Evaluating LLMs on Deceptive Text Across Cultures

Katerina Papantoniou[1], **Panagiotis Papadakos**[1] **and Dimitris Plexousakis**[1, 2]
[1] Institute of Computer Science, Foundation for Research and Technology - Hellas, Heraklion, Greece
[2] Computer Science Department, University of Crete, Heraklion, Greece
{papanton, papadako, dp}@ics.forth.gr, dp@csd.uoc.gr

Abstract

Deception is a pervasive feature of human communication, yet identifying linguistic cues of deception remains a challenging task due to strong context dependency across domains, cultures, and types of deception. While prior work has relied on human analysis across disciplines like social psychology, philosophy, and political science, large language models (LLMs) offer a new avenue for exploring deception due to their strong performance in Natural Language Processing (NLP) tasks. In this study, we investigate whether open-weight LLMs possess and can apply knowledge about linguistic markers of deception across multiple languages, domains, and cultural contexts, with language and country of origin used as a proxy for culture. We focus on two domains, opinionated reviews and personal descriptions about sensitive topics, spanning five languages and six cultural settings. Using various configurations (zero-shot, one-shot, and fine-tuning), we evaluate the performance of LLMs in detecting and generating deceptive text. In detection tasks, our results reveal cross-model and cross-context performance differences. In generation tasks, linguistic analyses show partial alignment with known deception cues in human text, though this knowledge appears largely uniform and context-agnostic.

1 Introduction

Researchers in many disciplines agree that deceptive behaviour is context-dependent, influenced by factors such as domain, culture, and the type of deception (Markowitz and Hancock, 2022). This also reflects to the linguistic markers used to spot and analyse deceptive language (Newman et al., 2003; Ott et al., 2011; Pérez-Rosas and Mihalcea, 2014; Taylor et al., 2017). LLMs are paving the way for the advancement of various NLP tasks and deception related tasks are no exception. Here, we investigate if LLMs can capture and utilize accordingly the context-dependency of deception. This is critical, since the utilization of LLMs in real-world and high-stake applications in a way that fails to capture deception nuances could have serious consequences.

Although related studies show promise, they exhibit several important limitations. Many rely on small or non-human annotated datasets that fail to capture the complexity and nuance of real-world deception across diverse and dynamic contexts. Consequently, the high performance reported in controlled settings may not generalize well to other domain-specific deceptive content. Moreover, the deployed approaches often lack transparency, making it difficult to understand the underlying mechanisms of deception detection. Cultural and contextual variability in deceptive behavior remains underexplored, despite being a well-established factor in deception studies, while the absence of a standardized ground truth hinders deeper insights into why certain texts are perceived as deceptive.

In this work, we aim to partially address these limitations by evaluating both deception detection and generation capabilities of LLMs across a diverse set of deception datasets and cultural contexts. We further ground our analysis in well-established linguistic cues associated with deception and explore the relationship between these cues and the models' self-assessed factuality and interpretability.

Our goals are to examine whether current and open-weight LLMs:

- can detect and generate deceptive language across varied contexts.
- possess knowledge of linguistic indicators of deception, and whether they can apply this knowledge in both detection and generation tasks.

Proceedings of Recent Advances in Natural Language Processing,pages 884–893
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-101

- can recognize and adapt to contextual variations in deceptive behavior across languages, domains, and cultures. For this task, we use language and country of origin as proxies for culture (Hofstede, 1980; Pérez-Rosas and Mihalcea, 2014).

To the best of our knowledge, this is the first work that examines LLMs' capabilities for deception-related tasks via this viewpoint.

2 Related Work

In this section, we review prior research on the linguistic cues of deception and the performance of LLMs in deception-related tasks.

2.1 Linguistic Cues of Deception

Cognitive complexity. Deceptive language is expected to be characterised by less complex utterances, due to the increased cognitive load. Cues like 'mean word length', 'mean sentence length', 'mean preverb length', 'syllables', LTR (lemma token ratio - lexical diversity-), 'conjunctions', 'subordinate clauses' and 'readability' fall in this case. 'Motion verbs' (e.g., walk) are considered less cognitively complex and associated with deceptive utterances (Newman et al., 2003). However, socially oriented studies (e.g., the Pinocchio effect) and culturally focused research suggest that deceptive utterances may involve an increase in the quantity of text or similar levels of verbal output as truthful statements (Swol et al., 2012).

Non-immediacy. Non-immediacy can be conveyed through self and group references. Deceivers from individualistic cultures often use fewer first-person and more third-person pronouns to create distance. In contrast, those from collectivistic cultures may use more first-person and fewer third-person pronouns to distance the group from the deceit, or show no significant difference (Taylor et al., 2017; Papantoniou et al., 2022). Demonstrative pronouns (e.g., this, that) also signal physical or psychological distance.

Present. The narration of an actual past event is recalled from memory so it occurs naturally in the past tense, while a fabricated one, originating from imagination, is often conveyed in the present tense, indicating that the narrator is mentally constructing it in real time. (Christiansen, 2021).

Sentiment & emotions. Negative sentiment and emotions have been linked with deception (Newman et al., 2003; Hauch et al., 2012) perhaps as a reflection of the negative emotions experienced by the deceiver. However, studies on various contexts (spam reviews, imaginary stories) have found deceivers to use more positive sentiment and emotions (Ott et al., 2011; Toma and Hancock, 2012).

Specificity. Less specificity as expressed in the usage of more 'vague words', 'hedges', 'boosters', 'adverbs', 'adjectives', 'rate of adjectives and adverbs' and less 'spatial details', 'time details', 'named entities', 'exclusion words', and 'negations' have also often observed (Burgoon et al., 2003) and it is in line with theories like Reality Monitoring (Johnson and Raye, 1981). From a cultural perspective Taylor et al. (2017) found that liars from individualist cultures provide fewer perceptual details and more social details, whereas collectivist cultures show the reverse trend.

2.2 LLMs & Deception Tasks

The task of automated text-based deception detection task is traditionally approached as a classification task. Various features such as psycholinguistic indicators drawn from prior work on deception, n-gram features (Ott et al., 2011), syntactic features (Feng et al., 2012), have been exploited in this task. Here we focus on related work that employs LLMs.

Several works have explored the capacity of large language models (LLMs) to detect or generate deceptive content. Azaria and Mitchell (2023) for instance, investigate whether LLMs are internally aware of the truthfulness of statements they produce or consume. In their approach, a classifier is trained using the hidden layer activations of an LLM while it reads or generates a statement. This classifier achieves an accuracy between 71% and 83% on short sentences. Loconte et al. (2023) exploits FLAN-T5 models to classify texts across a range of domains, such as personal opinions, autobiographical memories, and future intentions. The results highlight the impact of model size, with larger models having superior performance. Boumber et al. (2024) examines the deception detection capabilities of LLMs enhanced with Retrieval-Augmented Generation (RAG). While LLM-based methods performed competitively, parameter-efficient fine-

tuning (PEFT) adapter approaches were the most effective.

In the context of deceptive content generation, Chen and Shu (2023) explore how prompting LLMs to produce misinformation impacts human and automated detection. They find that both humans and LLM-based detectors often fail to identify LLM-generated misinformation, highlighting its deceptive potency. Complementing this, Ignat et al. (2024) present the MAiDE-up dataset, with 10,000 real and 10,000 LLM-generated hotel reviews in ten languages. Their linguistic analysis reveals stylistic and semantic differences between human and AI-generated texts. Despite the limited training data, the fine-tuned models perform well in detecting LLM-generated deception across languages.

3 Methodology

3.1 Tasks

In this work, we explore three core tasks related to deception and language, using open-source and open-weight LLMs. Task A is deception detection, where models are evaluated on their ability to classify texts as deceptive or truthful. We assess the performance under zero-shot, one-shot, and fine-tuned settings using a variety of available datasets. Fine-tuned models trained on the English UDC_{train} dataset, were also evaluated in non-English languages. Task B is the introspective reasoning, where the models are prompted to explain their decisions. This includes reflection on culturally and linguistically grounded deception cues, providing insights into the model's interpretability and internal reasoning processes. Finally, Task C is deceptive text generation, which includes both paraphrasing existing truthful or deceptive content and open-ended generation of new deceptive statements. This allows us to analyze how LLMs construct deception across different styles, topics, and cultural contexts.

For the linguistic analysis, we built upon the feature set introduced in Papantoniou et al. (2022), extending it to the Italian language. The features are primarily count-based, covering linguistic categories such as pronouns, verbs, sentiment, and emotion. The English set includes 75 features, while sets for other languages are slightly smaller due to limited resources (e.g., sentiment lexicons). Readability is measured using a Flesch-like score (Flesch, 1948) adapted for each language. To examine linguistic patterns in LLM-generated datasets, we conducted statistical analysis using the non-parametric Mann–Whitney U test (two-tailed) for each dataset, comparing the distributions of features between truthful and deceptive texts. We applied a stringent significance threshold, setting the confidence level at 99.9% ($\alpha = 0.001$), in order to provide stronger evidence against the null hypothesis and minimize the likelihood of false positives.

3.2 Datasets

We use a unification of ten, quite diverse English deception detection-related datasets, named the Unified Deception Dataset (UDC). The UDC comprises of Bluff the Listener (Skalicky et al., 2020), OpSpam (Ott et al., 2011), DeRev2014 (Fornaciari and Poesio, 2014), DeRev2018 (Fornaciari et al., 2020), DecOp (english part) (Capuozzo et al., 2020b), Real Life Deception Dataset (Pérez-Rosas et al., 2015), Miami University Deception Detection Database (Lloyd et al., 2018), Diplomacy (Peskov et al., 2020), Open Domain Deception Dataset (Pérez-Rosas and Mihalcea, 2015) and Box of Lies (Soldner et al., 2019)[1]. UDC is diverse in many aspects, such as origin of text, genres, domains, and annotation methods (self-reported -SR- and distant supervision -DS-). A stratified split of UDC was used to create training, testing, and validation subsets with an 80-10-10 ratio. UDC is imbalanced in favor of the truthful class.

In addition to the test subset of the UDC we incorporate several other deception detection datasets in various languages for evaluation. These supplementary datasets are annotated using SR labels provided by the individuals who generated the content. This allows us to explore deception detection performance across different cultural and linguistic contexts. Notice that cultures are grouped into two categories individualistic and collectivistic ones, one of the six dimensions of national culture identified in (Hofstede, 1980). Table 1 provides an overview of the datasets.

3.3 Models and Settings

We experiment with LLMs with similar model sizes for a fairer comparison. The selection of model size was driven by memory and storage restrictions. We locally run models and in all cases quantization was used. Different models from the same

[1]We do not include the MAiDE-up dataset since it has not been annotated by experts.

Dataset	Lang.	Culture	Domain	Annotation	#T	#D	Reference
UDC_{train}	en	Individual	multi	SR & DS	19471	7402	
UDC_{val}	en	Individual	multi	SR & DS	2438	921	-
UDC_{test}	en	Individual	multi	SR & DS	2369	991	-
restaurant	en	Individual	restaurant	SR	53	55	Abri et al. (2020)
4city	en	Individual	hotel	SR	314	319	Li et al. (2014)
cCult	en	Individual	friend, disputed	SR	300	300	Pérez-Rosas et al. (2014)
boulder	en	Individual	hotel, electronics	SR	451	1041	Salvetti et al. (2016)
CLiPS	nl	Individual	product	SR	644	644	Verhoeven et al. 2014
almela	es	Individual	friend, disputed	SR	300	299	Almela (2021)
decop	it	Individual	disputed	SR	1247	1248	Capuozzo et al. (2020a)
cCult	ro	Collective	friend, disputed	SR	432	432	Pérez-Rosas et al. (2014)
cCult	es-mx	Collective	friend, disputed	SR	172	174	Pérez-Rosas et al. (2014)

Table 1: Overview of the used datasets.

family were also examined to observe their evolution. Specifically, the following models were employed: mistral-7B-instruct-v0.1 (*mistral*) (Jiang et al., 2023), falcon2-11B (*falcon*) (Malartic et al., 2024), phi-3-medium-4k-instruct (*phi3*) (Abdin et al., 2024), phi-3.5-mini-instruct(*phi3.5*), gemma2-9B-instruct (*gemma*) (Team et al., 2024), llama-2-13b-chat (*llama2*) (Touvron et al., 2023), llama-3-8B-instruct (*llama3*), llama-3.1-8B-instruct (*llama3.1*), DeepSeek-R1-Distill-Llama-8 (*deepseek*) (DeepSeek-AI et al., 2025), cerbero-7b (*cerbero*) (Galatolo and Cimino, 2023) for Italian, eva-mistral-turdus-7b-spanish (*esmistral*) for Spanish/Mexican and roLlama3.1-8b-instruct (*rollama*) (Masala et al., 2024) for Romanian. There is no LLM specifically for Dutch. All are open-weight except of gemma that is also open-source.

3.3.1 Fine-tuning

We fine-tuned the following models: Llama-3-8B (*llama3_FT*), Llama-3.1-8B (*llama3.1_FT*), Phi-3 Medium (*phi3_FT*), phi3.5:3.8b (*phi3.5_FT*), gemma-2-9b (*gemma_FT*), and DeepSeek-R1-Distill-Llama-8B (*deepseek_FT*) models for sequence classification over UDC_{train} dataset. All models were loaded with 4-bit quantization and equipped with Low-Rank Adapters (LoRA) (Hu et al., 2021) on all linear layers. LoRA was configured with rank $r = 16$, $\alpha = 8$, dropout = 0.05, and Rank-Stabilized LoRA (Kalajdzievski, 2023) enabled. We used AdamW with an initial learning rate of 1e-4, a linear scheduler, and trained for one epoch. The experiments were conducted on a single NVIDIA GeForce RTX 4090.

3.3.2 Classification Setting

Prompts were written in English, except for language-specific models (e.g., esmistral), where the entire prompt was translated[2]. For both zero-shot and one-shot prompting, we set the temperature to 0 and top_p to 0.9 to ensure deterministic outputs. In the one-shot setting, we ran 10 trials, each time randomly selecting a pair of examples (one truthful and one deceptive) from the same dataset.

3.3.3 Generation Setting

Appropriate prompts were used to generate both open-ended and paraphrased texts. The open-ended texts include reviews for products and services (e.g., hotels, restaurants, books, hotel chains) and opinionated texts on eleven sensitive topics such as abortion, gun control, cloning, and human relationships, topics previously studied in related work (Newman et al., 2003; Ott et al., 2011). Each prompt instructs the model to generate a truthful and a deceptive version of a given text by applying the appropriate linguistic markers. The prompt requests texts of approximately 160 words and incorporates the comprehensive definition of deception provided by Masip et al. (2004). Texts were generated in five languages representing six cultures (Spanish are used for Spain and Mexico), aligned with those used for detection. To explicitly control for culture, prompts directed the LLM to impersonate native speakers from specific countries. For the review domain, item names were imaginary generated by using LLMs, except in the hotel and food chain categories were existing names of chains used. To promote diversity and creativity, we varied the temperature parameter. For the open-ended texts we had one example with 0, and 3 examples for each value in {0.6,0.7,0.8}, generating 480 texts per domain and language. For the paraphrases we selected a value from {0,0.1,0.2}, generating one paraphrase

[2] https://github.com/nidhaloff/deep-translator

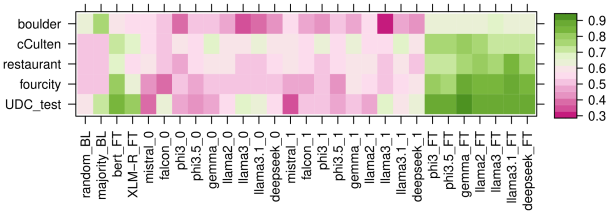

Figure 1: Accuracy (English)

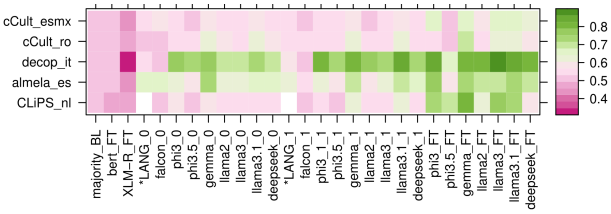

Figure 2: Accuracy (non-English)

for each dataset entry. Prompts were translated into the target languages, as ablation studies showed that translated prompts produced richer linguistic features. For paraphrasing, models were instructed to rewrite each text while preserving its original label (truthful or deceptive), explicitly guiding them to use appropriate linguistic markers.

4 Deception Detection Results

4.1 Classification Results

Model support[3] is the number of times the LLM (in 0-shot and 1-shot settings) responds appropriately to the prompt. While most models show high compliance (often near 100%), some fail to follow the template or refuse to respond due to perceived ethical concerns. Such exceptions include rollama_1-shot on Romanian (53%), llama2_1-shot on cCult-en (30%), falcon_1-shot on CLiPS (30%), and lamma3.1_1-shot on UDC_*test* (26%).

We define the following baselines: a majority-class (majority_BL) for all the datasets, a random baseline (random_BL) mainly for the imbalanced datasets (UDC_*test*, boulder) and two encoder-only architectures namely bert-base-uncased (bert_FT) and xlm-roberta-base (XLM-R_FT) that supports multiple languages. The last two baselines were fine-tuned in the same dataset (UDC) as the fine-tuned LLMs.

Figures 1 and 2 present the classification results for English and non-English test sets, respectively. The language specific models (cerbero, esmistral and rollama) are represented by the *LANG-0 and *LANG-1 labels (no specific LLM for Dutch was deployed).

Fine-tuned models consistently outperform others, with a particularly large margin in the English and CLiPS (nl) datasets. The same holds for the rest non-English datasets, although the performance gap is smaller. The highest overall accuracy is observed on the UDC_*test* dataset, which is an in-domain scenario for the fine-tuned models. Between the remaining settings, 1-shot slightly outperforms 0-shot, but the difference is marginal.

Instructed models fine-tuned on language-specific corpora consistently underperform. However, this cannot be solely attributed to the instruction tuning process, as their performance depends heavily on the underlying base model. For instance, both cerbero and esmistral are based on mistral and its instruct model do not perform well on the English dataset. Similarly, rollama, based on llama3, achieves results comparable to the instruct model in the 1-shot setting, and in the 0-shot setting performs near chance level (50.2% accuracy). These findings suggest that instruction tuning in a specific language does not necessarily degrade a model's general capabilities. Within the LLaMA family, newer models consistently outperform older ones, despite llama2 having more parameters. In the Phi model family, the larger phi3 model (14B parameters) generally outperforms its smaller variants. A exception is the boulder dataset, where phi3.5 (3.8B parameters) consistently achieves better results across all settings.

Among the fine-tuned models, the gemma model achieves the best performance across six datasets. The llama3.1 model performs best on the restaurant dataset, while llama3 leads in the boulder and decop datasets. The phi3 model stands out on the almela dataset. In the zero-shot setting, gemma again delivers strong results, achieving the best performance in five datasets. In the 1-shot setting, the results are more balanced. The *llama3.1* excels in four datasets (crossCult-esMx, CLiPS, crossCult-enUs, decop), *mistral* in two (restaurant, boulder), *gemma* in two (crossCult-ro, almela) and *llama3* in the UDC_*test*. We observe that in non-english datasets the *llama3.1* model is better compared to *llama3*, likely due to the improved multilingual support of *llama3.1*.

We consistently observe that at least one LLM

[3]Due to space restrictions, accompanying material is available at https://gitlab.isl.ics.forth.gr/papanton/evaluating-llms-on-deceptive-text

setting outperforms the baseline methods across all datasets. The only exception is the boulder dataset, where the majority_BL achieves the highest performance. This dataset is notably peculiar. The dataset is heavily imbalanced toward the deceptive class, an unrealistic scenario. It includes both fabrications and lies, leading to variation in knowledge and emotional expression, which challenges LLMs' generalization capabilities.

4.2 Introspection

In this task, we prompt the most robust models in the classification task, gemma and llama3.1, to explain their classification decision via linguistic cues that they deemed important, indicating which of these may carry cultural interpretations. We constrain their responses using the linguistic feature list from Papantoniou et al. (2022). Both models frequently referenced pronoun- and sentiment-related features. However, llama3.1 had difficulty adhering to the provided list, often returning hallucinated features (sentiment_compound). Neither model effectively used the features to discriminate between truthful and deceptive instances or achieved consistently high per-class success rates. Regarding culturally-relevant features, gemma produced a more concise list with an emphasis on pronouns, while llama 3.1, yielded more unstable and inconsistent outputs.

5 Deceptive Text Generation Results

In this task, we use gemma for text generation due to high template adherence. Classification was again done using gemma and llama 3.1. The support was high both for the paraphrased and open-ended texts.

Table 2 reports the results of the Mann-Whitney tests, highlighting statistically significant features with at least a moderate effect size ($r \geq 0.3$). • denotes deceptive features while ○ truthful. Due to space constraints, sentiment and emotion features from various lexicons have been aggregated into two composite features: pos_emo and neg_emo. Notably, in the original human-written datasets, the same statistical test yielded no significant features above the effect size threshold, except for the 4city dataset and the restaurant dataset. As a result in the table we report with orange color (•, ○) the features that are SS for the smaller effect size ($r \geq 0.1$) that overlap with the paraphrased SS. For the restaurant and 4city datasets, we also report SS with $r \geq 0.3$ overlapping or not.

We observe that the number of statistically significant (SS) features threshold is generally higher for paraphrased texts, especially in the English datasets, compared to open-ended generated texts. Interestingly, open-ended texts exhibit a broader range of distinct features across all datasets (20 in total), whereas paraphrased texts involve fewer (16). This might suggest that in open-ended generation, gemma applies linguistic markers more freely, while in paraphrased generation, the model is more constrained by the structure and style of the original human-written text. These distinctions guided our hypothesis during the analysis of open-ended and paraphrased text generation. Prior research on human-generated text (see the meta-analysis by Hauch et al. (2012)) has found only small but significant effect sizes for linguistic cues. This suggests that the differences between groups are amplified in LLM-generated texts, highlighting the stronger expression of such cues in model outputs.

The open-ended texts from sensitive domains consistently show the fewest SS features. This likely reflects the LLMs' training data, which more frequently includes review-style content about products and services than essay-like opinions on sensitive topics or friendships. As a result, LLMs are more adept at generating diverse texts in review domains. Similarly, in the highly diverse UDC_test dataset, the paraphrased texts yield only one SS feature ('vague words'), underscoring the strong context dependency of deception-related cues. In paraphrased texts for non-English languages, we observe reduced linguistic diversity compared to English, whereas open-ended texts show no notable cross-linguistic differences.

A closer look at specific cues reveals several noteworthy patterns, largely aligning with the findings discussed in Section 2.1. The deceptive LLM-generated texts are less specific and less cognitively complex. This is reflected in higher use of 'vague words', 'hedges', and 'adverbs'. In contrast, features associated with cognitive complexity like 'mean word length', 'mean sentence length', 'syllables', and 'conjunctions' all are more prevalent in truthful texts. Similarly, and in accordance to the related work, higher 'readability scores', which indicate simpler and more fluent texts, are linked to deception. This feature is important in paraphrased texts across languages. Sentiment and emotion cues, often context-dependent in prior research,

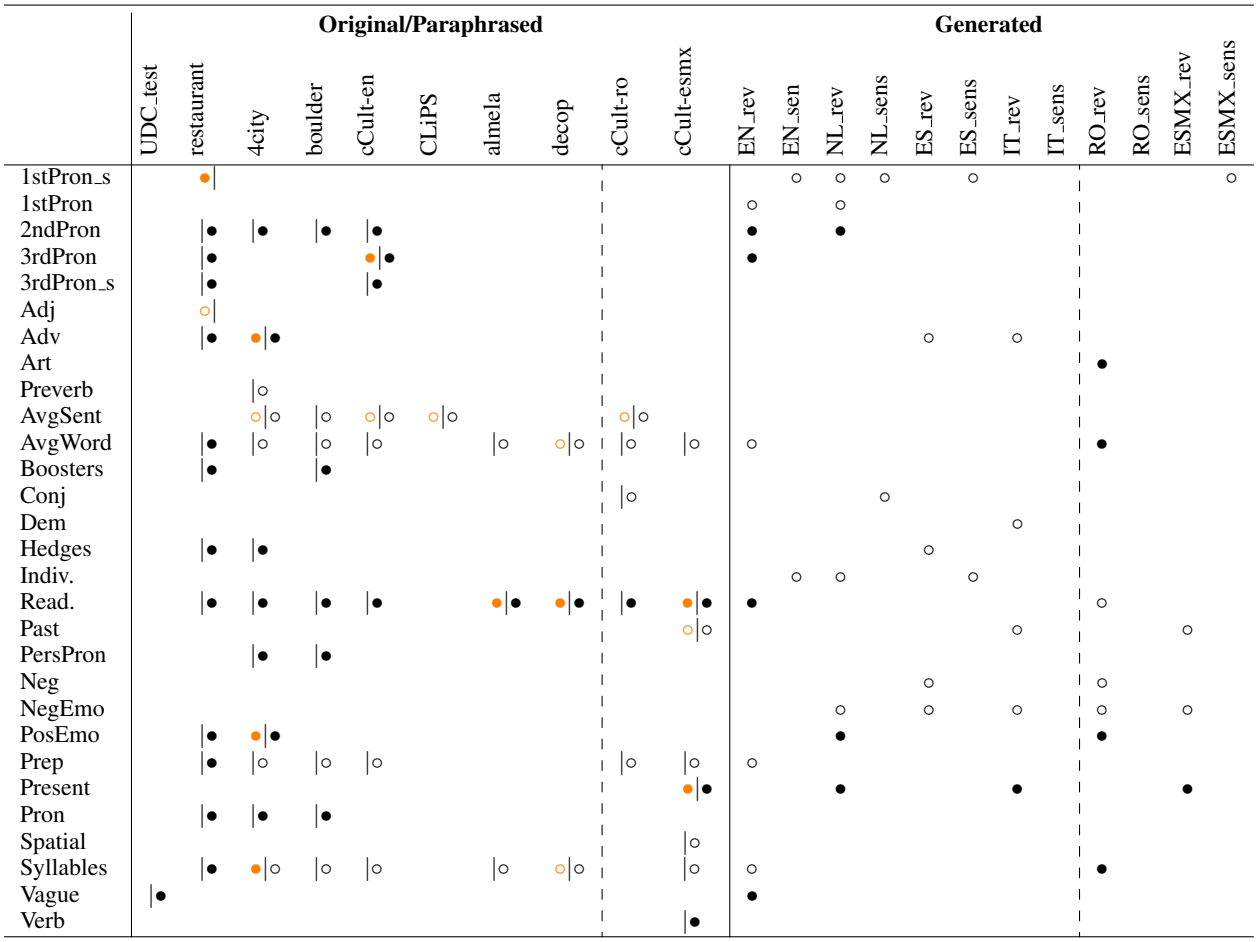

Table 2: Statistically significant features from Mann-Whitney tests (● for deceptive and ○ for truthful). The dashed lines divide the individualistic/collectivistic datasets. With orange SS features with r $\geq$ 0.1 that overlap with the paraphrased SS. For the restaurant and 4city datasets, we also report SS with r $\geq$ 0.3 overlapping or not. More details about the features are provided in the aforementioned repository.

show a consistent trend across languages and domains: truthful texts exhibit more negative sentiment and emotions (e.g., anger, sadness), while deceptive texts tend toward more positive language. Temporal cues follow established patterns, with deceptive texts favoring the present tense and truthful ones favoring the past tense. Pronoun usage also aligns with prior work, as third-person pronouns appear more frequently in deceptive texts and first-person pronouns in truthful ones. In general, we do not observe any noticeable variation across cultures.

One of the most unexpected findings is the association of 'second person personal pronouns' with deceptive texts in both the English and Dutch datasets. While some non-deception-focused studies suggest that second-person pronouns enhance personal engagement and conversational tone (Sun et al., 2024), the deception literature remains divided. Newman et al. (2003) argue that liars tend to avoid second-person pronouns, whereas Ickes et al. (1986) suggest the opposite. More recent studies do not consider this feature diagnostic for deception. An analysis of specific examples reveals expressions such as 'you should definitely try it!' 'you'll leave full', 'you won't regret it!', 'You'll discover', adding a positive sentimental and emotional tone.

5.1 Classification of the Generated Text

Figures 3 and 4 present the detection accuracy scores for the LLM-generated texts. In the paraphrased setting, gemma_FT consistently outperforms gemma_0-shot, particularly for English, highlighting both the alignment of paraphrased texts with human language and the model's stronger support for English. In contrast, the open-ended setting yields lower accuracies overall, with noticeable variation across domains and models. In the sensitive/best friend domain, all gemma-based mod-

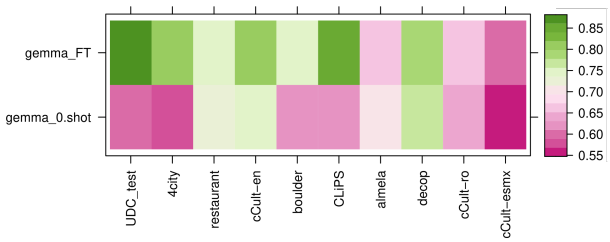

Figure 3: Accuracy for paraphrased texts

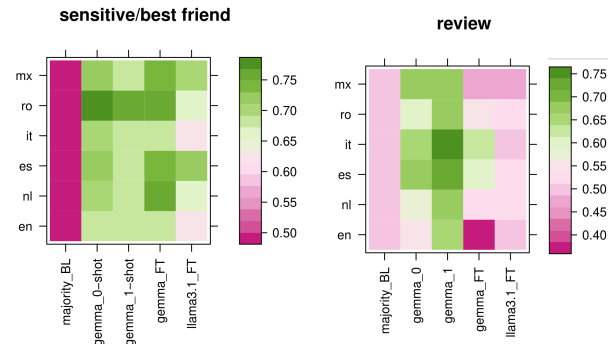

Figure 4: Accuracy for open-ended texts

els perform adequately, while llama3.1_FT shows significantly lower performance. In the review domain, fine-tuned models generally perform poorly, with accuracies near or below the baseline, except for the Italian dataset where llama3.1_FT performs well. Lastly, across review datasets, the 1-shot setting consistently outperforms the 0-shot setting. Overall, these results underscore the gap between LLM-generated and human text, particularly in the review domain.

6 Conclusion and Future Work

The results demonstrate that LLMs show partial success in identifying and generating deception-related content, especially when fine-tuned. Their performance is generally adequate in classification tasks. However, critical limitations remain. Open-ended text generation reveals inconsistencies and reduced accuracy, particularly in cross-cultural and sensitive domains. The findings highlight the importance of context, in terms of domain and culture. A key limitation of the current study is the use of country as a proxy for culture, that can oversimplify complex cultural identities in pluralistic societies. Future work should explore Retrieval-Augmented Generation (RAG) approaches to inject explicit knowledge of deception markers and improve interpretability. Experimentation with larger LLMs and the integration of domain and cultural metadata could enhance context-aware reasoning. Finally, addressing biases in underrepresented languages and improving explainability mechanisms is crucial for ensuring ethical and inclusive systems.

References

M. Abdin, J. Aneja, and etal. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

F. Abri, L. F. Gutiérrez, A. S. Namin, K. S. Jones, and D. R. W. Sears. 2020. Linguistic Features for Detecting Fake Reviews. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 352–359.

Á. Almela. 2021. A Corpus-Based Study of Linguistic Deception in Spanish. *Applied Sciences*, 11(19).

A. Azaria and T. Mitchell. 2023. The Internal State of an LLM Knows When It's Lying.

D. Boumber, B. E. Tuck, R. M. Verma, and Fatima Z. 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.

J. K. Burgoon, J. P. Blair, T. Qin, and J. F. Nunamaker. 2003. Detecting deception through linguistic analysis. In *Intelligence and Security Informatics*, pages 91–101, Berlin, Heidelberg. Springer Berlin Heidelberg.

P. Capuozzo, I. Lauriola, C. Strapparava, F. Aiolli, and G. Sartori. 2020a. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1423–1430, Marseille, France. European Language Resources Association.

P. Capuozzo, I. Lauriola, C. Strapparava, F. Aiolli, and G. Sartori. 2020b. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1423–30, Marseille, France. European Language Resources Association.

C. Chen and K. Shu. 2023. Can LLM-Generated Misinformation Be Detected?

Th. W. Christiansen. 2021. Linguistics and Deception Detection (DD): A Work in Progress. *Studies in Logic, Grammar and Rhetoric*, 66(2):169–200.

DeepSeek-AI, Daya Guo, and etal. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

S. Feng, R. Banerjee, and Y. Choi. 2012. Syntactic Stylometry for Deception Detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–75, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221–233.

T. Fornaciari, L. Cagnina, P. Rosso, and M. Poesio. 2020. Fake opinion detection: how similar are crowdsourced datasets to real data? *Lang. Resour. Eval.*, 54(4):1019–1058.

T. Fornaciari and M. Poesio. 2014. Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

F. A Galatolo and M. G.C.A Cimino. 2023. Cerbero-7B: A Leap Forward in Language-Specific LLMs Through Enhanced Chat Corpus Generation and Evaluation. *arXiv preprint arXiv:2311.15698*.

V. Hauch, I. Blandón-Gitlin, J. Masip, and S. L. Sporer. 2012. Linguistic cues to deception assessed by computer programs: A meta-analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 1–4, Avignon, France. Association for Computational Linguistics.

G. Hofstede. 1980. *Culture's consequences: International differences in work-related values*. Sage Publications.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models.

W. Ickes, S. Reidhead, and M. Patterson. 1986. Machiavellianism and self-monitoring: As different as "me" and "you". *Social Cognition*, 4(1):58–74.

O. Ignat, X. Xu, and R. Mihalcea. 2024. MAiDE-up: Multilingual deception detection of gpt-generated hotel reviews. *CoRR*, abs/2404.12938.

A. Q. Jiang, A. Sablayrolles, and etal. 2023. Mistral 7b.

M. K. Johnson and C. L. Raye. 1981. Reality Monitoring. *Psychological Review*, 88(1):67–85.

D. Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora.

J. Li, M. Ott, C. Cardie, and E. Hovy. 2014. Towards a General Rule for Identifying Deceptive Opinion Spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–76. Association for Computational Linguistics.

E. Paige Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman. 2018. Miami University deception detection database. *Behavior Research Methods*, 51(1):429–439.

R. Loconte, R. Russo, Pasquale Capuozzo, P. Pietrini, and G. Sartori. 2023. Verbal lie detection using Large Language Models. *Scientific Reports*, 13(1).

Q. Malartic, N. R. Chowdhury, and et al. 2024. Falcon2-11b technical report.

D. M. Markowitz and J. Hancock. 2022. Lies and Language: A Context-Contingent Approach to Verbal Cues of Deceit. *PsyArXiv*.

M. Masala, D. C. Ilie-Ablachim, and et al. 2024. "Vorbeşti Româneşte?" a recipe to train powerful romanian llms with english instructions.

J. Masip, E. Garrido, and C. Herrero. 2004. Defining deception. *Anales de Psicología / Annals of Psychology*, 20(1):147–172.

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–75. PMID: 15272998.

M. Ott, Y. Choi, C. Cardie, and J. 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 - Volume 1*, HLT '11, pages 309–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Papantoniou, P. Papadakos, T. Patkos, G. Flouris, I. Androutsopoulos, and D. Plexousakis. 2022. Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *Natural Language Engineering*, 28(5):545–606.

V. Pérez-Rosas, C. Bologa, M. Burzo, and R. Mihalcea. 2014. *Deception Detection Within and Across Cultures*, pages 157–75. Springer International Publishing, Cham.

V. Pérez-Rosas and R. Mihalcea. 2014. Cross-cultural Deception Detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–5, MD. Association for Computational Linguistics.

V. Pérez-Rosas and R. Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125, Lisbon, Portugal. Association for Computational Linguistics.

D. Peskov, B. Cheng, A. Elgohary, J. Barrow, C. Danescu-Niculescu-Mizil, and J. Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. 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. ACM.

F. Salvetti, J. B. Lowe, and J. H. Martin. 2016. A Tangled Web: The Faint Signals of Deception in Text - Boulder Lies and Truth Corpus (BLT-C). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

S. Skalicky, N. Duran, and S. A Crossley. 2020. Please, please, just tell me: The linguistic features of humorous deception. *Dialogue and Discourse*, 11(2):128–149.

F. Soldner, V. Pérez-Rosas, and R. Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Z. Sun, C. C. Cao, S. Liu, Y. Li, and C. Ma. 2024. Behavioral consequences of second-person pronouns in written communications between authors and reviewers of scientific papers. *Nature Communications*, 15(1).

L. M. Van Swol, M. T. Braun, and D. Malhotra. 2012. Evidence for the Pinocchio Effect: Linguistic Differences Between Lies, Deception by Omissions, and Truths. *Discourse Processes*, 49(2):79–106.

P. J. Taylor, S. L., S. M. Conchie, and T. Menacere. 2017. Culture moderates changes in linguistic self-presentation and detail provision when deceiving others. *Royal Society Open Science*, 4(6):170128.

Gemma Team, Morgane Riviere, and etal. 2024. Gemma 2: Improving Open Language Models at a Practical Size.

C. L. Toma and J. T. Hancock. 2012. What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles. *Journal of Communication*, 62(1):78–97.

Hugo Touvron, Louis Martin, and etal. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Ben Verhoeven and Walter Daelemans. 2014. CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).