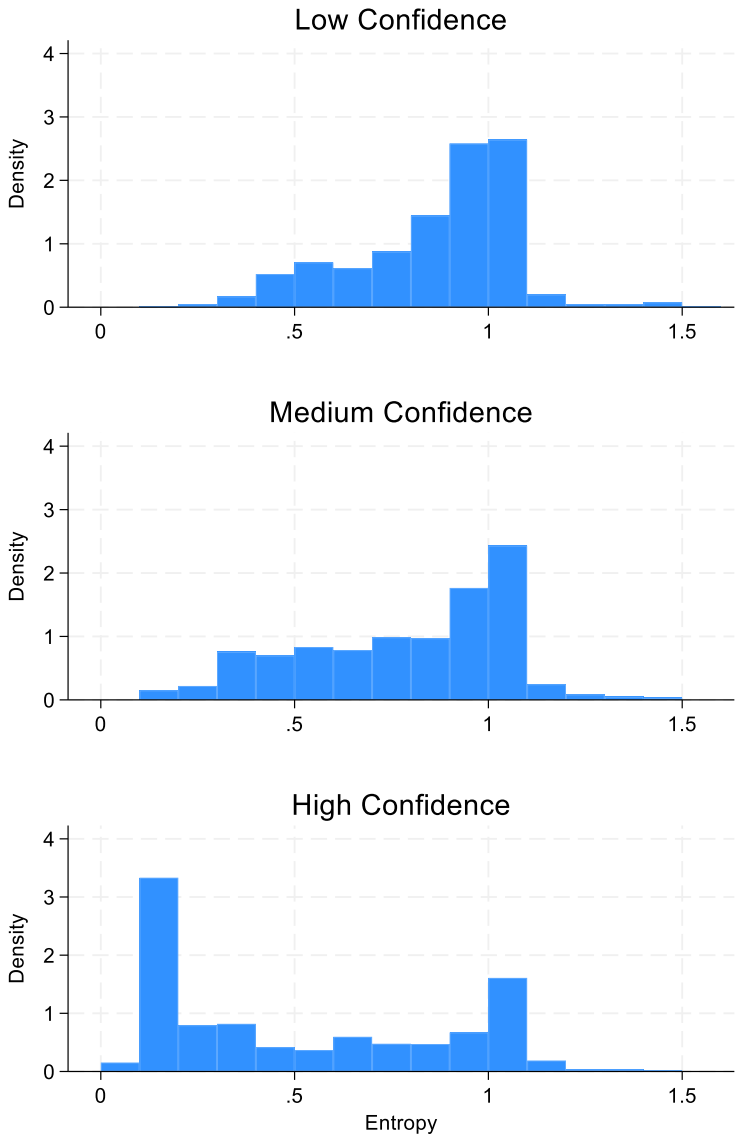
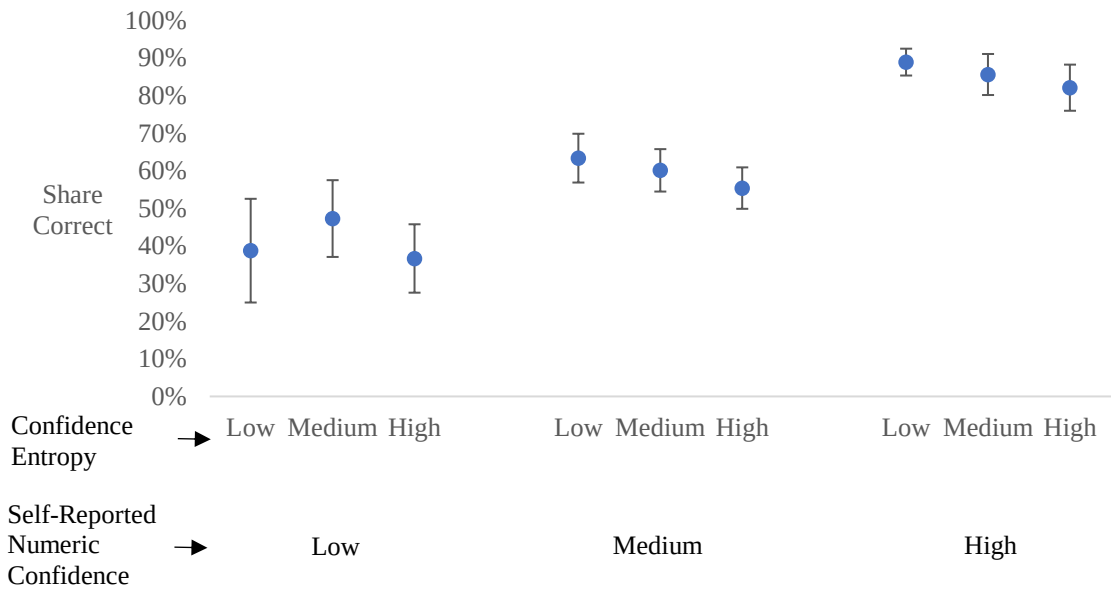


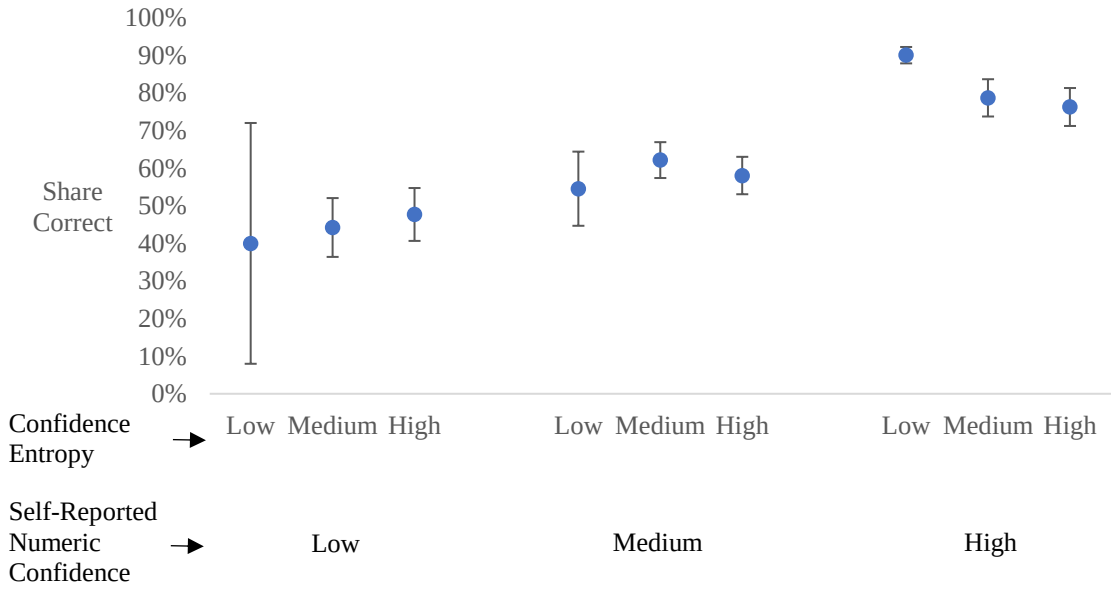
Figure 4A-B

Eyewitness Accuracy by Confidence Level and Confidence Entropy

A. Smalarz et al. (2021)



B. Grabman & Dodson (2022)



Note: Whiskers denote 95% confidence intervals for each accuracy level.