On Assessing the Faithfulness of LLM-generated Feedback on Student Assignments

Qinjin Jia, Jialin Cui, Ruijie Xi, Chengyuan Liu, Parvez Rashid, Ruochi Li, Edward Gehringer
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
qjia3, efg@ncsu.edu

ABSTRACT

Feedback on student assignments plays a crucial role in steering students toward academic success. To provide feedback more promptly and efficiently, researchers are actively exploring the use of large language models (LLMs) to automatically generate feedback on student artifacts. Although the generated feedback is highly fluent, coherent, and plausible sounding, LLMs are prone to hallucinating content that is unfaithful to the input document, hence discouraging the adoption of automated feedback systems in actual classes. In this paper, we analyze the limitations of data-driven and prompt-driven automated feedback systems in generating feedback for student project reports. We examined the feedback and found that both systems generate a considerable amount (> 23%) of hallucinated content. Furthermore, we explored various methods for measuring hallucinations and showed that instruction fine-tuned ChatGPT yields better alignment with human judgment. Our work sheds light on the hallucination problem in feedback generation while identifying several prominent challenges for future research.

Keywords

Automated feedback generation, automated review generation, hallucination, faithfulness, factuality

1. INTRODUCTION

Feedback is integral in guiding students through their learning process, offering valuable insights that enable them to strengthen or correct their understanding of knowledge and content [1, 17, 25, 43]. However, providing quality feedback on student assignments, particularly on open-ended and complex ones, presents a substantial challenge for educational resources and is often difficult to deliver in a timely manner. This demand for immediate and cost-efficient feedback has catalyzed interest in developing various automated feedback systems. Recent advancements in large language models (LLMs), exemplified by ChatGPT [50], have rendered LLM-based feedback systems the most prevalent medium for generating feedback [20, 22]. Previous studies have shown that the feedback generated by such LLM-based systems is highly fluent, coherent, and human-like [6, 21].

However, it is also evident that LLM-based automated feedback systems are susceptible to generating spurious text [19]. In other words, while these systems are capable of producing feedback that is often indistinguishable from instructor feedback in terms of fluency and coherence, they may generate prose that is erroneous, misleading, or entirely irrelevant to the input student assignments [6, 20, 21]. This phenomenon is commonly referred to as hallucination, which poses a substantial challenge that undermines the reliability of automated feedback systems [35]. Hallucinated content can potentially lead students to believe what is false, impacting their understanding of knowledge and even their motivation to learn [17, 20]. Therefore, concerns about hallucination impede the deployment of feedback systems in real classes.

Data-driven and prompt-driven are two state-of-the-art approaches to implementing LLM-based feedback generation systems. These orthogonal approaches correspond to two distinct paradigms for customizing LLMs for feedback generation tasks: *pre-training and then fine-tuning* [44, 52, 27] and *prompt engineering* [45, 46, 13, 6]. The former involves further training LLMs to capture underlying patterns from datasets of student work and instructor feedback to produce feedback [21, 14]. On the other hand, the latter entails leveraging human-crafted textual prompts to guide LLMs in generating desired feedback for student assignments [6, 42]. Interestingly, to the best of our knowledge, no prior study has investigated the prevalence of hallucination in feedback generated by both feedback systems on the same dataset.

In this paper, our primary objective is to gain insight into the hallucination problem in feedback generation by thoroughly examining hallucinated content and exploring methods for measuring hallucinations. Specifically, we first implement a BART-based [27] data-driven feedback system and an OpenAI ChatGPT4-based [50] prompt-driven feedback system, for automatically producing feedback on student project reports. Subsequently, we conduct a human evaluation of the feedback generated by the systems and analyze the types of hallucinated content in the feedback. Additionally, we investigate natural language inference (NLI)-based and OpenAI ChatGPT3.5-based approaches for measuring hallucinations in feedback. Lastly, we explore whether hallucination raises ethical concerns in practical applications.

Q. Jia, J. Cui, R. Xi, C. Liu, P. Rashid, R. Li, and E. Gehringer. On assessing the faithfulness of llm-generated feedback on student assignments. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 491–499, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

© 2024 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

https://doi.org/10.5281/zenodo.12729868

To investigate hallucination in feedback generation, our work seeks to answer the following research questions:
RQ1: Are both data-driven and prompt-driven automated feedback systems susceptible to hallucination? What is the prevalence of hallucinated content in generated feedback?
RQ2: Is the hallucinated content produced by the systems primarily intrinsic (i.e., contrary to the input information) or extrinsic (i.e., neither supported nor contradicted)?
RQ3: Are there effective methods for measuring the hallucinations? Can the accuracy of the measurements be improved by utilizing human annotations on hallucination?
RQ4: Does the hallucinated content in feedback raise any ethical concerns? For example, does the hallucinated content contain offensive language or private information?

The main conclusions are as follows: First, both the data-driven and prompt-driven automated feedback systems generate a significant amount of hallucinated content (27.1% and 23.5%, respectively). Second, we observe that intrinsic hallucinations are more common in our data-driven system, whereas extrinsic hallucinations are more prevalent in our prompt-driven system. Third, when measuring hallucinations, instruction fine-tuned ChatGPT demonstrates the best alignment with human judgment (F_1 score $\approx 72\%$), yet still leaves considerable scope for improvement. Lastly, we fail to find any ethical concerns directly associated with hallucinations. The findings highlight opportunities to further study the mitigation and measurement of hallucinations.

Our main contributions are: 1) we analyze hallucinations in the feedback generated by two state-of-the-art LLM-based automated feedback systems for student project reports; 2) we collect a new dataset of hallucinations to facilitate future research endeavors; 3) we also implement NLI-based and ChatGPT3.5-based models for measuring hallucinations; 4) we highlight several prominent challenges for future work.

The paper is organized as follows: Section 2 first presents related work on hallucination in feedback generation. Then, Section 3 describes the dataset utilized for this study. Following that, Section 4 elaborates our methodology for investigating hallucination in feedback and explains the methods for measuring hallucinations. Subsequently, Section 5 presents and discusses our experimental results. Finally, Section 6 concludes the paper, acknowledges the limitations of our work, and provides a discussion on future research.

2. LITERATURE REVIEW

In the following, we first review research on automated feedback generation. Then, Section 2.2 provides background on hallucinations in text generation. After that, Section 2.3 surveys methods for measuring hallucinations. Lastly, Section 2.4 reviews ethical concerns related to hallucination.

2.1 Automated Feedback Systems

In the realm of education, feedback is defined as information provided by an agent (e.g., teacher, peer) regarding the performance of learners. Feedback has been shown to significantly impact student learning outcomes [1, 17]. Prior research has focused on building various feedback-generation systems to produce feedback for students. These studies can be broadly classified into three categories based on the feedback-generation models (i.e., feedback engines) used.

2.1.1 Expert-driven methods

Expert-driven methods (e.g., [37, 38]) leverage pre-designed templates and rules that encode expert knowledge and experience to produce feedback . While such systems yield accurate feedback and are not data-hungry, they are not suitable for dealing with complex and open-ended assignment types because creating and maintaining a vast collection of expert-designed rules for such assignments is nearly impossible. For example, Nagata et al. [37] used this approach to diagnose preposition errors and produce feedback for English writing.

2.1.2 Data-driven methods

Instead of relying on manually designed feedback rules, data-driven methods implicitly derive the rules from data on paired student work and instructor feedback [7]. The dominant implementation of such methods involves utilizing data to fine-tune large language models (LLMs), such as BERT [23] and GPT [45]. For example, Jia et al. [20, 21] showed an BART-based system to provide feedback on project reports. MacNeil et al. [34] utilized GPT3 to help novices program.

2.1.3 Prompt-driven methods

With the development of LLMs, especially general-purpose LLMs such as ChatGPT [50], prompt-driven methods that require little or no paired student work and feedback data have emerged. Such systems rely on prompt engineering [31] (e.g., expert-designed prompts) to guide LLMs in generating feedback. For example, Dai et al. [6] studied the viability of using ChatGPT to generate feedback on student proposals. Liu et al. [32] utilized GPT4 to help students learn code.

2.1.4 Comparisons between methods

Expert-driven and prompt-driven methods both require expert input, but they fundamentally differ in the role of this input. Expert-driven methods directly encode expert knowledge into templates and rules. Thus, such methods entirely rely on knowledge in expert input to produce feedback. In contrast, prompt-driven methods use expert input indirectly by providing prompts that steer LLMs to generate feedback. Thus, the generation process does not completely rely on expert input, but rather more on internal knowledge of LLMs.

Data-driven and prompt-driven methods both harness LLMs, yet they diverge in the source of knowledge for producing feedback. Data-driven methods directly leverage the extensive knowledge contained within the training data, learning the underlying patterns to generate feedback. Conversely, prompt-driven methods primarily rely on internal knowledge of LLMs, which is constructed from massive pre-training corpora. Thus, prompt-driven systems are less data-intensive, but they depend heavily on the comprehensiveness of internal knowledge and the quality of human-crafted prompts.

2.2 Hallucination in Text Generation

Despite the encouraging performance of the aforementioned data-driven and prompt-driven automated feedback systems, they inevitably encounter a number of challenges in practical applications, including hallucination [19], ambiguity [49], incompleteness [55], under-informativeness [12], and bias [39]. Among these obstacles, hallucination is notably intractable and seriously impedes the actual deployment of the systems.

2.2.1 *Definition of hallucination*

Within the domain of text generation, *hallucination* is defined as an occurrence where the text generated by a model is nonsensical or unfaithful (i.e., inaccurate) to the information presented in the source content [11, 35, 41, 55]. In simpler terms, the system-generated text can appear fluent, coherent, and plausible sounding, yet it could be inconsistent or irrelevant to the corresponding student submission.

In addition to hallucination, the literature encompasses two related concepts, namely *faithfulness* and *factuality*, which are sometimes used interchangeably or conflated [19]. Both terms are antonyms of hallucination, referring to being actual or based on fact. However, they differ in what is considered as *fact*. For faithfulness, fact is the input content, whereas for factuality, fact refers to world knowledge [35].

2.2.2 *Categorization of hallucinations*

After defining hallucination and clarifying related concepts, we now further discuss two types of hallucinations [18, 19].

Intrinsic Hallucinations are characterized by the generated text directly contradicting the input. This implies that the text is inconsistent with the input information. For instance, an output may erroneously claim "your work is missing a test plan," even though the student has already provided one.

Extrinsic Hallucinations refer to cases where the faithfulness of generated text cannot be verified against source content (i.e., neither supported nor contradicted). For example, a model may fabricate text "your code screenshots have quite small text," despite the absence of any image in the input.

In addition, [55] further classifies intrinsic hallucinations into input-conflict and context-conflict. The former refers to the generated content that contradicts the input, while the latter refers to the content that is internally self-contradictory. However, our work does not make this further distinction.

2.2.3 *Sources of hallucination*

Lastly, we explore potential factors that may induce hallucinations in text generation to better understand the issue.

Data is a primary cause of hallucination, which mainly refers to the lack of pertinent knowledge or internalization of false knowledge [2, 55]. Inconsistencies, incompleteness, or biases in training data can directly lead to hallucinations. Moreover, LLMs are susceptible to misinterpreting spurious correlations, such as highly co-occurring associations, as faithful knowledge, thereby generating hallucinated outputs [28].

Training and Inference also contribute to hallucination [40], which refers to hallucinated generation stemming from the discrepancy in decoding between training and inference time (i.e., exposure bias [47]). During training, the next token is predicted conditioned on the ground-truth prefix sequences (teacher-forced [15]). However, in inference, LLMs generate the next token conditioned on their previous generation.

2.3 Measuring Hallucinations

In addition, measuring hallucinations in generated text is indispensable and serves two essential purposes. First, quantifying the level of hallucination empowers researchers and developers to assess the reliability of systems and compare different methods. Second, identifying hallucinated content is the primary step towards mitigating hallucinations through post-editing [19, 26]. Nevertheless, the free-form and open-ended nature of text generation poses a challenge in measuring hallucinations produced by LLM-based systems [55]. We now proceed to survey potential measurement methods.

2.3.1 *Reference–based methods*

Reference-based methods refer to a set of metrics that measure hallucinations by comparing generated texts with reference texts (e.g., ground-truth feedback provided by instructors) based on the content overlap, such as ROUGE [30], or semantic similarity, such as BERTScore [54]. However, such methods necessitate reference texts and have been shown to exhibit a weak correlation with human judgment [8, 35, 24].

2.3.2 *Question answering–based methods*

Question-answering (QA) based methods implicitly measure the consistency between generated texts and input texts [55]. These methods operate on the idea that if a piece of generated text is faithfully aligned with its source content, then similar answers should emerge from the same question [9]. However, QA-based methods rely heavily on the quality of the question-generation and question-answering models [19].

2.3.3 *Natural language inference–based methods*

Natural language inference (NLI) is the task of determining whether, given a premise, a hypothesis is true (entailment) or false (contradiction) [53]. Such methods are based on the idea that the source content should entail the information in faithful generated text [10, 35]. Thus, if the model tells us that the input content is likely to entail the output, we can conclude that the generated text is faithful and vice versa.

2.3.4 *Off-the-shelf LLM–based methods*

Off-the-shelf LLM-based methods refer to harnessing the capabilities of general-purpose LLMs (e.g., Meta's LLaMA [51] and OpenAI's ChatGPT [50]) to serve as hallucination evaluators. Such methods involve using prompts to query LLMs to judge the faithfulness of generated text [36, 26]. However, the effectiveness of such methods is highly dependent on the training corpora of LLMs and the quality of prompts [55].

2.4 Ethical concerns about Hallucination

The ethical implications pertaining to hallucination have received scant attention in the literature, despite the fact that hallucinations can potentially raise certain ethical concerns. Celikyilmaz et al. [4] highlight that generating offensive language is an ethical transgression that may appear in LLM-based systems. Carlini et al. [3] demonstrate that LLMs may recover and reproduce sensitive personal information, such as physical addresses, email addresses, and phone numbers, from the training corpora. Both of these phenomena are considered to be associated with hallucinations because the systems are generating unfaithful content [19]. Nevertheless, there is a lack of established methodologies for effectively assessing the tendency of LLM-based systems to be oblivious to ethical issues of hallucination. In this study, we analyze the ethical violations related to hallucinations by manually examining feedback generated by the systems.

Table 1: Samples of instructor and system-generated feedback for student project reports.

Feedback provided by the instructor: "The document is reasonably readable. The UI changes are described, but the code changes are not. The test plan is not adequately described. It should say what tests are to be performed, not just which functionality is to be tested. It is reasonable to test the UI manually, but model and controller tests should be automated."
Feedback from the data-driven system: "The writeup is very readable, but it does not explain the changes made to the code. It is not clear how the code is organized. The test plan is not described. This is a very important part of the design. I would have liked to see a description of how the tests work. Also, the test plan should have described how the tests work."
Feedback from the prompt-driven system: "The document outlines several improvements to the system's account request feature, including adding options for new institutions and introducing a user introduction text area. However, it lacks clarity in explaining how these changes are implemented and tested. The absence of detailed examples or code snippets makes it difficult to understand the modifications made to the system. Additionally, the document would benefit from a more organized structure, such as a table of contents, to improve readability and navigation."

3. DATA

In this section, we introduce the datasets used for this study. First, Section 3.1 describes a dataset of student project reports and feedback used for training or prompting models. Section 3.2 then explains how we implement data-driven and prompt-driven systems for generating feedback. Then, Section 3.3 presents the dataset used for evaluating the models for measuring hallucinations. Finally, Section 3.4 discusses how participants' privacy rights were respected during the data collection process. In addition, Table 1 above exhibits three samples of instructor and generated feedback texts.

3.1 Project reports and Feedback Dataset

In order to train the data-driven feedback system and provide examples for the prompt-driven feedback system, we utilize the dataset introduced in [20, 21], which consists of 484 student project reports and paired instructor feedback. The dataset is from a graduate-level object-oriented development course and it encompasses project reports for which students engaged in activities such as refactoring existing code, implementing new features and functionalities, or developing automated unit tests for software modules.

In their reports, students are instructed to document the work that has been completed, the methodologies that they have used, and how they tested their UI design and code. Following this, the instructor reviews each of these reports and provides textual feedback. The average number of words for each student project report in this dataset is 704, which corresponds to 951 subword tokens. For instructor feedback, the average number of words and the average number of subword tokens per review are 55 and 71, respectively.

3.2 Data-driven and Prompt-driven Systems

3.2.1 BART-based data-driven system

We employ a BART-based data-driven automated feedback system, as introduced in [20, 21], for generating feedback. The BART model is an encoder-decoder LLM [27], capable of effectively capturing underlying relationships from one sequence of text (e.g., project reports) to another (e.g., feedback). We use the "facebook/bart-large-cnn" checkpoint[1] to initialize the parameters and then fine-tune the model with 434 pairs of project reports and feedback. Subsequently, we generate feedback for 50 project reports from the test set.

3.2.2 ChatGPT-based prompt-driven system

We also implement a ChatGPT4-based prompt-driven automated feedback system, similar to the one described in [6]. To guide the model in generating feedback for reports in the test set, we craft a few-shot prompt – *"You are an instructor responsible for providing feedback on student project reports. I will provide you with project reports and your task is to provide feedback for each of them. Here are five examples of project reports and paired feedback . . . "* The five examples in the prompt are randomly selected from the training set.

3.3 Hallucination detection Dataset

To assess the performance of various methods in measuring hallucinations, we collect a new dataset consisting of feedback generated by the data-driven system. Specifically, by leveraging different combinations of decoding strategies and hyperparameters, we generate five sets of feedback for reports in the test set, thereby simulating a scenario that requires the measurement models for selecting hyperparameters. We manually label the faithfulness of each sentence.

The dataset consists of 250 feedback messages, with a total of 1430 sentences. The average word counts per feedback message and per sentence are 80 and 14, respectively. Approximately 29% of the sentences contain hallucinated content. The inter-annotator agreement between two annotators, measured by Cohen's κ coefficient, is 69.3%. This level of agreement is classified as substantial according to [5], which demonstrates that our annotations are reliable.

3.4 Privacy Protection

In this work, we have rigorously protected the privacy of student data. This research has been approved by the institutional review board (IRB) at our institution. Data is handled in a way that complies with the FERPA regulations[2]. The specific measures taken to protect privacy include: 1) anonymizing all identifying information in the raw student data using random identifiers; 2) employing regular-expression techniques and manual review to remove all potential sensitive information, such as document links that could reveal individual identities; 3) securely storing the de-identified data on a cloud service managed by the university.

[1] https://huggingface.co/facebook/bart-large-cnn

[2] The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99) is a Federal law that protects the privacy of student education records.

Table 2: Sample intrinsic and extrinsic hallucinations with explanations. Hallucinated content is indicated by wavy underlines.

Intrinsic hallucination example 1: "It is not clear what the changes are, and it would be helpful to have a more detailed description of them." (**Explanation:** This project report has included a list of changes, although it is not detailed.)

Intrinsic hallucination example 2: "Additionally, the document would benefit from a more organized structure, such as a table of contents, to improve readability and navigation." (**Explanation:** The project report has a table of contents.)

Extrinsic hallucination example 1 "Also, the screenshots are not very large, and the code snippets are not clearly separated." (**Explanation:** There is no image in the input doc, so the faithfulness of this text is neither supported nor refuted.)

Extrinsic hallucination example 2 "Finally, the document should consider providing examples or mock-ups of the proposed changes to facilitate better understanding and feedback from stakeholders." (**Explanation:** This is irrelevant text.)

4. METHODOLOGY

In this section, we detail our methodologies. Section 4.1 formally defines the problem. Section 4.2 then explains the process for evaluating the generated feedback. Finally, Section 4.3 introduces two methods for measuring hallucinations.

4.1 Problem and preliminaries

Automated feedback generation is often formulated as a *text-to-text generation* task, where the source text X represents input student work and the output target text is feedback Y. Then, feedback generation can be formally expressed as:

$$Y = \mathcal{F}_{\mathcal{M}}(X, \mathbb{C}) \tag{1}$$

where the model $\mathcal{F}_{\mathcal{M}}(\cdot)$ is usually powered by LLMs, and $\mathbb{C}$ is a group of desired properties (e.g., length). We consider that Y is hallucinated if it contradicts the information presented in X, or the faithfulness cannot be verified by X.

Our objective is to gain insights into hallucination by examining the feedback generated by data-driven and prompt-driven systems. Furthermore, we aim to explore and develop methods for detecting hallucinated sentences within Y.

4.2 Hallucinations in the feedback

To better understand hallucination in feedback generation, we conduct human evaluation of feedback generated by both systems and analyze the types of hallucinated content. In the assessment, two human annotators were presented with student project reports and system-generated feedback messages. They were instructed to only assess the hallucinations in the feedback and not focus on other quality dimensions. The annotators are assigned to the responsibilities of 1) evaluating whether each feedback sentence contains hallucinated content, and 2) for every occurrence of hallucinated content, discerning whether the hallucination is intrinsic or extrinsic.

It is worth noting that hallucinations are not necessarily erroneous [19, 35]. For certain extrinsic hallucinated content, although it is actually factual, it may not be faithful (i.e., it is hallucinated). For example, the feedback sentence *"This is a good example of a design doc, but it would have been better to show the code in a larger font."* is identified as an instance of extrinsic hallucination. While the font size of the code in the report is indeed small, there is no information regarding font size in the input to the feedback system. Thus, although it conforms to real-world facts (i.e., factual), this feedback sentence remains unfaithful to the input content.

Table 3: Percentages of hallucinated (intrinsic and extrinsic) content in generated feedback. Lower numbers for hallucinations and the higher number for faithfulness are boldfaced.

Method	Hallucination %		Faithfulness %
	Intrinsic	Extrinsic	
Data-driven	16.0	**11.1**	72.9
Prompt-driven	**9.8**	13.7	**76.5**

4.3 Measures for Hallucinations

4.3.1 NLI-based Method

We implement an NLI model using BART [27] and initialize all parameters with the "bart-large-mnli" checkpoint[3] [53], which is pretrained on the multi-genre NLI (MNLI) dataset. Using the pretrained weights provides us with an out-of-the-box NLI model for assessing faithfulness. In order to judge the faithfulness of generated feedback, we treat the source content (e.g., a student project report) as the premise, and each sentence from the feedback as hypothesis. We first test it as an off-the-shelf classifier, and then we further assess it after supervised fine-tuning using human-annotated data.

4.3.2 ChatGPT3.5-based Method

We also leverage a ChatGPT3.5-based model for measuring hallucinations. Specifically, we utilize the "ChatGPT3.5-turbo-1106" model.[4] We experiment with various prompts and ultimately select the one (shown in Appendix A) that demonstrates the best in terms of the macro F_1 score under the one-shot condition (i.e., with a single example provided in the prompt). Similar to the use of NLI-based model, we initially evaluate this model as an off-the-shelf classifier. Subsequently, we instruction fine-tune the model using human-annotated data and assess its performance again.

5. EXPERIMENTS AND RESULTS

5.1 Experimental setup

5.1.1 Training Details

The NLI model is trained on an NVIDIA A100 GPU (40GB) with a batch size of 4, a learning rate of 5e-5, epochs of 5 with early stopping, and the AdamW optimizer [33] with a weight decay of 0.01. The ChatGPT3.5 model is trained using the OpenAI ChatGPT API, with a temperature of 0 to reduce randomness and ensure deterministic outputs. The learning rate, batch size, and number of epochs are auto-set.

[3] https://huggingface.co/facebook/bart-large-mnli

[4] https://platform.openai.com/docs/models/gpt-3-5-turbo

Table 4: Macro-F_1 scores and their breakdown on F_1-Hallucination and F_1-non-Hallucination for models measuring hallucinations. For each score, we report a 95% confidence interval (CI). The highest $\mathbf{F_1}$ score in each column is boldfaced.

Method	NonHal	Hal	F_1
NLI	76.8±3.2	39.4±5.8	58.1±2.9
Fine-tuned NLI	85.4±3.5	54.0±6.3	69.8±3.2
ChatGPT3.5	70.8±3.7	48.8±4.9	59.9±3.6
Fine-tuned ChatGPT3.5	**86.9**±4.7	**57.2**±6.6	**72.1**±3.9

5.1.2 Evaluation metrics

We utilize the macro-F_1 score to evaluate models since we define the task of measuring hallucinations as a binary classification. The macro-F_1 score is the unweighted mean of F_1 scores calculated per class, so there is no distinction between highly and poorly populated classes. Thus, the macro-F_1 score provides a better picture of whether the model performs well on all classes than the micro-F_1 score [16, 29]. We report the macro-F_1 score as well as its breakdowns on hallucination and non-hallucination since the dataset is skewed.

5.2 Experimental Results and Discussion

In this section, we present experimental results and provide answers to the research questions (RQs) outlined earlier.

5.2.1 RQ1 – Prevalence of hallucinated content

The first experiment aims to determine whether both data-driven and prompt-driven automated feedback systems are prone to hallucination. Table 2 shows samples of hallucinations, and Table 3 presents the results of human evaluation. The prevalence of hallucinated content (intrinsic + extrinsic) in the feedback generated by data-driven and prompt-driven systems is 27.1% and 23.5%, respectively. That is, hallucinations occur in about a quarter of the sentences.

Overall, the results suggest that both systems are susceptible to generating hallucinated content. However, the prompt-driven system outperforms the data-driven system in terms of faithfulness, with a gap of 3.6%. It is also worth noting that the prompt-driven system tends to first summarize the project report. The summaries, while accurate, are not helpful to students, and may lead to an underestimation of the actual proportion of hallucinated content in the feedback.

5.2.2 RQ2 – Intrinsic or extrinsic hallucination

In the second experiment, we aim to explore the types (intrinsic or extrinsic) of hallucinated content in the feedback. As shown in Table 3, the percentages of intrinsic and extrinsic hallucinations in the feedback generated by the data-driven system are 16.0% and 11.1%, respectively. However, the feedback from the prompt-driven system contains 9.8% intrinsic hallucinations and 13.7% extrinsic hallucinations.

The results suggest that intrinsic hallucinations are more common in the data-driven system, while extrinsic hallucinations are more prevalent in the prompt-driven system. We speculate that the data-driven system misinterprets spurious correlations in the data, leading to more feedback that contradicts the input information. The prompt-driven system, on the other hand, relatively lacks domain knowledge, making it more susceptible to generating irrelevant content.

5.2.3 RQ3 – Measuring the hallucinations

We now turn to the experimental results of measuring the hallucinations. As shown in Table 4, the out-of-the-box NLI and ChatGPT3.5 models achieve $\mathrm{F_1}$ scores of 58.1% and 59.9%, respectively. However, the $\mathrm{F_1}$ scores for the fine-tuned NLI and ChatGPT3.5 are 69.8% and 72.1%, which are improvements of 11.7% and 12.2%, respectively. Both models perform better on the non-hallucinated samples.

The results indicate that existing out-of-the-box LLMs still struggle to effectively judge the faithfulness of feedback. However, with the provision of additional human annotations, the performance of these models in measuring hallucinations can be significantly improved. Nonetheless, their performance is still not reliable enough for practical application, warranting further exploration and study. Therefore, measuring hallucinations remains a research opportunity.

5.2.4 RQ4 – Ethical concerns about hallucination

The objective of the last experiment is to investigate whether the hallucinated content in feedback engenders any ethical concerns, particularly concerning the presence of offensive language or private information. As we mentioned, systematic methods are still lacking for assessing potential ethical issues arising from hallucinations. Consequently, we manually examined all system-generated feedback, and found no instances of the aforementioned ethical transgressions.

6. CONCLUSION

LLM-based automated feedback systems will play a crucial role in the future AI-powered educational ecosystem [48]. However, concerns about hallucination impede the deployment of feedback systems in actual classroom settings. In this paper, we have examined the feedback generated by both data-driven and prompt-driven systems for student project reports. The results demonstrate that both systems produce a considerable amount of intrinsic and extrinsic hallucinations. Moreover, the instruction fine-tuned ChatGPT achieves better alignment with human judgment in measuring hallucinations, yet more effective methods are still to be uncovered. This work contributes to a better understanding of hallucination in feedback generation, and our aspiration is to facilitate the practical deployment of feedback systems.

Limitations and future work: There are two main limitations to this study. First, while analyzing the types of hallucinated content in the feedback provided valuable insights, understanding the root causes of these hallucinations remains unexplored. Future research could deepen the understanding of hallucination issues by analyzing their sources. Second, we did not delve into how to effectively design prompts when using ChatGPT to measure hallucinations. Subsequent work could focus on creating more effective prompts to enhance the performance of hallucination detection. Addressing these limitations could significantly contribute to refining the methodologies for evaluating and mitigating hallucinations in automated feedback systems, thereby paving the way for their more responsible application in education.

7. REFERENCES

[1] W. Alharbi. E-feedback as a scaffolding teaching strategy in the online language classroom. *Journal of Educational Technology Systems*, 46(2):239–251, 2017.

[2] Y. Cao, A. M. Nair, E. Eyimife, N. J. Soofi, K. Subbalakshmi, J. R. Wullert II, C. Basu, and D. Shallcross. Can large language models detect misinformation in scientific news reporting? *arXiv preprint arXiv:2402.14268*, 2024.

[3] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[4] A. Celikyilmaz, E. Clark, and J. Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, and G. Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

[7] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerdt. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162:104094, 2021.

[8] B. Dhingra, M. Faruqui, A. Parikh, M.-W. Chang, D. Das, and W. Cohen. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, 2019.

[9] E. Durmus, H. He, and M. Diab. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, 2020.

[10] O. Dušek and Z. Kasner. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, 2020.

[11] K. Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, 2020.

[12] L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[13] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics (ACL), 2021.

[14] S. Gombert, A. Fink, T. Giorgashvili, I. Jivet, D. Di Mitri, J. Yau, A. Frey, and H. Drachsler. From the automated assessment of student essay content to highly informative feedback: a case study. *International Journal of Artificial Intelligence in Education*, pages 1–39, 2024.

[15] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks.

[16] M. Grandini, E. Bagli, and G. Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[17] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[18] Y. Huang, X. Feng, X. Feng, and B. Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.

[19] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[20] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, and E. Gehringer. Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. *International Educational Data Mining Society*, 2022.

[21] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, E. Gehringer, et al. Automated feedback generation for student project reports: A data-driven approach. *Journal of Educational Data Mining*, 14(3):132–161, 2022.

[22] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[23] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[24] K. Krishna, A. Roy, and M. Iyyer. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[25] S. Kusairi. A web-based formative feedback system development by utilizing isomorphic multiple choice items to support physics teaching and learning. *Journal of Technology and Science Education*, 10(1):117–126, 2020.

[26] D. Lei, Y. Li, M. Wang, V. Yun, E. Ching, E. Kamal, et al. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation,

translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[28] S. Li, X. Li, L. Shang, Z. Dong, C.-J. Sun, B. Liu, Z. Ji, X. Jiang, and Q. Liu. How pre-trained language models capture factual knowledge? a causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732, 2022.

[29] Z. Li, Y. Huang, M. Zhu, J. Zhang, J. Chang, and H. Liu. Feature manipulation for ddpm based change detection. *arXiv preprint arXiv:2403.15943*, 2024.

[30] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In M.-F. Moens and S. Szpakowicz, editors, *Text summarization branches out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[32] R. Liu, C. Zenke, C. Liu, A. Holmes, P. Thornton, and D. J. Malan. Teaching cs50 with ai. 2024.

[33] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[34] S. MacNeil, A. Tran, D. Mogil, S. Bernstein, E. Ross, and Z. Huang. Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2*, pages 37–39, 2022.

[35] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, 2020.

[36] N. Mündler, J. He, S. Jenko, and M. Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.

[37] R. Nagata, M. Vilenius, and E. Whittaker. Correcting preposition errors in learner english using error case frames and feedback messages. In H. W. Kristina Toutanova, editor, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–764, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[38] S. Narciss, S. Sosnovsky, L. Schnaubert, E. Andrès, A. Eichelmann, G. Goguadze, and E. Melis. Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, 71:56–76, 2014.

[39] R. Navigli, S. Conia, and B. Ross. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*.

[40] A. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, and D. Das. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, 2020.

[41] A. P. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, D. Das, and G. Tech. Totto: A controlled table-to-text generation dataset.

[42] T. Phung, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generating high-precision feedback for programming syntax errors using large language models. *arXiv preprint arXiv:2302.04662*, 2023.

[43] P. Race. Using feedback to help students to learn. *The Higher Education Academy*, 2001.

[44] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[45] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[46] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[47] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

[48] J. Roschelle, J. Lester, and J. Fusco. Ai and the future of learning: Expert panel report., 2020. https://circls.org/reports/ai-report.

[49] A. Tamkin, K. Handa, A. Shrestha, and N. Goodman. Task ambiguity in humans and language models. *arXiv preprint arXiv:2212.10711*, 2022.

[50] O. Team. Chatgpt: Optimizing language models for dialogue, 2022.

[51] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[52] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[53] W. Yin, J. Hay, and D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, 2019.

[54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[55] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

APPENDIX

A. PROMPT USED IN THE WORK

We followed the prompt-engineering guidance[5] published by OpenAI ChatGPT to design the prompt as below.

system_in = "You will be provided with a student project report (delimited by XML tags <doc> and </doc>) and the corresponding textual feedback on that project report (delimited by XML tags <feedback> and </feedback>, with sentences separated using </sen>).

However, the provided feedback is generated by large language models. And it may not be entirely faithful (i.e., some sentences in the generated feedback may be hallucinated - irrelevant, made-up, or inconsistent with the project report).

Your task is to evaluate whether each sentence of the feedback is faithful (i.e., not hallucinated) to the corresponding student project reports. To accomplish this, you should read and understand the content of the reports, and then read each sentence of the feedback to evaluate whether the sentence is faithful.

Use labels to mark, 1 if the sentence is likely to be faithful, and 0 if the sentence is likely to be unfaithful. Please judge each sentence and return the original sentence and its label.

<An example is inserted here.>

For example, the feedback contains 6 sentences: 1. This is a very good description of the changes made to the code. 2. The design doc is very readable. 3. It explains the changes well. 4. However, it does not explain how the code was refactored, or how it was tested. 5. Also, the code is not consistent with the design doc. 6. This is a good thing.

If you think the sentences 1, 2, 3, 5 are faithful to the project report. And the sentences 4, 6 are unfaithful to the project report. You should return the original sentences and labels like this:

1. This is a very good description of the changes made to the code. - 1

2. The design doc is very readable. - 1

3. It explains the changes well. - 1

4. However, it does not explain how the code was refactored, or how it was tested. - 0

5. Also, the code is not consistent with the design doc. - 1

6. This is a good thing. - 0

Please judge the sentences and follow the format above to return your answers."

[5] https://platform.openai.com/docs/guides/prompt-engineering