To Tell The Truth: Language of Deception and Language Models

Anonymous ACL submission

Abstract

Text-based false information permeates online discourses, yet evidence of people's ability to discern truth from such deceptive textual content is scarce. We analyze a novel TV game show data where conversations in a high-stake environment between individuals with conflicting objectives result in lies. We investigate the manifestation of potentially verifiable language cues of deception in the presence of objective truth, a distinguishing feature absent in previous text-based deception datasets. We show that there exists a class of detectors (algorithms) that have similar truth detection performance compared to human subjects, even when the former accesses only the language cues while the latter engages in conversations with complete access to all potential sources of cues (language and audio-visual). Our model, built on a large language model, employs a bottleneck framework to learn discernible cues to determine truth, an act of reasoning in which human subjects often perform poorly, even with incentives. Our model detects novel but accurate language cues in many cases where humans failed to detect deception, opening up the possibility of humans collaborating with algorithms and ameliorating their ability to detect the truth.

1 Introduction

Deception is pervasive in conversational dialogues. Individuals motivated by self-interest often feel compelled to embellish the truth to promote their interests at the expense of others. Misleading communication, such as false testimony (Tetterton and Warren, 2005), fake news (Shu et al., 2017), identity fraud in dating sites (Lazarus et al., 2022), sock puppetry (Kumar et al., 2017), and propaganda campaigns (Allcott and Gentzkow, 2017), abundant daily, impacts political, social, and economic outcomes. This exchange of information leading to the decision of who and what to believe necessitates the tacit development of truth detection capability during conversations (Bond and DePaulo, 2006).

In what follows, we explore if textual cues may increase the likelihood of fraud detection even in the presence of more overt visual or aural indicators. Consider the CEO scam, when fraudsters act as company executives to trick a victim into sending unauthorized wire transfers or divulging private information through email. In addition, textual cues may be crucial for an impartial observer to identify duplicity in social media conversations when audio and visual cues are often manipulated with little to no chance of face-to-face conversations (Rapoza, 2021).

This paper examines linguistic cues in a conversational exchange between contestants and judges participating in the TV game show, To Tell The Truth. In the game, the three contestants, under pretenses, mislead the four judges who attempt to infer the real central contestant (CC) via back-and-forth questioning. First, the game show offers a high-stake situation where contestants have a financial incentive to lie and deceive, and the judges are under pressure to perform in front of a crowd to detect the deception. Second, this data provides factual information to aid in assessing the contestant—a vital and distinguishing aspect of deception detection from the text than other datasets.

The study of deception detection using computational methods has traditionally focused on detecting 'what' is the truth using multimodal cues (Soldner et al., 2019). The definition of truth is often convoluted and depends on context; hence computationally detecting 'what is the truth' is challenging unless defined otherwise (Peskov and Cheng, 2020). Works that focused on exploring language cues (Fornaciari and Poesio, 2013; Ott et al., 2013) mainly restricted their analysis to psycholinguistic and hand-engineered linguistic features, which may not extend to scenarios where such cues are missing. Even though psycholinguistic features indicate the interlocutor's intention on a syntactic or token level, they may not demonstrate a deeper

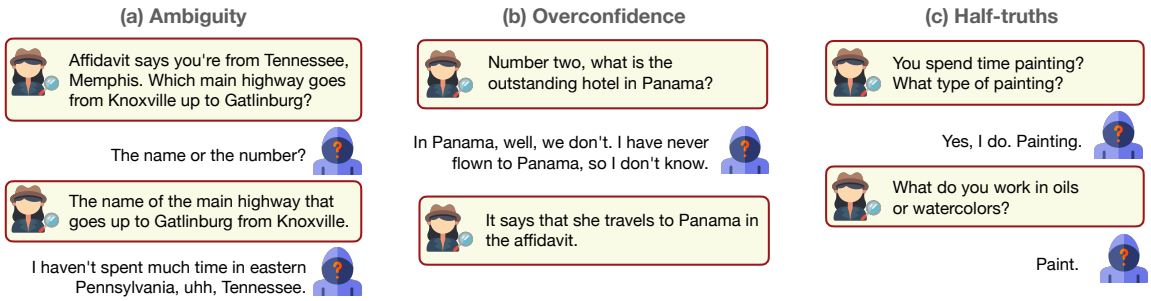

Figure 1: Examples of language cues for detecting deception: ambiguity, overconfidence, and half-truths from our dataset (T5). When used features, they can significantly enhance detection ability for both models and humans.

semantic understanding of the text in the discourse context. Recent progress in language models' ability to understand text prompted us to benchmark large language models' (LLMs) performance for the first time in detecting deception.

Armed with two main questions: 1. Do enough language cues exist to discern truth from deceptive conversations without other multimodal cues; and 2. Can a class of algorithmic detectors identify these cues, compose them in a valid chain of reasoning, and identify the truth?—in this paper, **we demonstrate a bottleneck framework** that progressively scans a deceptive conversation, analyzes each snippet by verifying utterances against objective truth, semantically understanding complex indicators of deception such as ambiguous responses, half-truths, and overconfidence, can satisfactorily reason its prediction for detecting deception. **We release a new conversational dataset**, To Tell The Truth from Text (T5), unique to the previous datasets, that contains a verifiable objective truth, forming the basis of lie detection. Our model can detect deception in cases where all judges failed to detect lies correctly, indicating its ability to uncover new reasoning chains that might be insightful to humans to learn better predictors for deception. Our model sometimes fails, where judges could correctly identify deception, leaving room for researchers to advance the frontier of the model performance in deception detection.

2 Related Work

Deception detection. Deception is an act emerging since the beginning of time with the Serpent and Eve in the Garden of Eden: *And the serpent said unto the woman, Ye shall not surely die*. However, humans are often not very good at spotting these lies (Belot et al., 2012; Gneezy, 2005), and are no better than making random decisions (Ockenfels and Selten, 2000). This raises the question of whether there are valid indicators of deception. According to studies (Wang et al., 2010; DePaulo et al., 2003; Ekman, 1997; Zuckerman et al., 1981; Wang et al., 2010), employing non-verbal signals such as visual, facial, and aural cues might significantly help distinguish sincere and opportunistic communication. Recent studies mention participants are significantly more accurate at spotting lies from both audio and videos (82%) or only videos (66%), compared to text (57%) (Wittenberg et al., 2021; Groh et al., 2022). In this paper, we investigate if a computational model can instead detect deception in text using language cues.

Deception + NLP datasets. Automated deception detection techniques so far predominantly utilized visual cues such as facial or eye movements to detect deception and time to response ((Meservy et al., 2005; Gonzalez-Billandon et al., 2019), or linguistic cues from transcriptions from court hearings (Fornaciari and Poesio, 2013), deceptive hotel reviews (Ott et al., 2013), news articles from Buzzfeed datasets (Potthast et al., 2017), and fact-checked tweets (Van Der Zee et al., 2022). Datasets that require assigning specific individuals roles (lier/truth-teller) include a multimodal conversational dataset Box of Lies (Soldner et al., 2019), Golden Globes (Darai and Grätz, 2013) differ from Diplomacy gameplay (Peskov and Cheng, 2020), and Real or Spiel (Ho et al., 2016) where one can *choose* to lie. In contrast, our derived dataset from this game show is conversational, grounded in a real deceptive environment with the presence of objective truth to detect the deception not present in existing datasets. The only work, (Banerjee et al., 2023) that investigated, To Tell The Truth gameshow neither built computational models nor analyzed cues from interactions.

Models detecting deception. Computational models that focus on language cues mainly use psy-

cholinguistic features (Gîrlea et al., 2016; Soldner et al., 2019), or syntactic parse of the texts (Soldner et al., 2019) to identify predictors for deception. Ho et al. (2016) utilized a power dynamics vocabulary to identify deception in long-term relationships. In our paper, we focus on more complex signals, such as ambiguity or half-truths, and benchmark the performance of LLMs for the first time in the context of deception detection from text.

3 To Tell The Truth

To examine conversational dynamics in deception detection, conversations among individuals laced with the intent to deceive given the prior unembellished truth is the key to our analysis. To Tell the Truth, season 1, is a game of deliberate misrepresentation and was aired on American TV weekly from 1956 to 1959. Every episode comprises multiple independent sessions. A regular session comprises a host, four judges, and three contestants. One of the three contestants was the central contestant (CC), while the other two were imposters.

3.1 A game walk-through

A session begins with the host asking every contestant's name (e.g., Jane Doe). The host publicly reads some factually true facts about the CC from a signed affidavit. The judges cross-question (for a fixed time) each contestant individually, by their respective numbers (Number one, two, or three).

The CC must answer truthfully to a question. The imposters, on the contrary, lie to impersonate the CC and deceive the judges. Following the definition of deception: "Typically, when [someone] lies [they] say what [they] know to be false in an attempt to deceive the listener" (Siegler, 1966), our setup is a perfect case of deception. At the end of cross-questioning, judges disentangle the facts from the fiction and independently (and simultaneously) submit their votes for the real CC.

For each incorrect identification of the CC, the entire group of contestants was paid $250 with a plausible individual maximum of $333. Given the fundamental tenet of preferences, all contestants ideally desire increased collective financial gain. This translates to the intention of all, including the real CC, to deceive the judges successfully.

3.2 Data Collection

For this paper, we derive a slightly different game, To Tell The Truth **from Text** (T5). We transcribe 150 such games using the Whisper, a state-of-the-art transcription model with a word error rate of 8.81% compared to human transcription's 7.61% (Radford et al., 2023).

During the early evaluation, we observed that all Whisper models (irrespective of size) often transcribed the proper names incorrectly. To address this, we manually review the automation-generated transcripts with the original video and corrected them for likely inconsistencies. Transcripts are cleaned for unnecessary noise or filler words in questions asked by judges (e.g., umm, uhh-hh) and any multi-lingual conversations beyond English. ("How do you pronounce your name in Russian?", "Please answer in French. I want to hear your accent."). We do not include irrelevant mockery and conversations in-between judges or with the host. Owing to noise and inconsistency in the rationale for judges' votes, we have refrained from including them in our dataset.

Comment on data leak in LLMs. Our dataset *does not* exist in its textual form on the internet. Hence we do not necessarily run the risk of direct data contamination when applying LLM on them. However, for extra caution, we randomly swapped the challenger identities (e.g., changed number one to number three and vice-versa), which means it is not possible to "copy" the answer from the internet, if available, as the labels are now swapped too. Additionally, we replaced the participant names with placeholders ('Participant_X'), where X is a random integer.

3.3 The T5 Dataset

Each data point in T5 entails three main components from an independent game session: the name of the real CC, the affidavit containing the objective truth about the CC, and the conversations (Q/As) between the judges and the contestants. Every data point, on average, has 12-15 Q/A pairs.

Our dataset is novel since it is based on conversations around the 1950s before social media and the internet existed. Unique interpersonal interactions exist in T5, for instance, Edmund Hillary, the first person to summit Mount Everest in 1953, was one of the CCs in a session, but his appearance was unknown within the US entertainment industry, analogous to contemporary online crimes with unknown scammers.

Ambiguity/Randomness. According to Ekman (1997), liars cannot keep their claims consistent,

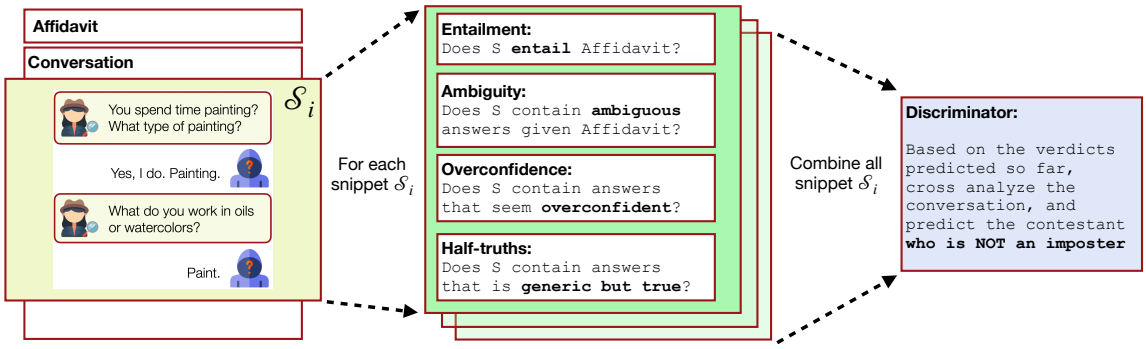

Figure 2: Pipeline of the bottleneck model deriving bottleneck controls and the discriminator collates them for final prediction. We use few-shot LLMs to extract such controls that outperform an end-to-end approach.

leading to ambiguity that exposes their lies. Figure 1 shows an example where the contestant mentions Pennsylvania while Tennessee is being discussed; this may have been an oversight on the contestant's part, or it may be a random interjection to stall a conversation; either way, it highlights the possibility that the contestant could be a liar.

Overconfidence. Overconfidence in the context of deception has been characterized in three ways: overestimating one's actual performance, overplacing of one's performance in comparison to others, and excessive precision in one's beliefs (Moore and Healy, 2008). This behavior is similar to the contestants' intent to mislead the judges in Figure 1 where the contestant lied too confidently (Serra-Garcia and Gneezy, 2021) and hence made a small but important factual error for judges to understand that the individual is an imposter.

Half-truths. In deceitful conversations, half-truths are less sinful than outright lies and explicit distortions (Carson, 2010). Given the constraint of the real CC to answer truthfully, uttering half-truths becomes a prominent strategy to suppress facts vital for identification (DePaulo et al., 2003). Figure 1 illustrates a contestant who does not mention what paint they use even after repeated questioning.

Dataset statistics. T5 is comprised of 150 data points with a volume of 86,746 words. There are 1546 utterances, including both judges and contestants as speakers. 450 unique contestants appeared in 150 sessions (datapoints), but judges reappeared from a unique set of size 56.

4 Approach

We investigate the possibility of using an LLM to uncover informative cues from the language of deception. Here we define the task, our base models, and a bottleneck model capable of reasoning the language cues for successful deception detection.

4.1 Task

We define the truth detection task as follows: given the name, affidavit($\mathcal{A}$), and a conversation ($\mathcal{C}$) as input x, predict the real contestant as y from an output label space of `Number One`, `Number Two`, `Number Three`. This is a discriminative task set up in the form of a 3-way classification problem. We use the terms *deception detection* and *truth identification* interchangeably in the rest of the paper.

4.2 Base Models

Our base model is an LLM. We initiate the base model with a **task prompt** that includes the brief description and rules of T5 such as all contestants are incentivized to lie with the constraint that real CC still has to stick to the truth. The input followed by the task prompt contains the name of the CC, the affidavit, and the conversation between the judges and the contestants. Please refer to the appendix for the complete prompt.

4.3 Bottleneck Models

Our base models process the input end-to-end to predict the real CC by performing necessary reasoning implicitly. However, implicit reasoning cannot be attributed to literature-backed linguistic cues, as discussed in Section 3.3. These cues can be explicitly extracted and used as a features. However, to ensure models restrict reasoning with these features, we use bottleneck models (Koh et al., 2020).

Identifying real CC involves assessing Q/A pairs addressed to a contestant at a time, as a **snippet** ($\mathcal{S}$) of the conversation and assessing the likelihood of the addressed contestant being the real CC. Our bottleneck models are employed through a set of

bottleneck controls, which are the high-level predictors for deception detection (from Section 3.3).

A bottleneck model takes the form of $f(g(\mathcal{S}))$; where g is a mapping function that maps the input snippet $\mathcal{S}$ to a bottleneck control, predictive for deception and f is the final discriminator that maps the intermediate bottleneck controls to the output label space. The success of the final prediction depends on the success of the intermediate functions generating bottleneck controls. We use LLMs for both f and g. For each g, we write a bottleneck prompt ($p_{\text{bottleneck}}$) for each control, which we discuss here (also see Figure 2). This simulates the systematic uncovering of these cues over the conversation, as the original judges would do.

Bottleneck controls.

- **Entailment:** As per game rules, each answer from the contestants should be verified in the light of the affidavit. We view this as an entailment task (Tafjord et al., 2022). Given a premise and a hypothesis, an entailment task would be to predict if the hypothesis entails, contradicts, or does not relate to the premise. We set the affidavit $\mathcal{A}$ as the premise and a snippet $\mathcal{S}$ as the hypothesis and predict one of these: entail, contradiction, or neutral.

- **Ambiguity/Randomness:** Each snippet contains ambiguous or deliberately random responses from the contestants, indicative of deception (see Figure 1). We develop a bottleneck prompt that takes a snippet $\mathcal{S}$ as an input to predict control values: ambiguous or unambiguous, in the light of the contestant being deceptive.

- **Overconfidence:** Similarly as above, the next bottleneck prompt ascertains if the responses reveal overconfidence in a contestant (e.g., Figure 1), indicating deception. The model predicts a verdict: overconfident or neutral.

- **Half-truths:** Finally, we develop a bottleneck prompt to decode an utterance as a half-truth (example in Figure 1) to predict if the snippet contains half-truths and hence is indicative of deception or not.

While the bottleneck controls are predicted for each conversation snippet, they can be derived either *independently* or *sequentially*. It is analogous to the original setting, where the snippets appear one by one, with the possibility that an older snippet may influence future questions from the judges and future answers from the contestants. For independent bottleneck controls, the mapping function is realized as $g(\mathcal{S}_i)$ for the i-th snippet. For sequential bottleneck control, the mapping function takes the form $g(\mathcal{S}_1, \cdots, \mathcal{S}_{i-1}, \mathcal{S}_i)$.

Discriminator. The discriminator function $f(\cdot)$ takes annotated part of the conversations with the derived bottleneck controls for every snippet, to predict the real CC.

We use OpenAI LLMs as the candidate base models and also for f and g: `text-davinci-003` (Brown et al., 2020), `gpt3.5-turbo-16k`, and `gpt4` (OpenAI, 2023).

4.4 Baselines and Evaluation

One of the primary baselines for our system is to compare the model's performance with human performance.

Base models. For zero-shot models, our primary baselines will be the base models with all LLM variants that do not break the decision-making process through bottlenecks. Kojima et al. (2022) show encouraging performance when a chain-of-thought (CoT) prompt is added to a zero-shot LLM: "`Let's think step by step`,"—becomes our baseline.

Supervised Models. For completeness, we also consider three supervised baselines where we only train the discriminator f using XGBoost classifier, mirroring (Soldner et al., 2019). For the features required for the XGBoost classifier, we consider two options: psycholinguistic features from Soldner et al. (2019) and `gpt-3` embeddings of derived bottleneck controls from our `gpt-4`-based bottleneck model. For the LIWC-supervised baseline, we generate LIWC features for responses given by each contestant and concatenate them for the complete feature vector for the classifier. Finally, we train a BERT model (Devlin et al., 2019). For a fair comparison, we evaluate supervised models by a leave-one-out scheme spanning the full T5 dataset.

Bottleneck Models. For variations of our bottleneck approach, we create all possible combinations for f and g with our LLM variants. `gpt-4` as both $f(\cdot)$ and $f(\cdot)$ is our model, and rest 8 are baselines. We ablate four bottleneck controls individually while keeping the rest the same to compare with our model. We also evaluate if independent or sequential bottleneck derivation affects model

Models	Acc (↑)	Acc@2 (↑)	% wins (↑)	κ (↑)
Human*	41.3	–	–	–
Random	33.3	66.6	–	–
Base Models (best setups)				
GPT-3, 2-shot	29.3	56.0	100	0.85
GPT-3.5, 2-shot	33.3	70.0	97	0.81
GPT-4, 2-shot	34.7	72.0	90	0.74
CoT Models (best setups)				
GPT-3, 2-shot	27.3	55.3	100	0.91
GPT-3.5, 2-shot	30.0	65.3	100	0.91
GPT-4, 2-shot	32.0	64.7	97	0.84
Bottleneck Models (best combinations)				
f: GPT-3, g: GPT-4, 0-shot	29.3	57.3	93	0.84
f: GPT-3, g: GPT-4, 2-shot	30.0	58.1	93	0.85
f: GPT-3.5, g: GPT-4, 0-shot	35.3	71.3	77	0.70
f: GPT-3.5, g: GPT-4, 2-shot	36.0	72.0	77	0.73
f: GPT-4, g: LIWC	33.3	71.3	100	0.91
f: GPT-4, g: GPT-4, 0-shot	**39.3**	**77.3**	73	0.68
f: GPT-4, g: GPT-4, 2-shot	**39.3**	**77.3**	–	–
Supervised Models (leave-one-out)				
BERT	35.3	71.3	–	–
GPT-3-emb. + XGBoost	35.3	68.0	–	–
LIWC + XGBoost	34.0	67.3	–	–

Table 1: Accuracy (Acc) and Accuracy@2 (Acc@2) across models. % wins indicate human pairwise evaluation for the explanations with Fleiss' κ (Fleiss and Cohen, 1973) scores as agreement.

performance. Finally, Soldner et al. (2019) suggests LIWC features (Ott et al., 2013) are effective in predicting hidden intents in deceptive communication. We use such LIWC features as $g(\cdot)$, an alternative to our bottleneck features, pairing them with a `gpt-4` based discriminator. For all baselines, prompts are provided in the appendix.

Evaluation To evaluate model performance, we use accuracy and accuracy@2; the latter denotes if the correct prediction appears in the top two guesses. We use the session-level macro-average accuracy for human performance, as every session (one datapoint in T5) has predictions from 4 judges. We use a pairwise comparison in AMT and an absolute metric to evaluate the quality of the generated explanations from the models. For pairwise comparison, we measure % of times explanations generated by our model are preferred by 3 human evaluators (in the majority) than explanations from a competing baseline. Following Majumder et al. (2021), we use the e-ViL score on explanation where the models predicted accurately. We ask the annotators *if an explanation is satisfactory* with four options: yes, partial-yes, partial-no, and no. This required us to take an intersection of samples when both comparing models generated correct predictions, an average of which was 31.

Models	Acc (↑)	Acc@2 (↑)	% wins (↑)	κ (↑)
Human*	40.0	–	–	–
Random	33.3	66.6	–	–
Bottleneck Zero-shot Model. f: GPT-4, g: GPT-4, sequential				
Full model	39	77	–	–
w/o entailment	34 (-5)	71 (-6)	97	0.91
w/o ambiguity	35 (-4)	71 (-6)	90	0.91
w/o overconfidence	34 (-5)	71 (-6)	93	0.91
w/o half-truths	32 (**-7**)	68 (**-9**)	100	0.96
independent-full	36 (-3)	75 (-2)	93	0.85

Table 2: Performance of the models under ablation. Numbers in parenthesis denotes performance drop from the 'Full model'.

5 Experiments

Our experiments are driven by the following hypotheses, and results are analyzed accordingly:

Can few/zero-shot models detect deception? Yes. Table 1 contains the results. The bottleneck model with GPT-4 as both g and f performs the best in accuracy and accuracy@2. GPT-4 as g consistently yields better performance than GPT-3.5 or GPT-3. Both GPT-3.5 and GPT-4 as f are capable of detecting deception better than random. Base models perform worse than bottleneck models irrespective of the choice of the LLM and yield accuracy the same as random. However, the CoT models further decrease accuracy across all LLMs, primarily attributing to the CoT models' attempt to reason the potentially irrelevant or distracting information present in the conversation.

Why bottleneck models are better than base models? Table 2 shows the results of our ablation study that investigates the effect of each bottleneck control. The largest drop (7 points for Accuracy) occurs when the half-truth control is omitted, which aligns with our dataset analysis where real CC deliberately utters half-truths to deceive the judges, which, unless understood, may confuse the judges as well as models. We also find the quality of controls is better when derived sequentially (with conversation history) as compared to independent derivations, mirroring the original game setting.

Are few-shot models better than zero-shot models? Depends. For smaller or earlier LLMs (GPT-3, GPT-3.5), few-shot examples helps to improve their performance from zero-shot setup. However, for bottleneck models (and even in base/CoT models), GPT-4 achieves similar performance both in few-shot and zero-shot setups. Adding more demonstrations is often difficult due to LLM's lim-

ited context length. Further summarization (Park et al., 2023) or selection of few-shot examples (Madaan et al., 2022) are possible, but we leave it as a future work.

Are supervised models better than few/zero-shot models? **No**. Despite training on almost the full dataset (leave-one-out), the state-of-the-art supervised models do not outperform our zero-shot bottleneck models, indicating the superior ability of the LLMs to derive better bottleneck controls and act as a better discriminator.

Can models explain their chain-of-reasoning to detect deception? **Yes**. Table 1 shows results for human evaluation. In the pairwise comparison, our best model (bottleneck, f: GPT-4, g: GPT-4) wins unanimously against all other competing baselines. Indeed, the quality of the explanations is distinctly worse for bottleneck models using GPT-3 or GPT-3.5 when compared to GPT-4, denoting GPT-4's ability to better bottleneck controls which form the basis of a good explanation. Similarly, base models cannot generate high-quality explanations mainly due to their unconstrained nature of generation. Figure 3b shows that our model has the highest e-ViL score, reflecting the trend from pairwise comparisons.

Qualitative analysis of the model generated explanations. Figure 4 exhibits model-generated examples. In Figure 4a, the model correctly identifies the CC, whereas *All* of the judges fail. They mistook detailed information and deceptive appearances of the imposters (here, a swimmer's tan, athletic body, etc.). The model identifies bizarre details and randomness, even in the incorrect information by the imposters, while trying to capture the essence of the half-truths (highlighted) given by the real CC. In similar other cases where the model outperforms all the judges, the model mentions humor, overconfidence, deliberate incompleteness in answers, and wordy descriptions as possible cues of deceptive challengers.

In Figure 4b, the model predicts CC correctly, and the success for the judges is divided into half. Here, the model accurately recognizes a mistakenly erroneous response provided by the real CC as well as an "overall consistent behavior," denoting honesty. The judges who were not duped might have noticed CC's constant behavior or the inconsistencies among the imposters, while the duped judges might have relied on CC's inaccurate response to

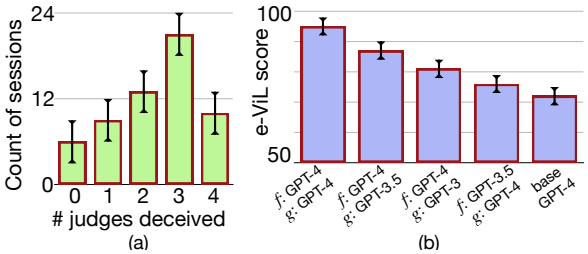

Figure 3: **(a)** Histogram of a number of judges who were deceived in session where our model predicted correctly. The distribution is skewed; skewness = -0.501, indicating our model predicts correctly, significantly ($p < 0.1$) more in sessions where more judges were deceived. **(b)** e-ViL scores for explanations from our top-5 accurate models, showing our best model generates more satisfactory explanations.

draw their conclusions. One of the judges who was not deceived in this case reappeared in the majority of the sessions, which indicates possible learning to support her claim (Banerjee et al., 2023).

Can models detect deception better than humans? **Depends**. Note that the human judges had access to all potential audio-visual and language cues in the original setting. Audio-visual cues are very important, when present, in deception detection primarily due to the ease and abundance of detecting them (DePaulo et al., 2003). Indeed, judges consistently use visual cues to explain their guesses, such as detecting the real swimmer/climber observing sun tan on the contestants or picking on the twinkle in the eyes for detecting a real joker. Considering this, our best model's performance is almost comparable with human performance—in absolute terms, our model incorrectly predicted 3/150 more examples compared to human judges. This is slightly different than results obtained in (Soldner et al., 2019), where they show linguistic features are the weakest predictors. However, it might be due to the fact that their linguistic features (part of them are LIWC features) were not expressive enough, or our dataset contains more linguistic cues than theirs. This also signifies the uniqueness of T5 with the existence of the objective truth (affidavit), the basis of deception detection, and cross-questions revealing important factual and language cues to detect the real CC.

Do models detect deception for a complementary set than humans? **Yes**. Among the 59/150 correct predictions by our best model, we analyze how human judges performed in those game ses-

(a) Model: Not deceived | Judges: All deceived | Correct answer: Number One

Number One: This contestant gave short, incomplete but true answers. For example, when asked about the boat she took to England, she said she didn't go by boat because she gets seasick. **This is a half-truth as she didn't specify how she actually got to England.** (...)
Number Two: This contestant gave long, detailed answers. (...) However, she was unsure about the name of the famous hotel in Montreal, which could indicate that she is an imposter.
Number Three: This contestant gave a random answer when asked about the time it took Gertrude Ederle to swim the English Channel. (...)

(b) Model: Not deceived | Judges: 2 deceived, 2 not deceived | Correct answer: Number One

Number One: This contestant gave generic and incomplete answers, which is a characteristic of the real person according to the hints. (..) They also struggled with the question about the size limit for a toy poodle, which could indicate unfamiliarity with the American measurement system, consistent with the affidavit stating that Maria is Brazilian. However, they failed to answer the question about candling an egg, which is a basic veterinary practice. **Despite their failure to answer one question, Number One's overall performance was most consistent.**
Number Two: This contestant gave **detailed and confident answers, which is a characteristic of an imposter.** They knew the correct (...) . However, they incorrectly stated (...).
Number Three: This contestant gave some correct answers, (...). However, they also gave some incorrect and vague answers. For example, they (...) suggested that the cure for a common chicken disease is to break the chicken's neck, **which is incorrect and unprofessional. Their answers are generally detailed and confident.**

(c) Model: Deceived | Judges: All not deceived | Correct answer: Number Three

Number One: He seems to be truthful and not overly detailed in his responses. **He doesn't know what a double clutch is**, which is a bit suspicious for a driving champion, but it's possible he never encountered it. His answer about the gas station address is generic but could be true. (...)
Number Two: He seems to be very detailed in his responses, which could indicate he's an imposter. (...) He also provides a specific address for his gas station, which could be made-up..
Number Three: He gives a random answer about parking near a hydrant, which could indicate he's making up facts. (...) His answer about double clutching is detailed and could be true, but it's also a bit overconfident. He provides a specific address for his gas station, which could be a made-up detail.
Based on these observations, I would **rank the contestants** in the following order of likelihood of not being an imposter: 1. Number One, **2. Number Three**, 3. Number Two.

Figure 4: Model generated explanations. Correct cues are colored in blue, and wrong cues are colored in red.

sions. Figure 3a shows the histogram of the discrete variable–the number of judges deceived per session, considering only 59 sessions said above. The plot is skewed towards a higher number of deceived judges per session, indicating that our model performed better on a set where humans detect deception poorly. This implies our model discovered novel reasoning pathways, purely based on language cues, for correct predictions, which human judges failed at, even with the presence of multimodal cues and incentives. We qualitatively examine the explanations generated by the model where all judges got deceived (10/59) and discovered that our model generated more informative and plausible reasoning chains than explanations by human judges mentioned during the original sessions (see Figure 4).

Can bottleneck features enhance human detection rate? Yes. We performed a correlation analysis between the presence of a bottleneck feature and human prediction jointly using a multivariate regression model on cases where having these features significantly helped the computational model to predict correctly. On the intersection where humans were wrong but the bottleneck model was correct, we find either no correlation ($p < 0.05$ for entailment and ambiguity) or negative correlation ($p < 0.05$ for overconfidence and half-truths), indicating humans when poor at detecting such linguistic cues suffers significantly in detecting lies.

Are some sessions more difficult than others? To the model, yes; to humans, mostly no. Banerjee et al. (2023) confirm no selection bias among the contestants across the sessions in the original game, indicating similar difficulty levels for the judges across sessions. They also highlight weakly significant selection bias among the judges, indicating better performance via learning and possible further selection as a judge in the session might have influenced their performance. Qualitatively, we see models' errors clustered around sessions where judges use non-language cues successfully; however, model predictions among themselves were highly correlated ($\rho = 0.7, p < 0.05$) for each underlying LLM variant. Figure 4c shows a case where the model fails to recognize the CC, but none of the judges were deceived. The model hallucinates in its reasoning to decide between Number One and Number Three as CC. It provides reasons for both to have incomplete, somewhat more detailed, a few factually accurate responses, and however forth—making the prediction at random. However, the real CC exists in the top two choices from the models, indicating judges picked up cues beyond language. Indeed, in this case, judges recognized the Southern accent of the CC and verified via the affidavit, a critical aural cue that was not available to our model.

6 Conclusion and Outlook

In this paper, we first showed the existence of a class of algorithmic detectors based on LLMs that can successfully identify language cues of deception without the presence of other visual or audio cues. We contribute a novel dataset T5 for deception detection in the presence of objective truth and achieve a model performance comparable to human performance. We further find that our best model performs well in cases where humans perform poorly and discover novel language models that could augment human reasoning to detect deception, opening up the possibility of human-LLM collaborations to combat misinformation.

7 Limitations

We acknowledge that online misinformation can be very different in nature than lies in our dataset; however, we find examples of false information in Quora, Reddit where non-experts with propaganda use strategies like half-truths to misguide people. T5 is a relatively small dataset; however, we showed that statistically significant analysis can be done with it. The human prediction data is derived from the original game show; hence the setup may not match exactly when we are evaluating text models. We are running additional human experiments to gather true human performance on T5.

8 Ethical Concerns

To Tell The Truth videos are freely available on YouTube, and we transcribed freely available videos using an open-source transcription model. The dataset is in English. The original sessions occurred in the 1950s; hence we do not observe an equitable diversity in gender when it comes to the gender of the challengers. For all sessions, there were two female judges and two male judges. We occasionally observe judges asking questions that are biased toward gender or race; hence any model that will be trained on this dataset may risk containing similar bias. In our paper, we do not train any generative model on this data minimizing that risk. We acknowledge the potential misuse of such truth-detection systems, and we are following up with controlled experiments to understand if humans would over-rely on such systems.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Priyodorshi Banerjee, Sanmitra Ghosh, and Sanchaita Hazra. 2023. Experience, learning and the detection of deception. *Journal of Economic Criminology*, page 100010.

Michèle Belot, V Bhaskar, and Jeroen Van De Ven. 2012. Can observers predict trustworthiness? *Review of Economics and Statistics*, 94(1):246–259.

Charles F. Jr. Bond and Bella DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10:214 – 234.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Thomas L Carson. 2010. *Lying and deception: Theory and practice*. OUP Oxford.

Donja Darai and Silvia Grätz. 2013. Attraction and cooperative behavior. *University of Zurich Department of Economics Working Paper*, (82).

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Paul Ekman. 1997. Deception, lying, and demeanor. *States of mind: American and post-Soviet perspectives on contemporary issues in psychology*, pages 93–105.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21:303–340.

Codruta Gîrlea, Roxana Girju, and Eyal Amir. 2016. Psycholinguistic features for deceptive role detection in werewolf. In *North American Chapter of the Association for Computational Linguistics*.

Uri Gneezy. 2005. Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Jonas Gonzalez-Billandon, Alexander M Aroyo, Alessia Tonelli, Dario Pasquali, Alessandra Sciutti, Monica Gori, Giulio Sandini, and Francesco Rea. 2019. Can a robot catch you lying? a machine learning system to detect lies during interactions. *Frontiers in Robotics and AI*, 6:64.

Matthew Groh, Aruna Sankaranarayanan, Andrew Lippman, and Rosalind Picard. 2022. Human detection of political deepfakes across transcripts, audio, and video. *arXiv preprint arXiv:2202.12883*.

Shuyuan Mary Ho, Jeffrey T Hancock, Cheryl Booth, and Xiuwen Liu. 2016. Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication. *Journal of Management Information Systems*, 33(2):393–420.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *ICML*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th international conference on world wide web*, pages 857–866.

Suleman Lazarus, Mark Button, and Richard Kapend. 2022. Exploring the value of feminist theory in understanding digital crimes: Gender and cybercrime types. *The Howard Journal of Crime and Justice*, 61(3):381–398.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. *ArXiv*, abs/2201.06009.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *International Conference on Machine Learning*.

Thomas O Meservy, Matthew L Jensen, John Kruse, Judee K Burgoon, Jay F Nunamaker, Douglas P Twitchell, Gabriel Tsechpenakis, and Dimitris N Metaxas. 2005. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43.

Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review*, 115(2):502.

Axel Ockenfels and Reinhard Selten. 2000. An experiment on the hypothesis of involuntary truth-signalling in bargaining. *Games and Economic Behavior*, 33(1):90–116.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *ArXiv*, abs/2304.03442.

Denis Peskov and Benny Cheng. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Kenneth Rapoza. 2021. Can "fake news" impact the stock market?

Marta Serra-Garcia and Uri Gneezy. 2021. Mistakes, overconfidence, and the effect of sharing on detecting lies. *American Economic Review*, 111(10):3160–3183.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Frederick A Siegler. 1966. Lying. *American Philosophical Quarterly*, 3(2):128–136.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Veronica S Tetterton and Amye R Warren. 2005. Using witness confidence can impair the ability to detect deception. *Criminal Justice and Behavior*, 32(4):433–451.

Sophie Van Der Zee, Ronald Poppe, Alice Havrileck, and Aurélien Baillon. 2022. A personal model of trumpery: linguistic deception detection in a real-world high-stakes setting. *Psychological science*, 33(1):3–17.

Joseph Tao-yi Wang, Michael Spezio, and Colin F Camerer. 2010. Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American economic review*, 100(3):984–1007.

Chloe Wittenberg, Ben M Tappin, Adam J Berinsky, and David G Rand. 2021. The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences*, 118(47):e2114388118.

Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier.

A Experiments and Results

Base Models For GPT-3.5 and GPT-4[1], we provide both the system and the user prompts. For GPT-3, there is only one prompt; hence the system and user prompts from above will be merged into one.

Bottleneck Models We have four bottleneck prompts for entailment, ambiguity, overconfidence, and half-truths.

Complete qualitative examples Figures 5 to 7 contain complete illustrations of the three qualitative examples discussed in the main paper.

B Datasheet

Motivation for Dataset Creation. To Tell The Truth from Text (T5) was created primarily to study the language of deception; by analyzing interlocutors' utterances with well-known predictors of deception and reasoning theories to design, develop, and evaluate models for a deception detection task purely based on language cues. Banerjee et al. (2023) explore the nature of the deception detection task in the original setting, while our task is a derivative of the original game, focusing on textual misinformation.

Recent seasons: Not only the show had been intermittently revived from 2016-2022 on ABC, but the new shows are also not as structured as the older ones. Few snippets of the full show that exist include more features of entertaining acts, unstructured questioning, and no compensation for the challengers (imposters), leading to uncertainties regarding the participants' true intent to deceive.

Dataset Composition. To Tell the Truth, season 1, is a game of deliberate misrepresentation and was aired on American TV weekly from 1956 to 1959[2]. Every episode comprises multiple independent sessions. A regular session comprises a host, four judges, and three contestants. One of the three contestants was the central contestant (CC), while the other two were imposters. The contestants come from all walks of life, including the US Mint director, an Olympic swimmer, a bachelor who served in World War II and lives in Long Island, and others. The judges are well-known figures in the Hollywood entertainment industry. In T5, there are 1546 utterances, including both judges and contestants as speakers. 450 unique contestants appeared in 150 sessions (datapoints), but judges reappeared from a unique set of size 56. The supplementary code folder has one sample datapoint from T5. Upon acceptance, we will release the full dataset.

Data Collection Process. For this paper, we derive a slightly different game, To Tell The Truth, **from Text** (T5). We transcribe 150 such games using the Whisper, a state-of-the-art transcription model with a word error rate of 8.81% compared to human transcription's 7.61% (Radford et al., 2023).

Data Preprocessing. During the early evaluation, we observed that all Whisper models (irrespective of size) often transcribed the proper names incorrectly. To address this, we manually review the automation-generated transcripts with the original video and corrected them for likely inconsistencies. Transcripts are cleaned for unnecessary noise or filler words in questions asked by judges (e.g., umm, uhh-hh) and any multi-lingual conversations beyond English. ("How do you pronounce your name in Russian?", "Please answer in French. I want to hear your accent."). We do not include irrelevant mockery and conversations in-between judges or with the host. Owing to noise and inconsistency in the rationale for judges' votes, we have refrained from including them in our dataset.

[1] https://platform.openai.com/docs/models

[2] https://en.wikipedia.org/wiki/To_Tell_the_Truth

System: You are an expert detective.

User: You are provided with a situation of deliberate misrepresentation. There are three contestants, Number One, Number Two, and Number Three, all claiming to be the same person. However, there is only one of them who is actually the person they are all claiming to be. There is an affidavit that includes truthful information about the claimed person. The affidavit is publicly available. However, the affidavit does not contain all the truth about the person, and it is possible that new information is obtained in the conversation that is actually true.

There is a judge who cross-questions all three contestants to uncover the truth. All the contestants will get monetary rewards if they can deceive the judge. However, the contestant, who is the real person, is sworn to speak truthfully. But they will most often utter half-truths to deceive the judges.

The name of the claimed person is given as Name:
The affidavit of the claimed person is given as Affidavit:
The conversation (in a question-answer form) between the judge and the contestants is given as Conversations:
If a question starts with addressing a particular contestant (e.g., Number One), then that question and all the following questions are addressed to that contestant unless a different contestant (e.g., Number Two) is explicitly addressed.

Based on the affidavit and the conversation, predict the contestant who is not an imposter. First, generate your rationale behind your prediction. Then, write ### followed by the single option from Number One, Number Two, Number Three as the answer.

<Input Name, Affidavit, Conversation>

Answer:

We will analyze the conversation snippet by snippet in the order it happened originally. For every snippet:

Entailment: Cross-check the answer with the affidavit (you can verify the information from your factual knowledge). Use your judgment to generate a verdict that the answer(s) make the addressed contestant a likely imposter or likely the true person.

Ambiguity: A random or unrelated answer may indicate an imposter as they might be making up the facts. Based on this, generate a verdict that the answer(s) make the addressed contestant a likely imposter or likely the true person.

Overconfidence: An overconfident answer may indicate an imposter as they will go overboard to prove themselves as real. Based on this, generate a verdict that the answer(s) make the addressed contestant a likely imposter or likely the true person.

Half-truths: The real person is sworn to speak truthfully. Hence in order to deceive, they may speak half-truths. Based on this, generate a verdict that the answer(s) make the addressed contestant a likely imposter or likely the true person.

Correct Prediction: Number One
Model Prediction: Correct
Judges Predictions: All Wrong

Model Explanation:
Based on the affidavit and the conversation, here's my analysis:
Number One: This contestant gave short, incomplete but true answers. For example, when asked about the boat she took to England, she said she didn't go by boat because she gets seasick. This is a half-truth as she didn't specify how she actually got to England. When asked about what's on her mind after 21 hours in the water, she said she didn't remember anything. This is a generic answer that could be true. She also correctly named the two most important department stores in Toronto.
Number Two: This contestant gave long, detailed answers. For example, when asked about the grease used by channel swimmers, she specified axle grease. She also correctly answered the question about the governor general of Canada. However, she was unsure about the name of the famous hotel in Montreal, which could indicate that she is an imposter.
Number Three: This contestant gave a random answer when asked about the time it took Gertrude Ederle to swim the English Channel. She also gave a detailed answer about the food she eats when distance swimming, which could indicate that she is an imposter. However, she was truthful when she said she was afraid of her legs getting tired and of eels in the water.
Based on this analysis, I would rank the contestants in the following order of likelihood of not being an imposter: 1. Number One 2. Number Two 3. Number Three

Figure 5: Qualitative example

Correct Prediction: Number One
Model Prediction: Correct
Judges Predictions: 2 Correct, 2 Wrong

Model Explanation:
Based on the affidavit and the conversation, we can make the following observations:
1. Number One: This contestant gave generic and incomplete answers, which is a characteristic of the real person according to the hints. For example, when asked about poultry pathology, they gave a very basic answer. They also struggled with the question about the size limit for a toy poodle, which could indicate unfamiliarity with the American measurement system, consistent with the affidavit stating that Maria is Brazilian. However, they failed to answer the question about candling an egg, which is a basic veterinary practice.
2. Number Two: This contestant gave detailed and confident answers, which is a characteristic of an imposter according to the hints. They knew the correct number of toes on a poodle and the correct term for a French poodle. They also knew the correct degree for a doctor of veterinary medicine. However, they incorrectly stated that there is only one variety of poodle, which is incorrect as there are three varieties: Toy, Miniature, and Standard.
3. Number Three: This contestant gave some correct answers, such as the life expectancy of a chicken and the meaning of "saudades". However, they also gave some incorrect and vague answers. For example, they failed to correctly name the three sizes of poodles and suggested that the cure for a common chicken disease is to break the chickenś neck, which is incorrect and unprofessional.
Based on these observations, the ranking from most likely to least likely to be the real Maria Angela Magalaich Dolovera is as follows:
1. Number One: Despite their failure to answer one question, their overall performance was most consistent with the hints about the real person's behavior.
2. Number Three: Although they gave some correct answers, their incorrect and vague answers suggest that they may be an imposter.
3. Number Two: Their detailed and confident answers, along with their incorrect statement about the varieties of poodles, strongly suggest that they are an imposter.

Figure 6: Qualitative example

Correct Prediction: Number Three
Model Prediction: Wrong
Judges Predictions: All Correct

Model Explanation:
Based on the affidavit and the conversation, here's my analysis:
Number One: He seems to be truthful and not overly detailed in his responses. He doesn't know what a double clutch is, which is a bit suspicious for a driving champion, but it's possible he never encountered it. His answer about the gas station address is generic but could be true. He correctly identifies the octagon-shaped road sign as a stop sign, which is basic knowledge for a driver.
Number Two: He seems to be very detailed in his responses, which could indicate he's an imposter. He doesn't know what a double clutch is either, but he correctly identifies the diamond-shaped road sign as a hazard and explains the meaning of a broken line on the road. He also provides a specific address for his gas station, which could be a made-up detail.
Number Three: He gives a random answer about parking near a hydrant, which could indicate he's making up facts. He correctly identifies the square-shaped road sign as providing information about the road. His answer about double clutching is detailed and could be true, but it's also a bit overconfident. He provides a specific address for his gas station, which could be a made-up detail.
Based on these observations, I would rank the contestants in the following order of likelihood of not being an imposter: 1. Number One 2. Number Three 3. Number Two

Figure 7: Qualitative example