LLMs as Lie Detectors: Criteria-Based Veracity Assessment of Verbal Statements in Swedish

Anonymous ACL submission

Abstract

Determining the veracity of a verbal statement is a frequent and challenging task for legal practitioners. The Swedish Supreme Court has proposed a set of truth-indicative reliability criteria for assessing verbal statements in the Swedish legal system. LLMs' growing ability to reason and handle complex tasks raises questions about how they perform veracity judgments and how they would implement criteria-based statement analysis. In this study, we explore LLMs' ability to perform veracity judgments on a novel Swedish mock-crime dataset, focusing on accuracy, truth bias, confidence-accuracy relations, and how criteria analysis affects performance. We compare simple prompting to a two-step method in which the LLM is first asked to assess the reliability criteria and then perform a criteria-informed veracity judgment. We find that accuracy is at chance level under both conditions – which is similar to human performance in previous investigations – but that the two-step method often leads to a more reasonable truth bias and reduces the over-confidence of the models. Further analysis indicates that criteria assessments do influence final judgments but do not correlate with ground truth.

1 Introduction

Analyzing verbal statements to determine their veracity constitutes a major challenge for practitioners in the legal system. Humans perform poorly at lie detection. Meta-analyses suggest that humans, including police officers, detectives, and judges, achieve an average accuracy of 54% (Bond and DePaulo, 2006; Aamodt and Custer, 2006). Humans also tend to overestimate their ability (Loconte et al., 2026). At the same time, verbal statements are important evidence in every criminal case. Especially when technical evidence is scarce or nonexistent, evaluating verbal statements becomes pivotal to the trial's outcome.

Research has shown that focusing on evaluating verbal cues – the language of the statement – is the most promising method for making veracity judgments more accurately (Brennen and Magnussen, 2023; Bogaard et al., 2025). Based on this notion, the Swedish Supreme Court (SSC onward) has set out a set of linguistic *reliability criteria* that say that a statement is more likely to be true if it is **clear**, **detailed**, **long**, or **logical**, and **vivid**. Other criteria-like concepts that are often used for arguing credibility in judicial decisions are whether the statement seems **reasonable** and **self-experienced**. These criteria stem from frameworks designed to aid veracity judgments, such as content-based criteria analysis (CBCA, Köhnken, 2004) and reality monitoring (RM, Johnson and Raye, 1981).

Traditionally, NLP research in the area of verbal lie detection has been centered around stylometric analysis of linguistic cues (Newman et al., 2003; Levitan et al., 2018; Loconte et al., 2023). However, with the rise of LLMs as complex reasoning analysts, the potential of LLM applications in this area increases dramatically. Previous work has shown potential in LLMs for text-based veracity judgments (Hazra and Majumder, 2024; Miah et al., 2025). There have also been signs of strong truth bias (Markowitz and Hancock, 2024) and lie bias (Markowitz and Levine, 2025). To the best of our knowledge, no previous study has explored structured criteria analysis for veracity judgments with LLMs. Furthermore, most previous studies have focused on English-language datasets of autobiographical memories (Loconte et al., 2023) or opinions about friendships (Markowitz and Levine, 2025), rather than on data that resembles a crime scenario.

In this study, we explore LLMs for veracity judgments in a forensic investigation context using the reliability criteria from the SSC, on a novel Swedish mock-crime dataset.[1] We pose the follow-

[1]The data set will be released upon publication.

ing research questions:

1. How do LLMs perform veracity judgments?
   (a) How accurate are the judgments?
   (b) To what degree do LLMs exhibit truth bias?
   (c) Do confidence and accuracy correlate?
2. How does criteria analysis affect performance?
3. How strong are the criteria assessments as predictors for final judgment and ground truth?

Our results indicate that LLMs resemble humans in performing at chance level while being overly confident. Adding criteria analysis does not improve accuracy but often leads to a better calibrated truth bias and somewhat lower confidence estimates. Our analysis also indicates that criteria assessments do influence final judgments but do not in themselves correlate with ground truth.

2 Background and Related Work

2.1 Veracity Judgments in Legal Contexts

Determining whether someone is lying or telling the truth is a complex task. Meta-analyses suggest that the accuracy of human lie detection lies around 54%, and that legal practitioners such as police officers, detectives, and judges are no better than laypeople (Bond and DePaulo, 2006; Aamodt and Custer, 2006), despite them often believing otherwise (Loconte et al., 2026).

Making veracity judgments constitutes one of the most challenging tasks that legal practitioners are destined to perform. To make judgments more fair and accurate, practitioners in Sweden are encouraged to evaluate the reliability of the statement rather than focusing on nonverbal behavior, such as eye gaze aversion, or other scientifically inaccurate indicators of veracity (Brennen and Magnussen, 2023; Bogaard et al., 2016), using the SCC *reliability criteria* (Sw. *tillförlitlighetskriterier*). These criteria are presented through precedents from cases in which the assessment of verbal statements played a significant role. These are criteria that, if fulfilled, are said to indicate that the statement is more likely to be true and that the person – plaintiff, witness, or defendant – is more likely to be telling the truth. The list includes **clear**, **detailed**, **long**,[2] and **logical**, and **vivid**,[3] among others.

[2]Precedent from the SSC, Case nr: B 4663-16 (I) and B 4648-16 (II), 2017.

[3]Precedent from the SSC, Case nr: B 2937-10, 2010.

The reliability criteria are associated with other criteria-based frameworks that were designed to aid veracity judgments, such as content-based criteria analysis (CBCA, Köhnken, 2004) and reality monitoring (RM, Johnson and Raye, 1981). The diagnosticity of linguistic criteria, like the reliability criteria, has been evaluated quite extensively using both manual coding of the criteria in verbal statements (Willén and Strömwall, 2012), and with computational stylometric analysis using Linguistic Inquiry and Word Count (LIWC, Pennebaker et al., 2001) or other software (Hauch et al., 2015). Meta-analyses show that the predictive power of linguistic criteria for veracity judgments is generally weak, and may point in different directions; the same criterion can be indicative of both truth and deception across studies (Hauch et al., 2015; Gancedo et al., 2021).

As with the majority of CBCA and RM criteria, the reliability criteria are solely truth-indicative. This can lead to increased truth bias if many criteria are met (Sporer and Masip, 2023). Truth bias is one of the main reasons people perform poorly in deception detection. Truth bias is described by Levine's truth-default-theory (TDT, Levine, 2014). TDT analyzes lying from a communicative perspective. The point is that humans are gullible in that we expect others to tell the truth. This relates to Grice's Cooperative Principle, especially the maxim of quality, which commands speakers to be truthful (Grice, 1975).

The correlation between confidence and accuracy in human decision-making is a complex question in lie detection research. There is evidence both against (DePaulo et al., 1997) and for a significant correlation that means humans are more confident when they are correct (Smith and Leach, 2019).

2.2 Veracity Judgments by LLMs

Previous research on LLMs for veracity judgments has explored traditional fine-tuning (Loconte et al., 2023), enhancing deception detection using RAG (Boumber et al., 2024), prompting with concept analysis (Hazra and Majumder, 2024), and vision-language models for multimodal classification (Miah et al., 2025). Miah et al. (2025) also explored LLMs in text-based classifications, where fine-tuning Llama-3.1-8b yielded 92% accuracy on an opinion spam dataset. Hazra and Majumder (2024) implemented a prompting strategy using

concept analysis to identify the liar among three participants in a game show. Their analysis showed that LLMs could correctly identify cases in which human judges had failed, demonstrating their potential for this task.

Previous research has also found truth bias in LLMs. Markowitz and Hancock (2024) experimented with GPT-4 and found it to be even more truth-biased than humans overall (99%). Prompting skepticism and providing base rates reduced the truth bias, but it remained higher than in humans. In a second study on Gemini-1.5-Flash, the truth bias was prominent for some datasets (e.g., opinions about friendships), but was replaced by strong deception bias in others (e.g., cheating in a game) (Markowitz and Levine, 2025). We have not found any previous research on the relation between confidence and accuracy for LLM veracity judgments.

A tangent research field is that of the detection of lies produced by LLMs. Like humans, LLMs can produce statements they know are false. There is research on methods for detecting lies in LLM output and behavior. One stream of research centers on using LLM activations train a classifier that then successfully separates truthful from deceptive activations (Zou et al., 2023; Bürger et al., 2024), achieving 94% accuracy in experiments. Another proposed method is to train a classifier on the LLM's responses to unrelated questions it gets prompted with after the lie (Pacchiardi et al., 2024). The fact that LLMs can lie in ways similar to humans underscores the importance of researching LLMs for lie detection, even as lie detectors.

3 Experimental Design

The experimental design of this study consists of prompt-based experiments on a selection of small and large LLMs. The purpose is to explore how LLMs perform veracity judgments and how their performance is affected when forced to consider the SSC's reliability criteria. We will therefore compare one-step prompting, which asks for a direct veracity judgment and confidence assessment, to two-step prompting, where the first prompt asks for criteria assessments and the second prompt asks for a veracity judgment and confidence assessment based on the criteria assessments. This design allows us to compare accuracy, truth bias, and correlation between confidence and accuracy under both conditions, and to investigate whether criteria assessments correlate with judgments or with ground truth.[4]

3.1 Data

The dataset used in this study is a novel Swedish mock-crime dataset of verbal statements, which constitutes a significant contribution. Upon publication, we will release the dataset under a Creative Commons license. The instructions and stimuli used to collect the data, along with the data splits, will be made available with the dataset.

The data consist of 251 statements, of which 40 were used for development and 211 in the main experiments. The statements are in Swedish and collected through an online questionnaire-based mock crime experiment hosted in Qualtrics.[5] Participants were reached by distributing the questionnaire to technical and administrative staff within the research organization, through social media and Prolific.[6] The participants were allotted one of two point-of-view video stimuli to watch. One video showed a person stealing a parcel from an open apartment post room, and the other showed a person noticing the parcel but not stealing it. The participants were told to imagine they were the person in the video. In the following, the police knock on their door and ask them to give their statement regarding the alleged crime. Half were instructed to lie, and the other half to tell the truth about their actions. This resulted in a dataset with 65 true confessions, 64 false confessions, 67 true denials, and 55 false denials.

3.2 Prompting Strategies

The prompts were developed and polished to ensure adequate output format and face-value compliance on Llama-3.3-70b via the Groq API, and on the held-out development set of 40 evenly distributed statements from the original dataset. The remaining 211 statements were used for the end experiments. The full prompts are available in Appendix A.1. On the note of prompting strategies, research has shown that imposing output-format constraints may affect the reasoning abilities of the LLM (Tam et al., 2024), Because of this, we let the model output its most comfortable format, which is markdown or plain text. The prompts are formalized with the intent not to constrain the model

[4] Code for experiments and evaluation will be published on GitHub.

[5] qualtrics.com

[6] prolific.com

too much regarding structure. The prompts were written in English, although the statements are in Swedish. We experimented with Swedish prompts in a pilot study on the development data but found no interesting differences beyond an increase in token consumption. For this reason, we opted for English prompts.

3.2.1 One-Step Prompting

The prompt for direct veracity judgments, which we view as our baseline, starts with role assignment (**You are a professional criminal investigator working on a theft case**) followed by background information about the case, and the statement to be judged. The model is then asked to make a binary veracity judgment **(T/F)**, provide a confidence score **0–1**, and provide a short rationale to support its judgment.

3.2.2 Two-Step Prompting

The two-step prompting strategy was inspired by Hazra and Majumder (2024)'s two-step analysis, in which the LLM is first asked to evaluate the statement based on a set of concepts and then to make a veracity judgment.

Step 1: Criteria Analysis In the criteria analysis step, the model is given the same role and background information as in the baseline prompt, but instead of making a veracity judgment, the model is asked to perform criteria analysis. The model is asked to read and analyze the statement and to provide a **0-100** fulfillment score for seven criteria: **clear, detailed, long, vivid, logical, self-experienced, reasonable**, followed by a brief rationale to support the scoring.

Step 2: Criteria-Based Judgment In the judgment step, the prompt is almost identical to the baseline prompt, except that the statement is accompanied by the criteria analysis. The criteria analysis is the full output from the criteria-analysis step: the criteria fulfillment scores and the rationale. For the judgment instruction, the model is told to read the statement and the criteria analysis, and decide whether the statement is true or false. As in the baseline setting, the model is also asked to provide a confidence score and a brief rationale to support its judgment.

3.3 LLM Choice and Hyperparameters

To increase the likelihood of producing generalizable results, we included a wide range of LLMs. The following models were used through the Groq API:[7] Llama-3.1-8b, Llama-3.3-70b (Meta AI, 2024), Qwen3-32b (Qwen Team, 2025), GPT-OSS-20b, GPT-OSS-120b (OpenAI, 2025). To extend the analysis to SoTA models, we employed Gemini-3.1-Flash-Lite (Google, 2026) through Google GenAI SDK:[8].

Gemini models are preview versions and were run in May 2026. For more details on specific dates, please contact the authors. For all models, max tokens was set to 4000 and `top_p` to 1. The temperature was set to 0.1 for direct veracity judgment and criteria-based veracity judgment, and 0.6 for the criteria analysis. GPT and Gemini models were used with low and high reasoning effort, resulting in nine different model setups for the end experiments. Groq models were run 5 times, Gemini models were run 3 times. GPT-OSS-20b-high suffered from degenerate generation and looping behavior, resulting in 293 analyzable samples – a significantly smaller sample than for the others. Llama-3.1-8b, Llama-3.3-70b and GPT-OSS-20b-low yielded 1055 analyzable samples each. Qwen3-32b yielded 1050, GPT-OSS-120b-low yielded 1043, and GPT-OSS-120b-high yielded 1044. Gemini models cleared 633 samples each over three runs.

3.4 Evaluation

The aim of this study is to explore how LLMs perform veracity judgments, whether criteria analysis affects performance, and whether criteria analysis predicts final judgments and ground truth. To assess LLM veracity judgment performance, we evaluate the accuracy and truth bias. Truth bias is measured as the relative frequency of the label **T** in model predictions, indicating how prone the model is to believing a statement. A truth bias below 50% indicates that the model preferred the label **F**. For correlation analyses of confidence and accuracy, we use point-biserial correlations with Holm-Bonferroni correction. To assess the effect of criteria analysis on performance, we compare the performance of the direct veracity judgments (one-step prompting) and the criteria-based veracity judgments (two-step prompting) using McNemar's test with Holm-Bonferroni correction.

To assess the relationship between criterion assessments and final judgments, we use multiple

[7] https://console.groq.com/home
[8] https://ai.google.dev/gemini-api/docs/libraries

logistic regression to estimate weights and pseudo R2 scores that indicate how well the final judgments are explained by the criterion scores. We perform the same analysis against the statements' ground truth to explore the diagnosticity of the criteria and whether the LLMs give adequate weight to each criterion, or whether there is an alignment gap between the criterion's actual effect size and how much the LLMs utilize it in their judgments.

4 Results

The experimental results are presented in the following three subsections, each corresponding to one of our three main research questions. All numerical results can be found in Table 1.

4.1 RQ1: Basic LLM Performance

The results from our direct judgment experiments (one-step prompting) suggest that LLMs perform veracity judgments with an accuracy around chance level, which is poor but comparable to human accuracy. As shown in Table 1, the accuracy scores range from 45 to 53 percent.

The truth bias varies widely across models. The highest degree of truth bias was observed in predictions from GPT-OSS-20b-high, where around 90% of the statements were deemed true. At the most skeptical extreme, we find Llama-3.1-8b, which identified only about 9% of the statements as truthful. A bias toward the label **F** is also found for Qwen-32b, GPT-OSS-20b-low, GPT-OSS-120b-low, and Gemini-3.1-Flash-Lite-low.

Average confidence scores range from 0.75 to 0.85, indicating that the models are overconfident relative to their accuracy. Two models show a statistically significant correlation between confidence and accuracy in direct judgments: GPT-OSS-20b-low (r= 0.1050, p=0.00383) and Gemini-3.1-flash-lite-low (r=0.1543, p=0.00068). Even though the correlation is quite weak in both cases, this suggests that the models may produce sound confidence scores that could potentially be meaningful when incorporating LLMs for decision-making.

4.2 RQ2: Impact of Criteria Analysis

Adding criteria analysis and two-step prompting has a significant effect on accuracy for only two models, in opposite directions. For Llama-3.3-70b, performance worsens (p=0.0289); for Gemini-3.1-flash-lite, the performance is instead significantly improved (p=0.00406). Two examples from the output from these models are available in the Appendix A.2. For other models, differences are non-significant and can go in either direction, which indicates that adding criteria analysis in general neither helps nor hurts performance, which remains at chance level.

Regarding truth bias, significant impacts from the criteria analysis and two-step prompting were found across five models. In four instances, a mild mitigating effect from the criteria analysis was observed. Llama-3.1-70b, Qwen3-32b, GPT-OSS-120b-high, and Gemini-3.1-flash-lite-high correctly calibrated their biases toward a more sound approach. In the case of Gemini-3.1-Flash-Lite-low, truth bias increased from 45% to 58%, which could seem like a flipped effect, going from slight **F**-bias to slight **T**-bias. However, given that the base rate of true statements is 53% for this dataset, a truth bias of 58% is arguably better than 45%, even though the latter is closer to 50%. The increase can therefore be viewed as an improvement. For the other models, the difference was non-significant.

Adding criteria analysis significantly lowered the confidence scores for all but three models. For GPT-OSS-20b-high and Gemini-3.1-Flash-Lite-high, the confidence significantly increased. GPT-OSS-120b-low showed no significant difference in average confidence. Significant correlations between accuracy and confidence were found for Qwen3-32b (r=0.12, p=0.00078) and Gemini-3.1-Flash-Lite-low (r=0.11, p=0.03309). Interestingly, Gemini-3.1-Flash-Lite-low moves from positive to negative correlation between accuracy and confidence. The same difference, although milder, was present for Gemini-3.1-Flash-Lite-high. This analysis suggests that the LLMs' confidence is not a robust indicator of their accuracy.

Some criteria were more similarly scored than others. To assess how different criteria relate to each other, we inspect overall correlations between criteria (see Figure 1). We find that **detailed** is strongly correlated with **vivid** and **long**, while **reasonable** is strongly correlated with **logical**. By contrast, **detailed** and **reasonable** were viewed as completely separate features by the LLMs.

4.3 RQ3: Criteria Scores as Predictors

Criteria scores had a considerable effect as predictors for the final judgments. As shown in Table 2, the logistic regression analysis indicates that the criteria scores drive the final judgments at a

Model	Accuracy			T-Bias			Confidence			r		
and reasoning effort	Dir	Cri	Diff.	Dir	Cri	Diff.	Dir	Cri	Diff.	Dir	Cri	Diff.
Llama-3.1-8b	0.48	0.46	−0.02	0.09	0.06	**−0.04***	0.802	0.745	**−0.057***	0.04	0.02	−0.02
Llama-3.3-70b	0.53	0.50	**−0.03***	0.58	0.54	**−0.04***	0.807	0.768	**−0.039***	0.00	−0.07	−0.07
Qwen3-32b	0.49	0.52	+0.03	0.22	0.37	**+0.15***	0.791	0.711	**−0.080***	0.03	0.12*	+0.09
GPT-OSS-20b-low	0.45	0.48	+0.03	0.41	0.38	−0.03	0.779	0.708	**−0.071***	0.11*	0.07	−0.04
GPT-OSS-20b-high	0.46	0.44	−0.02	0.89	0.92	+0.03	0.743	0.773	**+0.030***	0.06	0.01	−0.05
GPT-OSS-120b-low	0.47	0.47	−0.01	0.22	0.20	−0.03	0.692	0.692	0.000	−0.03	−0.02	+0.01
GPT-OSS-120b-high	0.52	0.52	0.00	0.79	0.73	**−0.05***	0.763	0.740	**−0.023***	−0.01	0.04	+0.05
Gemini-3.1-Flash-Lite-low	0.46	0.54	**+0.08***	0.45	0.59	**+0.14***	0.845	0.837	**−0.008***	0.15*	−0.11*	**−0.26***
Gemini-3.1-Flash-Lite-high	0.49	0.46	−0.04	0.60	0.56	−0.04	0.834	0.845	**+0.011***	0.08	−0.07	**−0.15***

Table 1: Classification performance for each model and reasoning effort setting for the direct judgments (Dir) and criteria-based judgments (Cri). Significant differences between direct and criteria-based judgments are in bold. **T-bias** = truth bias, **r** = correlation between accuracy and confidence, * indicates statistical significance. All scores are averaged over three or five runs, depending on the API used. Note that GPT-OSS-20b-high experienced significant difficulties producing usable output, resulting in only 293 samples across five runs.

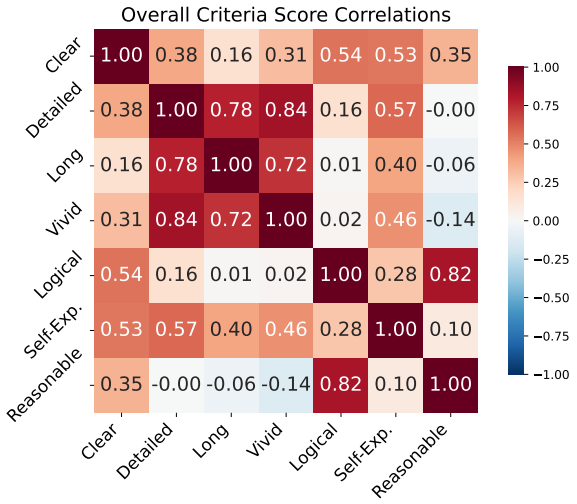

Figure 1: Correlations between criteria scores across all models and reasoning efforts.

Model	Final Judgments	Ground Truth
Llama-3.1-8b	0.312	0.031
Llama-3.3-70b	0.652	0.047
Qwen3-32b	0.235	0.048
GPT-OSS-20b-low	0.309	0.024
GPT-OSS-20b-high	0.183	0.040
GPT-OSS-120b-low	0.325	0.035
GPT-OSS-120b-high	0.336	0.037
Gemini-3.1-Flash-Lite-high	0.459	0.039
Gemini-3.1-Flash-Lite-low	0.479	0.040

Table 2: McFadden's Pseudo-R^2 values for final judgments and criteria-based judgments.

moderate to high level across all models, with McFadden's pseudo-R^2 values ranging from 0.235 to 0.652. High pseudo-R^2 values indicate that the predictions were highly contingent on the criteria analysis, suggesting that the criteria analysis effectively influenced model predictions. Model size appears to have influenced the extent to which criterion scores determine model predictions, as the lowest pseudo-R^2 values were observed for smaller models such as Llama-3.1-8b and Qwen3-32b, while larger models such as Llama-3.3-70b and Gemini-3.1-Flash-Lite showed higher pseudo-R^2 values.

Contrary to final judgments, when fitting the criteria scores to the ground truth, the relationship is weak. Pseudo-R^2 values range from 0.024 to 0.048, which shows that the criteria scores had a weak explanatory power for the ground truth of the statements. This corroborates the notion that linguistic criteria for veracity judgments are weak indicators.

The most criteria-influenced model was Llama-3.3-70b, with a pseudo-R^2 of 0.652. The same model was also among the least incompetent at estimating fulfillment of valid criteria, with a pseudo-R^2 of 0.047. Curiously, the criteria implementation and two-step prompting still yielded lower accuracy than in the direct judgments for this model.

Across all models, there are large discrepancies between how much the models relied on the criteria in their predictions and how much the criteria actually indicate whether a statement is true. Figure 2 shows the coefficients from logistic regression analysis comparing the criteria scores' relationship to final judgments and ground truth. A positive coefficient means that the criterion drove **T** predictions, while a negative coefficient means that the criterion drove **F** predictions.

According to the theory, all implemented criteria are truth-indicative, meaning they should drive **T** predictions and, if valid, yield positive coefficients

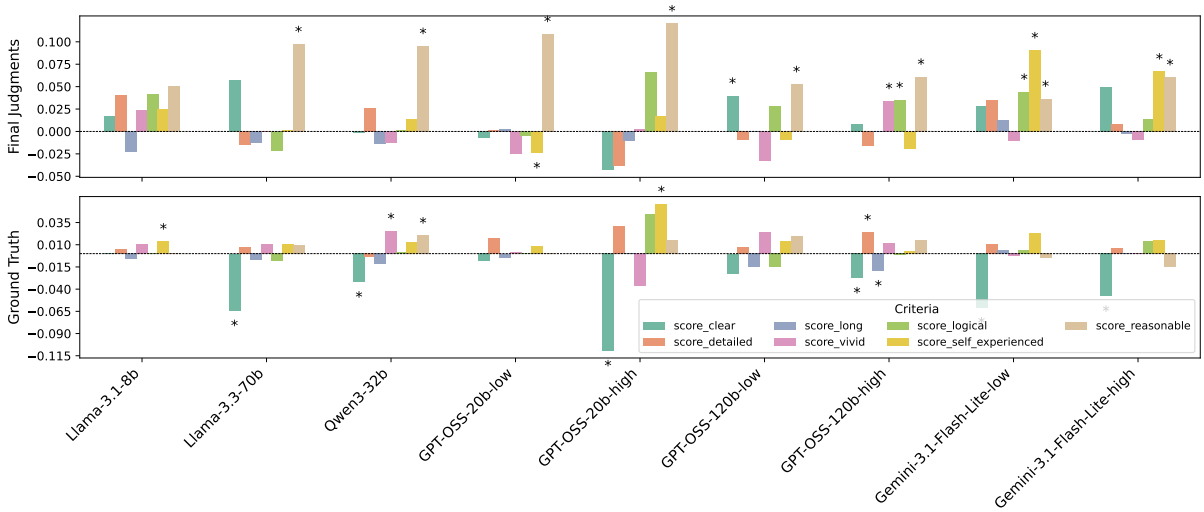

Figure 2: Coefficients from logistic regression on the criterion scores against final judgments (upper subplot) and ground truth (lower subplot). * indicates statistical significance with Holm-Bonferroni correction based on alpha=0.05. Note that sample sizes vary across models, with GPT-OSS-20b-high standing out at only 293 samples.

for ground truth. In Figure 2, we can see that this is not the case for most LLMs' criteria assessments, since coefficients are negative for various criteria. Models did not necessarily agree with each other on which criteria were the most important, and whether fulfillment of a certain criterion should entail that the statement is true or false. There is no model that continuously identifies all criteria as truth-indicative in the final judgment stage, and there is no model that continuously scores true statements higher than false statements.

There are, however, some observations that rhyme with theory. **Reasonable** is the one criterion that significantly drove **T** judgments for all models but Llama-3.1-8b, where it was insignificant but still positive. **Self-experienced** was positive for all models, although only significant for Llama-3.1-8b and GPT-OSS-20b-high, the latter of which only resembles a smaller sample of the data. Apart from these, no criteria were consistently interpreted or extracted as theory suggests. **Long** was regarded as deception-indicative by half of the models. **Detailed** drove judgments in both directions depending on the model, as did **vivid** and **logical**. **Clear** was the only criterion that continuously appears deception-indicative, where all coefficients are negative for ground truth, of which six were significant. Yet, all but one model (GPT-OSS-20b) believed that higher clarity scores indicated that the statement was true. GPT-OSS-20b-high was the only model that managed to capture and use the **clear** criterion in a close-to-adequate way, but underestimated its predictive value and should be interpreted with caution due to its lower sample size. Qwen3-32b shows the same pattern of underestimating. All the other models consistently misinterpreted the predictive strength and direction of their **clear** scorings.

5 Discussion

The aim of this study was threefold: to explore the performance of LLMs at veracity judgments, to explore the effect of criteria analysis on this performance, and to explore how criteria analysis influences final judgments and whether any criteria assessments correlate with ground truth.

Experiments have shown that LLMs perform at around chance level, underscoring the task's complexity. In the analysis, we have found that truth bias varies widely across models and between different reasoning efforts for the same model. The relationship between confidence and accuracy is non-significant for most models, but positively significant for two models in the direct judgments (GPT-OSS-20b-low and Gemini-3.1-Flash-Lite-low). Overall, results indicate that the confidence of LLMs is a poor predictor of their accuracy.

Adding criteria analysis, and asking LLMs to base veracity judgments on their own criteria assessment, has virtually no effect on performance, which remains at chance level. However, some effect was observed for the truth bias, which was significantly better calibrated for four models for the criteria-based judgments (Llama-3.3-70b, Qwen3-

32b, GPT-OSS-120b-high, and Gemini-3.1-Flash-Lite-low). The lowering of the already extremely low truth bias for Llama-3.1-8b from 9% to 6% is the only significant counterexample. In other cases, the truth bias was not significantly affected by the criteria analysis. From this, we conclude that criteria analysis may improve the soundness of predictions by affecting how skeptical or gullible the models behave.

Criteria analysis also affected the models' confidence estimation, which overall became less certain when making criteria-based judgments than when making direct judgments. Interestingly, for the confidence-accuracy correlations, Gemini-3.1-Flash-Lite significantly dropped from slightly positive to slightly negative. Considering the mild but significant decrease in confidence and increase in accuracy, this suggests the criteria made the model more cautious in issuing high confidence, leading to more careful decisions and higher correctness. That could explain the negative correlation: lower confidence led to more correct judgments.

Regarding the third research question, we found that LLMs' final judgments were moderately to highly influenced by the criteria analysis, providing McFadden's pseudo-R^2 scores between 0.235 and 0.652. Of the individual criteria, **reasonable** influenced the LLMs' judgments the most across models, with all models consistently believing that high scores for reasonable indicate that the statement is likely to be true. Larger models were observed to be more influenced than smaller models, although we did not analyze this further. Whereas criteria scores predicted final judgments to a moderate degree, criteria scores did not, in general, predict the ground truth of the statements. Even though some significant predictors were found, they were not significant for all models, and some appeared to indicate deception rather than truth. **Self-experienced** was the only significant truth-indicator for two models. Interestingly, the **clear** criterion was significant as a deception-indicator for six models. This completely clashes with the theoretical assumption that clarity entails a statement's truthfulness. In conclusion, there was a clear discrepancy: LLMs believed the criteria scores to be stronger predictors than they actually were.

6 Conclusion

In this study, we set out to explore how LLMs perform veracity judgments, whether adding reliability criteria analysis can inform LLM judgments, and how LLMs' interpretations of these criteria relate to the predictive power of the criteria. Based on our analysis, we conclude that LLMs perform veracity judgments at chance level, regardless of whether they have access to analyses of reliability criteria to aid their judgment. Criteria analysis merely affected the truth bias in the predictions and the LLM's confidence in its judgments. While LLMs appear able to use criteria scorings to influence their judgments, the criteria scorings themselves did not provide discriminative information for the judgment to be more accurate than if the LLM judged the statement directly. This entails that the criterion scores were not strongly indicative of either truth or deception. Given that previous research has shown that criterion effect sizes are small, we are inclined to conclude that our results corroborate this finding. We cannot know for certain, however, whether the inefficacy of the criteria analysis in this study was due to the quality of the criteria extraction or to the absence of effective criteria in the data.

However, given the strong influence of criteria on final judgments, our two-step prompting method presents promising avenues for future work. If only the relationship between the criteria scores and the ground truth were stronger, it would be possible, or even probable, to achieve much better results with this same method. In future work, we plan to explore other ways to quantify criteria, define additional potential indicators of truth or deception, and compare different criterion combinations.

Other findings worth further investigation include the effects of reasoning and reasoning effort. Raising reasoning effort from low to high increased accuracy for direct judgments (same pattern for GPT-20b, GPT-120b, and Gemini-Flash-Lite). Truth bias also increased for these same instances. However, for the criteria-based judgments, only GPT-120b showed higher accuracy with high reasoning effort than with low. Truth bias increased for all but one model: Gemini-3.1-Flash-Lite, when going from low to high reasoning. Qualitative analysis of the LLMs' rationales for their judgments, as well as internal reasoning, is also something we would like to investigate further in a future study. Since the evolution of LLMs as reasoning agents has been a fast-moving trajectory, it will be worth evaluating more and larger models for this task in the future.

Limitations

There are several limitations to our study. First and foremost, our dataset may lack significant features that perhaps would be more salient otherwise, considering that the statements were given in a low-stakes scenario and without time constraints, which enabled participants to plan and edit their statements as they pleased. Part of the dataset consists of very short statements, down to only a few words. This probably affects LLMs' ability to extract meaningful features from some statements. It is also likely that these statements are inherently more difficult to assess, which may lead to greater variability in their assessment. In continuation, assessments of these statements weigh down on overall performance. A more extensive error analysis could answer these questions.

We observed high correlations among some criteria, which may affect the reliability of our regression analysis of those specific criteria due to multicollinearity. Especially the ones that showed high correlation with another criterion, such as **reasonable** with **logical**. Furthermore, The point system for criterion evaluation is not a common practice and serves as a proxy for the internal weighting that practitioners use in their heads when deciding whether a statement meets a given criterion. This may introduce artifacts about

Regarding models, GPT-OSS-20b with high reasoning effort struggled with providing analyzable output, leading to a decreased sample size for this model configuration, which may bias comparisons between models. Also, results reflect specific model versions and API settings, and some are preview architectures, which poses a challenge for reproducibility.

Ethics Statement

Determining whether someone is telling the truth or not in a forensic context is a complex task that can have severe implications for all parties involved, plaintiffs and defendants, if done incorrectly. LLMs and AI agents are not mature enough to perform this task, as our research shows. Practitioners should take strong precautions when using AI tools, and we do not encourage the use of AI for making veracity judgments of verbal statements in real-world settings.

During data collection, all participants provided consent and received appropriate information about our research aims. They could end their participation at any time. For compensation, Prolific participants received payment through the Prolific platform. The technical and administrative staff at the research organization participated in exchange for an informative lecture. Participants gathered through social media did not receive compensation.

This research project has been approved by an ethics committee.

References

Michael G. Aamodt and H. Custer. 2006. Who can best catch a liar?: A meta-analysis of individual differences in detecting deception. *The Forensic Examiner*, 15(1):6–11.

Glynis Bogaard, Rafail Dimopoulos, and Conny W. E. M. Quaedflieg. 2025. Lying Under Pressure: Examining the Impact of Stress and Veracity on Verbal and Nonverbal Behaviors in Credibility Judgments. *Applied Cognitive Psychology*, 39(1):e70018.

Glynis Bogaard, Ewout H. Meijer, Aldert Vrij, and Harald Merckelbach. 2016. Strong, but Wrong: Lay People's and Police Officers' Beliefs about Verbal and Nonverbal Cues to Deception. *PLOS ONE*, 11(6).

Charles F. Bond and Bella M. DePaulo. 2006. Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3):214–234.

Dainis Boumber, Bryan E. Tuck, Rakesh M. Verma, and Fatima Zahra Qachfar. 2024. LLMs for Explainable Few-shot Deception Detection. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, pages 37–47, Porto Portugal. ACM.

Tim Brennen and Svein Magnussen. 2023. Lie Detection: What Works? *Current Directions in Psychological Science*, 32(5):395–401.

Lennart Bürger, Fred Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 138393–138431. Curran Associates, Inc.

Bella M DePaulo, Kelly Charlton, Harris Cooper, James J Lindsay, and Laura Muhlenbruck. 1997. The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review*, 1(4):346–357.

Yurena Gancedo, Francisca Fariña, Dolores Seijo, Manuel Vilariño, and Ramón Arce. 2021. Reality Monitoring: A Meta-analytical Review for Forensic Practice. *The European Journal of Psychology Applied to Legal Context*, 13(2):99–110.

Google. 2026. Gemini 3.1 flash lite preview [may 2026]. *https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite-preview*.

H. Paul Grice. 1975. Logic and conversation. In Donald Davidson, editor, *The logic of grammar*, pages 64–75. Dickenson Pub. Co.

Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer. 2015. Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, 19(4):307–342.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To Tell The Truth: Language of Deception and Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8506–8520, Mexico City, Mexico. Association for Computational Linguistics.

Marcia K. Johnson and Carol L. Raye. 1981. Reality monitoring. *Psychological Review*, 88(1):67–85.

Günter Köhnken. 2004. The Detection of Deception in Forensic Contexts: Statement Validity Analysis and the 'detection of the truth'. *Detection of deception in forensic contexts*, pages 41–63.

Timothy R. Levine. 2014. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, 33(4):378–392.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, New Orleans, Louisiana. Association for Computational Linguistics.

Riccardo Loconte, Merylin Monaro, Pietro Pietrini, Bruno Verschuere, and Bennett Kleinberg. 2026. Humans incorrectly reject confident accusatory ai judgments. *Computers in Human Behavior*, 182:109019.

Riccardo Loconte, Roberto Russo, Pasquale Capuozzo, Pietro Pietrini, and Giuseppe Sartori. 2023. Verbal Lie Detection using Large Language Models. *Scientific Reports*, 13(1).

David M. Markowitz and Jeffrey T. Hancock. 2024. Generative AI Are More Truth-Biased Than Humans: A Replication and Extension of Core Truth-Default Theory Principles. *Journal of Language and Social Psychology*, 43(2):261–267. _eprint: https://doi.org/10.1177/0261927X231220404.

David M Markowitz and Timothy R Levine. 2025. The (in)efficacy of ai personas in deception detection experiments. *Journal of Communication*, 75(6):459–469.

Meta AI. 2024. The llama 3 herd of models. *arXiv:2407.21783*. Full author list: Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia , Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal

Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi (Jack)Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu (Sid)Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, Zhiyu Mae.

Md Messal Monem Miah, Adrita Anika, Xi Shi, and Ruihong Huang. 2025. Hidden in plain sight: Evaluation of the deception detection capabilities of LLMs in multimodal settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31013–31034, Vienna, Austria. Association for Computational Linguistics.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv:2508.10925*. Full author list: OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman,

Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily (Xiaoxuan) Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, Shengjia Zhao.

Lorenzo Pacchiardi, Alex J Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y Pan, Yarin Gal, Owain Evans, and Jan Brauner. 2024. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. In *The Twelfth International Conference on Learning Representations*, 5212.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Qwen Team. 2025. Qwen3 technical report. Full author list: An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, Zihan Qiu.

Andrew M Smith and Amy-May Leach. 2019. Confidence can be used to discriminate between accurate and inaccurate lie decisions. *Perspectives on Psychological Science*, 14(6):1062–1071.

Siegfried L. Sporer and Jaume Masip. 2023. Millennia of legal content criteria of lies and truths: wisdom or common-sense folly? *Frontiers in Psychology*, 14:1219995.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Rebecca M. Willén and Leif A. Strömwall. 2012. Offenders' lies and truths: an evaluation of the Supreme Court of Sweden's criteria for credibility assessment. *Psychology, Crime & Law*, 18(8):745–758.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Prompts

In this section, the prompts from the experiments are outlined. The preamble describes role assignment and is put at the beginning of all prompts. The following boxes describe the three different instruction continuations.

Preamble: Role assignment and context

You are a professional criminal investigator working on a theft case.
The theft happened in an apartment building yesterday. The plaintiff, Leonard, has reported that a parcel addressed to him has disappeared from the post room. The parcel contained a MacBook with an approximate worth of $2 500. The package was stolen from the post room inside the entrance of the apartment building.
The police have interviewed Leonard's neighbours as potential suspects.

One-Step Prompting: Direct Veracity Judgment

(...) You are now going to read the statement given by one of the neighbours, and your task is to decide whether this neighbour seems to be telling the truth or not.
Please read the statement carefully and make your judgment. Answer with ###T### if you think the statement is true, and ###F### if you think the statement is false. Give a confidence score between 0 and 1 that describes how confident you are in your judgment.
Here is the statement: statement
Output template (MUST follow exactly):
CLASSIFICATION: ###<T/F>### Confidence score: <0-1>
RATIONALE: <3-6 sentences in English>

Two-Step Prompting: Step 1: Criteria Analysis

(...) You are now going to read the statement given by one of the neighbours, and your task is to analyse the statement based on the following criteria: clear, detailed, long, vivid, logical, self-experienced, and reasonable.
Please read the statement carefully and consider to what degree the statement fulfills each criterion. In your answer, list the criteria and assign a fulfillment score of 0-100 for each criterion.
Here is the statement: {statement}
Output template (MUST follow exactly):
CRITERIA SCORES: Clear: <0-100> Detailed: <0-100> Long: <0-100> Vivid: <0-100> Logical: <0-100> Self-experienced: <0-100> Reasonable: <0-100>
RATIONALE: <3-6 sentences in English>

Two-Step Prompting: Step 2: Criteria-Based Veracity Judgment

(...) You are now going to read the statement given by one of the neighbours, along with an analysis of the same statement, and your task is to decide whether this neighbour seems to be telling the truth or not.
Please read the statement carefully and consider its fulfillment of each criterion. The statement and the criteria-based analysis should guide you in deciding whether the neighbour is telling the truth or not. Be transparent about how confident you are in your decision. Answer with ###T### if you think the statement is true, and ###F### if you think the statement is false. Give a confidence score between 0 and 1 that describes how confident you are in your judgment.
Here is the statement and the criteria scores (based on a 1-100 scale): {statement} {criteria_analysis}
Output template (MUST follow exactly):
CLASSIFICATION: ###<T/F>### Confidence score: <0-1>
RATIONALE: <3-6 sentences in English>

A.2 Examples from LLM Judgments

In this section, we present examples from experiments that exemplify statement characteristics and model behavior in two cases:

- Where direct judgments outperformed criteria-based judgments (A.2.1)
- Where criteria-based judgments outperformed direct judgments (A.2.2)

We select statements that were continuously misclassified in one setting and correctly classified in the other over multiple runs. Error rates for direct and criteria-based judgments are calculated by subtracting the mean accuracy for the specific statement and model setting from 1. The statements are sorted by the difference between the error rates for direct and criteria-based judgments. The following examples have the highest delta scores, meaning they repeatedly fall into the same category: either they are frequently misjudged when criteria are present (see A.2.1), or they are better understood through the implementation of criteria (see A.2.2).

A.2.1 Example 1: Criteria Confuse

Model: Llama-3.3-70b

Statement: Jag såg ett paket i brevlådan men när jag gick in i min lägenhet. Jag tog den inte med in min lägenhet. Det var inte jag som tog bort paket. *[I saw a package in the mailbox but when I went into my apartment. I didn't take it into*

my apartment. I wasn't the one who removed the package.]

Baseline: "CLASSIFICATION: ###T### Confidence score: 0.8 The neighbour's statement seems to be truthful as it provides a straightforward denial of taking the package. The statement also mentions seeing the package in the post room, which suggests that the neighbour is aware of the package's existence and is not trying to feign ignorance. However, the statement is in Swedish, which may introduce some uncertainty due to potential translation errors. Additionally, the neighbour's claim of not taking the package is unsupported by any evidence, which slightly reduces the confidence in the statement's truthfulness. Nevertheless, the statement's simplicity and lack of obvious contradictions suggest that the neighbour is likely telling the truth. Overall, the confidence in the statement's truthfulness is relatively high, but not absolute.

Criteria Assessment: "Clear: 80 Detailed: 20 Long: 0 Vivid: 0 Logical: 60 Self-experienced: 80 Reasonable: 80

The statement is relatively clear as it conveys the neighbour's claim of not taking the package. However, it lacks detail, as the neighbour does not provide any specific information about the package or their actions beyond seeing it and not taking it. The statement is also very short, which contributes to its low score for the ""long"" criterion. The neighbour's account is not vivid, as it does not include any descriptive language or sensory details. The statement is somewhat logical, as it provides a straightforward denial of involvement in the theft. The neighbour's claim is based on their self-experienced event of seeing the package, which adds to the statement's self-experienced score. The statement is reasonable, as it is plausible that the neighbour saw the package and did not take it, but the lack of detail and corroborating evidence reduces its overall reasonableness."

Criteria-Based Judgment: "CLASSIFICATION: ###F### Confidence score: 0.7

RATIONALE: The neighbour's statement lacks detail and vividness, which raises suspicions about their involvement in the theft. Although the statement is clear and conveys a straightforward denial of involvement, its brevity and lack of descriptive language make it less credible. The neighbour's claim is based on their self-experienced event, but the absence of corroborating evidence or specific information about the package or their actions reduces the statement's reasonableness. Furthermore, the low score for the ""logical"" criterion suggests that the neighbour's account may not be entirely plausible. Overall, while the statement is not entirely implausible, its weaknesses and lack of detail lead me to believe that the neighbour may be hiding something, hence my decision to classify the statement as false."

A.2.2 Example 2: Criteria Help

Model: Gemini-3.1-Flash-Lite

Statement: Jag såg att det låg ett paket på Leonards hylla när jag kom hem. Sedan träffade jag Mats senare på kvällen när jag slängde sopor, han berättade att Leonards paket hade stulits. Hur han hade fått veta det sa han inte, men jag utgick från att Leonard hade sagt det till honom. Jag visste inte att det var en dator i paketet när jag såg det, personligen så hade jag nog sett till att vara hemma om jag skulle få en så dyr leverans... men jag har tyvärr inte sett eller hört något konstigt, ni borde prata med Mats som också bor här bredvid. *[I saw that there was a package on Leonard's shelf when I got home. Then I met Mats later in the evening when I was taking out the trash, he told me that Leonard's package had been stolen. He didn't say how he had found out, but I assumed Leonard had told him. I didn't know that there was a computer in the package when I saw it, personally I would have probably made sure to be home if I was going to get such an expensive delivery... but unfortunately I haven't seen or heard anything strange, you should talk to Mats who also lives next door.]*

Baseline: "CLASSIFICATION: ###F### Confidence score: 0.75

RATIONALE: The witness claims they saw a package on Leonard's shelf, but then immediately deflects suspicion by suggesting the police talk to Mats, who supposedly knew about the theft before anyone else. The statement includes a subtle, judgmental remark about Leonard's choice to have an expensive item delivered to a post room, which often serves as a psychological tactic to shift blame or minimize the victim's credibility. Furthermore, the witness claims they didn't know the package contained a computer, yet they immediately reference the value of the item (""if I were to receive such an expensive delivery""), suggesting they knew more about the contents than they admitted. This combination of deflection and in-

consistent knowledge about the package's value makes the statement highly suspicious."

Criteria Assessment: "Clear: 90 Detailed: 60 Long: 40 Vivid: 50 Logical: 85 Self-experienced: 90 Reasonable: 80

The statement is very clear and easy to follow, providing a straightforward timeline of events from the witness's perspective. While it lacks deeper detail regarding the specific times of the sightings or descriptions of the package, it remains internally logical and clearly grounded in the witness's own experience. The inclusion of the unsolicited advice about how Leonard should have handled his delivery adds a touch of personal bias, which makes the statement feel authentic but slightly less objective. Overall, it is a helpful, albeit brief, account that effectively points the investigation toward another potential witness, Mats."

Criteria-Based Judgment:
"CLASSIFICATION: ###T###
Confidence score: 0.85

RATIONALE: The statement is internally consistent and follows a logical chronological order, which is a strong indicator of truthfulness. The witness admits to seeing the package, which places them at the scene, but they also provide a plausible explanation for how they learned of the theft through a third party, Mats. While the unsolicited advice about how Leonard should have handled his delivery is slightly judgmental, it feels like a genuine, albeit biased, personal reaction rather than a calculated deception. The witness's willingness to direct the investigation toward Mats suggests they are not trying to hide information or deflect suspicion from themselves in a defensive manner."

A.3 AI Statement

During the preparation of this work, Grammarly was used to improve clarity and flow in the text. Github Copilot was used for drafting and reviewing code.