Robustness Evaluation in Natural Language Understanding: A Survey and Perspective in the Era of Large Language Models

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) increasingly serve as the backbone of modern Question Answering (QA) systems, ensuring their robustness to input variation has become a critical concern. In this paper, we survey the trajectory of robustness evaluation for QA, with a particular focus on perturbation-based methods applied to textual input. We first review synthetic perturbation approaches developed for earlier neural models and discuss their continued relevance and adaptation to recent LLMs. We then examine natural perturbations, which originate from real-world language variation and provide a more realistic basis for evaluating robustness in practical scenarios. Based on our analysis, we identify key limitations in current robustness research and advocate for a shift toward evaluation methodologies that emphasize natural linguistic variability. We also outline future research directions, including the need for systematic evaluation protocols, a deeper understanding of robustness in the context of LLM-based QA, and explicit consideration of benchmark leakage when evaluating the robustness of LLMs.

1 Introduction

Large Language Models (LLMs) have achieved remarkable progress in natural language understanding, as evidenced by their strong performance on a wide array of academic benchmarks in Natural Language Processing (NLP) (Bang et al., 2023; Team et al., 2025; Yang et al., 2025) and their widespread adoption in real-world applications, such as conversational agents including ChatGPT (OpenAI et al., 2024a) and DeepSeek (DeepSeek-AI et al., 2025). As these models are increasingly deployed across both routine information-seeking contexts and critical domains such as healthcare, law, and finance, concerns regarding their robustness and generalization capabilities have become increasingly prominent (Wang et al., 2024; Zhu et al., 2024; Nalbandyan et al., 2025). Even subtle variations in input can lead to erroneous outputs, which may have significant consequences, particularly in high-stakes environments. Accordingly, evaluating and improving the robustness of LLMs has become a pressing challenge for the NLP community (Zhang et al., 2025).

Among the many applications of LLMs, Question Answering (QA) over unstructured textual information stands out as both a practical task and a diagnostic benchmark for evaluating authentic language understanding. Although recent studies suggest that modern LLMs exhibit improved robustness compared to earlier neural QA systems (e.g., DeBERTaV3 (He et al., 2023)) (Fang et al., 2023), there is growing concern that, despite achieving seemingly impressive results on simplistic evaluations involving static test instances, the performance of LLMs deteriorates in a rapidly evolving world where diverse textual variations are common (Wu et al., 2025). This limitation becomes even more apparent when contrasted with the remarkably robust language processing systems observed in humans, for whom variations in text rarely impede comprehension (Reicher, 1969; Rawlinson, 1976; McCusker et al., 1981; Mayall et al., 1997)—a property that reflects linguistic competence (CANALE and SWAIN, 1980) as opposed to mere performance (Chomsky, 1969, 2006).

Figure 1 illustrates such brittleness, where altering the syntactic structure of the question can lead state-of-the-art (SOTA) reasoning LLMs–OpenAI's `o3` to generate an incorrect response, even when the underlying semantic meaning remains unchanged—an error pattern rarely observed in human comprehension (*CASE 1: Question paraphrasing*). In *CASE 2: Knowledge update*, appending a contextual update stating that the Summer Olympic Games have been held four times in East Asia, and modifying the question to inquire "As of 2025" makes it impossible to infer the referent

Instruction: Read the paragraphs and answer the question.

Paragraph A: 2008 Summer Olympics *"The 2008 Summer Olympic Games, officially known as [...] It was the third time that the Summer Olympic Games were held in East Asia and Asia, after Tokyo, Japan, in 1964 and Seoul, South Korea, in 1988. This was later followed by the 2020 Summer Olympics, also held in Tokyo, which were postponed due to the COVID-19 pandemic and ultimately took place from 23 July to 8 August 2021, marking the fourth time the Games were hosted in East Asia."*

Paragraph B: Summer Olympic Games *"The Summer Olympic Games (French: "Jeux olympiques d'été") or the Games of the Olympiad, first held in 1896, is an international multi-sport event that is hosted by a different city every four years. [...] The Winter Olympic Games were created due to the success of the Summer Olympics."*

Question: *"As of 2008, when did the game which was held only three times in East Asia first held?"*

Original Prediction: It refers to the Summer Olympic Games, which were first held in 1896 (in Athens, Greece).

CASE 1: Question paraphrasing: *"As of 2008, when was the first time the game—held only three times in East Asia—took place?"* It first took place in 1964, when Tokyo, Japan hosted the Summer Olympic Games. ✗

Prediction by a human: 1896

CASE 2: Knowledge update: *"As of 2025, when did the game which was held only three times in East Asia first held?"* The Summer Olympic Games were first held in 1896. ✗

Prediction by a human: Unanswerable

Figure 1: Illustrative cases of OpenAI's o3 failures on test instances reflecting real-world textual variation. The original QA example is taken from the HotpotQA dataset (Yang et al., 2018). We slightly modify the original question (*When did the game which held three times in in East Asia first held*) to make it grammatically correct and eliminate ambiguity.

of "the game" with certainty. While a human annotator appropriately recognise that the question becomes unanswerable given the revised paragraph and the question, o3 continues to provide an answer 1896, thereby failing to account for the contextual shift. These cases demonstrate the brittleness of LLMs in handling linguistic variations in real-world scenarios and challenge claims regarding their human-level reading comprehension and language understanding (Shojaee*† et al., 2025; Rajeev et al., 2025).

In this paper, we review the trajectory of robustness evaluation in QA, from its application to early neural language models to more recent developments involving LLMs. We begin by examining synthetic perturbation approaches developed for earlier models and then discuss how these methods have been adapted and applied to LLMs. Then, we turn to a relatively underexplored yet critical category—natural perturbations—which refer to variations arising from real-world language evolution, and provide a survey of related work. Natural perturbations are increasingly viewed as more representative of real-world robustness challenges, as they reflect the types of linguistic variability that LLMs are likely to encounter in practical use (Wu et al., 2025). In contrast, synthetic approaches rely on predefined manipulation strategies, which may not fully capture the complexity and diversity of natural language variation. Finally, we identify key limitations in existing robustness research and propose promising directions for future work. The structure of our survey paper is shown in Figure 2.

To summarize, our contributions are threefold:

1. We survey existing work on applying synthetic perturbation approaches to LLMs as a complement to the survey presented in (Ho et al., 2023), and demonstrate that techniques originally developed for earlier QA systems remain relevant and effective in the context of LLMs.

2. We provide a comprehensive review of studies that employ natural perturbations for robustness evaluation, organized by the source of the perturbation, and argue for a critical shift in focus from synthetic perturbations to real-world, naturally occurring variations.

3. We identify key limitations in current QA

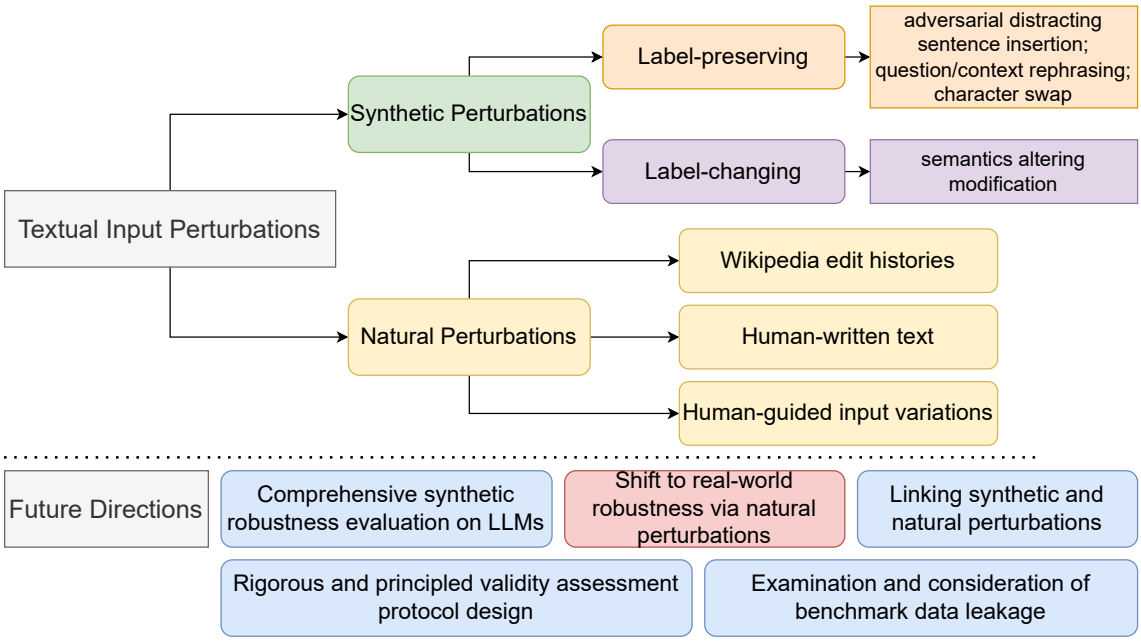

Figure 2: Overview of synthetic and natural perturbation methods for robustness evaluation surveyed in this paper, along with proposed future directions.

robustness evaluation research and propose promising directions for future work, particularly in light of the evolving capabilities and deployment scenarios of LLMs.

2 Textual Input Perturbations

An effective way to assess model robustness is through input perturbations, whereby textual inputs are deliberately modified to observe how model predictions change under such conditions. In this paper, we broadly categorise existing robustness evaluation work into two types—**synthetic** and **natural**—based on fundamental differences in the methodologies used to construct the perturbations.

2.1 Synthetic Perturbations

A significant body of work on NLP robustness assessment relies on synthetic perturbations—deliberate modifications based on predefined input transformation strategies. These methods assume that the gold label is either preserved or altered under bounded perturbations. Accordingly, synthetic perturbation techniques can be broadly classified into two categories: label-preserving and label-changing. In the following, we briefly trace the trajectory of synthetic perturbation-based robustness evaluation work in a representative QA task. For a more detailed and comprehensive survey on robustness evaluation in QA and broader coverage across other NLP tasks, we refer readers to (Ho et al., 2023) and (Wang et al., 2022b; Schlegel et al., 2023), respectively.

Label-preserving The majority of existing work adopts the label-preserving [1] assumption, employing synthetic textual perturbations like the insertion of adversarial distracting sentence (Jia and Liang, 2017; Wang and Bansal, 2018; Chen et al., 2022a; Tran et al., 2023), the rephrasing of the question (Gan and Ng, 2019) or the reading paragraph (Wu et al., 2021, 2023), the addition of misinformation (Pan et al., 2023) and character-level manipulations such as character swaps (Si et al., 2021).

Label-changing Another line of work introduces small but meaningful input perturbations that intentionally alter the gold label, with the expectation that the model should adapt its prediction to reflect the change (Gardner et al., 2020; Schlegel et al., 2021; Geva et al., 2022).

Synthetic perturbations introduced in earlier work have primarily been applied to pre-LLM neural QA models. A consistent finding is that, despite achieving strong, human-comparable performance on held-out test sets (Devlin et al., 2019; Liu et al., 2019; He et al., 2021), these models exhibit varying degrees of performance degradation under

[1] Note that label-preserving does not necessarily imply semantics-preserving.

synthetic perturbation settings—highlighting their reliance on statistical shortcuts to bypass genuine task requirements and exposing a lack of robustness (Ho et al., 2023).

More recently, SOTA LLMs have demonstrated superior natural language understanding across a wide range of NLP tasks—including QA (OpenAI et al., 2024b; OLMo et al., 2025; Yang et al., 2025)—and outperform their fine-tuned pre-trained language model predecessors in terms of robustness (Fang et al., 2023). Nevertheless, synthetic perturbation approaches developed for earlier QA systems remain relevant, as LLMs have also been shown to be vulnerable to such perturbations. For instance, Levy et al. (2023) adapted the adversarial distracting sentence injection technique originally proposed in (Jia and Liang, 2017), prompting a strong GPT-4 model to generate a distractor sentence that answers the question similar to the original but with one critical detail changed—referred to as the "almost detail". The instruction encouraged the model to reuse much of the original question's phrasing while omitting the actual answer. This perturbation strategy was later found to successfully mislead less proficient GPT-TURBO, GPT3.5, and even GPT-4 itself. Fang et al. (2023) empirically investigated the effects of diverse synthetic perturbations (e.g., neighboring character swaps, synonym replacements, and combinations of multiple attack methodologies) on other LLMs such as LLAMA (Touvron et al., 2023a), and observed similar patterns of robustness failure. Besides, recent work by Bhuiya et al. (2024) revealed that, despite not requiring downstream task-specific fine-tuning like earlier QA systems, leading LLMs including GPT, LLAMA 2 (Touvron et al., 2023b) and MIXTRAL 8X7B (Jiang et al., 2024) still tend to exploit simplifying cues to circumvent the requirement to perform multi-hop reasoning. This is evidenced by their poor generalisability under a controlled challenge setting, in which distractor paragraphs were introduced to present seemingly plausible yet incorrect alternative reasoning paths, while ensuring that the correct final answer remained unchanged.

2.2 Natural Perturbations

Unlike synthetic perturbations, which typically rely on hypothesised manipulation strategies (Le et al., 2022), natural perturbations originate from authentic variations observed in real-world scenarios and are therefore considered more relevant for evaluating real-world robustness (Wu et al., 2025). However, this aspect has long been significantly neglected, and natural perturbations remain comparatively underexplored relative to their synthetic counterparts. Table 1 summarises a non-exhaustive body of literature on natural perturbation methods applied across diverse NLP tasks. In the following, we categorise these works by the sources of natural perturbations.

Wikipedia edit histories Wikipedia's revision histories provide a rich source of human-authored textual changes over time, offering a valuable corpus for studying real-world text variations. As one of the earlier efforts, Belinkov and Bisk (2018) explored robustness in Neural Machine Translation (NMT) by applying single-word perturbations to non-English source-side sentences. They built a lookup table of lexical errors, such as typos and misspellings, extracted from French (Max and Wisniewski, 2010) and German (Zesch, 2012) Wikipedia edit histories. Words in the source sentences were then replaced with corresponding errors from the table, where applicable. Eger and Benz (2020) extended the same approach to POS tagging, Natural Language Inference (NLI), and Toxic Comment Classification (TC), leveraging revision histories from English Wikipedia. Building on this line of research, natural perturbations were further generalised to various Question Answering (QA) tasks (Wu et al., 2025). Instead of perturbing individual tokens, they substituted entire reading paragraphs with their edited counterparts retrieved from English Wikipedia revision histories, enabling evaluation under naturally occurring, context-level perturbations. The study assessed the sensitivity of neural language models ranging from early BERT-based architectures to SOTA LLMs, revealing that robustness issues persist across model generations.

Human-written text Textual content authored by human writers may contain a diverse range of errors and thus serve as a potential source of natural perturbations, as exemplified by essays written by non-native Czech speakers (Šebesta et al., 2017; Belinkov and Bisk, 2018) and by more than 18 million sentences produced by internet users across nine real-life datasets (Le et al., 2022).

Human-guided input variations Some studies involve recruiting human annotators to manually craft or verify input variations. We categorise such variations as natural when annotators are not informed that their modifications will be used to

Task/Reference	Natural Perturbation Method	Level	Validity	Defense Strategies
NMT (Belinkov and Bisk, 2018)	Replace each word in the source-side sentence with available edits mined from French and German *Wikipedia edit histories* and *human-written essays* in Czech	Word	Unclear	Average character embedding; Training on perturbed data
POS tagging, NLI, TC (Eger and Benz, 2020)	Replace words with natural human errors from the *Wikipedia edit history*	Word	Unclear	Training on perturbed data
QA (Wu et al., 2025)	Replace the reading passage with its counterparts based on available English *Wikipedia edit histories*	Context	Human	Adversarial training with perturbed data; In-context demonstrations
Toxic Comments/Hate Speech/Online Cyberbullying Texts Detection (Le et al., 2022)	Retrieve and substitute words using perturbations extracted from *a large corpus of over 18M sentences written by netizens*, based on phonetic similarity and edit distance	Word	Human	Sound-Invariant CNN; Adversarial training with perturbed data
Over 80 unique tasks from MMLU and BIG-BENCH LITE (Sun et al., 2024)	Recruit *36 NLP graduate students* to compose novel instructions that are appropriate for a given task but superficially different from those seen during instruction fine-tuning	Instruction	Unclear	Aligning representations of equivalent instructions
LLMs Code Generation (Chen et al., 2025)	Apply perturbations from 21 specific categories to the original natural language prompt. These categories, which may occur in real-world scenarios, were suggested by *experienced practitioners from the open-source community, industry, and academia* through the online survey	Mix	Human	–

Table 1: A non-exhaustive summary of existing literature on natural perturbations. For each work, we list the studied task, method for generating naturally perturbed test data (with *italicized underlined text* indicating the corpus source), and the perturbation level. We also indicate whether the work verifies the validity of the adversarial examples and whether it proposes defense strategies. "Unclear" denotes that no systematic experiments were conducted, though some works discuss or qualitatively assess validity.

fool the model. In this setting, edits are guided by the annotators' own judgments of necessity and tend to reflect real-world scenarios, rather than being made with the explicit goal of inducing model failure (Wallace et al., 2019; Bartolo et al., 2020)—thus preserving their naturalness. In (Sun et al., 2024), researchers examined the robustness of instruction-tuned models to instruction rephrasing across more than 80 NLP tasks drawn from MMLU (Hendrycks et al., 2021) and BIG-BENCH LITE (Srivastava et al., 2023). A total of 36 NLP graduate researchers were recruited to write novel instructions they believed would best elicit the desired behavior for each task. These newly

crafted (unobserved) instruction phrasings, while differing superficially from those seen during instruction fine-tuning, were shown to consistently degrade model performance—highlighting limitations in the models' generalisability. Rather than having human annotators directly propose perturbations, Chen et al. (2025) introduced 21 types of real-world scenario variations targeting natural language descriptions in LLM-based code generation tasks, derived from survey responses collected from professionals in industry and research institutions with experience using LLMs for code generation. Similarly, interview responses from employees at 16 British, German, and American NGOs, whose work directly involves online hate, were used to design 29 real-world functional tests aimed at revealing specific weaknesses in hate speech detection models (Röttger et al., 2021).

Note that the term "natural" is overloaded in NLP literature, where it can also refer either to the extent to which synthetically modified text preserves linguistic characteristics such as fluency, coherence, grammaticality, and clarity, i.e., its naturalness (Jin et al., 2020; Li et al., 2020; Schlegel et al., 2021; Qi et al., 2021; Wang et al., 2022a; Dyrmishi et al., 2023), or to naturally occurring out-of-distribution data shift (Wang et al., 2022b). This contrasts with our focus, where "natural" pertains to perturbations arising from real-world scenarios rather than those engineered artificially. Some works also propose that a natural synthetically perturbed sample should be imperceptible to human judges (Li et al., 2020; Garg and Ramakrishnan, 2020) or convey the impression of human authorship (Dyrmishi et al., 2023). However, this proposition remains a subject of debate (Zhao et al., 2018; Wang et al., 2022b; Chen et al., 2022b).

3 Future Directions

Looking ahead to robustness evaluation in the LLMs era, we outline key limitations in existing QA robustness research and propose promising directions for future work, based on the literature survey presented in Section 2.

Systematic evaluation of LLM robustness under a comprehensive range of synthetic perturbations. While numerous synthetic perturbation strategies have been proposed for earlier QA systems, their impact on modern LLMs remains underexplored, particularly for more intricate perturbation types. A thorough and unified evaluation covering a diverse range of perturbation strategies would help identify systematic weaknesses and better inform the development of robust LLM-based QA systems.

Shift toward robustness evaluation under natural perturbations. As LLMs are increasingly deployed in real-world applications, ensuring their robustness in practical settings becomes critically important. This calls for more realistic evaluation methods that reflect the challenges LLMs are likely to encounter post-deployment. Natural perturbation approaches offer a promising direction by leveraging input variations that arise organically from real-world use cases. Such evaluations provide deeper insight into how LLMs handle naturally occurring linguistic shifts and user-generated content, thereby offering a more reliable measure of practical robustness.

Deeper investigation into the relationship between natural and synthetic perturbations. Future research should systematically examine the extent to which synthetic perturbations approximate natural variations, and where they fall short. Such inquiry could illuminate the limitations of synthetic methods in capturing the complexity and diversity of real-world text evolution. Prior studies consistently showed that real-world natural perturbations exhibit more diverse and nuanced linguistic phenomena, which are difficult to replicate through synthetic strategies (Belinkov and Bisk, 2018; Wu et al., 2025), and often result in more valid adversarial examples (Le et al., 2022). However, findings on adversarial training remain mixed. For example, Belinkov and Bisk (2018) reported that, in the context of a NMT task, training on synthetic perturbations did not improve robustness against natural ones. In contrast, Wu et al. (2025) found that, for QA tasks, training on synthetic perturbations could improve robustness to natural perturbations, and in some cases be even more effective than training on natural examples—an outcome observed across both encoder-only models and LLMs in few-shot settings. These contradictory findings underscore the need for a more unified understanding of the relationship between natural and synthetic perturbations in robustness evaluation across various NLP tasks and LLMs.

More rigorous and principled design of validity assessment protocols. Assessing the validity of adversarial examples is crucial for disentangling model limitations from degradation in input qual-

ity. Most existing studies rely on human annotators to perform validity checks; however, these annotations sometimes suffer from unsatisfactory inter-annotator agreement (Wu et al., 2023). This highlights the need for a well-defined theoretical framework for human answerability, which is essential for accurately evaluating the validity of perturbed or adversarial inputs. We further advocate for greater transparency in the human annotation process and the development of carefully designed methodologies to measure human performance (Tedeschi et al., 2023). Such practices would provide more precise insights into the true validity of perturbations and the reliability of robustness evaluations.

Consideration of benchmark leakage in robustness assessment. As LLMs and their training corpora continue to scale, the risk of benchmark leakage has become an increasingly pressing concern (Sainz et al., 2024). Instances from held-out evaluation datasets may often be inadvertently included in training data, and in more severe cases, this exposure may extend to ground-truth labels (Dodge et al., 2021; Li et al., 2024). This issue is further exacerbated by the lack of transparency regarding the training data used for most frontier LLMs. Such contamination compromises the integrity of robustness assessments by conflating memorisation with genuine generalisation. Future research should explicitly account for and rigorously examine the impact of benchmark leakage to ensure that robustness claims accurately reflect model behaviour under truly unseen conditions.

4 Conclusion

Robustness remains a critical challenge for QA systems, especially as LLMs are increasingly deployed in real-world and high-stakes applications. In this survey, we reviewed the trajectory of robustness evaluation methods, with a focus on perturbation-based approaches applied to textual input. We first examined synthetic perturbations, outlining their development in earlier models and their continued relevance for evaluating LLMs. We then surveyed work on natural perturbations, which provide a more realistic perspective on model behavior in the face of genuine linguistic variability.

Drawing from these findings, we identified key limitations in existing robustness research and advocated for a shift toward evaluation approaches grounded in naturally occurring variations. We also highlighted future directions, including the need for more systematic evaluation protocols and a deeper understanding of the relationship between synthetic and natural perturbations. As LLMs become central to modern NLP systems, advancing robust QA methods will be essential to ensuring reliability and trustworthiness in practical deployments.

Limitations

This survey primarily focuses on robustness evaluation techniques that involve perturbations to the textual input in QA and broader NLP tasks. In particular, we review methods that apply synthetic or natural variations to the question or context inputs, as commonly studied in the evaluation of language understanding capabilities.

We do not include studies that focus on perturbations to prompts or instructions, which represent a distinct line of research aimed at understanding how LLMs respond to variation in task framing or instruction phrasing. Additionally, we exclude work on jailbreak attacks that are designed to circumvent safety mechanisms in LLMs. While such work is relevant to issues of safety and alignment, it falls outside the scope of this survey, which is centered on robustness in the context of QA performance.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. 2024. Seemingly plausible distractors in multi-hop

reasoning: Are large language models attentive readers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2514–2528, Miami, Florida, USA. Association for Computational Linguistics.

MICHAEL CANALE and MERRILL SWAIN. 1980. Theoretical bases of communicative approaches to second language teaching and testing*. *Applied Linguistics*, I(1):1–47.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022a. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Junkai Chen, Li Zhenhao, Hu Xing, and Xia Xin. 2025. Nlperturbator: Studying the robustness of code llms to natural language variations. *ACM Trans. Softw. Eng. Methodol.* Just Accepted.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. 2022b. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11222–11237, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Noam Chomsky. 1969. Aspects of the theory of syntax.

Noam Chomsky. 2006. *Language and Mind*, 3 edition. Cambridge University Press, Cambridge.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Salijona Dyrmishi, Salah Ghamizi, and Maxime Cordy. 2023. How do humans perceive adversarial text? a reality check on the validity and naturalness of word-based adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8822–8836, Toronto, Canada. Association for Computational Linguistics.

Steffen Eger and Yannik Benz. 2020. From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 786–803, Suzhou, China. Association for Computational Linguistics.

Jingliang Fang, Hua Xu, Zhijing Wu, Kai Gao, Xiaoyin Che, and Haotian Hui. 2023. Robustness-eva-mrc: Assessing and analyzing the robustness of neural models in extractive machine reading comprehension. *Intelligent Systems with Applications*, 20:200287.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Mor Geva, Tomer Wolfson, and Jonathan Berant. 2022. Break, perturb, build: Automatic perturbation of reasoning paths through question decomposition. *Transactions of the Association for Computational Linguistics*, 10:111–126.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}.

In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Xanh Ho, Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2023. A survey on measuring and mitigating reasoning shortcuts in machine reading comprehension. *Preprint*, arXiv:2209.01824.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Thai Le, Jooyoung Lee, Kevin Yen, Yifan Hu, and Dongwon Lee. 2022. Perturbations in the wild: Leveraging human-written text perturbations for realistic adversarial attack and defense. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2953–2965, Dublin, Ireland. Association for Computational Linguistics.

Mosh Levy, Shauli Ravfogel, and Yoav Goldberg. 2023. Guiding LLM to fool itself: Automatically manipulating machine reading comprehension shortcut triggers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8495–8505, Singapore. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.

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

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

K. Mayall, G. W. Humphreys, and A. Olson. 1997. Disruption to word or letter processing? the origins of case-mixing effects. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 23:1275–1286.

L. X. McCusker, P. B. Gough, and R. G. Bias. 1981. Word recognition inside out and outside in. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):538–551.

Grigor Nalbandyan, Rima Shahbazyan, and Evelina Bakhturina. 2025. SCORE: Systematic COnsistency and robustness evaluation for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 470–484, Albuquerque, New Mexico. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd*

Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 525–539, Nusa Dua, Bali. Association for Computational Linguistics.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meghana Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudan, James Zou, and Nazneen Rajani. 2025. Cats confuse reasoning llm: Query agnostic adversarial triggers for reasoning models. *Preprint*, arXiv:2503.01781.

G. E. Rawlinson. 1976. *The significance of letter position in word recognition*. Phd thesis, University of Nottingham.

G. M. Reicher. 1969. Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, 81(2):275–280.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D'Amico-Wong, Melissa Dell, Run-Ze Fan, Shahriar Golchin, Yucheng Li, Pengfei Liu, Bhavish Pahwa, Ameya Prabhu, Suryansh Sharma, and 9 others. 2024. Data contamination report from the 2024 CONDA shared task. In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 41–56, Bangkok, Thailand. Association for Computational Linguistics.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2021. Semantics altering modifications for evaluating comprehension in machine reading. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13762–13770.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2023. A survey of methods for revealing and overcoming weaknesses of data-driven natural language understanding. *Natural Language Engineering*, 29(1):1–31.

Karel Šebesta, Zuzanna Bedřichová, Kateřina Šormová, Barbora Štindlová, Milan Hrdlička, Tereza Hrdličková, Jiří Hana, Vladimír Petkevič, Tomáš Jelínek, Svatava Škodová, Petr Janeš, Kateřina Lundáková, Hana Skoumalová, Šimon Sládek, Piotr Pierscieniak, Dagmar Toufarová, Milan Straka, Alexandr Rosen, Jakub Náplava, and Marie Poláčková. 2017. CzeSL grammatical error correction dataset (CzeSL-GEC). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Parshin Shojaee*†, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 634–644, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. In *The Twelfth International Conference on Learning Representations*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Son Quoc Tran, Phong Nguyen-Thuan Do, Uyen Le, and Matt Kretchmar. 2023. The impacts of unanswerable questions on the robustness of machine reading comprehension models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1543–1557, Dubrovnik, Croatia. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. 2022a. Distinguishing non-natural from natural adversarial samples for more robust pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 905–915, Dublin, Ireland. Association for Computational Linguistics.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2024. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *IEEE Data Eng. Bull.*, 47(1):48–62.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022b. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 575–581, New Orleans, Louisiana. Association for Computational Linguistics.

Winston Wu, Dustin Arendt, and Svitlana Volkova. 2021. Evaluating neural model robustness for machine comprehension. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2470–2481, Online. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2023. Are machine reading comprehension systems robust to context paraphrasing? In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–196, Nusa Dua, Bali. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2025. Pay attention to real world perturbations! natural robustness evaluation in machine reading comprehension. *Preprint*, arXiv:2502.16523.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Torsten Zesch. 2012. Measuring contextual fitness using error contexts extracted from the Wikipedia revision history. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 529–538, Avignon, France. Association for Computational Linguistics.

Kun Zhang, Le Wu, Kui Yu, Guangyi Lv, and Dacao Zhang. 2025. Evaluating and improving robustness in large language models: A survey and future directions. *Preprint*, arXiv:2506.11111.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, page 57–68, New York, NY, USA. Association for Computing Machinery.