Rethinking Robustness Evaluation for Question Answering: From Synthetic Stress Tests to Natural Language Variation

Anonymous ACL submission

Abstract

In this position paper, we contend that prevailing robustness evaluation practices for Question Answering (QA) do not adequately capture system behavior under real-world conditions. Current evaluations predominantly rely on synthetic perturbations defined by idealized assumptions about linguistic validity and label preservation, whose relevance to deployment scenarios is often unclear. Consequently, robustness assessed in such settings may provide a distorted view of the reliability of QA systems. This limitation becomes increasingly salient as Large Language Models (LLMs) are deployed in interactive and agent-based applications, where language variation emerges organically and compounds across multiple interactions. We examine commonly adopted synthetic perturbation paradigms, analyze their limitations, and contrast them with emerging efforts that evaluate robustness using naturally occurring perturbations. Building on this analysis, we advocate for a community-wide shift toward robustness evaluation grounded in real-world language variation and more reliable evaluation protocols.

1 Introduction

Despite strong benchmark performance on Question Answering (QA) (Yang et al., 2025), Large Language Models (LLMs) can often exhibit brittle behaviour when faced with variations in textual input (Zhang et al., 2025). These failures highlight a fundamental challenge for deploying LLM-based QA systems in real-world environments, where language is inherently variable, noisy, and continuously evolving (Markov et al., 2023; Kumar and Mishra, 2025).

Robustness evaluation has therefore become a central concern. The dominant paradigm relies on *synthetic perturbations*, which apply predefined input transformations such as adversarial sentence insertion (Jia and Liang, 2017; Tran et al., 2023), character swaps (Fang et al., 2023), or paraphrasing (Wu et al., 2023), to stress-test QA systems. More recently, a smaller but growing body of work has shifted attention toward *natural perturbations* that arise organically from real-world language variation (Wu et al., 2025b,a), rather than from artificially imposed input manipulations.

In this position paper, we argue that **current QA robustness evaluation paradigms are misaligned with how LLMs fail in practice, and that meaningful progress requires a shift from synthetic, assumption-driven perturbations toward evaluations grounded in naturally occurring language variation**. While synthetic perturbations offer control diagnostics and yield valuable insights into model behaviour (Wang et al., 2022b; Ho et al., 2023), their relevance to real-world deployment remains unclear, raising questions about their ecological validity as practical robustness probes (Kumar and Mishra, 2025). Moreover, synthetic approaches typically rely on strong assumptions regarding label preservation, human answerability, and linguistic validity; in practice, however, these assumptions are often violated by the perturbations themselves, thereby undermining the reliability of robustness evaluation (Dyrmishi et al., 2023). In contrast, natural perturbations better capture real-world conditions and tend to preserve adversarial validity more reliably (Le et al., 2022; Wu et al., 2025b); yet, they remain underexplored and rarely integrated into existing robustness evaluation protocols.

The paper first examines dominant synthetic perturbation approaches used for QA, tracing their application from early language models to modern LLMs, and analyses their limitations. It then presents natural perturbations that emerge from naturally occurring linguistic variation in practice. Based on this analysis, we outline concrete directions for rethinking robustness evaluation protocols for LLM-based QA.

2 Synthetic Perturbations: What Have We Learned?

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 task—Question Answering. 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 ground truth 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 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, state-of-the-art (SOTA) LLMs have demonstrated superior natural language understanding across a wide range of NLP tasks—including QA (OpenAI et al., 2024; OLMo et al., 2025; Yang et al., 2025); they also 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.

3 Natural Perturbations: Reflecting Real-World Failure Modes

Unlike synthetic perturbations, 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., 2025b). 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 naturality (Jin et al., 2020; Li et al.,

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

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). 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., 2025b). 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 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 naturality. Sun et al. (2024) 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).

4 Call for Action

We call on the NLP community to reposition robustness evaluation around naturally occurring language variation as the primary testbed for assessing LLM reliability. Concretely, this entails building and maintaining evaluation benchmarks derived from real-world linguistic evolution (e.g., edits, revisions, and user-driven reformulations), treating synthetic perturbations as auxiliary diagnostics rather than the main evaluation signal. Robust-

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., 2025b)	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.

ness studies should explicitly verify perturbation validity at scale, reporting not only performance degradation but also rates of invalid, ambiguous, or unanswerable cases. As LLMs are increasingly deployed as interactive agents, robustness evaluation should further account for perturbation accumulation across multi-step interactions, where small natural variations can compound into systemic failures. Finally, the community should adopt evaluation protocols that explicitly control for benchmark leakage and prioritize reliability-oriented metrics such as consistency, calibration, and recovery behavior over single-point accuracy. Together, these actions aim to align robustness evaluation with the conditions under which LLM-based systems operate in practice.

Limitations

Our analysis focuses primarily on robustness evaluation for QA and related agentic applications. The extent to which our conclusions generalize to other NLP tasks, such as generation-heavy or multimodal settings, requires further investigation and is beyond the scope of this position paper. While we argue for prioritizing naturally occurring perturbations, constructing large-scale, high-quality natural perturbation benchmarks poses significant practical challenges, including data collection, annotation cost, and potential domain bias. We therefore call for further discussion and research on these challenges.

References

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.

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.

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.

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, 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.

Pankaj Kumar and Subhankar Mishra. 2025. Robustness in large language models: A survey of mitigation strategies and evaluation metrics. *Preprint*, arXiv:2505.18658.

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.

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.

Ilia Markov, Kseniia Kharitonova, and Elena L. Grigorenko. 2023. Language: Its origin and ongoing evolution. *Journal of Intelligence*, 11:61.

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).

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, 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. 2024. 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.

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.

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).

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*.

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.

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. 2025a. Natural context drift undermines the natural language understanding of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1248–1259, Suzhou, China. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2025b. Pay attention to real world perturbations! nat-

ural 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.

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*.