Hidden in Plain Sight: Evaluation of the Deception Detection Capabilities of LLMs in Multimodal Settings

Md Messal Monem Miah[1], Adrita Anika[2*], Xi Shi[1], Ruihong Huang[1]
[1]Texas A&M University
[2]Amazon
{messal.monem, xishi, huangrh}@tamu.edu, adritani@amazon.com

Abstract

Detecting deception in an increasingly digital world is both a critical and challenging task. In this study, we present a comprehensive evaluation of the automated deception detection capabilities of Large Language Models (LLMs) and Large Multimodal Models (LMMs) across diverse domains. We assess the performance of both open-source and proprietary LLMs on three distinct datasets—real-life trial interviews (RLTD), instructed deception in interpersonal scenarios (MU3D), and deceptive reviews (OpSpam). We systematically analyze the effectiveness of different experimental setups for deception detection, including zero-shot and few-shot approaches with random or similarity-based in-context example selection. Our findings indicate that fine-tuned LLMs achieve state-of-the-art performance on textual deception detection, whereas LMMs struggle to fully leverage multimodal cues, particularly in real-world settings. Additionally, we analyze the impact of auxiliary features, such as non-verbal gestures, video summaries, and evaluate the effectiveness of different prompting strategies, such as direct label generation and post-hoc reasoning generation. Experiments unfold that reasoning-based predictions do not consistently improve performance over direct classification, contrary to the expectations.

1 Introduction

Deception detection—the ability to identify intentionally misleading statements or behaviors—plays a critical role in safeguarding security, justice, and societal trust. Traditionally, its primary applications have been in criminalistics, particularly in interrogation and law enforcement settings such as suspect interrogations and security screenings. However, its relevance has expanded beyond these domains to border security (Sánchez-Monedero and Dencik, 2022), healthcare (Taylor et al., 2017), social media platforms (Qureshi et al., 2022), and consumer protection (Ott et al., 2011).

Despite its significance, deception detection remains inherently difficult, as human accuracy in detecting deception is only slightly above chance, $\sim 54\%$ (Charles F. Bond and DePaulo, 2006). **Cognitive Load Theory** (Vrij et al., 2008) suggests that lying demands greater mental effort, which can lead to detectable inconsistencies, but deceivers often mitigate this by rehearsing or simplifying their fabrications (Vrij et al., 2017). **Interpersonal Deception Theory** (Buller and Burgoon, 1996) highlights deception as an adaptive process, where deceivers adjust their behavior based on audience reactions, reducing the reliability of static detection methods. Levine (2014) further explains humans' bias toward assuming truthfulness, making them prone to overlooking deceptive cues. These challenges have driven the development of automated deception detection systems that systematically analyze linguistic, acoustic, and visual cues to improve reliability and scalability.

Researchers have increasingly explored automated approaches that combine advances in computer vision, natural language processing, and deep learning for deception detection. Early computational models in deception detection often relied on handcrafted features (Rill-Garcia et al., 2019; Zhang et al., 2020; Thannoon et al., 2018; Fan et al., 2015; Bai et al., 2019), drawing from facial micro-expressions, acoustic descriptors, and linguistic markers. With the emergence of deep learning, end-to-end architectures can directly learn deception-related patterns from raw multimodal data—text, audio, and video—leading to improved deception detection performance while reducing reliance on laborious feature engineering (Guo et al., 2023; Rani et al., 2023; Guo et al., 2024). Despite these advances, existing deception detection systems still face challenges related to generalization, as deception cues vary across individuals, cultures, and

*Work done outside of role at Amazon

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31013–31034
July 27 - August 1, 2025

contexts. Additionally, many deep learning models operate as black-box systems, making it unclear whether they genuinely capture deception-related patterns or rely on statistical shortcuts.

Recently, **Large Language Models (LLMs)** have demonstrated strong cognitive reasoning capabilities, excelling in tasks, such as emotion recognition (Cheng et al., 2024; Lei et al., 2024; Zhang et al., 2024a), sentiment analysis (Zhang et al., 2024b), and fact verification (Zhang and Gao, 2023). These models leverage large-scale pretraining and in-context learning to adapt to new tasks with minimal labeled data. LLMs' ability to identify subtle linguistic cues, integrate multimodal inputs (Chu et al., 2023; Liu et al., 2023; Zhang et al., 2023), and and generate step-by-step reasoning behind the judgment through chain-of-thought prompting (Wei et al., 2022) makes them promising candidates for automated deception detection. However, empirical evidence on LLM-driven deception detection, particularly in real-world multimodal settings, remains limited.

In this work, we take a comprehensive step toward filling this gap by challenging state-of-the-art LLMs with multiple deception detection tasks spanning three well-established datasets- **Real-life Trial Dataset, RLTD** (Pérez-Rosas et al., 2015), **Miami University Deception Detection Database, MU3D** (Lloyd et al., 2018), and **Opinion Spam Dataset, OpSpam** (Ott et al., 2011). These datasets cover deception across online, controlled, and real-world legal settings, collectively capturing diverse deception strategies and manifestations. The key contributions of this work are:

- We benchmark several state-of-the-art open-source and proprietary LLMs for deception detection on three datasets, providing a large-scale comparison of these models on diverse deception detection scenarios. Additionally, We assess the performance of open-source large multimodal models on the two multimodal (RLTD, MU3D) datasets, offering insights into how visual and acoustic cues can impact deception detection performance.

- We explore various fine-tuning and inference setups, including zero-shot prompting, random and similarity-based example selection for few-shot learning. We further investigate how different prompting strategies (direct label generation vs. post-hoc reasoning generation) affect deception detection results, shedding light on the best strategies for designing LLM-driven deception detection pipelines.

- We incorporate additional features, such as non-verbal gestures for RLTD and video summaries for RLTD and MU3D, to evaluate the influence of auxiliary features on model performance.

By presenting a thorough empirical study of LLM-based deception detection across multiple domains and modalities, we contribute a holistic perspective on the efficacy and limitations of these models.

2 Related Works

Early research on automated deception detection leveraged handcrafted linguistic, syntactic, and lexical features, including Linguistic Inquiry and Word Count (LIWC) indicators, part-of-speech distributions, and n-gram features, to capture linguistic, psychological and stylistic patterns indicative of deception. These features were utilized in statistical models such as logistic regression, decision trees, and support vector machines (SVM) to classify deceptive and truthful statements (Ott et al., 2011; Pérez-Rosas et al., 2015; Levitan et al., 2018; Rill-Garcia et al., 2019; Mathur and Matarić, 2020; Kamboj et al., 2021). Audio-based deception detection has relied on Mel-frequency cepstral coefficients (MFCCs) and prosodic cues, such as pitch and speaking rate, to distinguish deceptive from truthful speech (Hirschberg et al., 2005; Levitan et al., 2018; Bai et al., 2019; Gupta et al., 2019; Chebbi and Jebara, 2021). Additionally, research in nonverbal deception detection has focused on facial Action Units (AUs) extracted from video data, which capture microexpressions and facial muscle movements associated with deceptive behavior (Rill-Garcia et al., 2019; Belavadi et al., 2020; Bai et al., 2019; Mathur and Matarić, 2020; Mathur and Matarić, 2021). These approaches, though effective in constrained settings, often struggle with generalization across datasets and speaker variations, necessitating the exploration of more robust deep learning techniques.

Recent advances in deep learning have led to an increasing adoption of CNNs and LSTMs for deception detection tasks across both textual and multimodal domains (Karimi et al., 2018; Ding et al., 2019; Karnati et al., 2022; Sehrawat et al., 2023; Prome et al., 2024). Transformer based models and attention mechanisms have also been applied in recent deception detection research, leveraging contextual embeddings to capture subtle deception cues (Ilias et al., 2022; Hsiao and Sun, 2022; Gao et al., 2024). Guo et al. (2023) presents a novel method called Parameter-Efficient Crossmodal Learning (PECL) that uses a temporal adapter to capture temporal attention and a fusion module to merge audio and visual cues for audio-visual deception detection. Building on these developments, emerging research is now harnessing the capabilities of LLMs—whose success across diverse cognitive tasks underscores their potential—to capture intricate linguistic nuances and further enhance deception detection. In their study, Loconte et al. (2023) employ variants of the FLAN-T5 model (Chung et al., 2022) to detect deception across a range of textual contexts. Boumber et al. (2024) investigates the effectiveness of LLMs in deception detection using a Retrieval Augmented Generation (RAG) framework for few-shot learning in various textual domains. Our work advances this line of research by investigating the application of LLMs in real-world multimodal scenarios.

3 Background

3.1 Problem Definition

Deception detection is the task of identifying whether a statement or behavior is deliberately misleading. We define this task as a binary classification problem, where the goal is to predict $y \in \{\texttt{Truthful}, \texttt{Deceptive}\}$ given an input processed by a large language model. Formally, let p denote a task-specific prompt that instructs the model to process the input content and generate the classification label as either truthful or deceptive, and let t represent the textual content under analysis (for instance, a speech transcript or an online review). In the simplest setting, the input is $x = p \odot t$, where $\odot$ denotes concatenation, and the model generates the prediction via $y = f_\theta(x)$, where f_θ represents the LLM parameterized by θ.

Although textual cues can be highly informative for detection deception, additional cues may arise from non-verbal or multimodal sources. To account for such signals, we allow the input to be augmented by auxiliary features u, which could include descriptive text of facial expressions and body movements, or a textual summary of the observed video content or speech characteristics. In that case, the model processes $x = p \odot t \odot u$. Furthermore, when employing large multimodal models (LMMs) with the capacity of handling audio or video, the input can incorporate raw audio or video directly, denoted by a and v respectively, such that $x = p \odot t \odot [a, v]$

3.2 Datasets

Real-life Trial Dataset (RLTD) Pérez-Rosas et al. (2015) is constructed from publicly available courtroom trial recordings. Labels are assigned based on trial outcomes, with guilty verdicts indicating deception and non-guilty verdicts or exoneration indicating truthfulness. In some cases, the same individual contributes both deceptive and truthful statements, capturing within-subject deception variations. The dataset includes 121 video clips (60 truthful and 61 deceptive) with transcripts. The videos are also annotated for non-verbal features using the MUMIN multimodal coding scheme (Allwood et al., 2007), focusing on facial expressions, gaze, head, and hand movements.

Miami University Deception Detection Database (MU3D) Lloyd et al. (2018) is a controlled deception dataset capturing instructed deception in interpersonal scenarios. Participants were asked to describe individuals they liked or disliked while alternating between truthfulness and deception. The dataset comprises 320 (160 truthful and 160 deceptive) videos with metadata, including trustworthiness ratings, anxiety ratings, demographic details, and full speech transcriptions.

Opinion Spam Dataset (OpSpam) Ott et al. (2011) focuses on deception in online reviews and consists of 1600 reviews evenly split between truthful and deceptive opinions about hotels. Deceptive reviews were artificially generated by paid participants instructed to write persuasive but fabricated reviews, while truthful reviews were collected from genuine user feedback on platforms like TripAdvisor and Yelp. The dataset presents a linguistic deception challenge where fabricated narratives must be distinguished from authentic experiences.

Together, these datasets provide a rigorous benchmark for evaluating LLMs and LMMs in deception detection across legal, interpersonal, and online domains, ensuring a comprehensive assessment of their effectiveness.

3.3 Baselines

We evaluate the LLM based approaches against several deep-learning and transformer based baselines for text-only and multimodal deception detection. For the baselines, we extract modality-specific features using state-of-the-art pre-trained encoders. We obtain textual features from the final hidden states of the RoBERTa-base (Liu et al., 2019) model. For acoustic features, we use the final encoder hidden states of the Whisper-base (Radford et al., 2022) model, which has demonstrated robust performance in various audio tasks (Miah et al., 2023; Feng and Narayanan, 2023). For visual features, we sample the input video at 30 fps and encode each frame using CLIP (Radford et al., 2021).

We consider four baselines to compare against LLM-based approaches. First, we implement a text-only baseline by fine-tuning RoBERTa with a two-layer MLP classification head. Second, we follow Venkatesh et al. (2019) to employ a Bi-LSTM with attention network on the multimodal features described previously. We concatenate the resulting representations from each modality and use linear layers for multimodal classification. In unimodal scenarios, we simply predict the label from unimodal representations. Third, we follow Krishnamurthy et al. (2023); Karnati et al. (2022) to use CNN with global average pooling for feature encoding. Again, we concatenate the features across all modalities for multimodal deception detection. Finally, we replicate the Parameter-Efficient Cross-modal Learning (PECL) model proposed in Guo et al. (2023), which uses a 1D-convolution-based temporal adapter to learn modality-specific temporal attention alongside pre-trained Wav2Vec2 and ViT backbone models, supplemented by a Plug-in Audio-Visual Fusion (PAVF) module for cross-modal attention. This design enables PECL to achieve strong performance in the audio-visual setting. We conduct all experiments using stratified **4-fold cross-validation** across all three datasets.

4 Experimental Setup

We evaluate three Large Language Models (LLMs) for their deception detection capabilities: LLaMA3.1-8B (Grattafiori et al., 2024), Gemma2-9B (Team et al., 2024), and GPT-4o (OpenAI et al., 2024). Additionally, we assess the performance of various Large Multimodal Models (LMMs), categorized based on their modality specialization. For video-language models, we consider LLaVA-NEXT-Video (Zhang et al., 2024c) and Qwen2VL (Wang et al., 2024), while MERaLiON-AudioLLM (He et al., 2025) and Qwen2-Audio (Chu et al., 2024) serve as the audio-language models. These models represent state-of-the-art architectures in language and multimodal understanding, offering a diverse perspective on deception detection across textual, audio, and visual modalities.

4.1 Experimental Configurations

We evaluate both zero-shot and few-shot inference setups. In zero-shot evaluation, the model receives only a task description prompt and input data without labeled examples. In few-shot evaluation, the model is provided with a set of labeled examples for in-context learning. Specifically we have experimented with $n = \{2, 4, 6, 8, 10\}$, as number of in-context examples. Under the zero-shot and few-shot setups, we experiment with various strategies and configurations, outlined below.

4.1.1 Response Generation Strategies

To systematically assess deception detection performance, We investigate two different response generation strategies: **direct label prediction**, where the model directly generates the label for the input as either `Truthful` or `Deceptive` without additional reasoning, and **post-hoc reasoning generation**, where the model is prompted to first generate the classification label y and then provide a justification r, such that: $(y, r) = f_\theta(x)$, where x is the input and f_θ represents the model parameterized by θ. The generated reasoning r serves as a justification for the classification decision, allowing for better interpretability of deception detection outcomes. We also evaluate the chain-of-thought prompting for reasoning generation. However, post-hoc reasoning generation is eventually adapted for better performance and interpretability, with further analysis provided in Appendix H.

4.1.2 In-Context Example Selection Strategies

For the few-shot prompting setup, we explore different strategies for selecting in-context examples. Similar to the baselines, we employ a 4-fold split for in-context example selection. The random selection approach involves choosing an equal number of truthful and deceptive examples randomly from the other 3 splits. In contrast, the similarity-based selection methods employ sentence-transformers to encode the target input and dataset samples, allowing for similarity-based retrieval. Within this method, we examine

Model	Config	Modality	RLTD		MU3D		OpSpam	
			Acc	F1	Acc	F1	Acc	F1
Baselines								
RoBERTa-ft	-	t	76.31	76.22	**67.92**	**67.81**	88.10	88.09
BiLSTM+Attention	-	t	69.42	69.37	65.94	65.76	<u>90.45</u>	<u>90.45</u>
		a	68.04	68.02	62.19	61.82	-	-
		v	75.48	75.38	55.21	55.11	-	-
		t, a, v	77.14	77.04	62.29	62.19	-	-
CNN	-	t	64.46	64.39	64.06	63.91	86.37	86.36
		a	60.88	60.13	60.42	60.26	-	-
		v	<u>82.09</u>	<u>82.09</u>	54.06	53.95	-	-
		t, a, v	**83.47**	**83.44**	60.41	60.38	-	-
PECL	-	a, v	80.17	80.13	56.56	56.56	-	-
LLM Inference								
LLaMA 3.1	Few shot	t	71.69	77.11	57.03	56.84	62.93	62.47
Gemma 2	Few shot	t	71.69	71.38	55.08	53.75	64.81	64.51
GPT-4o	Few shot	t	79.55	79.49	55.70	53.87	74.50	74.00
LLaVA-NEXT-Video	Zero shot	v	52.06	43.55	50.00	33.33	-	-
		t, v	64.46	61.79	50.31	50.31	-	-
Qwen2VL	Zero shot	v	51.24	37.95	50.31	39.94	-	-
		t, v	63.64	60.32	52.50	52.35	-	-
MERaLiON-AudioLLM	Zero shot	a	66.94	66.11	49.06	33.97	-	-
		t, a	66.12	63.57	49.38	34.12	-	-
LLM Finetuning								
LLaMA 3.1	-	t	69.63	69.62	57.74	57.58	**92.25**	**92.24**
		t, u	72.72	72.46	-	-	-	-
Gemma 2	-	t	75.21	75.19	<u>66.56</u>	<u>66.55</u>	90.25	90.18
		t, u	75.21	75.17	-	-	-	-
Qwen2VL	-	v	57.85	57.10	52.20	51.43	-	-
		t, v	71.90	71.90	56.25	53.51	-	-

Table 1: Comparison of Baselines and LLM Results Across Modalities (t: text, a: audio, v: video, u: non-verbal features)

two variants: **sim-top**, which selects the most similar examples irrespective of their label, and **sim-pair**, which ensures a balanced selection of truthful and deceptive examples based on similarity ranking.

4.1.3 Auxiliary Features

We incorporate additional auxiliary features on top of the textual contents in the multimodal datasets, that provide valuable non-verbal and contextual information. As a first set of features for the RLTD dataset, we include a curated selection of 16 non-verbal features, capturing facial expressions and body movements indicative of deceptive behavior. The features names are listed in Appendix B. These features allow the model to leverage fine-grained behavioral cues that are often imperceptible in textual analysis alone. In addition to non-verbal gestures, we experiment with video and audio summaries as auxiliary inputs. A video-language model, LLaVA-NeXT-Video is employed to generate summaries of the visual content, extracting key information regarding speaker posture, facial expressions, and body movements indicative of stress or deception. Similarly, an audio-language model, Qwen2-Audio is used to summarize the tonal and acoustic features of the speech, identifying variations in pitch, intonation, and vocal stress patterns. These summaries provide a higher-level contextual representation of the non-verbal elements within the dataset, aiding in deception detection by supplying a multimodal understanding of deceptive cues for the RLTD and MU3D datasets.

4.1.4 Fine-Tuning

To further enhance model performance, we fine-tune open-source LLMs using the LLaMA-Factory (Zheng et al., 2024) framework. We specifically fine-tune LLaMA3.1-8B, Gemma2-9B, and Qwen2-VL-7B. This fine-tuning process allows the models to better adapt to the nuances of deception detection by learning from domain-specific patterns and optimizing their ability to process multimodal cues effectively.

5 Results & Analysis

In Table 1, we focus on the best configurations for LLM inference across RLTD, MU3D, and OpSpam, leaving a more detailed analysis to subsequent sections. While, the text-only LLMs, GPT-4o, LLaMA 3.1, and Gemma 2, manage to narrow some of the gap with the baselines on RLTD and MU3D datasets, their few-shot configurations do not consistently outperform the strongest baselines. GPT-4o reaches an F1 score of 79.49 on RLTD and 74.00 on OpSpam, signaling modest gains over other LLMs in the few-shot setup. On the contrary, zero-shot variants of LLaVA-NEXT-Video and Qwen2VL on RLTD and MU3D datasets remain less effective, especially when relying solely on video features, indicating a limited capacity to exploit visual cues without additional training. Even in the multimodal setup, they fail to surpass the CLIP-based video-only baselines. A similar pattern emerges for MERaLION-AudioLLM, which exhibits moderate zero-shot performance on RLTD using audio or multimodal inputs, yet still lags behind the Whisper-based audio-only baselines. These results suggest that LMMs fail to extract necessary cross-modal information for deception detection, unlike their multimodal baseline counterparts.

When fine-tuned, LLaMA 3.1 achieves state-of-the-art performance on the OpSpam dataset. Additionally, fine-tuning using non-verbal features boosts performance over just using the transcripts for RLTD. Gemma 2 raises its MU3D F1 score to 66.55 and achieves 90.18 F1 on OpSpam. Likewise, Qwen2VL experiences a performance boost on RLTD once text and video features are fine-tuned jointly. Nevertheless, even these tailored LLMs do not consistently match or surpass the strongest baselines for the multimodal datasets.

5.1 Comparison of CNN Baselines and vision LLMs

Experimental results demonstrate that the CNN baselines perform the best when video features are used alone or fused with text and audio, underscoring the importance of visual information on RLTD's unrehearsed deception, where deception-related micro-expressions and body movements are depicted in the video. By contrast, MU3D contains scripted deception, enabling actors to mask deception-related acoustic and visual cues while they are on record. As a result, fine-tuned RoBERTa and Gemma-2 outperform CNNs on this dataset. This observation also explains why text-only LLMs underperform compared to multimodal CNN baselines on RLTD, as they cannot utilize the nuanced visual and acoustic cues.

During training, CNNs learn to align and fuse temporal cues across modalities, allowing them to attend to deception-relevant patterns like micro-expressions and movement trajectories. By contrast, vision-language models like LLaVA-NEXT-Video and Qwen2VL rely on zero-shot pre-training, focused on captioning, object tracking, and OCR, and thus lack inherent deception-specific cognitive knowledge. Inspection of their generated video summaries further reveals why they miss critical deception cues. An example video summary from LLaVA-NEXT-Video using the prompt presented in Appendix C.3 - *In the video, a woman is seated at a table, wearing glasses and a red blouse, engaging in a conversation or an interview. Her facial expressions are calm and composed, with minimal micro-expressions, and her eye movements are steady, suggesting a controlled demeanor. Her body language is relaxed, with minimal hand gestures and head movements, indicating a composed and collected demeanor. There are no visible stress signs or fidgeting patterns, and her posture remains consistent throughout the video...*

It is evident from the generated summary that the vision–language models such as LLaVA-NEXT-Video describe the scenes and the objects well, yet they consistently miss the fine-grained behavioural cues annotated in the dataset, e.g. *raised eyebrows, gaze at interlocutor, downward lip movement, repeated nods, bilateral hand movement, complex hand trajectories* for this particular video. Consequently, these LMMs often report contradictory observations (e.g., 'minimal hand gestures') where, in fact complex hand movements are present in the video. They trail CNN baselines even after fine-tuning on transcripts and video. A key reason is their limited temporal resolution: Qwen2VL is pre-trained at **2 fps**, and LLaVA-NEXT recommends **16 frames per video**, whereas our CNN baselines operate on **10 fps** streams, capturing and tracking

LLM	Dataset	Cues			
		Details	**Vagueness**	**Filler Words**	**Justification**
LLaMA 3.1	RLTD	86.20% (29)	78.87% (14)	84.00% (25)	43.37% (23)
	MU3D	57.70% (52)	31.25% (16)	60.7% (28)	69.23% (13)
	OpSpam	65.26% (1091)	63.25% (117)	-	42.10% (27)
Gemma 2	RLTD	76% (25)	63.63% (22)	83.33% (12)	100% (7)
	MU3D	75.0% (8)	66.67% (9)	50% (4)	62.5% (8)
	OpSpam	68.88% (50)	70.83% (24)	-	-
GPT-4o	RLTD	73.07% (26)	80.0% (20)	66.67% (3)	85.71% (7)
	MU3D	61.90% (21)	100% (2)	100% (1)	100% (3)
	OpSpam	58.75% (80)	77.77% (9)	-	-

Table 2: Accuracy percentages for different models and cues. The number of total data points is in paranthesis.

subtler micro-expressions. Raising the frame rate for LMMs increases latency and GPU memory requirements, curbing scalability. Taken together, the performance gap of CNN baselines and large vision-language models reflects domain-specific temporal limitations, pre-training biases, and practical resource constraints of current LMMs.

5.2 Interpreting LLMs' Reasoning

Table 6 in Appendix D shows that direct label prediction and post-hoc reasoning generation often lead to similar performance. In RLTD, generating reasoning lowers performance across all models. However, for MU3D and OpSpam, we occasionally observe some improvements when reasoning is generated. Considering marginal and occasional gains from generating reasoning and associated additional costs, we adopt direct label prediction for further experiments. However, reasoning remains valuable for understanding the LLM's decision-making, helping to identify biases and patterns in deception detection. We analyze both correctly classified and misclassified instances, examining patterns based on linguistic cues to understand LLMs strengths and limitations.

Specificity and Detail. To quantify the use of *specificity and detail* as a cue for deception detection, we identified instances where the model explicitly referenced 'specific detail' in its reasoning and assessed accuracy based on correctly classified samples. As shown in Table 2, models consistently used this cue, with accuracy ranging from 57% to 86%. Notably, for the RLTD dataset, which consists of courtroom trials, accuracy was higher across all three models. This suggests that specific details are more informative in legal contexts, where testimonies often contain detailed accounts of events, locations, and actions, making specificity a stronger indicator of truthfulness. Emotional deception, as in MU3D, may not always involve factual inconsistencies, making reliance on details less effective. Similarly, in the case of online reviews, deceptive reviewers can fabricate highly detailed experiences, while genuine reviewers may provide concise feedback without elaborate narratives. To further investigate this behavior, we analyzed 86 randomly selected OpSpam samples where the LLaMA model referenced specificity in its reasoning. Of these, 67 lacked detail and were all classified as deceptive, misclassifying 13 truthful reviews. In contrast, 19 were classified as truthful due to specific details, yet 7 were actually deceptive (Figure 5 *Example 12*). This bias toward treating specificity as a truth cue aligns with Reality Monitoring Theory (RMT), which links truthfulness to sensory-rich statements (Vrij, 2008). However, in online reviews, deceptive writers may create vivid narratives, while truthful reviewers might be concise. This over-reliance on specificity exposes a key limitation of LLM's reasoning process.

Vagueness. We examine the models' reliance on *vagueness* as a deception cue. Table 2 shows LLMs consistently use this cue, with GPT-4o demonstrating the highest accuracy. Analyzing LLaMA's behavior, we found that in MU3D, all 16 vagueness-based classifications were deceptive, misclassifying 11 truthful cases, suggesting that the model struggles to distinguish between genuine uncertainty and deceptive ambiguity in interpersonal communication. In RLTD, 14 instances were flagged as deceptive, with 11 correctly classified, indicating a slightly better alignment with deception patterns in courtroom testimonies. In OpSpam, 92 of 117 flagged cases were classified as deceptive (63.25% accuracy). This bias toward associating vagueness with deception often leads to overgeneralization and misclassification as vagueness can naturally occur in truthful statements due to memory recall limitations or subjective expression. For instance, in MU3D example (Appendix, Figure 5 *Example 7*), the speaker expresses strong negative emotions about a peer, saying, "He's gotten my friends in trouble," and "we stopped hanging out with him just because the, that whole reason," without clearly specifying what "that whole reason" entails. This conversational vagueness i.e. the use of non-specific phrases led the model to classify the statement as deceptive. This misclassification highlights how the model may over-rely on surface-level ambiguity as a deception signal, failing to account for the emotional and informal nature of interpersonal speech. In emotionally charged dialogue, vague references can reflect genuine uncertainty or conversational style rather than intent to deceive.

Hesitation and Filler Words. We investigate LLM's reliance on *hesitation and filler words* as deception cues. LLMs frequently associates verbal disfluencies (e.g., 'uh,' 'um') with deception, aligning with Cognitive Load Theory (Vrij, 2008), which suggests that lying requires greater mental effort, leading to pauses and hesitations. As shown in Table 2, GPT-4o relies on filler words less compared to LLaMa 3.1 and Gemma 2. Note, this cue was not used in OpSpam, as it comprises written reviews. We find that reliance on this cue sometimes leads to correct classifications—such as in Figure 4 *Example 1, 2*, where hesitation appeared alongside vagueness or contradictions. However, misinterpretations also occur, as seen in Figure 4 *Example 3*, where hesitation in a truthful statement resulted in a false deception label. Hesitation paired with detailed responses is often assumed to indicate truthfulness, correctly classified in Figure 4 *Example 4* but misapplied in *Example 5*.

Justification. To assess the LLM's use of *justification* as a cue, we identified instances where 'justify,' 'justifies,' or 'justification' appeared in its reasoning and reported accuracy in Table 2. The LLM often links justifications and indirect answers to deception, aligning with Criteria-Based Content Analysis (CBCA) (Vrij, 2008), which associates evasiveness with deception. Gemma applied this cue effectively in RLTD, correctly classifying 6 out of 7 cases. However, in MU3D, it consistently associated justification with deception, predicting all 8 instances as deceptive with 62.5% accuracy.

LLM	Example selection	RLTD Acc	RLTD F1	MU3D Acc	MU3D F1	OpSpam Acc	OpSpam F1
LLaMA 3.1	*random*	68.87	68.14	51.72	51.18	59.62	59.19
	sim-pair	<u>71.69</u>	<u>71.11</u>	54.76	54.14	58.04	57.78
	sim-top	71.28	70.25	**57.03**	**56.84**	<u>62.93</u>	<u>62.47</u>
Gemma 2	*random*	69.63	69.52	54.22	52.34	57.70	57.59
	sim-pair	<u>71.69</u>	<u>71.38</u>	54.92	53.69	60.14	59.98
	sim-top	71.07	70.68	<u>55.08</u>	<u>53.75</u>	<u>64.81</u>	<u>64.51</u>
GPT-4o	*random*	71.69	71.39	53.20	46.86	68.40	67.58
	sim-pair	**79.55**	**79.49**	55.08	50.35	71.87	71.23
	sim-top	77.69	77.69	<u>55.70</u>	<u>53.87</u>	**74.50**	**74.00**

Table 3: Performance comparison of example-selection strategies across RLTD, MU3D, and OpSpam. The best overall results are in **bold**, while model-specific best performances are <u>underlined</u>.

This suggests the model struggles to differentiate between genuine explanations and intentional deflection.

Emotions. We analyzed how LLMs interpret *emotions* in deception detection, finding that they often associate strong emotional reactions with truthfulness. This aligns with Statement Validity Analysis (SVA) (Vrij, 2008), which considers spontaneous emotions a sign of genuine experiences. While this assumption sometimes led to correct classifications (Figure 5 *Example 10*), it also resulted in misclassifications. For instance, the model mistakenly labeled a deceptive statement as truthful when exaggerated emotions were used to appear credible (Figure 4 *Example 6*) and failed to recognize playful language, misinterpreting emotional expression (Figure 5 *Example 11*). In MU3D, genuine expressions of admiration and affection were frequently misclassified as deception (Figure 5 *Example 8, 9*). This indicate that LLM often lacks the ability to accurately interpret emotions.

5.3 Impacts of In-context Examples

Table 3 presents a comparison of three in-context example selection strategies—random, sim-top, and sim-pair under a few-shot prompting setup. In general, both similarity-based methods (sim-pair and sim-top) surpass random selection, demonstrating the importance of carefully curating in-context examples. The principal distinction between sim-top and sim-pair lies in label balancing: sim-top selects the most similar examples regardless of their labels, whereas sim-pair enforces a balanced set of truthful and deceptive instances among those most similar. In terms of the LLMs, GPT-4o exhibits the highest average improvement (7.18% F1 score) when similarity-based few-shot examples are introduced, demonstrating more robust in-context learning capabilities relative to LLaMA3.1-8B and Gemma2-9B.

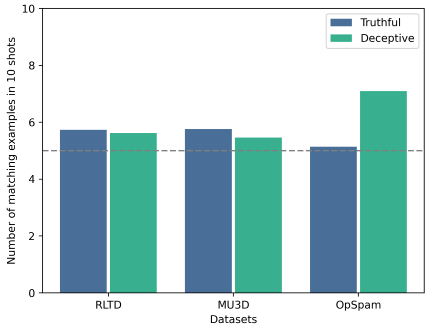

Figure 1: Average number of matching examples in 10-shot for sim-top strategy.

Looking closely at RLTD, the sim-pair approach slightly outperforms sim-top. In contrast, on MU3D and OpSpam, sim-top provides superior results, particularly on OpSpam, where the average F1 score improvement, $\sim 4\%$, is notably higher than that observed on MU3D $\sim 2\%$. Figure 1 further illuminates these findings by illustrating the average number of 'matching' examples (i.e., truthful examples for a truthful query and deceptive examples for a deceptive query) retrieved in a 10-shot setup using sim-top. For RLTD and MU3D, this number hovers around five, effectively mirroring the label balance of sim-pair. However, in OpSpam, especially for the deceptive queries, the average number of retrieved deceptive examples rises to about 7.1, enabling a substantial boost in performance. Concretely, this increase in label-specific examples elevates the deceptive-class F1 score from approximately 67% under sim-pair to 73% under sim-top. This finding also suggests that deceptive reviews in the OpSpam dataset exhibit a higher degree of semantic similarity compared to the other datasets, hence easily identifiable by the retriever. This OpSpam dataset specific bias and the implications are further discussed in Appendix I.

5.4 Impacts of Auxiliary Features

In Table 4, we compare three types of auxiliary features: non-verbal gestures, LLM-generated audio and video summaries under both zero-shot and few-shot settings. Each model uses randomly selected in-context examples when operating in the few-shot configuration. From these results, we observe that including non-verbal gestures alongside the transcript yields a modest improvement for GPT-4o ($\sim$ 1.4 points in F1 score on the RLTD dataset). This gain is consistent with GPT-4o's demonstrated strengths in in-context learning. In contrast, the inclusion of non-verbal features negatively impacts LLaMA 3.1 and Gemma 2: their tendency to over-

LLM	Feats	Config	RLTD		MU3D	
			Acc	F1	Acc	F1
LLaMA 3.1	*nv*	*zs*	51.24	35.34	-	-
		fs	63.02	62.79	-	-
	vs	*zs*	52.07	38.40	50.63	49.67
		fs	62.60	61.72	51.88	49.44
	as	*zs*	57.85	57.56	49.06	49.00
		fs	64.74	63.12	51.57	51.25
Gemma 2	*nv*	*zs*	52.07	38.40	-	-
		fs	66.67	64.85	-	-
	vs	*zs*	52.07	38.40	51.25	44.16
		fs	66.94	66.92	**53.44**	**51.56**
	as	*zs*	66.94	66.28	49.38	45.82
		fs	68.60	68.59	51.56	50.98
GPT-4o	*nv*	*zs*	65.29	61.71	-	-
		fs	**72.93**	**72.84**	-	-
	vs	*zs*	65.29	65.00	52.19	47.98
		fs	69.42	69.00	54.69	52.48
	as	*zs*	66.12	64.73	49.38	42.31
		fs	67.69	66.85	51.63	45.21

Table 4: Comparison of auxiliary features for the multimodal datasets. *zs*: zero shot; *fs*: few shot; *nv*: non-verbal; *vs*: video summaries, *as*: audio summaries.

predict the `Deceptive` label suggests that limited in-context examples are insufficient for these models to learn patterns of non-verbal gestures across truthful and deceptive scenarios. Turning to video summaries, GPT-4o again exhibits relative gains on MU3D, although the improvements for other models and datasets remain negligible or even degrade performance. A similar pattern holds for audio summaries: while certain configurations see a slight boost, many are on par with or slightly below the corresponding transcript-only results. Overall, additional features do not universally enhance predictive accuracy without fine-tuning.

6 Conclusion

Our comprehensive evaluation reveals that LLMs and LMMs exhibit promising capabilities for deception detection across diverse contexts. While fine-tuning significantly enhanced performance, improvements on multimodal datasets are still lagging, highlighting persistent challenges in capturing nuanced cross-modal deception cues in LMMs. Moreover, incorporating reasoning generation to explain predictions did not consistently improve overall accuracy over straightforward label prediction, emphasizing that the inherently ambiguous nature of deception cues makes it harder for the models to reason successfully. These findings underscore the importance of careful prompt design and in-context example selection while pointing to the need for further methodological refinements in practical deception detection applications.

Limitations

While our study shows promising results, it has several limitations that pave the way for future research. First, our experiments are limited to English-language datasets, which may not fully capture the linguistic diversity or cultural nuances necessary for broader applicability. Second, we focus exclusively on human deception, leaving the detection of AI-generated deceptive behaviors as an area for further exploration. Third, the reliance on a limited range of publicly available datasets and controlled scenarios may affect the generalizability of our findings to more varied, real-world contexts. Additionally, the deployment of deception detection systems involves ethical, privacy, and interpretability challenges that must be carefully managed, especially in legal or interpersonal settings. Finally, the computational cost—approximately 300 USD for experiments with GPT-based models—and the significant GPU resources required for open-source models highlight practical considerations for real-world implementation.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation via the award IIS-1942918 as well as support from the Texas A&M Institute of Data Science via an internal grant. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3/4):273–287.

Chongyang Bai, Maksim Bolonkin, Judee K. Burgoon, Chao Chen, Norah E. Dunbar, Bharat Singh, V. S. Subrahmanian, and Zhe Wu. 2019. Automatic long-term deception detection in group interaction videos. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1600–1605.

Vibha Belavadi, Yan Zhou, Jonathan Z. Bakdash, Murat Kantarcioglu, Daniel C. Krawczyk, Linda Nguyen, Jelena Rakic, and Bhavani Thuriasingham. 2020. Multimodal deception detection: Accuracy, applicability and generalizability. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 99–106.

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*, IWSPA '24, page 37–47, New York, NY, USA. Association for Computing Machinery.

David B. Buller and Judee K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory*, 6(3):203–242.

Jr. Charles F. Bond and Bella M. DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234. PMID: 16859438.

Safa Chebbi and Sofia Ben Jebara. 2021. Deception detection using multimodal fusion approaches. *Multimedia Tools and Applications*, 82:13073–13102.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 37, pages 110805–110853. Curran Associates, Inc.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *ArXiv*, abs/2311.07919.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Mingyu Ding, An Zhao, Zhiwu Lu, Tao Xiang, and Ji-Rong Wen. 2019. Face-focused cross-stream network for deception detection in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803.

Cheng Fan, Heming Zhao, Xueqin Chen, Xiaohe Fan, and Shuxi Chen. 2015. Distinguishing deception from non-deception in chinese speech.

Tiantian Feng and Shrikanth Narayanan. 2023. PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches For Speech Emotion Recognition Using Pre-trained Speech Models . In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Los Alamitos, CA, USA. IEEE Computer Society.

Tommaso Fornaciari, Leticia Cagnina, Paolo Rosso, and Massimo Poesio. 2020. Fake opinion detection: how similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, 54.

Shuai Gao, Lin Chen, Yuancheng Fang, Shengbing Xiao, Hui Li, Xuezhi Yang, and Rencheng Song. 2024. Video-based deception detection via capsule network with channel-wise attention and supervised contrastive learning. *IEEE Open Journal of the Computer Society*, 5:660–670.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Chichung Kot. 2023. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22078–22088.

Xiaobao Guo, Zitong Yu, Nithish Muthuchamy Selvaraj, Bingquan Shen, Adams Wai-Kin Kong, and Alex C. Kot. 2024. Benchmarking cross-domain audio-visual deception detection. *Preprint*, arXiv:2405.06995.

Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. 2019. Bag-of-lies: A multimodal dataset for deception detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 83–90.

Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, and Ai Ti Aw. 2025. Meralion-audiollm: Bridging audio and language with large language models. *Preprint*, arXiv:2412.09818.

Julia Hirschberg, Stefan Benus, Jason M. Brenier, Fiona Enos, Sarah Friedman, Susan Gilman, Cynthia Girand, Marc Graciarena, Andreas Kathol, Laura Michaelis, Blanton L. Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing deceptive from non-deceptive speech. In *Proceedings of Interspeech 2005*, pages 1833–1836.

Shun-Wen Hsiao and Cheng-Yuan Sun. 2022. Attention-aware multi-modal rnn for deception detection. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3593–3596.

Loukas Ilias, Felix Soldner, and Bennett Kleinberg. 2022. Explainable verbal deception detection using transformers. *Preprint*, arXiv:2210.03080.

Manvi Kamboj, Christian Hessler, Priyanka Asnani, Kais Riani, and Mohamed Abouelenien. 2021. Multimodal political deception detection. *IEEE MultiMedia*, 28(1):94–102.

Hamid Karimi, Jiliang Tang, and Yanen Li. 2018. Toward end-to-end deception detection in videos. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1278–1283.

Mohan Karnati, Ayan Seal, Anis Yazidi, and Ondrej Krejcar. 2022. Lienet: A deep convolution neural network framework for detecting deception. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3):971–984.

Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. 2023. A deep learning approach for multimodal deception detection. In *Computational Linguistics and Intelligent Text Processing*, pages 87–96, Cham. Springer Nature Switzerland.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *Preprint*, arXiv:2309.11911.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

E. Paige Lloyd, Jason Deska, Kurt Hugenberg, Allen McConnell, Brandon Humphrey, and Jonathan Kunstman. 2018. Miami university deception detection database. *Behavior Research Methods*, 51.

Riccardo Loconte, Roberto Russo, Pasquale Capuozzo, Pietro Pietrini, and Giuseppe Sartori. 2023. Verbal lie detection using large language models. *Scientific Reports*, 13(1):22849. Published on 2023/12/21.

Leena Mathur and Maja J. Matarić. 2020. Introducing representations of facial affect in automated multimodal deception detection. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 305–314, New York, NY, USA. Association for Computing Machinery.

Leena Mathur and Maja J Matarić. 2021. Affect-aware deep belief network representations for multimodal unsupervised deception detection. *Preprint*, arXiv:2108.07897.

Md Messal Monem Miah, Adarsh Pyarelal, and Ruihong Huang. 2023. Hierarchical fusion for online multimodal dialog act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7532–7545, Singapore. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long

Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 59–66, New York, NY, USA. Association for Computing Machinery.

Shanjita Akter Prome, Md Rafiqul Islam, Md. Kowsar Hossain Sakib, David Asirvatham, Neethiahnanthan Ari Ragavan, Cesar Sanin, and Edward Szczerbicki. 2024. Ldnet: A robust hybrid approach for lie detection using deep learning techniques. *Computers, Materials and Continua*, 81(2):2845–2871.

Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. 2022. Deception detection on social media: A source-based perspective. *Knowledge-Based Systems*, 256:109649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Anku Rani, Dwip Dalal, Shreya Gautam, Pankaj Gupta, Vinija Jain, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023. Sepsis: I can catch your lies - a new paradigm for deception detection. *ArXiv*, abs/2312.00292.

Rodrigo Rill-Garcia, Hugo Jair Escalante, Luis Villasenor-Pineda, and Veronica Reyes-Meza. 2019. High-level features for multimodal deception detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Puneet Kumar Sehrawat, Rajat Kumar, Nitish Kumar, and Dinesh Kumar Vishwakarma. 2023. Deception detection using a multimodal stacked bi-lstm model. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 318–326.

Felix Soldner, Bennett Kleinberg, and Shane D. Johnson. 2022. Confounds and overestimations in fake review detection: Experimentally controlling for product-ownership and data-origin. *PLOS ONE*, 17(12).

Javier Sánchez-Monedero and Lina Dencik. 2022. The politics of deceptive borders: 'biomarkers of deceit' and the case of iborderctrl. *Information, Communication & Society*, 25(3):413–430.

John B. Taylor, Scott R. Beach, and Nicholas Kontos. 2017. The therapeutic discharge: An approach to dealing with deceptive patients. *General Hospital Psychiatry*, 46:74–78.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Harith H. Thannoon, Wissam H. Ali, and Ivan A. Hashim. 2018. Detection of deception using facial expressions based on different classification algorithms. In *2018 Third Scientific Conference of Electrical Engineering (SCEE)*, pages 51–56.

Sushma Krupa Venkatesh, Raghavendra Ramachandra, and Patrick A. H. Bours. 2019. Video based deception detection using deep recurrent convolutional neural network. In *International Conference on Computer Vision and Image Processing*.

Aldert Vrij. 2008. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.

Aldert Vrij, R. Fisher, Samantha Mann, and Sharon Leal. 2008. A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, 5(1-2):39–43.

Aldert Vrij, Ronald P. Fisher, and Hartmut Blank. 2017. A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Jiaxuan Zhang, Sarah Ita Levitan, and Julia Hirschberg. 2020. Multimodal deception detection using automatically extracted acoustic, visual, and lexical features. In *Interspeech*.

Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. 2024a. Visual prompting in LLMs for enhancing emotion recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4484–4499, Miami, Florida, USA. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. Llava-next: A strong zero-shot video understanding model.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Models

All the models except GPT-4o are hosted on huggingface. GPT-4o model is used via OpenAI API.

B List of Non-verbal Features

RLTD dataset comes with a set of 40 manually annotated non-verbal features. These features are broadly categorized into *facial displays* and *hand movements*. The original annotation provides a binary value for each of these features with respect to whether these attributes were demonstrated by the primary speaker in the video. We filter the most relevant 16 features and use the feature names directly for generating LLM predictions. A list of these features is - `Both Hands Movement, Complex Hands Trajactory, Downwards Lip Movement, Eyes Closing Repeatedly, Frown, Gaze Down, Gaze Side, Gaze at Interlocutor, Head Down, Mouth Closed, Mouth Opened, Raise Eyebrows, Repeated Nods, Scowl, Single Hand Movement, Upwards Lip Movement.`

C Prompts

C.1 Prompt for Direct Label Prediction

`You are an expert judge, specializing in deception detection. Your task is to evaluate transcripts from courtroom trials or interviews with exceptional accuracy. Based on the provided transcript, assess whether the speaker's statements are 'Truthful' or 'Deceptive'. Utilize linguistic cues, contextual clues, and logical consistency in the transcript to make your judgment. Respond with only one word: either 'Truthful' or 'Deceptive'. Do not provide additional explanation, reasoning, or commentary in your response.`

`Below are some reference examples:` **`<few shot examples>`**

`Carefully analyze the transcript below, paying close attention to the content and context, and provide your judgment.`

C.2 Prompt for Reasoning Generation

`You are an expert judge, specializing in deception detection. Your task is to evaluate transcripts from courtroom trials or interviews with exceptional accuracy. Based on the provided transcript, assess whether the speaker's statements are 'Truthful' or 'Deceptive'. Utilize linguistic cues, contextual clues, and logical consistency in the transcript to make your judgment. Respond with both your judgment and the reasoning behind your judgment. Provide your response in the format- Judgment: <your judgment in one word 'Truthful' or 'Deceptive'>, Reasoning: <your reasoning for the judgment based on the transcript and your analysis>.`

`Below are some reference examples:` **`<few shot examples>`**

`Carefully analyze the transcript below, paying close attention to the content and context, and provide your answer in the desired format.`

C.3 Prompt for Video Summary Generation

`Analyze this video and provide a comprehensive and detailed summary based on: - Speaker details (appearance, positioning); if there are more than one person, focus only on the person being interviewed or who is answering questions. - Facial expressions (micro-expressions, eye movements, eye-brow movements, lip`

Model Name	Model ID	License
LLaMA 3.1	`meta-llama/Llama-3.1-8B-Instruct`	llama 3.1
Gemma 2	`google/gemma-2-9b-it`	gemma
GPT-4o	`gpt-4o-2024-08-06`	proprietary
LLaVA-NEXT-Video	`llava-hf/LLaVA-NeXT-Video-7B-hf`	llama2
Qwen2VL	`Qwen/Qwen2-VL-7B-Instruct`	apache-2.0
MERaLiON-AudioLLM	`MERaLiON/MERaLiON-AudioLLM-Whisper-SEA-LION`	meralion-public-license
Qwen2-Audio	`Qwen/Qwen2-Audio-7B-Instruct`	apache-2.0

Table 5: Model Information

```
and mouth movements etc.) - Body language (hand gestures, head movements, posture changes etc.) - Physical indicators (visible stress signs, fidgeting patterns etc.) - Key scene descriptions (describing the most crucial moments from the video) Describe any notable behavioral patterns or changes over time. Focus on any observable visual cues. The final summary should be a paragraph containing all the important information extracted from the input video according to the instructions provided.
```

C.4 Prompt for Audio Summary Generation

```
Analyze the input audio and provide a summary of the pitch and tone of the speaker in the audio recording. Describe any notable acoustic patterns briefly.
```

D Response Generation Strategies

The results in Table 6 offer a comparative analysis between direct label prediction and post-hoc reasoning generation across the three datasets. We systematically evaluate whether generating reasoning after the label contributes positively to the model's predictive performance, under both zero-shot and few-shot prompting settings. Across most of the settings, particularly on RLTD dataset, direct label prediction tends to yield higher accuracy and F1 scores. For example, GPT-4o achieves an F1 score of 71.39 on RLTD with few-shot direct label prediction, outperforming its label+reasoning counterpart (69.63 F1 score). However, this trend does not hold universally. In the MU3D dataset, which involves scripted deception, post-hoc reasoning occasionally matches or slightly improves performance. LLaMA 3.1, for instance, reaches its best F1 score of 56.15 on MU3D using few-shot post-hoc reasoning. In the OpSpam dataset, dominated by textual content, the advantage again leans toward direct label prediction. GPT-4o in particular shows a noticeable drop in F1 score from 67.58 (few-shot label) to 61.04 (few-shot label + reasoning), suggesting that the inclusion of generated explanations may introduce noise or ambiguity, especially when no visual or behavioral cues are available to ground the reasoning. While post-hoc reasoning generation provides interpretability of model predictions, it does not consistently improve classification performance, and in many cases, leads to modest degradation.

E Performance Trends in Few-Shot Learning

In our study, we examined how large language models (LLMs) perform across various datasets using few-shot prompting. We calculated the F1 score by averaging results across all seeds to ensure consistent measurement as illustrated in Figure 3. Our findings reveal that models like GPT-4o initially improve with more few-shot examples, demonstrating their ability to use additional data effectively. However, this improvement subsequently declined when too many examples were provided, likely due to the increased complexity of the prompts complicating the model's reasoning ability. LLaMA 3.1 consistently showed significant gains with an increased number of examples for OpSpam and MU3D, indicating strong adaptability to more extensive data inputs. Gemma 2's performance improved on the OpSpam dataset with more examples but declined on the MU3D and RLTD datasets after a certain point. This pattern suggests a possible optimization ceiling, where additional examples no longer contribute to performance enhancements and may instead hinder the model's effectiveness due to prompt saturation.

F Evaluating the Efficacy of Beam Search in Reasoning

We performed beam search to evaluate its potential to enhance performance in label generation with reasoning for the LLaMA and Gemma models

LLM	Config	Response Generation	RLTD Acc	RLTD F1	MU3D Acc	MU3D F1	OpSpam Acc	OpSpam F1
LLaMA 3.1	zero shot	label	54.67	51.00	49.61	48.51	52.35	52.33
		label + reasoning	52.07	50.27	48.44	47.95	51.18	51.17
	few shot	label	68.87	68.14	51.72	51.18	59.62	59.19
		label + reasoning	65.71	65.61	**56.49**	**56.15**	61.43	60.83
Gemma 2	zero shot	label	67.77	66.67	52.35	48.20	49.28	48.45
		label + reasoning	66.12	64.20	55.63	54.42	50.28	47.00
	few shot	label	69.63	69.52	54.22	52.34	57.70	57.59
		label + reasoning	68.18	68.00	53.91	50.73	59.68	57.75
GPT-4o	zero shot	label	67.63	67.62	52.42	43.98	58.53	53.72
		label + reasoning	64.46	64.31	51.41	41.89	59.04	53.63
	few shot	label	**71.69**	**71.39**	53.20	46.86	**68.40**	**67.58**
		label + reasoning	69.63	69.07	52.17	42.67	64.20	61.04

Table 6: Comparison of different response generation strategies (direct label prediction vs. post-hoc reasoning generation) under zero-shot and few-shot settings. The few-shot examples are randomly selected. The best and second best results are indicated by **bold** and underline respectively.

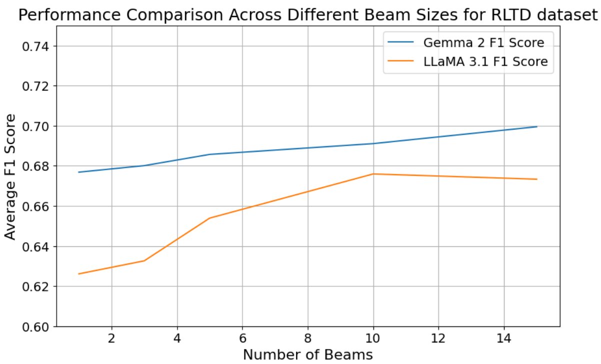

Figure 2: F1 score across different beam sizes on RLTD dataset

LLM	Order	RLTD Acc	RLTD F1	MU3D Acc	MU3D F1	OpSpam Acc	OpSpam F1
LLaMA 3.1	$l \to r$	65.71	65.61	56.49	56.15	61.43	60.83
	$r \to l$	58.13	55.31	52.08	49.50	58.60	57.62
Gemma 2	$l \to r$	68.18	68.00	53.91	50.73	59.68	57.75
	$r \to l$	58.68	54.39	50.63	50.39	61.62	59.36

Table 7: Comparison of LLM performances under different reasoning generation orderings across datasets. $l \to r$: label $\to$ reasoning; $r \to l$: reasoning $\to$ label.

on the RLTD dataset. We calculated the average F1 scores for 10-shot experiment across all seeds for varying beam sizes, specifically [1, 3, 5, 10, 15], with the results detailed in Figure 2. Our analysis reveals that as the number of beams increases, the performance generally increases. However, even the best outcomes obtained through the beam search failed to surpass the performance levels achieved via direct label generation for the RLTD dataset.

G LLM Reasoning Analysis Examples

We have illustrated several examples on the basis of deception cues across three datasets in Figure 4, Figure 5.

H Post-hoc Reasoning Generation vs Chain-of-Thought Reasoning

While generating additional reasoning for predicted labels, we adopt a **post-hoc reasoning generation** strategy, where the model first outputs the classification label, followed by a justification. An alternative is **chain-of-thought reasoning**, where the model first reasons over the input before predicting the final label. We chose the post-hoc approach based on empirical evidence. Specifically, we conducted a controlled comparison of the two prompting strategies — (a) label $\to$ reasoning (post-hoc) and (b) reasoning $\to$ label (chain-of-thought)—under a few-shot setting across three datasets. As shown in Table 7, the post-hoc reasoning generation strategy generally outperforms chain-of-thought.

In addition to the performance improvements, the post-hoc reasoning reasoning strategy offers better interpretability for analyzing LLM's decision making process. When reasoning is generated prior to the label, LLMs often produce explanations that weigh cues supporting both possible classes. This makes it difficult to use the generated reasoning to interpret the model's decision, as it lacks a clear alignment with the predicted label. An example of LLM prediction through chain-of-thought - `Reasoning: 1. The speaker starts with "Ah, I don't remember," which is a common phrase used when someone is unsure`

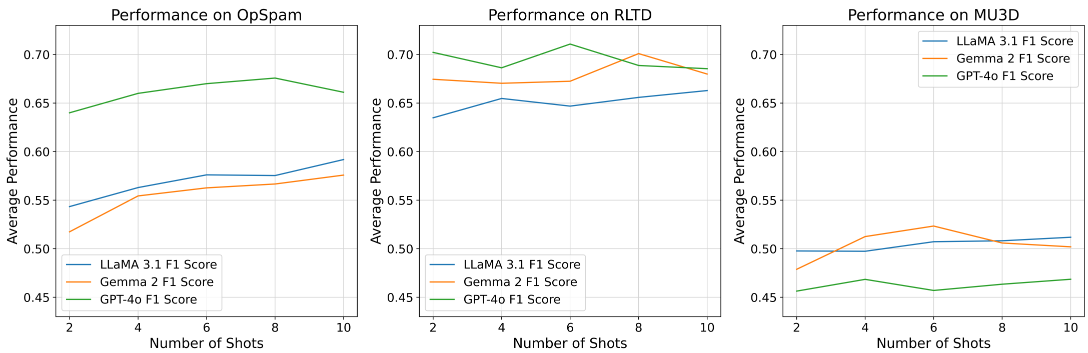

Figure 3: F1 score across n-shots in few shot learning

Example	Transcript	LLM's reasoning	Label & Prediction	Comments
Example 1: RLTD trial_lie_011	Uh, I had just been on a cruise with my husband. We hadn't been on a vacation for years, alone, without the children. (pause, clears throat) I had a project I needed to complete, and Hemy insisted it had to be completed that ... day, and so the only way in which to -- and it required his input, and the only way in which to finish it was to work on it that night.	The speaker **pauses and clears their throat** before continuing, which can be a sign of nervousness or hesitation. Additionally, the phrase "I had a project I needed to complete" is vague and doesn't provide specific details about the project, which raises suspicions.	Label: Deceptive Prediction: Deceptive	Hesitation and vagueness
Example 2: RLTD trial_lie_059	...um person in the congregation was involved in activities like that. The elders would sit down with the individual and try to determine what's going on. The objective is as Jesus said, \if you have a hundred sheep, and one's lost, leave the ninety-nine and go find the one\". So the objective is to help that individual. Whether it's to er, er, worshiping er, someone or anything like that or celebrate Christmas or fornication or whatever it happens to be. The elder would mate with the individual based on person's um attitude at that time will have direct relevance to what they do. "	the speaker's use of the phrase "mate with the individual based on person's um attitude at that time will have direct relevance to what they do" is **vague and open to interpretation**. …. The speaker's **hesitation** and awkward phrasing ("er, er, worshiping er, someone") also suggest that they are trying to carefully choose their words to avoid revealing the true nature of the church's activities.	Label: Deceptive Prediction: Deceptive	Hesitation and vagueness
Example 3: RLTD trial_truth_015	just talked to everybody, you know, you know he was flirtatious, I mean he, ah, you know, um, I don't know ... I don't know who told me that. I don't know enough about it. I just knew he was a flirt and that ... You know, he was ... I remember Clancy telling me, you know, at times, you know, he was just him. You know, he would just flirt with everybody, just goofing around. Nobody really took it seriously.	The use of filler words ("um," "ah," "you know") and **hesitation phrases ("I don't know," "I don't know who told me that")** indicate that the speaker is struggling to provide a straightforward answer.	Label: Truthful Prediction: Deceptive	Hesitation
Example 4: RLTD trial_truth_028	I ... It was while back but I believe, I did. I ask hm ... I asked about the battery, it will take time for battery to charge, You won't be able to use it out of the box. So, by that time that's when I asked she would like one that is corded.	There is **no indication of evasiveness, hesitation**, or inconsistency in the speaker's narrative, further supporting the conclusion that the statement is truthful.	Label: Truthful Prediction: Truthful	Hesitation
Example 5: RLTD trial_lie_040	Yes, he said he did. And he told me to take care of my wounds as necessary. No he said there was several wounds. Um he advised me to take Tylenol when I get the gel and um asked him for something to put on, put on my face.	There is **no indication of hesitation**, evasiveness, or inconsistency in the speaker's narrative, which further supports the conclusion that the speaker is telling the truth.	Label: Deceptive Prediction: Truthful	Hesitation
Example 6: RLTD trial_lie_036	... brother, sister and his children...I know that he was a fun loving Uncle and I'm ... can't continue to experience things in his life ... for a kid ... there is certainly no way to explain ... make sense of the situation and I'm truly sorry for that.	The speaker's tone and language suggest a genuine expression of sorrow and regret. The use of phrases such as "I'm truly sorry" and "can't continue to experience things in his life" **convey a sense of empathy and emotional pain.**	Label: Deceptive Prediction: Truthful	Strong emotion

Figure 4: Examples of LLM Reasoning on RLTD Dataset

Example	Transcript	LLM's reasoning	Label & Prediction	Comments
Example 7: MU3D (WM005_3NT)	He's a horrible guy. Ever since I met him he's been screwing up. He's gotten my friends in trouble. He's gotten me in trouble. Um. You know, we stopped hanging out with him just because the, that whole reason. Um. None of my friends like him anymore. They will not put a good word in for him. Um. I don't know why the kid does it. He does it to himself. But, uh, we're not good friends with him at all anymore. We've known him for three years. Um. You know, he's gotten us in trouble before and then blamed us for getting him in trouble. So I hope that, you know, I never see that kid again.	The speaker claims that the person has been "screwing up" and getting them and their friends in trouble, but also states that the reason they stopped hanging out with him is **"that whole reason," which is not specified**. This **vagueness** suggests that the speaker may be withholding information or being evasive.	Label: Truthful Prediction: Deceptive	Vagueness
Example 8: MU3D (BF003_1PT)	We met at school and she was the sweetest person I've met in the whole entire world. We instantly clicked. Um. We always hang out with each other whether it was at her house or my house, just walking around the neighborhood or playing games online or Facebook, anything. We just always together, always having fun. Um. She's super nice. She's really pretty. She's super sweet. She's like the best person in the world. If you met her you'd probably try to steal her from me as best friend. She's amazing.	The speaker's **excessive use of superlatives** ("the sweetest person I've met in the whole entire world", "super nice", "super sweet", "the best person in the world") creates a sense of hyperbole, which can be a linguistic cue for deception.	Label: Truthful Prediction: Deceptive	Admiration
Example 9: MU3D (BF002_1PT)	My roommate is a really good person. She's very sweet. She's honest with me and I can appreciate that, um, she makes me laugh a lot. Very funny. Always has a joke always very sassy and. Um has good comedic timing. Um she's also very (sigh), she's very sweet like I said. Um, she is important part in my life. Uh. She's the first person I met when I got here and we've been best friends ever since. She's really a great person. I enjoy hanging out with her a lot.	The speaker's expressions of **affection and admiration** for their roommate seem overly rehearsed and insincere. The repetition of phrases such as "she's a really good person,"and "she's really a great person" suggests that the speaker is trying to convince themselves and others.	Label: Truthful Prediction: Deceptive	Admiration
Example 10: MU3D (BM009_2NL)	My ex-girlfriend, um, she's, she's pretty mean, um. In the beginning the relationship started off well and then it started going downhill. Um, one time in our relationship that stood out to me was, um, was she was supposed to go to California for my cousin's wedding. Um, we, we bought the flight tickets and everything and she just didn't show up.	the speaker **does not provide any further details or emotions about how this incident affected them**, which is unusual given the severity of the situation.	Label: Deceptive Prediction: Deceptive	Lack of emotion
Example 11: OpSpam (positive_fold2_t_hardrock_15)	so my sweetie lures me out under the guise of dinner at the China Grill, upon arrival he tells me his boys are watching the game in a suite upstairs and want to say hi...we arrive at the 35th floor and my oh my!!! SURPRISE!!! ladies and gentlemen THIS is something to be admired...spacious and roomy, yet warm, cozy and inviting...great bed, comfortable furniture, great 42 inch plasmas, and the views are UNBELIEVABLE...my only complaint? a shower should have a door to keep the heat IN...other than that no complaints, room service was great, morning coffee was yummo...overall a fabulous experience!!!	the reviewer is trying to **convey a sense of excitement and surprise that is not typical of a genuine review.** … Furthermore, the reviewer mentions that the surprise was a "ladies and gentlemen THIS is something to be admired", which suggests that the reviewer is trying to make the experience sound more impressive than it actually was.	Label: Truthful Prediction: Deceptive	Playful language/ humor
Example 12: OpSpam (negative_fold5_d_intercontinental_2)	My stay at the Ambassador East Hotel was a pleasant one. I loved being able to take a quick stroll to Michigan Ave for some shopping. I also loved the Pump Room. What a fun place to hang out and enjoy a Blood Mary. The front desk staff was wonderful with my check in since my reservation was lost. They found me a room immediately even though the city was book up. I thought the rooms were clean and well maintained. The shampoo was great.	The reviewer mentions several specific amenities and services that they enjoyed, such as the location, the Pump Room, and the front desk staff's help with their lost reservation. They also **provide specific details** about the cleanliness and maintenance of the rooms, as well as of the shampoo.	Label: Deceptive Prediction: Truthful	Specific detail

Figure 5: Examples of LLM Reasoning on MU3D and OpSpam Dataset

Model Type	Model	Config	Accuracy	F1
Baselines	RoBERTa-ft	-	73.16	72.88
	BiLSTM+Attention	-	71.44	71.19
	CNN	-	69.92	69.70
LLM Inference	LLaMA 3.1	zero shot	59.76	59.72
	LLaMA 3.1	few shot	64.32	64.16
	Gemma 2	zero shot	57.31	55.95
	Gemma 2	few shot	62.36	61.95
	Gpt-4o	zero shot	64.19	64.10
	Gpt-4o	few shot	64.41	64.17
LLM Finetuning	LLaMA 3.1	-	73.40	73.31
	Gemma 2	-	**73.55**	**73.54**

Table 8: LLM and baseline model performances Comparison on Prolific dataset.

`or trying to downplay their memory of an event. This could be a sign of deception, but it's not conclusive on its own. 2. The speaker then says "evidently I, I - I tried to call Hemy," which is a vague statement. The use of "evidently" suggests that the speaker is not entirely sure about their actions, and the repetition of "I" creates a sense of hesitation. [...] 6. Overall, the speaker's statements seem inconsistent and vague, which could be a sign of deception. However, the timeline could suggest truthfulness.`
`Judgment: Deceptive`

In this case, the reasoning incorporates both pro-truthful and pro-deceptive cues before arriving at a final judgment. Hence, it becomes challenging to systematically identify which cues LLMs rely on most frequently for decision-making at scale. In contrast, the post-hoc reasoning generation strategy leads to more focused, label-aligned justifications that are easier to interpret and trace back to specific cues -

`Judgment: Deceptive`
`Reasoning: The speaker claims they tried to call Hemy (presumably their boss) to inform them that they had left the office and something had happened to their husband. However, the speaker immediately follows this statement with "Which seemed - was a very normal thing for me to have done." This phrase suggests an attempt to justify or rationalize their actions, which is a common linguistic cue for deception. Additionally, the speaker's initial claim "I don't remember" followed by a detailed account raises a red flag, indicating they may be filling in memory gaps with plausible details.`

Based on both empirical evidence and interpretability advantages, we adopt the post-hoc reasoning generation approach over chain-of-thought prompting.

I Data-Origin Confound in OpSpam Dataset

OpSpam is one of the most widely used benchmarks in textual deception detection and thus provides a relevant testbed within the scope of our study. However, contrasting Mechanical Turk-generated deceptive reviews with scraped genuine ones introduces a data-origin confound, potentially inflating model performance by encouraging reliance on stylistic artifacts rather than true deception cues, as discussed in Fornaciari et al. (2020); Soldner et al. (2022). Our primary goal is to evaluate the behavior of LLMs and LMMs under zero and few-shot settings with limited in-domain supervision. As shown in Table 6, LLMs don't appear to exploit OpSpam dataset bias in the zero-shot setup because they cannot infer such confounding factors from single input data points. In the case of few-shot results with randomly selected examples, we do observe a performance improvement from zero-shot to few-shot for OpSpam but that is consistent with other datasets. For instance, LLaMA 3.1 on RLTD experiences a 17.14% improvement, whereas the imrpovement on the OpSpam dataset is 6.86%. However, Table 3 reveals that when using semantically similar (sim-top) few-shot examples, the average performance improvement across 3 models is 5.54% over random example selection in the OpSpam dataset. This gain is higher than that of RLTD (3.19%) and MU3D (4.69%), which suggests that with as few as 10 carefully curated in-context examples, LLMs may begin to pick up on underlying dataset-specific patterns, including potential biases. Understanding how LLMs leverage these biases offers valuable insights for designing more robust deception detection systems. To expand our evaluation beyond crowdsourced reviews, we have conducted additional experiments with the dataset from Confounds and Overestimations in Fake Review Detection (Soldner et al., 2022), specifically under its “Pure Veracity” setting. this setting is particularly challenging since both the truthful and deceptive reviews are coming from real-world owners of the smartphones gathered via the Prolific platform. This setting is particularly challenging since both the truthful and deceptive reviews are coming from real-world owners of the smartphones gathered via the Prolific platform. The results on Prolific dataset using our text-based baselines and the same LLM approaches as discussed in the paper, are presented in Table 8. Experimental results indicate that fine-tuned model performance drops notably compared to OpSpam, but the overall comparative trend remains similar, with fine-tuned LLaMA 3.1 and Gemma 2 models outperforming the baselines. In zero and few-shot setups, we observe similar performance on both OpSpam and the Prolific dataset, further confirming, LLMs cannot pick up the nuanced platform-specific biases very well with limited in-domain examples.

J Hyper-parameters and Budgeting

J.1 Baselines

We use a learning rate of 4e-5 for training the baselines and the models are trained for 20 epochs. The models are trained on 1 A6000 GPU.

J.2 LLMs

For few-shot examples, we explore 2, 4, 6, 8, 10 examples and report the best results for few shot performance. All the results reported are an average of 3 seeds.

K AI Assistance

We have used ChatGPT for writing assistance in the paper writing.