Text-Based Veracity Assessment of Verbal Statements in Swedish

Clara Gustafsson* **Moa Lidén†** **Joakim Nivre***

*Dept. of Linguistics and Philology, Uppsala University, first.last@lingfil.uu.se
†Department of Law, Uppsala University, moa.liden@jur.uu.se

ACL Anthology ID 2026.findings-emnlp.2438 / revision 2 / 10 Nov 2026

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 of verbal statements. In this study, we explore this question on a novel Swedish mock-crime dataset. 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, but that the two-step method often leads to a more reasonable truth bias and reduces models' overconfidence. Further analysis indicates that criteria assessments predict final judgments, but not ground truth. We also find that the statement category (true or false and confession or denial) matters to overall performance and the effect of two-step prompting.

1 Introduction

Analyzing verbal statements to determine their veracity is a major challenge for practitioners in the legal system. Humans, including trained police officers, detectives, and judges, perform poorly at lie detection, reaching an average accuracy of 54% in meta-analyses (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.

Research has shown that focusing on evaluating verbal cues 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 presented a set of linguistic *reliability criteria* that deem a statement 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 are inspired by 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 on 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 reasoning agents, 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 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 following research questions:

1. How do LLMs perform veracity judgments?
   (a) How accurate are the judgments?
   (b) To what degree do LLMs exhibit truth bias?

[1]Dataset available on Huggingface: https://huggingface.co/datasets/clagu/MockCrimeTheftSwedish

(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?
4. How does performance vary across true and false confessions and denials?

Our results indicate that LLMs perform at chance level while exhibiting bias and overconfidence. Adding criteria analysis does not improve accuracy but often leads to a better calibrated truth bias and confidence. Regarding the predictive value of criteria assessments, our analysis indicates that they do predict final judgments but not ground truth. We also find that overall performance and the effect of criteria analysis depend on the type of statement.

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 is one of the most challenging tasks legal practitioners face. 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 where verbal statement assessments played a significant role. These are criteria that 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]. Statements are also argued to be reliable on the basis that they appear **self-experienced**, and **reasonable**.[4] These are generally viewed as over-arching; a statement is viewed as self-experienced if it is detailed and vivid, and reasonable if it is clear and logical.

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 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 indicate both truth and deception across studies (Hauch et al., 2015; Gancedo et al., 2021).

Truth bias is one of the main reasons humans perform poorly in deception detection, as is described by Levine's truth default theory (TDT, Levine, 2014). TDT analyzes lying from a communicative perspective and argues that lies are hard to detect because humans 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). 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).

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% accu-

[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.

[4]See for example precedent from the SSC, Case nr: B4003-14, 2015.

racy 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, and 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 prior research on the relationship between confidence and accuracy in LLM veracity judgments.

A tangent research field is 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 to train a classifier that then successfully separates truthful from deceptive activations (Zou et al., 2025; 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 LLMs. The purpose is to explore how LLMs make veracity judgments and how their performance is affected when given a criteria analysis as a basis for those judgments. We will therefore compare one-step prompting, which asks for a direct veracity judgment and confidence assessment, to two-step prompting, in which the first prompt asks for criteria assessments and the second asks for a veracity judgment and confidence assessment based on those 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.[5]

3.1 Dataset collection and preparation

The dataset used in this study is a novel Swedish mock-crime dataset of verbal statements, which constitutes a significant contribution. The dataset is available on Hugging Face, together with an extensive report of the instructions and stimuli.[6]

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. Participants were reached by distributing the questionnaire to technical and administrative staff at Uppsala University, through social media and Prolific. 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. Subsequently, the police knock on their door and ask them to give a statement regarding the alleged crime. Half were instructed to lie, and the other half to tell the truth about their actions.

After collecting the statements, we manually assessed each one and removed those where it appeared the participant had misunderstood the task. This resulted in a dataset with 65 true confessions, 64 false confessions, 67 true denials, and 55 false denials. All participants were located in Sweden, with Swedish as their native language. The age range was 18 to 73, with a median of 33 (mean = 35.77, st. dev. = 12.84). Of all participants, 131 were men, 120 were women, and three identified as non-binary. In total, 84 participants were reached through social media, and among the TeA group. The rest, 167 participants, participated through Prolific.

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 on the held-out development set of 40 evenly distributed

[5]Code can be found on Github: https://github.com/claraagu/veracity-swedish

[6]https://huggingface.co/datasets/clagu/MockCrimeTheftSwedish

statements. The remaining 211 statements were used for the end experiments. The full prompts are available in Appendix A.2. Since research has shown that imposing output-format constraints may affect the reasoning abilities of LLMs (Tam et al., 2024), we let the model output in its most comfortable format, markdown or plain text, instead of using a JSON schema. We did, however, use a plain text output template to ease parsing. The prompts were written in English. We experimented with Swedish prompts in a pilot study on the development data. Results from this ablation can be found in the Appendix C. Apart from an increase in token consumption, especially for prompt processing, there were no general patterns of performance gains or losses.

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. Numerical scores are used as a proxy for the degree of criteria fulfillment. Although numerical assessments are not a common legal practice, this allows us to quantitatively analyze criteria and model behavior. The accompanying rationale approximates the realization of the analysis as a qualitative argument.

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.

Criteria Analysis Content Ablation To assess the impact of the rationale vs. the scores in the analysis content used for classification, we conducted an ablation study using Gemini-3.1-Flash-Lite and GPT-OSS-120b with low reasoning effort. The ablation study revealed no significant difference in accuracy when using only scores, only rationale, or both. For more details, see Appendix B.

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 Gemini Team, 2026) through Google GenAI SDK.[8]. Experiments were run in May 2026. Gemini-3.1-Flash-Lite was a preview model at the time. For exact execution dates and model version names, see our GitHub repository.

For all models, max tokens was set to 4000 and top p to 1. The temperature was set to 0.1 for the judgment tasks and 0.6 for the criteria analysis. GPT and Gemini models were used with low and high reasoning effort, resulting in nine distinct model setups. Groq API models were run 5 times, Gemini models were run 3 times. GPT-OSS-20b-high exhibited degenerate generation and looping behavior, resulting in only 293 analyzable samples. 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

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

yielded 1044. Gemini models cleared 633 samples each over three runs.

3.4 Evaluation

The aim of this study is to explore how LLMs make veracity judgments, whether criteria analysis affects their performance, and whether it predicts final judgments and the ground truth. To assess LLM veracity judgment performance, we evaluate the accuracy and truth bias. We also examine how performance varies across the four categories in the data: true confessions, false confessions, true denials, and false denials. 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 direct veracity judgments (one-step prompting) and 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 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.

Finally, to examine how performance varies across statement categories and how it is affected by two-step prompting with criteria analysis, we conduct Generalized Estimating Equations (GEE) regression analyses on accuracy, truth bias, and confidence.

4 Results

The experimental results are presented in the following four subsections, each corresponding to one of our four main research questions. The main 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 the level of human accuracy reported in meta-analyses. 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, in which 89% of statements were deemed true. At the most skeptical extreme, we find Llama-3.1-8b, which identified only 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 yet weak 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 ($r = 0.1543$, $p = 0.00068$).

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.3. 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-3.3-70b, 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

Model and reasoning effort	Accuracy			T-Bias			Confidence			r		
	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.

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.

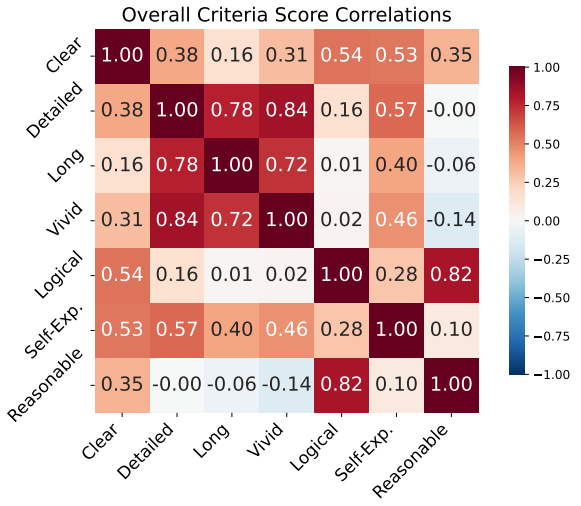


Figure 1: Correlations between criteria scores across all models and reasoning efforts.

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 moderate to high level across all models, with McFadden's pseudo-R^2 values ranging from 0.183 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: lower pseudo-R^2 values were observed for smaller models such as Llama-3.1-8b, GPT-OSS-20b-high, and Qwen3-32b, whereas larger models such as Llama-3.3-70b and Gemini-3.1-Flash-Lite showed higher pseudo-R^2 values. The low result for GPT-OSS-20b-high should be interpreted with caution due to the smaller sample size, which was due to poor compliance.

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

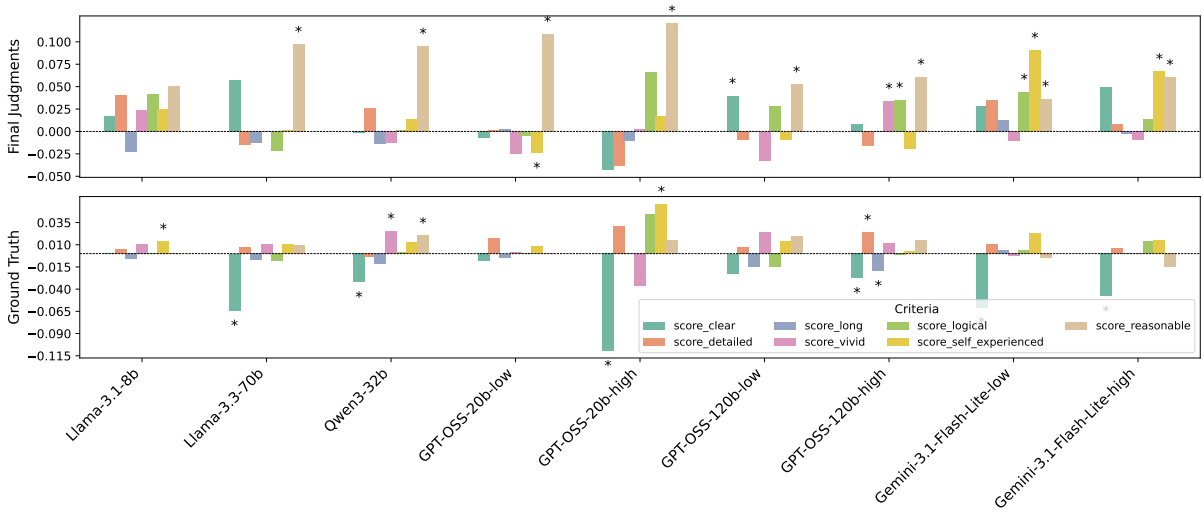


Figure 2: Coefficients from logistic regression on the criterion scores against final judgments (upper subplot) and ground truth (lower subplot). * indicates statistical significance with Holm-Bonferroni correction based on alpha=0.05. Note that sample sizes vary across models, with GPT-OSS-20b-high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

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-low	0.479	0.040
Gemini-3.1-Flash-Lite-high	0.459	0.039

Table 2: McFadden's Pseudo-R^2 values for logistic regression of criteria scores against final judgments and ground truth, respectively.

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 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 consistently identifies all criteria as truth-indicative in the final judgment stage, and no model that consistently scores true statements higher than false ones.

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.

Given the moderate-to-high correlations among some criteria (as seen in Figure 1), there is a risk of multicollinearity, which limits the strength of conclusions about individual criteria.

4.4 RQ4: Statement Category Analysis

While overall accuracy was close to chance, the four categories in the dataset were not assessed equally in our experiments. Figure 3 displays accuracies, truth bias, and confidence for direct judgments and criteria-based judgments across statement categories, pooled across all LLMs. Firstly, we can observe that true confessions were the hardest category to judge accurately, followed by false denials, true denials, and, lastly, false confessions, which models accurately disbelieved most of the time. A GEE regression analysis revealed that the classification patterns differed between confessions and denials, and true and false statements. Denials were harder than confessions ($p<.001$), and true statements were harder than false ones ($p<.001$). The GEE analysis also revealed a significant interaction between veracity and statement type for accuracy. Correctly judging a statement as true was significantly easier when it was a denial rather than a confession ($p<.001$).

Although the effect of the prompting strategy was insignificant overall ($p=.945$), and the interaction with statement type was insignificant ($p=.888$), the GEE analysis revealed significant interactions between prompting strategy and veracity ($p=.011$) and a three-way interaction between prompting strategy, veracity, and statement type ($p<.001$). This means that the criteria-based prompting altered the classification patterns differently across statement categories. In Figure 3, it is salient that with the criteria analysis at hand, models were overall more skeptical towards true confessions and more inclined to believe true denials, as there was a decrease in accuracy for true confessions, and an increase for true denials.

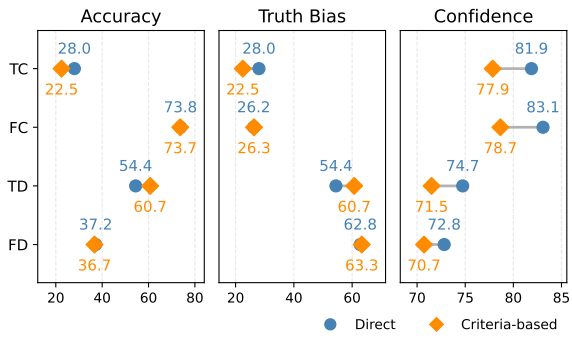


Figure 3: Percentage point values for accuracy, truth bias, and confidence across statement categories. Values are pooled across the 9 model setups. Scores from direct judgments are marked by blue circles, and criteria-based judgments are marked by orange diamonds. For true statements, accuracy and truth bias are identical since the mean of correct predictions and **T** judgments is the same.

The effects on truth bias are largely similar to those on accuracy, as expected, since truth bias for true categories is the same as accuracy. For false categories, truth bias equals the incorrect judgments, the inverse of accuracy. It was therefore expected to see a significant three-way interaction $p<.002$), an insignificant overall effect from prompting alone ($p=.945$), and a significant effect from statement type. Denials were more often classified as true ($p<.001$), which can also be seen in Figure 3. Veracity alone did not yield a significant difference ($p=.333$). Veracity and prompting showed a significant interaction ($p=0.008$), while prompting and statement type showed no significant interaction ($p=.880$). This means that the effect of prompting differed between true and false statements, and that statement type affected the relationship. Figure 3 shows this clearly as true confessions and denials are differently affected by the prompting, and the direction of the effect is decided by whether it is a confession or a denial.

Regarding confidence, GEE confirmed an effect clearly salient in Figure 3, which is that direct prompting yielded higher confidence overall ($p<.001$). A three-way interaction was observed in this case as well, indicating that the effect of prompting on confidence differed across the four statement categories ($p=.046$). Direct prompting vs. criteria-based prompting affects confidence differently for confessions and denials ($p<.001$). Denials were consistently dealt with at lower confidence. Veracity alone yielded no significant difference.

5 Discussion

The aim of this study was to explore the performance of LLMs at veracity judgments, whether it is affected by criteria analysis, how criteria analysis predicts final judgments, whether criteria assessments correlate with ground truth, and lastly, how performance differs across statement categories in the data.

Experiments have shown that LLMs perform at around chance level, underscoring the task's complexity. In our analysis, we have found that truth bias varies widely across models and within the same model across different reasoning efforts. The relationship between confidence and accuracy is non-significant for most models, but 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.

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 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). Criteria analysis also mitigated the models' overconfidence. From this, we conclude that criteria analysis may improve the soundness of predictions by affecting how skeptical or gullible the models behave and by reducing their overconfidence. However, we want to underscore that we do not observe any overall improvement in prediction accuracy.

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 0.183 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.

Finally, our statement category analysis revealed that statement category matters. Overall, models were better at detecting false confessions and recognizing true denials, but worse at accepting true confessions and detecting false denials. Using criteria also affected categories differently. These effects are important considerations for future work and highlight the complexity of detecting deceit in different forms.

6 Conclusion

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 did, however, affect the truth bias in the predictions and the LLM's confidence in its judgments. It also affected different statement categories differently. 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.

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

Future work should also take into consideration the role of statement type, whether the lie is a confession or a denial, since this had a significant effect on the overall performance and effect of the criteria analysis. Another finding for further exploration is the effect of reasoning effort on truth bias. 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 features that would otherwise be more salient, given that the statements were provided in a low-stakes scenario 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. On a broader note, the mock theft scenario used in the data collection does not reflect typical cases in which statement analysis, including the Swedish Supreme Court's reliability criteria, is applied. Statement analysis is more commonly applied when dealing with higher-profile cases such as rape or domestic violence. Since there are apparent difficulties with creating a scenario for a crime of that caliber, we resigned ourselves to the theft case. This affects the possibility of making inferences about policy from our results.

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 only as a proxy for the internal weighting that practitioners do when deciding whether a statement meets a given criterion. Even though we included a textual rationale paired with the numerical scores, the rationale is supposedly guided by the scores. This may introduce artifacts due to LLMs' behaviors when handling numerical information, both as input and output.

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.

Prior to data collection, ethical approval was obtained from the Swedish Ethics Authority, case no 2023-06788-01 "The International Centre for Evidence-Based Criminal Law". 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 Uppsala University participated in exchange for an EB-CRIME lecture. Participants gathered through social media did not receive monetary compensation.

Acknowledgements

We want to thank Fredrik Wahlberg for contributing to the design of the data collection and our reviewers for their valuable feedback and comments. This study was primarily funded by Knut and Alice Wallenberg Foundation, Wallenberg Academy Fellow, "Evidence-Based Criminal Law", with additional support from The Swedish Research Council, grant no. 2022-02056 "The International Centre for Evidence-Based Criminal Law". Gemini models were run using Google API credits gifted to the Computational Linguistics group at Uppsala University.

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 Peo-

ple'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 Gemini Team. 2026. Gemini 3.1 Flash Lite Preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite-preview`. Accessed May 2026.

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.

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.

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.

Lorenzo Pacchiardi, Alex Chan, Sören Mindermann, Ilan Moscovitz, Alexa 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 *International Conference on Learning Representations*, volume 2024, pages 54567–54604.

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

Qwen Team. 2025. Qwen3 Technical Report. ArXiv:2505.09388.

Andrew M Smith and Amy-May Leach. 2019. Confidence can be used to discriminate between accurate and inaccurate lie decisions. *Perspectives on Psychological Science : a journal of the Association for 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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. 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, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2025. Representation engineering: A top-down approach to ai transparency. ArXiv:2310.01405.

A Main Appendix

A.1 AI Statement

During the preparation of this work, Grammarly was used to improve clarity and flow. GitHub Copilot was used to draft and review code.

A.2 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.3 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.3.1)
- Where criteria-based judgments outperformed direct judgments (A.3.2)

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.3.1), or they are better understood through the implementation of criteria (see A.3.2).

A.3.1 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.3.2 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 inconsistent 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."

B Criteria Analysis Content Ablation

In this section, we present an ablation study to examine the roles of the scores and the rationale in

the criteria analysis.

We want to know how much the final judgments depend on the scores versus the generated rationale. This analysis is important for our third research question, which is primarily answered by analyzing the relationship between the criteria scores and the final judgment. This relationship was found to be moderate to high for various model setups (see Table 2 in section 4.3.

B.1 Ablation Design

We chose to perform this ablation study on Gemini-3.1-Flash-Lite and GPT-OSS-120b. The reason for including Gemini-3.1-Flash-Lite was that it performed the best and had the largest effect from the criteria analysis. For comparison, we also included an open-weight model and chose GPT-OSS-120b, since it had shown good compliance and was still available through Groq. The ablation was run in August 2026, which means the preview version of Gemini-3.1-Flash-Lite had been deprecated and replaced by the official version.

We used the same criteria analyses as in the original experiments, and ran the experiment the same number of times (3 runs for Gemini and 5 runs for GPT).

B.2 Results and Discussion

Figure 3 shows mean accuracy, truth bias, and confidence for the three ablation settings: both rationale and scores, rationale only, or scores only. For both Gemini-3.1-Flash-Lite-low and GPT-OSS-120b-low, there was no significant difference in accuracy between the three content variants. We can, however, observe signs that the models behave differently: the Gemini model seems to have benefited from including both rationale and scores, whereas the GPT model became less accurate with both.

Regarding truth bias, the content of the analysis had a significant effect on both models. For Gemini, we observe a synergy effect: the truth bias is highest when both rationale and criteria are included. For GPT, the opposite happens: truth bias drops dramatically. This behavior of GPT-OSS-120b-low is especially interesting, given that it was observed to be similarly false-biased in the direct judgment setting as in the two-step setting (0.22 in the direct and 0.19 in the criteria-based scenario).

For confidence, the effect for Gemini was insignificant. For the GPT model, we observe a significant decrease in confidence when both rationale and scores are included in the analysis. However, the effect size is small.

Model	Ablation type	Accuracy	Truth-Bias	Confidence
Gemini	**Both**	0.539	0.588	83.709
	Rationale only	0.512	0.513	82.638
	Scores only	0.528	0.539	84.697
	p-value	6.29e-01	2.62e-02*	6.54e-02
	Effect size	0.022	0.062	0.002
GPT	**Both**	0.466	0.194	69.262
	Rationale only	0.491	0.573	73.813
	Scores only	0.500	0.352	73.935
	p-value	2.70e-01	8.43e-72*	1.93e-45*
	Effect size	0.029	0.322	0.064

Table 3: Mean scores for Gemini-3.1-Flash-Lite and GPT-OSS-120b, with low reasoning effort. The p-values are from Chi-Square tests for accuracy and truth bias, and Kruskal-Wallis tests for Confidence. Effect sizes are reported as Cramérs V for accuracy and truth bias and Eta-squared for confidence.

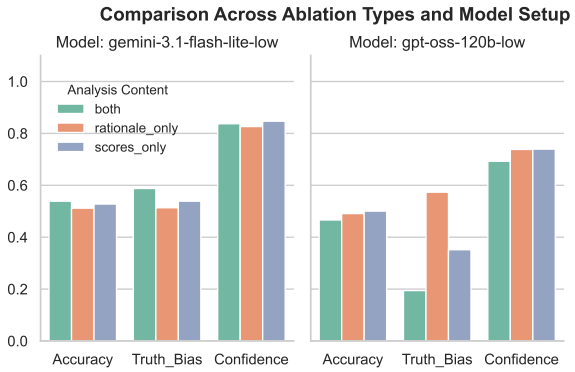


Figure 4: Comparison of final judgments from Gemini-3.1-Flash-Lite and GPT-OSS-120b-low across ablation types where the criteria analysis contains both rationale and scores, only rationale, or only scores.

From this analysis, we conclude that using both rationale and scores in the main experiments does not appear to be overall problematic for interpreting the relationship between the criteria scores and final judgments. We base this on the observation that accuracy is unaffected.

C Prompt Language Ablation from Pilot Study

During our pilot study, we performed an ablation, shifting the prompt language between English and Swedish. The purpose was to explore whether using Swedish or English in the prompt would have a considerable impact on the results and whether that would constitute evidence that Swedish prompts should be used.

Given that Swedish is the primary language in the Swedish legal context, this research would naturally focus on Swedish to the greatest extent possible. The criteria are themselves formulated in Swedish, and their English translations may have slight semantic differences. However, evidence suggests that multilingual LLMs may resort to an English-centered concept framework while processing other languages (Wendler et al., 2024). This would suggest that switching the prompts to Swedish would not necessarily make the assessment more nuanced and Swedish-oriented; it might just move the boundary of the language switch, since the LLM has been more conditioned to work with English from the beginning and might "think" in English regardless (Wendler et al., 2024).

Further, using English in the prompts may mean that models better understand the task and follow the output template better when the instructions are in English. Using English as the prompt language would also enable the same pipeline to be applied to further investigation of other datasets, for example, in English, which we plan to do in a future study. Lastly, since English is the language of presentation, using English makes the experiment more accessible and modular – more readily implementable with datasets in other languages.

C.1 Experimental Setup for the Pilot Study Prompt Language Ablation

We used the dev set of 40 statements, ten of each class (true confessions, false confessions, true denials, false denials). We then wrote prompts in both languages. This was done by having a native Swedish speaker correct automatic translations from Google Translate. More details about the prompts can be found under section C.3.

We ran the experiments on Llama-3.1-8b, Llama-3.3-70b, GPT-OSS-20b, and GPT-OSS-120b. All were run one time. For the analysis, we looked at differences in accuracy and token consumption. We separated completion tokens and prompt tokens for a more fine-grained analysis. We then performed a simple Student's t-test with Holm-Bonferroni correction using Scipy to assess whether there were differences in mean accuracy and token consumption between Swedish and English prompts.

C.2 Results from Prompt Language Ablation

Regarding accuracy, the analysis showed two instances of significant change when using Swedish prompts, compared to English ones. For Llama-3.1-8b, the accuracy increased significantly in the baseline setting (direct veracity judgments). For GPT-OSS-120b, Swedish prompts led to worse performance. See Figure 5.

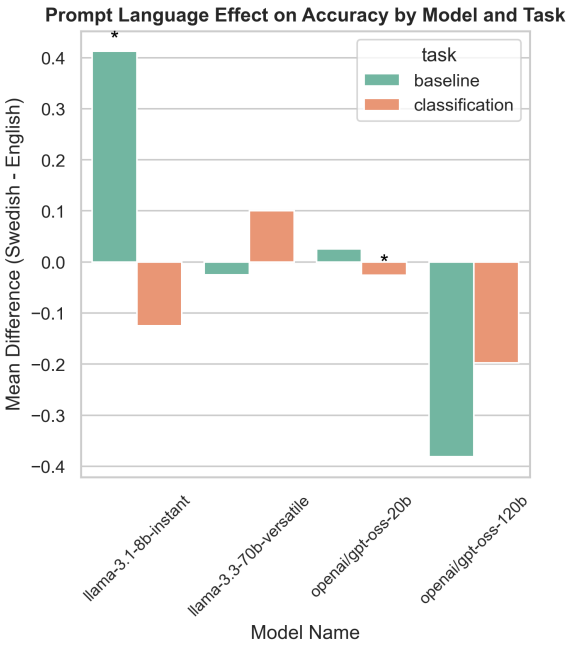


Figure 5: Difference in accuracy between Swedish and English prompts.

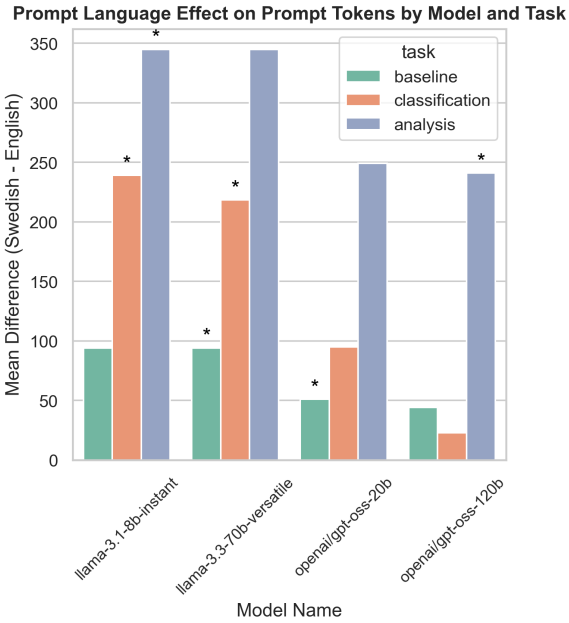


Figure 6: Difference in prompt tokens used when using Swedish and English prompts.

Prompt token consumption was more affected by changing the prompt language. The Swedish prompts yielded higher prompt token consumption across all models and tasks, as shown in Figure 6.

Statistically significant differences were found for all models for at least one task. Overall, larger differences are observed for the Llama model family.

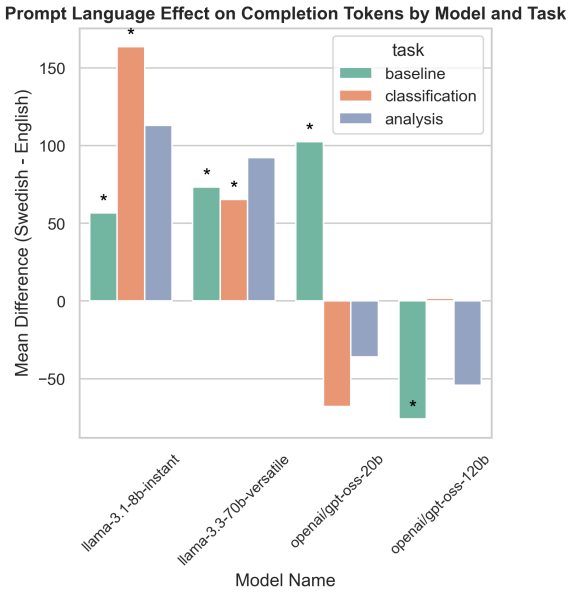


Figure 7: Difference in completion tokens used when using Swedish and English prompts.

For completion token counts, the effect of Swedish prompts on higher token counts was also observed for the Llama models, as shown in Figure 7. Notably, the language family appeared to have a greater impact on completion tokens than on prompt tokens. For GPT-OSS models, English prompts yielded higher completion token rates across all tasks except the baseline task for GPT-OSS-20b.

In summary, we observe that token consumption was overall more affected by the prompt language than was the accuracy. We also observe signs that models from different families may handle Swedish as prompt language to different degrees. The difference was overall larger for the Llama models. Based on these observations, we decided to use English prompts in the main experiments, as no clear gain was observed with Swedish.

C.3 Swedish Prompts from the Pilot Study

In this section, we present the Swedish prompts used in the pilot study. The prompts were written by having a native Swedish speaker adjust the output from Google Translate's automatic translation of the English prompts into Swedish. After conducting the pilot study, we made slight changes to the prompts. Some artifacts, therefore, exist in the pilot-study prompts that may affect comparisons with the final English prompts we use in our main experiments. For example, the order between classification (klassificering) and rationale (förklaring) is different. In the pilot study, the rationale was requested before classification, in both direct and criteria-based judgments. This may affect the transferability of the conclusions we draw from the pilot study regarding the effect of prompt language. We assume the covariance will be minor; however, we cannot be definite about it.

Pilot Study: Swedish Baseline Prompt

Du är en professionell brottsutredare som arbetar med ett stöldärende.
Stölden inträffade i ett hyreshus igår. Målsäganden, Leonard, har anmält att ett paket adresserat till honom har försvunnit från postrummet. Paketet innehöll en MacBook till ett ungefärligt värde av 25 000 SEK. Paketet stals från postrummet innanför entrén till hyreshuset.
Polisen har förhört Leonards grannar som potentiella misstänkta. Du ska nu läsa utsagan från en av grannarna, och din uppgift är att avgöra om denna granne verkar tala sanning eller inte.
Läs uttalandet noggrant och gör din bedömning. Svara med ###T### om du tror att utsagan är sann, och ###F### om du tror att utsagan är falsk.
Här är utsagan: {statement}
Mall för utdata (MÅSTE följas exakt):
Förklaring: <3-6 meningar på svenska>
Klassificering: ###<T/F>### Konfidensnivå: <0-1>

Pilot Study: Step 1: Swedish Analysis Prompt

Du är en professionell brottsutredare som arbetar med ett stöldärende.
Stölden inträffade i ett hyreshus igår. Målsäganden, Leonard, har anmält att ett paket adresserat till honom har försvunnit från postrummet. Paketet innehöll en MacBook till ett ungefärligt värde av 25 000 SEK. Paketet stals från postrummet innanför entrén till hyreshuset.
Polisen har intervjuat Leonards grannar som potentiella misstänkta. Du ska nu läsa ut-

sagan från en av grannarna, och din uppgift är att analysera utsagan baserat på följande kriterier: klar, detaljerad, lång, levande, logisk, självupplevd och rimlig.
Läs utsagan noggrant och fundera över i vilken grad utsagan uppfyller varje kriterium. Lista kriterierna i ditt svar och ge ett uppfyllelsepoäng på 0–100 för varje kriterium.
Här är utsagan: {statement}
Mall för utdata (MÅSTE följas exakt):
ANALYS: Klar: <0-100> Detaljerad: <0-100> Lång: <0-100> Levande: <0-100> Logisk: <0-100> Självupplevd: <0-100> Rimlig: <0-100>
Förklaring: <3-6 meningar på svenska>

Pilot Study: Step 2: Swedish Criteria-Based Veracity Judgment

Du är en professionell brottsutredare som arbetar med ett stöldärende.
Stölden inträffade i ett hyreshus igår. Målsäganden, Leonard, har anmält att ett paket adresserat till honom har försvunnit från postrummet. Paketet innehöll en MacBook till ett ungefärligt värde av 25 000 SEK. Paketet stals från postrummet innanför entrén till hyreshuset.
Polisen har förhört Leonards grannar som potentiella misstänkta. Du ska nu läsa utsagan från en av grannarna, tillsammans med en analys av samma utsaga, och din uppgift är att avgöra om denna granne verkar tala sanning eller inte.
Läs utsagan noggrant och överväg dess poäng för varje kriterium. Utsagan och den kriteriebaserade analysen ska vägleda dig i att avgöra om grannen talar sanning eller inte. Var transparent med hur säker du är på ditt beslut. Svara med ###T### om du tror att påståendet är sant, och ###F### om du tror att påståendet är falskt, och signalera hur säker du är med ett värde på 0–1.
Här är utsagan och kriteriepoängen: {statement} {criteria_analysis}
Mall för utdata (MÅSTE följas exakt):
Förklaring: <3-6 meningar på svenska>
Klassificering: ###<T/F>### Konfidensnivå: <0-1>