Automatic Hallucination Assessment for Aligned Large Language Models via Transferable Adversarial Attacks

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Although remarkable progress has been achieved preventing LLMs hallucinations, using *instruction tuning* and *retrieval augmentation*, it is currently difficult to measure the reliability of LLMs using available static data that is often not challenging enough and could suffer from data leakage. Inspired by adversarial machine learning, this paper aims to develop an automatic method for generating new evaluation data by appropriately modifying existing data on which LLMs behave faithfully. Specifically, this paper presents `AutoDebug`, an LLM-based framework for using prompt chaining to generate transferable adversarial attacks (in the form of question-answering examples). We seek to understand the extent to which these trigger hallucination behavior in LLMs.

We first implement our framework using ChatGPT and evaluate the resulting two variants of a popular open-domain question-answering dataset, Natural Questions (NQ) on a collection of open-source and proprietary LLMs under various prompting settings. Our generated evaluation data is human-readable and, as we show, humans can answer these modified questions well. Nevertheless, we observe pronounced accuracy drops across multiple LLMs including GPT-4. Our experimental results confirm that LLMs are likely to hallucinate in two categories of question-answering scenarios where (1) there are conflicts between knowledge given in the prompt and their parametric knowledge, or (2) the knowledge expressed in the prompt is complex. Finally, the adversarial examples generated by the proposed method are transferable across all considered LLMs, making our approach viable for LLM-based debugging using more cost-effective LLMs.

1 Introduction

Because of their superior capability in generating coherent and convincing outputs, large language models (LLMs), such as ChatGPT (OpenAI, 2022), GPT4 (OpenAI, 2023), Claude (Anthropic, 2023) and Palm (Anil et al., 2023), have been extensively applied as foundations for language technologies and interactive agents for assisting humans or carrying out autonomous explorations for general problem-solving.Although being more capable of *following instructions* (Ouyang et al., 2022), those *aligned* LLMs (open-source or proprietary) are still found to produce fabricated responses, also known as hallucinations (Ji et al., 2023). Specifically, hallucinations with instruction-following represent *faithfulness* issues, where the response is inconsistent with or even contradicting the task context, *e.g.,* instructions, dialog history, evidence and memories.

In addition to better instruction-tuning, another prominent approach found to be effective in reducing hallucination is to augment LLMs with retrieved external information, *i.e.,* retrieval-augmented

Submitted to R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023. Do not distribute.

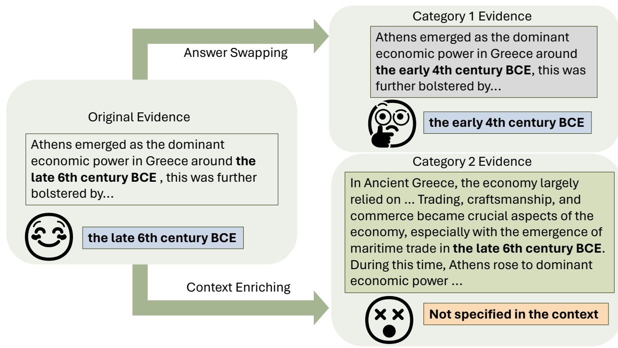

Figure 1: An example of how the original evidence is edited (answer swapping and context enriching) by `AutoDebug`. The question is “when did athens emerges as wealthiest greek city state?". “the late 6th century BCE” and “the early 4th century BCE” is the original and fake answer respectively. ChatGPT answers are next to the emoji.

LLMs (Shi et al., 2023). For example, most recent LLM-based information-seeking assistants (*e.g.,* BingChat[1], ChatGPT Plugins[2]) are capable of searching from the web so that they can respond more accurately to users’ queries. However, it is unclear whether those aligned LLMs augmented with external knowledge are reliable enough to be immune from hallucinations. Given LLMs’ wide adoption, how to *measure, detect* or *mitigate* those hallucinations is becoming increasingly important for achieving trustworthy and safe AI with broad scientific and societal impacts. Specifically, this paper aims to help developers measure the reliability of prompting with aligned LLMs.

Manually creating test cases for assessing hallucination in LLMs is hard to scale, because it is costly to identify cases where the LLMs are likely to fail. Moreover, as LLM-based applications are constantly adapting (*e.g.,* improved prompt engineering and backbone LLMs), those previously useful tests can soon become outdated. Motivated by the long line of work designing adversarial attacks to trigger undesirable behaviors in machine learning models (Madry et al., 2018; Goodfellow et al., 2014), we explore perturbing the prompts for measuring the reliability of LLMs. Unlike recent work on black-box LLMs that focuses on triggering jail-breaking behaviors (Zou et al., 2023; Carlini et al., 2023), we are interested in cases with benign users, who typically aim to interact with LLMs to finish legitimate tasks, and those inputs are *natural* to (understandable by) humans. Following Nie et al. (2020); Iyyer et al. (2018); Jia & Liang (2017), we aim to generate new probing data by *making edits* on the existing one where LLMs can already faithfully fulfill the intended requests.

In this work, we focus on the question-answering (QA) scenario where an LLM agent is designed to answer users’ information-seeking questions regarding a provided document, which is a simplified form of existing commercial LLM-based conversational assistants (*e.g.,* BingChat). As those LLMs are mostly not up-to-date, we propose a framework, `AutoDebug`, including two ways of synthesizing evaluation datasets, both aiming at editing the grounding evidence (Figure 1): 1) *answer swapping*, where the original answer is swapped to another valid answer while the remaining context is intact; 2) *context enriching*, where more relevant information is added to the provided document while the original supportive information is kept. The former simulates the scenario where only answer relevant part of the documents is corrected while the latter represents the evolving document where more relevant information is added leading to more complex documentation of specific topics. We then instantiate `AutoDebug` by designing *prompting chaining* with black-box LLMs, *i.e.,* using LLMs to generate new test cases that are more likely to trigger hallucinations in LLMs.

To verify the effectiveness of the proposed framework, we apply it to a popular open-domain QA dataset, Natural Questions (NQ) (Kwiatkowski et al., 2019), and generate two probing datasets, Category 1 and Category 2.First, human studies are conducted to verify the naturalness of the generated datasets, *i.e.,* the updated document is still understandable by humans and supportive of answering the corresponding question. We then evaluate our generated datasets on one open-source (Alpaca (Taori et al., 2023)) and four propriety (ChatGPT, Claude, Palm and GPT-4) LLMs under

[1]https://bing.com/chat
[2]https://openai.com/blog/chatgpt-plugins

various prompting scenarios, zero-shot, few-shot, and more enhanced prompting techniques designed to improve the reliability of prompting with LLMs. Although natural and supportive in the eyes of humans, both probing datasets trigger LLMs to produce incorrect answers, regardless of their model sizes and instruction-tuning data. We find that the self-attacks are more effective but attacking test examples generated by our method is transferable across all considered LLMs. This enables the possibility of debugging LLMs using test cases generated by more cost-effective LLMs. Lastly, our case study finds that simply using adversarial examples as in-context demonstrations is not effective in reducing hallucination, which calls for future research.

2 `AutoDebug` Framework

Assessing the hallucination of LLMs is challenging as we often do not know what changes in the prompt would trigger LLMs to hallucinate. In this paper, we present our approach `AutoDebug` for automatically constructing a large number of test cases that can surface hallucination issues. Given a pivot LLM, we first prompt it to identify *seed test cases* from a pool of existing data. Then we prompt the pivot LLM again to generate *attacking test cases* based on individual seed test cases. These attacking test cases are used to evaluate the performance of the pivot LLM (self-attack) as well as other LLMs (cross-attack). While `AutoDebug` is a general framework, we focus on the QA scenario where the LLMs to be evaluated need to answer open-domain questions based on their supporting evidence. The pipeline is illustrated in Figure 2 of Appendix.

Example Category		Knowlege Source	
Open-book	Closed-book	Memory	Evidence
Correct	Correct	✔	✔
Correct	Wrong	✘	✔
Wrong	Correct	✔	✘
Wrong	Wrong	✘	✘

Table 1: Classification of QA examples using the LLM behaviors and knowledge sources.

To identify seed test cases, we categorize QA examples into four types (Table 1) based on the condition of whether the pivot LLM can answer the question correctly under the open-book and closed-book settings in a zero-shot fashion (See Table 11 for examples). In the closed-book setting, only the question itself is given and the pivot LLM has to use its internal memory as the main knowledge source, whereas in the open-book setting, the associated supporting evidence is provided as well. If the LLM can answer the question in the closed-book setting, it indicates that the specific piece of knowledge is stored in its internal memory and can be successfully recalled. When the LLM gives different answers under the two settings, it suggests a potential conflict between the internal memory and the evidence. In this paper, the specific hallucination behavior of interest is that **an LLM can answer the question correctly with the original evidence but gives an incorrect answer when the evidence is perturbed**.[3] Therefore, we use the first two types of QA examples in Table 1 as the seed test cases and generate attacking test cases by perturbing the evidence and updating the answers if necessary. In other words, the pivot LLM would have 100% accuracy on the seed test cases.

To generate viable attacking test cases, we consider the following two perturbation approaches. **1) Update** the evidence using a new answer that may lead to a knowledge conflict. In the top-right example of Figure 1, we replace *"the late 6th century BCE"* with *"the early 4th century BCE"* in the evidence and test whether the LLM can update its answer accordingly. textbf2) Enrich the evidence using extra relevant facts that may dilute the information. In the bottom-right example of Figure 1, the evidence becomes much more dense though the answer is unchanged, and we test whether the LLM can still produce the original answer.

For the first approach, we keep both types of seed test cases. For the second approach, we exclude cases where the pivot LLM can answer correctly under the closed-book setting since perturbing the evidence for such cases may not surface the hallucination issue, *i.e.,* the LLM may simply use its internal memory to answer the question correctly and completely ignore the evidence. To assess the hallucination of LLMs, we can simply measure the accuracy of the predicted answers for the attacking test cases. If an LLM is less prone to hallucinate, it should be immune to these perturbations and maintain a high accuracy score. The evaluation considers both zero-shot and few-shot prompting. The zero-shot prompt for evaluation is identical to the one used for seed test selection above. The few-shot version inserts the demonstrations of evidence-question-answer triplets.

[3]Note the original answer may no longer be correct with the perturbed evidence.

Category 1: LLM-Proposed Alternative Answer Here, we present the first approach to generate test cases by updating the original evidence with alternative answers. Specifically, those alternative answers are proposed by an LLM via prompting. Note that the considered seed test cases are open-book correct with the pivot LLM. For each question, given the original answer and supportive evidence, we first ask the model to generate an alternative answer that is factually wrong using the following prompt. We then instruct the LLM to replace all the occurrences of the original answer with the alternative one.[4] Since most context is kept, the newly generated evidence is likely to support the alternative answer for most questions (as verified in §3.2). All used prompts are listed in Table 8.

Category 2: LLM-Enriched Evidence Our second strategy aims to enrich the original evidence with more relevant context, leading to a more complex context for answer reasoning. Unlike Category 1 discussed above, we only keep seed cases that are open-book correct but closed-book wrong to ensure that certain comprehension of the evidence is required to answer the question correctly. To ensure that the newly generated evidence still provides support for the question, we first extract the supporting sentence from the original evidence. We then gather relevant information from an external database to be used for composing the new evidence. Here, we consider two ways of retrieving passages from Wikipedia for fusing with the supporting sentence above, *i.e.,* evidence-focused expansion and question-focused expansion, where the former uses the original evidence as the query and the question is used for the latter case. As those two expansions bring in different types of relevant information, we create two corresponding copies of new evidence. To make the information more diverse, we select the top-k passages from different Wikipedia pages. To merge these passages into a single passage, we first ask the LLM to summarize the information of the retrieved set, and then merge the supporting sentence into the summary. The pivot LLM needs to extract and summarize key information so that the new evidence is human-readable and still supports the original answer. The corresponding prompts can be found in Table 9.

3 Experiments

Evaluation Metrics. Three evaluation metrics are reported, *i.e.,* exact match (EM) accuracy, token-level F1, and entailment accuracy. The first two metrics are traditionally used for evaluating QA models. However, they tend to be too strict for evaluating LLM-generated responses, since LLMs often produce long and verbose sequences to explain the answers (partially due to their alignment procedure). The entailment accuracy is a more lenient metric that checks whether "Question + LLM Output" can entail "Question + Answer". In this paper, we use a SOTA entailment model `nli-deberta-v3-base`[5] trained using Sentence-BERT (Reimers & Gurevych, 2019).

Source Data. We use the MRQA version (Fisch et al., 2019) of Natural Questions (Kwiatkowski et al., 2019) and conduct the following filtering steps: 1) remove duplicated Question-Evidence-Answer triplets and only keep one unique instance, 2) remove all evidence passages that are shorter than 10 words, 3) remove all cases with answers longer than 5 words. After this, 7189 instances are kept.For questions with multiple answers, if the answers are overlapping (*e.g.,* "1871" and "1871 A.D."), we randomly keep one, otherwise, the corresponding examples are removed.Note the same question may still appear in multiple instances because the supporting evidence can be different.

Generated Data. Unless otherwise specified, ChatGPT (`gpt-3.5-turbo-0301`) is the pivot LLM for identifying seed test cases and generating attacking test cases. When identifying seed test cases, we treat an answer produced by the pivot LLM as correct if it matches the reference answer exactly or can entail the reference answer in the same way as we compute the entailment accuracy. The retriever used for generating Category 2 cases is based on `all-mpnet-base-v2`[6]. In total, we obtain **3,539** and **2,211** attacking test cases in Category 1 and Category 2, respectively.

We evaluate five popular LLMs using the generated attacking test cases: Alpaca-7B (Taori et al., 2023), ChatGPT (`gpt-3.5-turbo-0301`), Claude2, PaLM, and GPT-4 (`gpt-4-0613`). In the few-shot setting, 5 static demonstration examples are used.

[4]Although a simple string match can also do the job, it can make the answer occurring sentences inconsistent with the neighboring context, *e.g.,* mismatched pronouns and aliases.

[5]`https://huggingface.co/cross-encoder/nli-deberta-v3-base`

[6]`https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

3.1 Main Results

Models	Method	Zero-shot			Few-shot		
		EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Closed-Book	0.28	5.44	4.86	1.13	6.64	4.94
	Open-Book	18.71	36.04	56.65	21.50	38.46	57.30
	Faithful Prompt	27.80	43.64	58.75	33.74	51.10	65.41
ChatGPT	Closed-Book	1.14	6.72	4.29	0.93	7.28	4.55
	Open-Book	43.71	59.99	77.31	40.44	54.58	65.33
	Faithful Prompt	44.73	40.04	42.98	40.04	52.75	62.11
Claude 2	Closed-Book	2.12	7.10	6.22	0.82	5.79	4.58
	Open-Book	44.62	56.37	59.08	20.32	34.09	69.77
	Faithful Prompt	52.95	65.05	71.80	39.28	50.97	71.83
Palm	Closed-Book	1.72	1.67	6.02	1.67	7.68	5.54
	Open-Book	57.50	65.75	74.71	65.75	75.74	78.41
	Faithful Prompt	64.17	68.41	79.20	68.41	78.61	81.46
GPT-4	Closed-Book	0.82	7.26	4.92	1.10	7.51	5.00
	Open-Book	54.11	68.50	81.29	58.94	72.58	81.01
	Faithful Prompt	58.49	71.70	82.51	63.49	75.72	82.25

Table 2: Evaluation of LLMs on Cat1 attack.

Models	Method	Zero-shot			Few-shot		
		EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Closed-Book	0.18	10.57	14.34	2.67	13.45	13.30
	Open-Book	9.27	39.35	42.79	14.52	45.56	47.40
	Faithful Prompt	15.06	43.65	42.65	20.58	53.40	50.88
ChatGPT	Closed-Book	0.09	10.66	0.27	9.81	25.02	22.03
	Open-Book	25.51	57.15	61.78	27.32	58.94	51.15
	Faithful Prompt	24.69	53.49	50.38	24.20	56.26	44.10
Claude 2	Closed-Book	8.01	19.89	15.97	6.24	19.49	22.75
	Open-Book	29.99	58.69	43.46	12.12	39.83	57.26
	Faithful Prompt	35.78	64.89	52.60	27.45	54.31	54.68
Palm	Closed-Book	10.58	25.67	22.89	11.99	25.23	21.26
	Open-Book	44.78	71.76	66.76	50.84	75.23	66.53
	Faithful Prompt	44.78	70.18	58.75	47.35	72.03	61.78
GPT-4	Closed-Book	18.32	36.17	37.04	20.76	38.04	36.14
	Open-Book	37.68	67.27	68.39	46.27	74.17	73.04
	Faithful Prompt	33.60	62.78	58.25	45.59	72.83	67.57

Table 3: Evaluation of LLMs on Cat2 attack.

We evaluate the five LLMs on the Category 1 and Category 2 data generated by ChatGPT, including both self-attack and cross-attack scenarios. In addition to vanilla zero-shot and few-shot promptings, we consider the recently proposed faithfulness prompting, *i.e.,* the opinion-based prompt by Zhou et al. (2023). For each model, we evaluate its performance of closed-book, open-book, and open-book with faithful prompting settings. The full list of various prompts can be found in Appendix.

Category 1. Here, the model is expected to predict the fake answer proposed by ChatGPT. Given that, the closed-book performance of all the models is expected to be near 0. We report the closed-book performance to validate the generation quality. The results are summarized in Table 2. As expected, the model resistance towards our attack is mostly correlated with its model size and capability. Specifically, larger and more capable models are more robust, *e.g.,* GPT-4 is more reliable than Alpaca-7B, which suggests that recent efforts in aligning LLMs is promising for developing more trustworthy models. Although GPT-4 is the most powerful model, it is not still immune to our attacks, indicating the effectiveness of our approach to trigger hallucination in SOTA LLMs. Though using the human-designed faithful prompt or using in-context examples helps the performance in some cases, there are no consistent improvements compared with zero-shot in general.

Category 2. We require the model to understand both the question-focused expansion and evidence-focused expansion cases, and one question is considered correct only when both are answered correctly. We report the merged result in Table 3, and we also report the few-shot performance on each case separately in Table 13 of Appendix. As we can see, there are large performance drops for all models, suggesting they fail to identify the relevant evidence information regardless of prompting techniques (the faithful prompting and in-context examples). It is worth noting that all the questions in Category 2 are closed-book wrong and open-book correct based on ChatGPT performance, which explains why the closed-book accuracies of other models are better. Similar to Category 1, the faithful prompt is observed to have no consistent benefits, which calls for future work to develop more reliable prompting techniques.

3.2 Human Evaluations

To evaluate whether the evidence generated by `AutoDebug` is supportive and human-readable, we randomly sample 500 cases from Category 1, 1000 cases from Category 2 with 500 examples for question-focused expansion, and 500 for evidence-focused expansion. We use Amazon Mechanical Turk to collect human judgments on this set. Each question is judged by three annotators, who are asked to read the evidence and decide whether it could support them to get the correct answer. To prevent annotators from randomly submitting "Yes" or "No", 10% of the data are used as validation checks where we know whether the evidence supports the answer. We only accept annotations from the annotators with at least 90% accuracy on the validation check. For each question, if the majority of the annotators think the generated evidence is supportive, it is then counted as human-readable. For all three categories, around 90% of the cases are human readable, supporting the quality of `AutoDebug`, with 90.8, 92.4 and 88.8 human-readable ratios for Category 1, Category 2 question-focused and evidence-focused, respectively.

3.3 Case Study: Is `AutoDebug` sensitive toward backbone LLMs?

Models	Method	ChatGPT			GPT-4			Alpaca-7B		
		EM	F1	Entail.	EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Closed-Book	0.8	4.69	5.80	2.60	7.37	8.60	2.20	9.86	9.60
	Evidence	25.00	40.57	61.20	26.8	43.88	68.2	26.00	43.95	65.80
	Faithful	37.20	53.46	72.20	39.60	57.49	76.00	36.60	53.93	70.80
ChatGPT	Closed-Book	0.40	4.79	4.40	1.60	5.72	5.80	1.00	7.19	6.00
	Evidence	43.00	54.88	66.20	49.60	61.55	71.60	38.40	51.56	61.40
	Faithful	42.80	53.25	61.80	51.40	61.53	70.40	40.00	52.57	61.20
Palm	Closed-Book	2.40	7.10	7.00	4.60	10.07	8.60	3.80	10.32	8.60
	Evidence	70.80	78.51	81.40	75.80	82.58	86.00	67.00	74.55	79.00
	Faithful	74.20	82.00	84.40	78.80	85.28	89.00	69.20	77.73	82.80
GPT-4	Closed-Book	0.6	5.77	4.20	1.20	6.36	5.60	0.20	8.54	6.00
	Evidence	65.20	76.66	84.00	59.20	69.18	76.40	57.00	67.23	73.80
	Faithful	69.80	79.04	84.80	67.40	75.98	81.80	59.60	70.15	78.40

Table 4: Cat1 attack using various LLMs.

Models	Method	ChatGPT			GPT-4		
		EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Closed-Book	1.80	7.57	8.00	2.00	7.61	8.80
	Evidence	17.80	44.85	52.20	9.00	37.16	42.40
	Faithful	22.40	53.96	57.00	16.00	46.28	43.80
ChatGPT	Closed-Book	3.20	12.63	8.40	3.20	12.75	8.60
	Evidence	29.40	57.12	50.80	23.20	50.76	46.20
	Faithful	24.40	54.61	41.60	23.20	52.80	43.20
Palm	Closed-Book	6.20	16.15	12.00	7.60	16.68	13.00
	Evidence	54.40	76.84	69.60	52.20	73.62	66.40
	Faithful	53.40	75.93	68.60	48.4	71.91	62.60
GPT-4	Closed-Book	12.20	24.71	20.20	13.60	24.49	22.60
	Evidence	49.40	74.38	74.20	24.00	47.18	37.60
	Faithful	51.80	73.68	71.00	35.00	62.04	52.40

Table 5: Cat2 attack using various LLMs.

To do that, we use alternative LLMs to generate attacking test cases other than ChatGPT. We consider both Alpaca-7b and GPT-4 for Category 1 and only GPT-4 for Category 2 given the task is more demanding. Due to the limitation of budget, we randomly sample 500 examples for this study. All prompts are similar to those used previously. The few-shot performances of Category 1 and Category 2 are reported in Table 4 and Table 5, respectively. As shown in Table 4, compared with ChatGPT and Alpaca, GPT-4 does not generate stronger attacks. It is probably because the alternative answers from GPT-4 are more receptive to all models. On the other hand, compared with ChatGPT, GPT-4 can generate more stronger attacks for Category 2 (Table 5). We find that GPT-4 is better at summarizing multiple pieces of information, leading to more complex evidence. Although all three models are most vulnerable to self-attacks, all `AutoDebug` attacks are transferable, making it possible to generate attacks using more cost-effective models.

4 Related Work

There is a long line of research in generating adversarial examples to trigger errors or undesirable behaviors from machine learning models (Szegedy et al., 2014; Goodfellow et al., 2014). To improve the robustness of machine learning models, there are also a number of methods proposed to defend against such attacks (Madry et al., 2018; Zhu et al., 2020; Li & Qiu, 2020; Cheng et al., 2021). However, models trained with adversarial learning are found to have at-odd generalization Tsipras et al. (2019); Zhang et al. (2019), *e.g.,* improving the accuracy on adversarial attacks can compromise the model performance on clean examples. Despite being more challenging due to its discrete nature, different text adversarial attacks with perturbed inputs imperceptible to humans have been proposed for question answering (Jia & Liang, 2017), natural language inference (Nie et al., 2020), and sentiment classification (Iyyer et al., 2018). One surprising phenomenon is that many adversarial examples are *transferable* (Papernot et al., 2016; Wallace et al., 2021). For example, Wallace et al. (2021) show that adversarial prefix optimized for one particular model can also transfer to models of different architectures and sizes. In addition to replying on white-box access to generate effective adversarial examples, recent work even reports that it is difficult to generate reliable examples via automatic search (Carlini et al., 2023). Our work is highly motivated by this long line of work, *i.e.,* making evidence edits while keeping the input legitimate for the targeted task so that the LLMs can not reliably answer the question. Here, we do not assume any model access except its text outputs, *i.e.,* black-box. We show that our proposed approach of generating adversarial test cases from a pivot LLM can trigger hallucination behaviors across state-of-the-art open-source and proprietary LLMs.

5 Conclusion

In this paper, we present `AutoDebug` that generates transferable adversarial attacks and successfully triggers hallucination behaviors of existing prominent LLMs. We believe `AutoDebug` could be used to help assess the hallucination of future LLMs, and potentially help mitigate hallucinations.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

Anthropic. Claude 2, 2023. URL `https://www.anthropic.com/index/claude-2`.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2023.

Hao Cheng, Xiaodong Liu, Lis Pereira, Yaoliang Yu, and Jianfeng Gao. Posterior differential regularization with f-divergence for improving model robustness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1078–1089, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.85. URL `https://aclanthology.org/2021.naacl-main.85`.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of 2nd Machine Reading for Reading Comprehension (MRQA) Workshop at EMNLP*, 2019.

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

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1875–1885, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1170. URL `https://aclanthology.org/N18-1170`.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL `https://doi.org/10.1145/3571730`.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL `https://aclanthology.org/D17-1215`.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL `https://aclanthology.org/Q19-1026`.

Linyang Li and Xipeng Qiu. Textat: Adversarial training for natural language understanding with token-level perturbation. *arXiv preprint arXiv:2004.14543*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL `https://aclanthology.org/2020.acl-main.441`.

OpenAI. ChatGPT, 2022. URL `https://openai.com/blog/chatgpt`.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `http://arxiv.org/abs/1908.10084`.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Replug: Retrieval-augmented black-box language models, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SyxAb30cY7`.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp, 2021.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/zhang19p.html`.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*, 2023.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BygzbyHFvB`.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Appendix

A.1 Pipeline Illustration

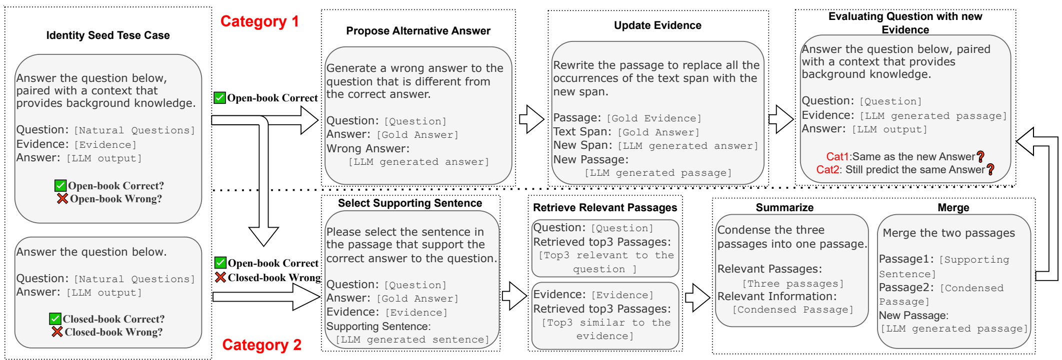

Figure 2: The pipeline of `AutoDebug`, including identifying seed cases, generating new tests, and hallucination evaluation.

A.2 Demonstration Instance

Question: who sings what lovers do with maroon 5 Evidence: “ What Lovers Do ” is a song by American pop rock band Maroon 5 featuring American R&B singer SZA . It was released on August 30 , 2017 , as the lead single from the band ’s sixth studio album Red Pill Blues (2017) . The song contains an interpolation of the 2016 song “ Sexual ” by Neiked featuring Dyo , therefore Victor Rådström , Dyo and Elina Stridh are credited as songwriters . Answer: American R&B singer SZA
Question: who plays lead guitar on i want you she ’s so heavy Evidence: John Lennon – lead and harmony vocals , multi-tracked lead guitar , Moog synthesizer Paul McCartney – harmony vocals, bass George Harrison – harmony vocals , multi-tracked lead guitar Ringo Starr – drums , congas , wind machine Billy Preston – Hammond organ Answer: John Lennon
Question: a long chain of amino acids linked by peptide bonds is a Evidence: The covalent chemical bonds are formed when the carboxyl group of one amino acid reacts with the amino group of another . The shortest peptides are dipeptides , consisting of 2 amino acids joined by a single peptide bond , followed by tripeptides , tetrapeptides , etc . A polypeptide is a long , continuous , and unbranched peptide chain . Hence , peptides fall under the broad chemical classes of biological oligomers and polymers , alongside nucleic acids , oligosaccharides and polysaccharides , etc . Answer: polypeptide
Question: when does the school year start in france Evidence: In Metropolitan France , the school year runs from early September to early July . The school calendar is standardised throughout the country and is the sole domain of the ministry . Answer: early September
Question: which city is selected under hriday scheme in karnataka Evidence: With a duration of 4 years (completing in November 2018) and a total outlay of 500 crore (US $78 million) , the Scheme is set to be implemented in 12 identified Cities namely , Ajmer , Amaravati , Amritsar , Badami , Dwarka , Gaya , Kanchipuram , Mathura , Puri , Varanasi , Velankanni and Warangal . Answer: Ajmer

Table 6: Five Randomly Selected Demo Instances from NQ Training Data for Few-shot Experiments.

Question: who sings what lovers do with maroon 5 Evidence: “ What Lovers Do ” is a song by American pop rock band Maroon 5 featuring British pop singer Adele. It was released on August 30 , 2017 , as the lead single from the band ’s sixth studio album Red Pill Blues (2017) . The song contains an interpolation of the 2016 song “ Sexual ” by Neiked featuring Dyo , therefore Victor Rådström , Dyo and Elina Stridh are credited as songwriters . Answer: British pop singer Adele
Question: who plays lead guitar on i want you she ’s so heavy Evidence: Paul McCartney – harmony vocals, bass George Harrison – harmony vocals , multi-tracked lead guitar Ringo Starr – drums , congas , wind machine Billy Preston – Hammond organ Answer: Paul McCartney
Question: a long chain of amino acids linked by peptide bonds is a Evidence: The covalent chemical bonds are formed when the carboxyl group of one amino acid reacts with the amino group of another. The shortest peptides are dipeptides, consisting of 2 amino acids joined by a single peptide bond, followed by tripeptides, tetrapeptides, etc. A lipid is a long, continuous, and unbranched peptide chain. Hence, peptides fall under the broad chemical classes of biological oligomers and polymers, alongside nucleic acids, oligosaccharides and polysaccharides, etc Answer: lipid
Question: when does the school year start in france Evidence: In Metropolitan France, the school year runs from late August to early July. The school calendar is standardised throughout the country and is the sole domain of the ministry Answer: late August
Question: which city is selected under hriday scheme in karnataka Evidence: With a duration of 4 years (completing in November 2018) and a total outlay of 500 crore (US $78 million) , the Scheme is set to be implemented in 12 identified Cities namely , Mumbai, Amaravati, Amritsar, Badami, Dwarka, Gaya, Kanchipuram, Mathura , Puri , Varanasi , Velankanni and Warangal . Answer: Mumbai

Table 7: Five Randomly Selected Demo Instances from NQ Training Data with altenative answers and generated evidence for Few-shot Counter Experiments.

A.3 Prompts

Generate Alternative Answer Prompt	A question and its correct answer is below. Generate a wrong answer to the question that is different from the correct answer. Make sure the wrong answer is short, and has the same type as the correct answer. Question: {Question} Answer: {Answer} Wrong Answer:
Replace Old Answer Prompt	A passage and a text span inside the passage is shown below. Rewrite the passage to replace all the occurrences of the text span with the new span. Passage: {Passage} Text Span: {Answer} New Span: {Alternative Answer} New Passage:

Table 8: Prompts for Cat1 Data Generation.

Select Supporting Sentence Prompt	A question, the answer, and a passage are shown below. Please select the sentence in the passage that supports to answer the question correctly. Question: {Question} Answer: {Answer} Passage: {Passage} Sentence:
Summarize Relevant Passages Prompt	Three relevant passages are shown below. Please condense the three passages into one passage. Relevant Passages: [1]: {Passage 1} [2]: {Passage 2} [3]: {Passage 3} Relevant New Information:
Merge Prompt	Two passages and a span are shown below. Please merge the two passages, and make sure to keep the span in the new passage. Passages: [1]: {Supporting Sentence} [2]: {Summarized Passage} Span: {Answer} New Passage:

Table 9: Prompts for Cat2 Data Generation.

Alpaca-7B	Below is an instruction that describes a task. Write a response that appropriately completes the request. Only output the answer without other context words. ### Instruction: {Question} ### Response:
PaLM	You are a helpful and informative bot that answers questions Be sure to respond in a complete sentence, being comprehensive, including all relevant background information. However, you are talking to a non-technical audience, so be sure to break down complicated concepts and strike a friendly and convers- tional tone. Only output the answer without other context words. QUESTION: {Question} ANSWER:
Claude 2	Human: Answer the question below. Only output the answer without other context words. Question: {Question} Assistant:
ChatGPT & GPT-4	system: You are a helpful assistant. user: Answer the question below. Only output the answer without other context words. Question: {Question} Answer:

Table 10: Closed-Book QA prompts for all considered models following their corresponding recommendations.

Alpaca-7B	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. Only output the answer without other context words. ### Instruction: {Question} ### Input: {Evidence} ### Response:
PaLM	You are a helpful and informative bot that answers questions using text from the reference passage included below. Be sure to respond in a complete sentence, being comprehensive, including all relevant background information. However, you are talking to a non-technical audience, so be sure to break down complicated concepts and strike a friendly and convers- tional tone. If the passage is irrelevant to the answer, you may ignore it. Only output the answer without other context words. QUESTION: {Question} PASSAGE: {Evidence} ANSWER:
Claude 2	Human: Answer the question below, paired with a context that provides background knowledge. Only output the answer without other context words. Context: {Evidence} Question: {Question} Assistant:
ChatGPT & GPT-4	system: You are a helpful assistant. user: Answer the question below, paired with a context that provides background knowledge. Only output the answer without other context words. Context: {Evidence} Question: {Question} Answer:

Table 11: Open-Book Inference Prompts for Different Models Following their Official Instructions.

Alpaca-7B	Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. ### Instruction: Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text? ### Response:
PaLM	Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text?
Claude 2	Human: Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text? Assistant:
ChatGPT & GPT-4	system: You are a helpful assistant. user: Instruction: read the given information and answer the corresponding question. Only output the answer without other context words. Bob said, “{Evidence}” Q: {Question} in Bob’s opinion based on the given text?

Table 12: Opinion-based Inference Prompts for Different Models Following Zhou et al. (2023)

A.4 Additional Results

Models	Method	Few-shot Question Only			Few-shot Evidence Only		
		EM	F1	Entail.	EM	F1	Entail.
Alpaca-7B	Closed-Book	2.67	13.45	13.30	2.40	13.35	12.89
	Open-Book	23.38	44.94	60.65	24.56	46.18	62.87
	Faithful Prompt	30.94	51.88	63.50	33.06	54.93	66.21
ChatGPT	Closed-Book	9.81	25.02	22.03	9.45	24.78	21.66
	Open-Book	40.93	59.10	67.89	40.66	58.78	67.03
	Faithful Prompt	40.89	57.59	64.22	38.22	54.94	60.88
Claude 2	Closed-Book	6.24	19.49	22.75	6.11	19.39	22.70
	Open-Book	22.16	39.63	71.73	22.21	40.03	73.95
	Faithful Prompt	38.13	53.17	68.70	39.35	55.45	70.78
Palm	Closed-Book	11.99	25.23	21.26	11.99	25.23	21.26
	Open-Book	58.44	72.89	73.45	61.96	77.58	78.11
	Faithful Prompt	55.63	70.15	70.28	58.48	73.90	73.32
GPT-4	Closed-Book	20.76	38.04	36.14	20.62	37.98	35.55
	Open-Book	54.23	72.85	80.69	56.54	75.48	83.31
	Faithful Prompt	54.95	71.76	77.25	57.08	73.89	78.79

Table 13: Few-shot result of Question-based Cat2 data and Evidence-based Cat2 data.