Why is "Chicago" Predictive of Deceptive Reviews? Using LLMs to Discover Language Phenomena from Lexical Cues

Jiaming Qu[*]
Amazon
Seattle, USA
qjiaming@amazon.com

Mengtian Guo
UNC Chapel Hill
Chapel Hill, USA
mtguo@email.unc.edu

Yue Wang
UNC Chapel Hill
Chapel Hill, USA
wangyue@email.unc.edu

ABSTRACT

Deceptive reviews mislead consumers, harm businesses, and undermine trust in online marketplaces. Machine learning classifiers can learn from large amounts of training examples to effectively distinguish deceptive reviews from genuine ones. However, the distinguishing features learned by these classifiers are often subtle, fragmented, and difficult for humans to interpret. In this work, we explore using large language models (LLMs) to translate machine-learned lexical cues into human-understandable language phenomena that can differentiate deceptive reviews from genuine ones. We show that language phenomena obtained in this manner are empirically grounded in data, generalizable across similar domains, and more predictive than phenomena either in LLMs' prior knowledge or obtained through in-context learning. These language phenomena have the potential to aid people in critically assessing the credibility of online reviews in environments where deception detection classifiers are unavailable.

1 INTRODUCTION

Online reviews are an essential source of information for consumers to make purchasing decisions and businesses to understand their customers. Although many platforms have implemented machine learning techniques to detect and block deceptive reviews [1], these systems are far from universally available, and users may still face deceptive content without algorithmic assistance. Unlike platforms that can leverage metadata or behavioral signals to identify deception [20], users rely almost exclusively on review text—and prior work shows that untrained people perform poorly at this task [3].

Therefore, it is important to help users develop keen eyes that can spot potentially deceptive reviews where algorithmic interventions are absent. A natural starting point is to educate users with the most salient signals identified by machine learning classifiers [14, 25, 27]. However, while classifiers can learn predictive lexical cues for accurately distinguishing deceptive reviews from genuine ones from large amounts of data, such text features are often subtle and fragmented. For example, prior work found that the word "Chicago" is predictive of deceptive hotel reviews [10]. These predictive words are often not random artifacts but reflect *underlying language phenomena* in deceptive text instead.

In this work, we view such words—predictive yet unintuitive lexical cues—as surface manifestations of underlying language phenomena and explore whether large language models (LLMs) can discover such phenomena. Our goal is *not* to build a stronger deception detector, but to evaluate whether LLMs can reliably translate machine-learned lexical cues into human-interpretable phenomena. In particular, we ask LLMs to answer questions in the spirit of "why is lexical feature X predictive of deceptive or genuine reviews?" On the one hand, LLMs would readily provide fluent, plausible explanations to such questions (which we call *conjectures* before they are *validated*), and their natural language outputs are human-comprehensible. On the other hand, plausibility alone does not guarantee correctness: LLM-generated conjectures would be useless if they are sheer hallucinations and fail to correspond to any real pattern that actually distinguishes deception from genuine reviews. To evaluate whether such conjectures reflect meaningful linguistic patterns, we use deception detection in hotel reviews as a case study and investigate three research questions (RQs):

- **RQ1 (Predictiveness)**: Can these phenomena distinguish deceptive reviews from genuine ones?
- **RQ2 (Generalization)**: Are these phenomena generalizable to new data in similar domains (i.e., similar products or services)?
- **RQ3 (Alternatives)**: Instead of explaining predictive words, can these phenomena be obtained by prompting LLMs to conjecture based on its prior knowledge or labeled examples?

Our results show that these LLM-conjectured language phenomena have predictive power, are generalizable to similar-domain data, and outperform phenomena obtained from LLMs' prior knowledge or in-context learning alone. Together, these findings indicate that it is feasible to use LLMs to translate subtle deception cues learned by statistical classifiers into human-understandable, predictive language phenomena. These language phenomena could ultimately be useful in helping people increase their acumen in assessing the credibility of online reviews and decrease their reliance on algorithmic filters when such filters are unavailable.

2 RELATED WORK

Feature importance explanations, which highlight input features that are most influential to the prediction output, are a popular approach to explaining machine learning model predictions [14, 25, 27]. While empirical studies have shown such explanations can improve end-users' decision-making and understanding of AI systems [11, 22], predictive features are not always self-explanatory. For example, studies applying such explanations on text data have found that the word "problems" is predictive of positive sentiment [23], and that the word "Chicago" is predictive of deceptive reviews [10]. Given large and representative data, such words are often not the result of overfitting but reflect underlying language phenomena that give rise to these word-label relations.

Prior studies have explored different approaches for making such phenomena more explicit, such as showing nearby words of the

---

arXiv:2511.13658v1 [cs.CL] 17 Nov 2025

predictive word [7, 23, 26] or considering interactions with other words in the input [4, 8, 30]. While these algorithms are effective at revealing context-related phenomena (e.g., negations and colloquial expressions like "no problems"), they may fail to capture complex phenomena that go beyond local contexts. For instance, the word "Chicago" is associated with deceptive hotel reviews because fake reviews often reinforce branding by emphasizing a hotel's full name, including the city name. In this paper, we use deception detection as a case study and leverage LLMs to explain the language phenomena behind deception cues. Our approach was inspired by recent research that prompted LLMs to verbalize predictive features into natural language narratives [15, 19, 32]. However, these approaches mainly *paraphrase* predictive words for better readability. Our approach is crucially different in that it goes beyond observed lexical cues to unobserved phenomena-based explanations.

3 METHODOLOGY

Problem Formulation: Consider a labeled text dataset $\{(x, y)\}$, where x represents a piece of text and y represents a label (e.g., genuine or deceptive review). Feature importance explanations identify words $\{w \in x\}$ that are predictive of a label y. In this work, we focus on a subset $\{w' \in x\} \subseteq \{w \in x\}$: words that are the most significant signals for the classifier but appear unintuitive to humans. Our goal is to leverage LLMs to *translate* these salient lexical cues into more human-interpretable language phenomena that plausibly give rise to the cues. Formally, we prompt an LLM to *conjecture* a candidate phenomenon associated with the cue. The LLM-conjectured phenomena may or may not reflect actual patterns in the underlying data. Thus, we *validated* whether the LLM-conjectured phenomena are predictive of genuine or deceptive reviews (RQ1), generalizable beyond the original data (RQ2), and dependent on classifier-learned predictive words (RQ3).

Dataset: We used two datasets in our study: a dataset containing 800 genuine and 800 deceptive reviews for Chicago hotels [17, 18] (denoted as $D_{Chicago}$) and another dataset containing 240 genuine and 240 deceptive reviews for Houston, New York, and Los Angeles hotels [12] (denoted as $D_{three-cities}$). Genuine hotel reviews were collected from verified travelers while deceptive reviews were written by crowd workers.

Identifying Predictive Words: To investigate whether LLMs can translate predictive lexical cues into meaningful language phenomena, we first trained logistic regression classifiers on $D_{Chicago}$ and identified predictive words. It is notable that we do not treat classification performance as the object of optimization. We deliberately used a simple model (logistic regression) solely as a *feature-discovery tool*, and predictive words can also be extracted using other sophisticated approaches [14, 25, 27].

Given the limited size of the dataset, we performed 10-fold cross validation with balanced training and testing folds. Each time, we trained a classifier with unigram TF-IDF features with lowercasing and stop word removal. The classifiers achieved an average F1-score of 0.88. For each classifier, we identified 25 words that were the most predictive of genuine or deceptive reviews through regression coefficients (50 predictive words in total). All selected words passed a Wald test for the significance of predictiveness. To obtain particularly stable lexical cues, we selected only words that were

predictive *across all 10 folds*, yielding 16 words for genuine reviews and 14 for deceptive reviews. As shown in Table 1, we expect that most laypeople without expertise in deception detection would naturally wonder about the underlying language phenomena.

RQ1 Experiment: Our RQ1 investigated whether using LLMs can be a generally reliable approach to discover language phenomena from lexical cues in deception detection. To this end, we used a *conjecture-then-validate* pipeline with two separate steps.

First, we prompted the LLM to conjecture the underlying phenomena for words predictive of hotel reviews being genuine or deceptive using the prompt below. We provided all 30 predictive words in the prompt and did not restrict the number of phenomena to be conjectured. All predictive words and their associated phenomena (conjectured by LLMs) are shown in Table 1.

> **System Prompt for Conjecturing Phenomena**
>
> You are an expert in deception detection. You will be provided a list of words that are most predictive of genuine or deceptive Chicago hotel reviews. Your task is to identify language phenomena or psycholinguistic patterns that appear in genuine and deceptive hotel reviews and are related to these words.

While the conjectured phenomena are plausible, all LLMs are prone to hallucinations. Prior studies have mainly used benchmark datasets [6, 28] or recruited human evaluators [5, 31] to evaluate LLM-generated contents. However, there is no ground truth for these language phenomena in this task. Therefore, we took an algorithmic evaluation approach: we prompted the same LLM to classify reviews as genuine or deceptive with the conjectured phenomena. The rationale is that if these phenomena reflect actual patterns in the data (i.e., non-hallucinated), they should improve the LLM's predictive performance. Otherwise, fabricated phenomena could mislead the LLM's reasoning.

Hence, we tested two prompt conditions. First, we use a *phenomena-in-the-prompt* condition, where we provided all the conjectured phenomena associated with both genuine and deceptive reviews in the prompt. Only phenomena but not predictive words were provided. Here, our main objective was to investigate whether LLMs can reliably discover language phenomena *at all*. Therefore, we prioritized an aggregated evaluation and leveraged the full set of phenomena to test their collective utility. Second, we tested a *zero-shot prompt*, where no auxiliary information was provided.

We tested the pipeline with four LLMs: `GPT-5-mini` from OpenAI [16], `Haiku 4.5` from Anthropic [21], `Nova Pro` from Amazon [2], and `Gemini-2.5-flash` from Google [13]. For all LLM calls in this work, we used the corresponding APIs with default settings. Two authors independently designed initial prompts and collaboratively revised the final versions for each prompting scenario. We interacted with LLMs through DSPy [9], a framework for modular construction of LLM applications by explicitly defining prompt components. This design improves the reproducibility of our experiments. All prompts are available in our codebase[1].

RQ2 Experiment: In addition to validating whether the LLM-conjectured phenomena are non-hallucinated, we investigated the generalizability of these phenomena. To this end, we trained logistic

---

[1]https://jiamingqu.com/deception_detection.zip

Table 1: LLM-conjectured phenomena from words predictive of genuine and deceptive reviews.

Predictive Words	LLM-conjectured Phenomena (aggregated over four LLMs with rephrasing)
small, large, quiet, bathroom, floor, breakfast	Genuine reviews use measurable, verifiable adjectives and specific room features.
location, street, river, michigan	Genuine reviews often provide concrete geographic references and local landmarks.
rate, priceline	Genuine reviews frequently mention pricing, booking platforms, or value for money.
reviews, helpful, us	Genuine reviews refer to other reviews or their desire to contribute useful information.
luxury, luxurious	Deceptive reviews often overemphasize the luxurious aspects and high-end services.
hotel, chicago, millennium	Deceptive reviews often use generic category terms and prominent place/brand names.
seemed, recently, definitely	Deceptive reviews tend to use words expressing certainty or vague temporal framing.
husband, vacation, staying, experience	Deceptive reviews use family roles or staged narratives to fabricate a plausible story.
food, towels	Deceptive reviews list desirable amenities or sensory cues without specific details.

regression classifiers on $D_{Chicago}$ and tested on $D_{three-cities}$ using three feature sets: (1) unigram TF-IDF features only, (2) phenomena-based features only, and (3) both features. This simulated a scenario where machine learning models are applied to out-of-distribution data: the model learns to differentiate between genuine and deceptive reviews for Chicago hotels and then applies this knowledge to classify hotel reviews from other cities. The rationale is that if these phenomena are generalizable across hotel reviews in different cities, the classifier using phenomena-based features should have better performance than the one using unigram TF-IDF features.

To generate phenomena-based feature values for both training and testing sets, we need to obtain a score $f(r, p)$ for each review r and each phenomenon p. Here, $f(r, p)$ measures the extent to which a review r reflects a phenomenon p. We approach this problem by computing $P(r|p)$, i.e., the probability that a review r is generated given phenomenon p as the prompt. This can be a proxy for phenomena-based feature values.

We followed a generative scoring approach using next-token probabilities: $P(r|p) = \prod_{i=1}^{n} P(w_i|p, w_1, \cdots, w_{i-1})$, where p is a phenomenon,[2] and r is a sequence of words $[w1, \cdots, w_n]$. Each next-word probability $P(w_i|p, w_1, \cdots, w_{i-1})$ was obtained by prompting the LLM with the word sequence $[p, w_1, \cdots, w_{i-1}]$ and reading out the probability of w_i as predicted by the LLM. An open-source LLM that provides a complete probability distribution over *all* possible next words is needed, as w_i might rank at a very low position. We used Gemma-7b [29] for this experiment.

We used this approach to generate all nine phenomena-based feature values for reviews in $D_{Chicago}$ and $D_{three-cities}$. It is notable that this approach is one way (and not the only way) to generate such phenomena-based feature values. Even if these approximated feature values are not perfectly accurate, as long as they can improve a deception detection model's performance, it already serves our goal: to show that the phenomena-based features can improve a subsequent deception detector's out-of-distribution performance. The results in Table 3 confirm that this is indeed the case.

RQ3 Experiment: We investigated whether LLMs can effectively discover language phenomena from predictive lexical cues in RQ1. A natural follow-up question is whether these phenomena could also be obtained by prompting LLMs in other ways. Therefore, RQ3 examined whether LLMs can conjecture meaningful phenomena from their prior knowledge or in-context learning. We used

Haiku 4.5 given its best performance on deception detection (Table 2) and tested two prompt conditions to conjecture phenomena. In the first condition, we prompted it to conjecture phenomena solely based on its prior knowledge. In the second condition, we randomly sampled 30 genuine and 30 deceptive reviews from $D_{Chicago}$, and prompted it to conjecture phenomena from sampled reviews. We used the same system prompt structure and only changed the input from predictive words to sampled reviews or to nothing. To compare the quality of the phenomena conjectured under each prompt condition (i.e., predictive words, sampled examples, and prior knowledge), we used the same experimental design as in RQ1: we measured Haiku 4.5's deception detection performance on $D_{Chicago}$ with the conjectured phenomena inserted into the *phenomena-in-the-prompt* condition.

4 RESULTS

RQ1: Table 2 shows evaluation results of the four LLMs' predictive performance under different prompt conditions on $D_{Chicago}$. The evaluation results are with respect to the "deceptive" label. Across all four LLMs, the *phenomena-in-the-prompt* condition consistently outperformed the *zero-shot* condition. These improvements indicate that **the conjectured phenomena capture real patterns in the data rather than being entirely hallucinated**.

It is notable that unlike other tasks (e.g., fact-checking and question answering) where ground truth facts exist, there are *no definitive answers* for the conjectured phenomena in this task. Therefore, we define "hallucinations" as cases in which LLMs come up with phenomena that are entirely fabricated and not related to any detectable patterns in the data[3]. Under this definition, the goal of RQ1 was not to assess whether the LLM-conjectured phenomena represent the exhaustive set of language patterns. Rather, this validity check was to assess whether the LLM-conjectured phenomena are meaningfully associated with real-world examples of genuine or deceptive reviews that any user might see on social platforms. The performance gains LLMs achieved with phenomena provided in the prompt over the zero-shot prompt verified those conjectured phenomena to be relevant to the deception detection task.

While our goal was *not* to benchmark different LLMs' performance in deception detection or engineer the most effective prompt, these results did reveal two important trends. First, deception detection remained a challenging task for LLMs. Under the zero-shot prompt condition, the four LLMs achieved an average F1-score

---

[2]We used the following prompt: "Write a hotel review that {phenomenon}:", where {phenomenon} represents one of the nine conjectured phenomena in Table 1.

[3]For instance, a fabricated phenomena might be "Deceptive reviews always indicate positive sentiment to promote business"—this is contradicted by the $D_{Chicago}$ corpus.

Table 2: RQ1 Results. We evaluated four LLMs' predictive performance on $D_{Chicago}$ under different prompt conditions.

Model	Prompt	Acc.	Prec.	Recall	F1
GPT-5 mini	zero-shot	0.6512	0.8601	0.3612	0.5088
	phenomena	0.6775	0.8272	0.4488	0.5818
Haiku 4.5	zero-shot	0.6688	0.7679	0.4838	0.5936
	phenomena	0.6962	0.7532	0.5838	**0.6577**
Nova Pro	zero-shot	0.5512	0.6798	0.1938	0.3016
	phenomena	0.5962	0.7601	0.2812	0.4106
Gemini-2.5	zero-shot	0.6388	0.8101	0.3625	0.5009
Flash	phenomena	0.6600	0.8033	0.4238	0.5548

Acc.: Accuracy. Prec.: Precision.

of 0.4762 (min = 0.3016, max = 0.5936). Since the dataset is balanced, their performance was no better than random guess. Second, although all LLMs achieved reasonable performance gains with phenomena provided in the prompt, their performance was still worse than that of a logistic regression classifier using unigram TF-IDF features. This traditional model performed surprisingly well on the deception detection task with an F1-score of 0.88.

RQ2: Table 3 shows evaluation results of the three logistic regression classifiers' predictive performance. The classifiers used different feature sets and were trained on $D_{Chicago}$ and tested on $D_{three-cities}$. Compared to the classifier using unigram TF-IDF features only, both classifiers using phenomena-based features and combined features achieved better predictive performance on hotel reviews from other cities. These results suggest that **the LLM-conjectured phenomena captured patterns that are more transferable than unigram lexical cues**.

Table 3: RQ2 Results. We trained logistic regression classifiers on $D_{Chicago}$ and tested on $D_{three-cities}$ using three different feature sets: unigram only, phenomena only, and both.

Feature Sets	Acc.	Prec.	Recall	F1
unigram	0.7396	0.8363	0.5958	0.6959
phenomena	0.7167	0.6688	0.8583	**0.7518**
unigram+phenomena	0.7792	0.8602	0.6667	0.7512

This result is not surprising: while the word "Chicago" is the most salient cue of deception in reviews of Chicago hotels, it does not necessarily hold true in hotel reviews from other cities—fabricated review writers might mention hotel names with "New York", "Houston", and "Los Angeles" instead. In fact, regression coefficients of a logistic regression classifier trained on $D_{three-cities}$ corroborated that city names are among the strongest predictors of deceptive hotel reviews within those locales as well. Hence, compared to the unigram text feature "Chicago", the corresponding phenomena-based feature—deceptive reviews tend to name-drop hotel and city names—offers a more generalizable insight into deceptive hotel reviews across cities. Compared to lexical features, such language phenomena enable the classifier to learn transferable insights and achieve better predictive performance on out-of-distribution data. Future work could similarly examine whether hotel brand names are a strong signal of deception in other hotel review corpora.

RQ3: Table 4 shows the predictive performance of `Haiku 4.5` when predicting with phenomena conjectured from three different

sources: predictive words, sampled reviews, and prior knowledge. Its performance was highest when phenomena were derived from predictive words, and was noticeably lower when derived from either sampled reviews or prior knowledge alone. These results indicate that **`Haiku 4.5` discovered accurate phenomena more effectively when guided by predictive lexical cues than when relying on examples or prior knowledge**. One possible explanation is that predictive words identified by the logistic regression classifier serve as discriminative signals distilled from large amounts of training data. Although these words may appear unintuitive to humans, they anchor the LLM's reasoning toward subtle psycholinguistic patterns that are difficult to infer from a small sample of reviews or from its prior knowledge alone.

Table 4: RQ3 Results. We evaluated `Haiku 4.5`'s predictive performance on $D_{Chicago}$ under the phenomena-in-the-prompt condition with different conjectured phenomena sets.

Phenomena from	Acc.	Prec.	Recall	F1
sampled reviews	0.6269	0.6862	0.4675	0.5561
prior knowledge	0.5988	0.6113	0.5425	0.5748
predictive words	0.6962	0.7532	0.5838	**0.6577**

5 DISCUSSION AND CONCLUSION

Summary of Findings: In this paper, we studied the feasibility of using LLMs to translate machine-learned lexical cues into higher-level, human-understandable language phenomena in deception detection. We introduced a conjecture-then-validate pipeline: an LLM first *conjectures* the underlying phenomena associated with predictive words, and these phenomena are then *validated* algorithmically by testing whether they meaningfully reflect patterns in the data. Our experimental results show that these conjectured phenomena have predictive power for distinguishing genuine and deceptive reviews (RQ1), are generalizable to new, unseen data in the same domain (RQ2), and are most accurate when derived from predictive words rather than prior knowledge or in-context learning samples (RQ3). Together, these findings indicate that LLMs can reliably translate nuanced lexical cues into human-understandable language phenomena in deception detection—a capability that may extend to other text classification tasks where predictive words may appear unintuitive and represent underlying linguistic patterns.

Human Evaluation: In a separate work, we conducted a crowd-sourced user study to evaluate the effect of using LLM-conjectured language phenomena to explain classifier-learned predictive words when training humans to detect deceptive hotel reviews [24]. The study found that participants who saw the conjectured phenomena became *significantly better* at detecting deception without algorithmic assistance than participants who saw the predictive words only. This is empirical evidence that the LLM-conjectured phenomena are easier for humans to assimilate than the predictive words.

Limitations and Future Work: While algorithmic evaluation results verified that the LLM-conjectured phenomena are largely non-hallucinated and generalizable, two main limitations remain. First, the conjecturing step currently relies on a single-pass prompt. An iterative prompting process may lead to more detailed phenomena. Second, our validation was limited to algorithmic measures (and crowd workers in [24]). Future work could incorporate expert evaluations to assess the quality of LLM-conjectured phenomena.

REFERENCES

[1] About Amazon Team. 2023. *Amazon, Booking.com, Expedia Group, Glassdoor, Tripadvisor, and Trustpilot launch first global Coalition for Trusted Reviews*. About Amazon EU. https://www.aboutamazon.eu/news/policy/amazon-booking-com-expedia-group-glassdoor-tripadvisor-and-trustpilot-launch-first-global-coalition-for-trusted-reviews

[2] Inc. Amazon Web Services. 2025. *Amazon Nova Foundation Models*. https://aws.amazon.com/ai/generative-ai/nova/ Accessed: 2025-11-04.

[3] Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10, 3 (2006), 214–234.

[4] Vadim Borisov and Gjergji Kasneci. 2022. Relational Local Explanations. *arXiv preprint arXiv:2212.12374* (2022).

[5] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[7] Alon Jacovi, Hendrik Schuff, Heike Adel, Ngoc Thang Vu, and Yoav Goldberg. 2023. Neighboring Words Affect Human Interpretation of Saliency Explanations. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11816–11833.

[8] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research* 22, 104 (2021), 1–54.

[9] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714* (2023).

[10] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[11] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[12] Jiwei Li, Myle Ott, and Claire Cardie. 2013. Identifying manipulated offerings on review portals. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1933–1942.

[13] DeepMind (Google LLC). 2025. *Gemini 2.5 Flash*. https://deepmind.google/models/gemini/flash/ Accessed: 2025-11-04.

[14] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[15] David Martens, James Hinns, Camille Dams, Mark Vergouwen, and Theodoros Evgeniou. 2023. Tell me a story! narrative-driven xai with large language models. *arXiv preprint arXiv:2309.17057* (2023).

[16] OpenAI. 2025. *Introducing GPT-5*. https://openai.com/index/introducing-gpt-5/ Accessed: 2025-11-04.

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

[18] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557* (2011).

[19] Bo Pan, Zhen Xiong, Guanchen Wu, Zheng Zhang, Yifei Zhang, and Liang Zhao. 2024. TAGExplainer: Narrating Graph Explanations for Text-Attributed Graph Learning Models. *arXiv preprint arXiv:2410.15268* (2024).

[20] Himangshu Paul and Alexander Nikolaev. 2021. Fake review detection on online E-commerce platforms: a systematic literature review. *Data Mining and Knowledge Discovery* 35, 5 (2021), 1830–1881.

[21] Anthropic PBC. 2025. *Introducing Claude Haiku 4.5*. https://www.anthropic.com/news/claude-haiku-4-5 Accessed: 2025-11-04.

[22] Jiaming Qu, Jaime Arguello, and Yue Wang. 2021. A Study of Explainability Features to Scrutinize Faceted Filtering Results. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1498–1507.

[23] Jiaming Qu, Jaime Arguello, and Yue Wang. 2024. Why is" Problems" Predictive of Positive Sentiment? A Case Study of Explaining Unintuitive Features in Sentiment Classification. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 161–172.

[24] Jiaming Qu, Jaime Arguello, and Yue Wang. 2025. Understanding the Effects of Explaining Predictive but Unintuitive Features in Human-XAI Interaction. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 296–311.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[28] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4149–4158.

[29] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).

[30] Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems* 33 (2020), 6147–6159.

[31] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291.

[32] Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. 2024. LLMs for XAI: Future Directions for Explaining Explanations. *arXiv preprint arXiv:2405.06064* (2024).