Exploiting Prompt-induced Confidence for Black-Box Attacks on LLMs

Meina Chen, Yihong Tang, Kehai Chen*
Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
{chenmeina2002@gmail.com, neuqtoyhom@gmail.com, chenkehai@hit.edu.cn}

Abstract

Large language models (LLMs) are vulnerable to adversarial attacks even in strict black-box settings with only hard-label feedback. Existing attacks suffer from inefficient search due to lack of informative signals such as logits or probabilities. In this work, we propose **Prompt-Guided Ensemble Attack (PGEA)**, a novel black-box framework that leverages *prompt-induced confidence*, which reflects variations in a model's self-assessed certainty across different prompt templates, as an auxiliary signal to guide attacks. We first demonstrate that confidence estimates vary significantly with prompt phrasing despite unchanged predictions. We then integrate these confidence signals in a two-stage attack: (1) estimating token-level vulnerability via confidence elicitation, and (2) applying ensemble word-level substitutions guided by these estimates. Experiments on LLaMA-3-8B-Instruct and Mistral-7B-Instruct-v0.3 on three classification tasks show that PGEA improves the attack success rate and query efficiency while maintaining semantic fidelity. Our results highlight that verbalized confidence, even without access to probabilities, is a valuable and underexplored signal for black-box adversarial attacks. The code is available at https://github.com/cmn-bits/PGEA-main.

1 Introduction

Large language models (LLMs) have achieved remarkable success across diverse NLP tasks such as text generation (Li et al., 2024), classification (Kostina et al., 2025), and question answering (Tan et al., 2023; Wu et al., 2024). However, recent studies reveal that LLMs remain vulnerable to adversarial attacks, especially in black-box settings where only the final label is observable and internal states such as logits or probabilities are inaccessible (Xu et al., 2024; Roshan and Zafar, 2024).

Most existing black-box attacks rely on hard-label feedback and heuristic token substitutions, which often result in inefficient search and limited precision (Ma et al., 2024; Li et al., 2025). Meanwhile, LLMs can generate *self-assessed confidence* in the form of explicit expressions of certainty about their outputs when prompted accordingly (Dong et al., 2024; Chen et al., 2024). However, the consistency and practical value of the confidence elicited in adversarial settings remain largely unexplored.

In this paper, we explore whether prompt-induced confidence variation can serve as an informative auxiliary signal to guide black-box attacks. We find that slight changes in prompt wording can induce significant variation in confidence estimates without changing model predictions. Based on this insight, we propose the Prompt-Guided Ensemble Attack (PGEA), which consists of two components: (1) a prompt-guided confidence estimation module that identifies vulnerable tokens, and (2) a word-level substitution attack that integrates these signals to generate more targeted adversarial examples.

We evaluate the effectiveness of PGEA based on LLaMA-3-8B-Instruct and Mistral-7B-Instruct-v0.3 across three benchmarks (SST-2, AG-News, StrategyQA). Experimental results show that our method improves attack success rates, reduces query counts, and better preserves semantic coherence compared to strong hard-label baselines.

Our contributions are threefold:

- We conduct the first systematic study of prompt-induced confidence variability in LLMs and its utility for adversarial guidance.
- We propose a novel black-box attack framework that integrates elicited confidence into ensemble word-level substitutions.
- We demonstrate that prompt-guided confidence reliably identifies token vulnerability, enabling more efficient and transferable attacks.

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12897–12903
November 4-9, 2025 ©2025 Association for Computational Linguistics

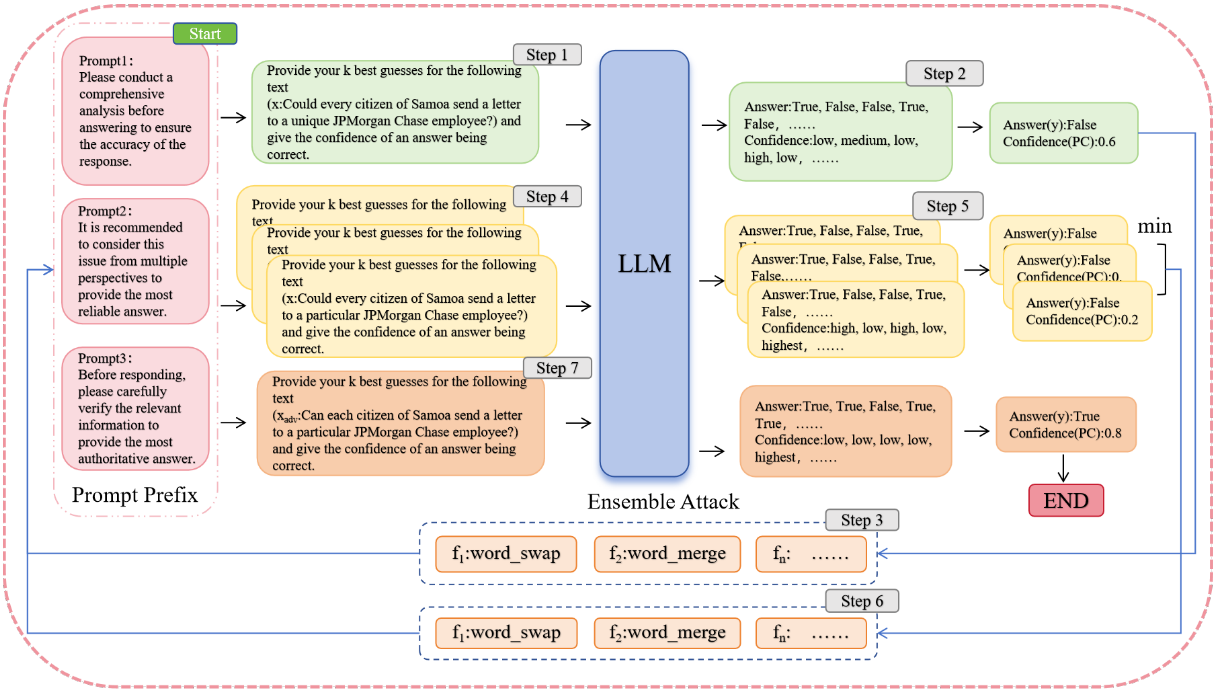

Figure 1: Overview of the Prompt-Guided Ensemble Attack framework. The process starts with multiple prompt prefixes to elicit confidence estimates from the LLM (Steps 1, 4, 7). These confidence scores guide an ensemble of adversarial perturbation methods (word-swap, word-merge, etc., Steps 3 and 6) to generate candidate adversarial examples. The model's self-assessed confidence on these candidates is then used to select effective perturbations, iteratively improving the attack success until termination conditions are met.

2 Preliminaries

Despite extensive pre-training (Brown et al., 2020; Touvron et al., 2023) and alignment efforts (Ouyang et al., 2022; Deng et al., 2025) aimed at bringing LLMs closer to human-level performance, recent work has shown that they still exhibit vulnerabilities against adversarial attacks. We briefly review relevant concepts on adversarial attacks in NLP, token-level perturbations, and confidence elicitation in large language models.

Adversarial Attacks in NLP. Adversarial attacks aim to minimally alter inputs while inducing incorrect predictions (Goodfellow et al., 2014). While continuous-space attacks such as FGSM (Goodfellow et al., 2014), BIM (Kurakin et al., 2016), PGD (Madry et al., 2017) are standard in vision, text inputs are discrete, making gradient-based optimization non-trivial.

Token-Level Perturbations. To craft natural adversarial texts, NLP attacks often rely on synonym substitution, paraphrasing, or character-level edits (Zhang et al., 2020; Ren et al., 2019; Gao et al., 2018; Pruthi et al., 2019). These modifications must preserve meaning and fluency, posing unique challenges, especially under black-box constraints where gradients and confidence scores are inaccessible.

Black-Box Constraints. Black-box attacks only access output labels, requiring repeated queries and heuristic search to discover effective perturbations (Shrotri et al., 2022; Yu et al., 2024). When targeting LLMs via API, query budgets and limited response formats further complicate the process.

Confidence Elicitation. Recent studies show that LLMs can verbally express confidence when prompted (Xiong et al., 2023; Tian et al., 2023; Formento et al., 2025). We refer to this as confidence elicitation. Even in the absence of logits, such verbal signals offer a soft proxy for the uncertainty of the model. In this work, we explore how to leverage elicited confidence to guide black-box adversarial perturbations more effectively.

3 Proposed Method

We introduce PGEA, a black-box adversarial framework that leverages prompt-induced confidence estimation combined with multiple perturbation strategies to improve attack effectiveness and transferability on LLMs.

3.1 Prompt-Guided Confidence Estimation

Prompt engineering can affect not only LLM outputs but also intermediate behaviors like style and reasoning. We investigate whether prompt phrasing influences the model's self-assessed confidence, even when the predicted label remains unchanged.

Our experiments reveal that subtle prompt variations, such as changing "How confident are you?" to "Please assess your certainty on a scale from 0 to 100," can significantly impact the model's confidence scores. More cautious prompts generally lead to better-calibrated confidence estimates, suggesting that prompt design plays a critical role in modulating model self-evaluation.

Leveraging this, we employ prompt-guided confidence elicitation to pinpoint tokens with lower confidence, which serve as effective proxies for vulnerability and help guide where perturbations are most likely to succeed.

3.2 Ensemble-Based Black-Box Attack

We integrate prompt-guided confidence with standard word-level substitution attacks to form a robust ensemble approach.

Given an input x, we generate a set of perturbed samples $\{x_1, x_2, \ldots, x_k\}$ using diverse adversarial strategies. For each perturbed input, a confidence-querying prompt P elicits self-assessed confidence scores $\{c_1(x_i), c_2(x_i), \ldots, c_k(x_i)\}$, which are aggregated into token-level uncertainty measures. Tokens are then ranked by vulnerability, guiding targeted replacements via black-box methods such as BERT-Attack (Li et al., 2020b) or TextFooler (Jin et al., 2020).

To enhance transferability, we ensemble multiple perturbation methods, each generating candidates from different perspectives (e.g., embedding, syntactic, semantic). We select adversarial examples that (a) flip the model prediction, and (b) correspond to regions with lowest confidence per prompt-guided estimation.

Formally, we optimize:

$$\delta^* = \arg\min_{\delta \in \mathcal{B}(x)} \sum_{k=1}^{K} c_k(x + \delta) \tag{1}$$

where $\mathcal{B}(x)$ is the set of candidate adversarial variants and $c_k(\cdot)$ the confidence under prompt P.

This ensemble strategy produces adversarial examples that are (1) semantically coherent, (2) guided by model uncertainty, and (3) transferable across model variants and prompts.

4 Experiments

4.1 Experimental Setup

We evaluate our prompt-based black-box integrated attack on the Meta-Llama-3-8B-Instruct model (Touvron et al., 2023) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) using three benchmark datasets: SST-2 (Socher et al., 2013), AG-News (Zhang et al., 2015), and StrategyQA (Geva et al., 2021). To improve the quality of confidence estimation, we design three diverse prompt prefixes to elicit self-assessed confidence from the models. Our attack combines two complementary perturbation strategies: word-swap and word-merge, to increase the likelihood of successful adversarial examples. Attack success is measured by the reduction in classification accuracy before and after the attack. We also compare our approach against the guided word substitution attack (CEAttack) to validate effectiveness.

4.2 Implementation Details

Our method follows a two-step prompting scheme. First, the model generates k guesses per input text, each with an associated confidence level. We set $k = 20$ for SST-2 and AG-News, and $k = 6$ for StrategyQA. Confidence levels are categorized into five discrete scores: *Highest*, *High*, *Medium*, *Low*, and *Lowest*, mapped respectively to scores from 5 to 1. The final confidence score is derived via distribution aggregation.

Model configurations follow prior work (Formento et al., 2025), employing a Dirichlet distribution to model confidence thresholds through parameters k_{pred} and k_{conf}. To reduce generation randomness and stabilize output consistency, we apply a temperature setting $\tau = 0$, consistent with previous adversarial attack studies.

Our integrated attack leverages two strategies:

- **Word-swap attack**: Inspired by the counterfitting approach (Formento et al., 2025), we perform greedy search to identify optimal word replacement positions, considering up to 10 candidate substitutions per position.

- **Word-merge attack**: We adopt the adversarial text generation technique from (Li et al., 2020a), which uses a mask-and-fill process to generate fluent and grammatically coherent adversarial samples, also limited to 10 replacement candidates per position.

Prompt prefix	Dataset	LLaMa-3-8B-Instruct ECE ↓	AUROC ↑	AUPRC Pos ↑	AUPRC Neg ↑	Mistral-7B-Instruct-v0.3 ECE ↓	AUROC ↑	AUPRC Pos ↑	AUPRC Neg ↑
Without prefix	SST2	0.1264	0.9696	0.9730	0.9678	0.1542	0.9537	0.9616	0.9343
	AG-News	0.1376	0.9293	-	-	0.1216	0.8826	-	-
	StrategyQA	0.0492	0.6607	0.6212	0.6863	0.1295	0.6358	0.6421	0.6185
Prefix prompt1	SST2	0.1212	0.9663	0.9690	0.9635	0.1456	0.9399	0.9499	0.9112
	AG-News	0.1293	0.9315	-	-	0.1404	0.8664	-	-
	StrategyQA	0.0464	0.6838	0.6404	0.7288	0.1016	0.6046	0.5984	0.6272
Prefix prompt2	SST2	0.0915	0.9646	0.9690	0.9632	0.1669	0.9493	0.9531	0.9415
	AG-News	0.1169	0.9388	-	-	0.1392	0.8653	-	-
	StrategyQA	0.0488	0.6562	0.6154	0.6950	0.1085	0.6124	0.6243	0.5876
Prefix prompt3	SST2	0.0727	0.9709	0.9732	0.9691	0.1670	0.9284	0.9468	0.9050
	AG-News	0.1274	0.8826	-	-	0.1115	0.8900	-	-
	StrategyQA	0.0389	0.6907	0.6646	0.7278	0.1807	0.5361	0.5123	0.5579

Table 1: Calibration results (ECE, AUROC, AUPRC) on SST2, AG-News and StrategyQA.

Model	Dataset	CA [%] ↑ Vanilla	AUA [%] ↓ Self-Fool	Text Hoaxer	SSP	CE	PGEA	ASR [%] ↑ Self-Fool	Text Hoaxer	SSP	CE	PGEA
LLaMa-3 8B-Instruct	SST2	90.56±0.14	88.35	82.93	81.93	72.69	**59.44**	2.22	8.43	9.73	19.73	**33.93**
	AG-News	61.62±0.38	61.17	49.30	45.27	43.06	**39.44**	0.33	19.41	26.71	30.74	**35.95**
	StrategyQA	60.22±0.17	59.52	45.29	42.28	32.67	**13.83**	1.66	24.67	29.67	45.67	**77.00**
Mistral-7B Instruct-v0.3	SST2	87.87±0.39	84.73	74.27	75.31	71.76	**61.51**	3.57	16.08	14.08	17.94	**30.17**
	AG-News	65.99±0.27	-	48.69	52.48	40.82	**30.32**	-	26.43	20.00	38.33	**54.19**
	StrategyQA	59.92±0.32	59.61	44.33	41.13	36.21	**14.29**	1.22	26.23	30.99	39.26	**75.73**

Table 2: Results of Prompt-Guided Ensemble Attack. **CA**: clean accuracy, **AUA**: accuracy under attack, **ASR**: attack success rate. Bold numbers indicate best results.

Prompt Prefixes	
Prompt1	Please conduct a comprehensive analysis before answering to ensure the accuracy of the response.
Prompt2	It is recommended to consider this issue from multiple perspectives to provide the most reliable answer.
Prompt3	Before responding, please carefully verify the relevant information to provide the most authoritative answer.

Table 3: Three different prompt prefixes

During the ensemble attack, both word-swap and word-merge methods are applied iteratively to each input, generating a diverse pool of adversarial candidates. This diversity enables more effective selection of perturbations that reduce classification confidence, thereby improving attack success.

4.3 Experiment Result

Prompt-guided Confidence Estimation Table 1 reports the calibration performance of verbal confidence elicitation across three datasets: SST-2, AG-News, and StrategyQA. We evaluate the expected calibration error (ECE) (Guo et al., 2017), the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPRC) for both positive and negative classes. Results demonstrate that our prompt designs consistently improve calibration metrics compared to the prompt in the method of CEAttack, with `prompt3` achieving the lowest ECE and highest AUROC on SST-2 and StrategyQA for LLaMa-3-8B-Instruct. Due to dataset-specific label availability, some metrics are not reported (marked as "-"). The three prompt prefixes we added are shown in Table 3.

Ensemble Attack Results To comprehensively assess the effectiveness of our proposed Ensemble Attack strategy, we evaluate it across three representative datasets—SST2, AG-News, and StrategyQA—using LLaMA-3-8B-Instruct and Mistral-7B-Instruct-v0.3 as the base models. We compare ensemble performance against four representative baselines: Self-Fool (SF; Xu et al., 2024), Text Hoaxer (TH; Ye et al., 2022), SSP Attack (Liu et al., 2023), and CE Attack (Formento et al., 2025). The results are summarized in Tables 2.

Effectiveness Table 2 shows the classification accuracy (CA), area under the confidence curve (AUA), and attack success rate (ASR). Across all datasets, the Ensemble Attack achieves the lowest AUA and highest ASR, indicating its superior ability to elicit overconfident predictions from the model. Notably, for the StrategyQA dataset on LLaMA-3-8B-Instruct, the ensemble yields

Model	Dataset	SemSim ↑					Original Perplexity ↓					After-Attack Perplexity ↓				
		SF	TH	SSP	CE	PGEA	SF	TH	SSP	CE	PGEA	SF	TH	SSP	CE	PGEA
LLaMa-3 8B-Instruct	SST2	0.87	0.89	0.87	0.88	**0.88**	73.75	76.51	69.04	69.81	**58.95**	82.95	113.0	143.81	111.16	**108.71**
	AG-News	0.86	0.94	0.88	0.93	**0.92**	354.12	78.62	66.31	72.01	**71.76**	320.06	99.02	193.16	98.90	**94.12**
	StrategyQA	0.87	0.89	0.89	0.89	**0.88**	281.38	104.83	115.63	105.42	**99.52**	220.73	182.15	232.31	206.23	**198.88**
Mistral-7B Instruct-v0.3	SST2	0.87	0.90	0.87	0.88	**0.88**	79.06	63.03	63.44	61.68	**59.63**	91.85	85.27	118.67	95.85	**108.96**
	AG-News	-	0.94	0.88	0.93	**0.92**	-	86.47	74.76	73.20	**70.30**	-	103.25	188.83	97.19	**90.04**
	StrategyQA	0.89	0.90	0.89	0.90	**0.88**	74.04	85.20	95.43	97.30	**92.93**	93.57	140.08	195.33	177.94	**193.85**

Table 4: Quality of attack inputs. Only successful perturbations are considered.

a significant AUA reduction from 32.67 (CE Attack) to 13.83 and boosts ASR from 45.67% to 77.00%, demonstrating its ability to amplify attack transferability. CA degradation is also most severe under the ensemble, confirming its stronger disruptive effect on model calibration.

To assess the generalizability of PGEA, we further evaluated our attack method on additional models, specifically Gemma2-9B-Instruct (Team et al., 2024) and Qwen2.5-7B-Instruct (Team, 2024). The experimental setup remained consistent with that described in the previous section, and the results are presented in the Table 5.

Model	Dataset	AUA [%]	ASR [%]
Gemma2-9B-Instruct	SST2	50.21	42.45
	AG-News	34.78	43.86
	StrategyQA	57.82	12.37
Qwen2.5-7B-Instruct	SST2	56.29	37.53
	AG-News	66.94	15.28
	StrategyQA	43.17	23.21

Table 5: Results from additional experiments on Gemma2-9B-Instruct and Qwen2.5-7B-Instruct.

Quality Preservation As shown in Table 4, semantic similarity (SemSim) between original and perturbed inputs remains consistently high (e.g., >0.88 on SST2), suggesting that ensemble-based perturbations maintain linguistic plausibility. Meanwhile, both the original and post-attack perplexities remain within acceptable bounds, with ensemble perturbations typically inducing only a moderate increase in perplexity (e.g., 94.12 and 90.04 on AG-News), confirming the fluency of adversarial examples.

Our ensemble strategy exhibits the strongest attack performance across all metrics. These findings indicate that confidence vulnerabilities are multifaceted in nature, and utilizing a combination of complementary perturbation strategies represents an effective approach to exploiting such vulnerabilities.

5 Conclusion

We propose Prompt-Guided Ensemble Attack (PGEA), a black-box adversarial method that leverages prompt-induced confidence signals to guide ensemble perturbations. By combining verbal uncertainty elicitation with word-level attacks, PGEA effectively identifies vulnerable tokens and enhances attack transferability. Experiments on three datasets show consistent improvements over baseline methods. Our findings also reveal that prompt phrasing plays a crucial role in shaping model confidence, highlighting new opportunities for prompt-based control and evaluation of large language models. Future work includes extending this framework to multilingual models and exploring adaptive prompting strategies for real-time adversarial defense.

Limitations

While PGEA shows promising attack performance, it has several limitations. The effectiveness of confidence elicitation depends on carefully designed prompts, which may require domain-specific tuning and could be sensitive to phrasing variations beyond the tested templates. The ensemble process introduces computational overhead due to multiple forward passes through substitute models, although this remains manageable compared to full gradient-based white-box attacks. Currently, the method focuses on English text perturbations, leaving cross-lingual vulnerability exploration for future work. Additionally, the attack success rate is limited by differences between substitute and target model architectures, especially when targeting commercially hardened APIs with unknown defense mechanisms. These limitations emphasize the challenges of balancing the power of attack and computational efficiency for real-world deployment.

Acknowledgments

We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions. This work was supported in part by the National Natural Science Foundation of China (62276077, U23B2055, 62350710797), in part by the Shenzhen Science and Technology Program (ZDSYS20230626091203008, KQTD2024072910 2154066), and in part by the Shenzhen College Stability Support Plan (GXWD20220817123150002, GXWD20220811170358002).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. DUAL-REFLECT: enhancing large language models for reflective translation through dual learning feedback mechanisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 693–704. Association for Computational Linguistics.

Qiyuan Deng, Xuefeng Bai, Kehai Chen, Yaowei Wang, Liqiang Nie, and Min Zhang. 2025. Efficient safety alignment of large language models via preference re-ranking and representation-based reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31156–31171, Vienna, Austria.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*.

Brian Formento, Chuan-Sheng Foo, and See-Kiong Ng. 2025. Confidence elicitation: A new attack vector for large language models. In *The Thirteenth International Conference on Learning Representations*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. *arXiv preprint arXiv:1801.04354*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8018–8025.

Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020a. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56:1–39.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

Zelin Li, Kehai Chen, Lemao Liu, Xuefeng Bai, Mingming Yang, Yang Xiang, and Min Zhang. 2025. Tf-attack: Transferable and fast adversarial attacks on large language models. *Knowledge Based Systems*, 312:113117.

Han Liu, Zhi Xu, Xiaotong Zhang, Xiaoming Xu, Feng Zhang, Fenglong Ma, Hongyang Chen, Hong Yu, and Xianchao Zhang. 2023. Sspattack: A simple and sweet paradigm for black-box hard-label textual adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13228–13235.

Yiqing Ma, Kyle Lucke, Min Xian, and Aleksandar Vakanski. 2024. Semantic-aware adaptive binary search for hard-label black-box attack. *Computers*, 13:203.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Khushnaseeb Roshan and Aasim Zafar. 2024. Black-box adversarial transferability: An empirical study in cybersecurity perspective. *Computers & Security*, 141:103853.

Aditya A Shrotri, Nina Narodytska, Alexey Ignatiev, Kuldeep S Meel, Joao Marques-Silva, and Moshe Y Vardi. 2022. Constraint-driven explanations for black-box ml models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8304–8314.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2024. An llm can fool itself: A prompt-based adversarial attack. In *12th International Conference on Learning Representations, ICLR 2024*.

Muchao Ye, Chenglin Miao, Ting Wang, and Fenglong Ma. 2022. Texthoaxer: Budgeted hard-label adversarial attacks on text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3877–3884.

Zhen Yu, Zhenhua Chen, and Kun He. 2024. Query-efficient textual adversarial example generation for black-box attacks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 556–569.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2020. Generating fluent adversarial examples for natural languages. *arXiv preprint arXiv:2007.06174*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems-Volume 1*, pages 649–657.