

Dean of LLM Tutors: Exploring Comprehensive and Automated Evaluation of LLM-generated Educational Feedback via LLM Feedback Evaluators

Keyang Qian, Yixin Cheng, Rui Guan, Wei Dai, Flora Jin, Kaixun Yang, Sadia Nawaz, Zachari Swiecki, Guanliang Chen, Lixiang Yan, Dragan Gašević

Centre for Learning Analytics, Faculty of Informaiton Technology, Monash University
Melbourne, VIC, Australia

{keyang.qian, yixin.cheng, rui.guan, wei.dai1, flora.jin, kaixun.yang1, sadia.nawaz, zach.swiecki, guanliang.chen, lixiang.yan, dragan.gasevic}@monash.edu

Abstract

The use of Large Language Model (LLM) based tutors to provide automated educational feedback to students on student assignment submissions has received much attention in the AI in Education (AIED) field. However, the stochastic nature and tendency for hallucinations in LLMs can undermine both quality of learning experience and adherence to ethical standards. To address this concern, we propose a method that uses LLM feedback evaluators (*DeanLLMs*) to automatically and comprehensively evaluate feedback generated by LLM tutor for submissions on university assignments before it is delivered to students. This allows low-quality feedback to be rejected and enables LLM tutors to improve the feedback they generated based on the evaluation results. We first proposed a comprehensive evaluation framework for LLM-generated educational feedback, comprising six dimensions for feedback content, seven for feedback effectiveness, and three for hallucination types. Next, we generated a virtual assignment submission dataset (n=200 assignment submissions) covering 85 university assignments from 43 computer science courses using eight commonly used commercial LLMs. We labelled and open-sourced the assignment dataset to support the fine-tuning and evaluation of LLM feedback evaluators. Our findings show that o3-pro demonstrated the best performance in zero-shot labelling of feedback (Accuracy 74.3%, F1-score 74.5%) while o4-mini demonstrated the best performance in few-shot labelling of feedback (Accuracy 74.9%, F1-score 75.2%). Moreover, GPT-4.1 achieved human expert level performance after fine-tuning (Accuracy 79.8%, F1-score 79.4%; human average Accuracy 78.3%, F1-score 82.6%). Finally, we used our best-performance model to evaluate 2,000 assignment feedback instances generated by 10 common commercial LLMs, 200 each, to compare the quality of feedback generated by different LLMs. Gemini 2.5 Pro demonstrated the highest feedback quality and 0-detection of hallucination issues. Small LLMs like GPT-4.1 nano and Gemini 2.0 Flash-Lite underperformed in our evaluation. Our LLM feedback evaluator method advances our ability to automatically provide high-quality and reliable educational feedback to students.

Introduction

Since feedback plays a crucial role in student success, delivering it in a timely and high-quality manner poses a substantial logistical problem (Middleton et al. 2023), particularly in large computing classes facing strict resource con-

straints (Nguyen et al. 2014; of Sciences et al. 2018). Consequently, the application of Large Language Models (LLMs) to generate automated feedback has emerged as a key focus areas in educational technology research (Kim et al. 2025; Rudolph, Tan, and Tan 2023; Dai et al. 2024a), especially in the field of computer science education (Denny et al. 2024; Balse et al. 2023; Hellas et al. 2023; Liffiton et al. 2023; Leinonen et al. 2023; Pankiewicz and Baker 2023; Phung et al. 2023a,b; Prather et al. 2023).

While the development of LLMs shows great promise for generating educational feedback, researchers have raised concerns about hallucinations often found in LLM-generated content (Yan et al. 2024). Hallucinations refer to the phenomenon where LLMs generate content that is irrelevant, fabricated, or inconsistent (Lakera AI 2024). A heated debate is ongoing within the GenAI research community whether hallucinations are an inherent aspect of LLM generation (Sangaré 2023). One prevalent perspective (Xu, Jain, and Kankanhalli 2024; Banerjee, Agarwal, and Singla 2024) suggest that hallucinations are unavoidable in LLMs. However, limited studies have focused on mitigating this issue in LLM-generated educational feedback. Furthermore, LLM generation may produce content that lacks relevance or meaning even in the absence of hallucinations (Zhao et al. 2023). Therefore, it is necessary to develop robust methodological approaches to automatically assess the quality of feedback generated by an LLM tutor (i.e., a LLM who serves like a human tutor to give educational feedback to students) before presenting it to students. Such methods can be utilised to establish a Dean system for LLM tutors. The Dean screens out low-quality and hallucinated feedback, prompts regeneration by the LLM tutor, and selects high-quality feedback from multiple generated options, to ensure that only high-quality feedback is delivered to students.

Despite extensive research on feedback generation by LLM-based tutors (Kim et al. 2025; Dai et al. 2024a; Denny et al. 2024), limited research has focused on automated evaluation of feedback generated by LLM tutors based on the feedback text itself. Lin et al. (2024) and Scarlatos et al. (2024) utilised LLMs to evaluate LLM-generated feedback for single dimensions, specifically praise components and correctness. Osakwe et al. (2022) employed four binary classifiers to assess feedback effectiveness across task, pro-

cess, self-regulation, and self levels. Dai et al. (2024b) and Dai et al. (2025) expanded evaluation frameworks to include readability and content-related dimensions, leveraging LLMs for assessment. However, they overlooked hallucination dimensions (Zhang et al. 2023b; Rawte, Sheth, and Das 2023), crucial for reliability, safety, and trustworthiness in educational feedback (Alfredo et al. 2024). Detecting hallucinations is feasible for LLMs (Zhang et al. 2023a; Rawte, Sheth, and Das 2023; Kaddour et al. 2023; Dhuliawala et al. 2023). The current study extends previous work on feedback quality evaluation principles by proposing a comprehensive evaluation framework for LLM tutor educational feedback, and labelled a feedback dataset according to the framework, which results will become open-source. Based on that, we explored different methods of letting LLMs serve as the Dean of LLM tutors to evaluate LLM tutor feedback.

Our contributions can be summarised as follows:

- We proposed a comprehensive evaluation framework for LLM tutor educational feedback with 16 dimensions, comprising six dimensions for feedback content, seven for feedback effectiveness, and three for hallucination types.
- According to this evaluation framework, we labelled feedback generated by 8 common commercial LLMs for 200 assignment submissions spanning 85 university assignments from 43 computer science courses. Since the AAAI26 anonymous submission does not allow insertion of hyperlinks, the dataset and codes are put in supplementary materials, and will be made publicly available upon acceptance.
- We found that a fine-tuned few-shot prompting GPT-4.1 model can achieve the same accuracy and F1 performance as human experts in evaluating LLM tutor feedback.
- Using the model, we comprehensively compared feedback quality of LLM tutors with difference common commercial LLMs.
- The 'Dean of LLM Tutors' framework represents a significant advance in our ability to deliver reliable and high-quality educational feedback to students automatically in a scalable way.

Method

Comprehensive LLM Feedback Evaluation Framework

We propose our framework that aims to comprehensively evaluate the generated feedback, with the following dimensions that are widely recognised evaluation dimensions in the feedback and LLM literature. The supporting literature are included for each dimension as follows.

- **Feedback content dimensions.** This aspect evaluate features of feedback content that are proved to be good feedback features for students in the literature. These dimensions include:
 - 1). **Alignment with goals.** This evaluates to what extent most comments are aligned to specific learning goals (Nicol and Macfarlane-Dick 2006).

- 2). **Specificity.** This evaluates to what extent all comments offer detailed, actionable guidance, and if the feedback includes any concrete example(s) of how to improve (Shute 2008; Goodman, Wood, and Hendrickx 2004).
- 3). **Motivational Tone.** This evaluates to what extent the tone of feedback is consistently positive, encouraging, and respectful (Nicol and Macfarlane-Dick 2006; Lin et al. 2024; Dai et al. 2025).
- 4). **Strength.** This evaluates to what extent the feedback mentions strength of specific aspect(s) of the student submission (Ryan et al. 2023; Brookhart 2017; Dai et al. 2024b).
- 5). **Weakness.** This evaluates to what extent the feedback mentions strength of specific aspect(s) of the student submission (Ryan et al. 2023; Brookhart 2017; Dai et al. 2024b).

Human coders graded these dimensions using 3-point Likert scale in the labelling (Fang et al. 2011).

- **Feedback effectiveness dimensions.** This aspect detects the effectiveness components in the feedback that are regarded as important components of effective feedback. These include:
 - 1). The three feedback dimensions: Feed forward, Feed up and Feed back (Hattie and Timperley 2007; Dai et al. 2024b, 2025).
 - i. **Feed forward.** Feed forward is about the next steps in learning and the next tasks and activities.
 - ii. **Feed up.** Feed up is about the student's understanding of the learning goal, clarifying the goal. Feed up refers to reminders about the goals and judgements about the success in goal attainment (Hattie and Timperley 2007).
 - iii. **Feed back.** Feed back is about the student's progress towards achieving the learning goal, and responding to student work.
 - 2). The four feedback levels: Feedback on task, Feedback on process, Feedback on self-regulation, and Feedback on self (Bennett and Kell 1989; Cavalcanti et al. 2020; Derham, Balloo, and Winstone 2022; Hattie and Timperley 2007).
 - i. **Feedback on task.** This level includes feedback comments about how well a task is being accomplished or performed, such as distinguishing correct from incorrect answers, acquiring more or different information, and building more surface knowledge (Hattie and Timperley 2007). Feedback can be about a task, such as whether the job is correct or incorrect, can include instructions for more or different information (Cavalcanti et al. 2020).
 - ii. **Feedback on process.** This refers to directive developmental guidance in relation to future work (Derham, Balloo, and Winstone 2022). Feedback can be directed to the process used to create a product or complete a task, is more directed to information processing, or learning processes that require under-

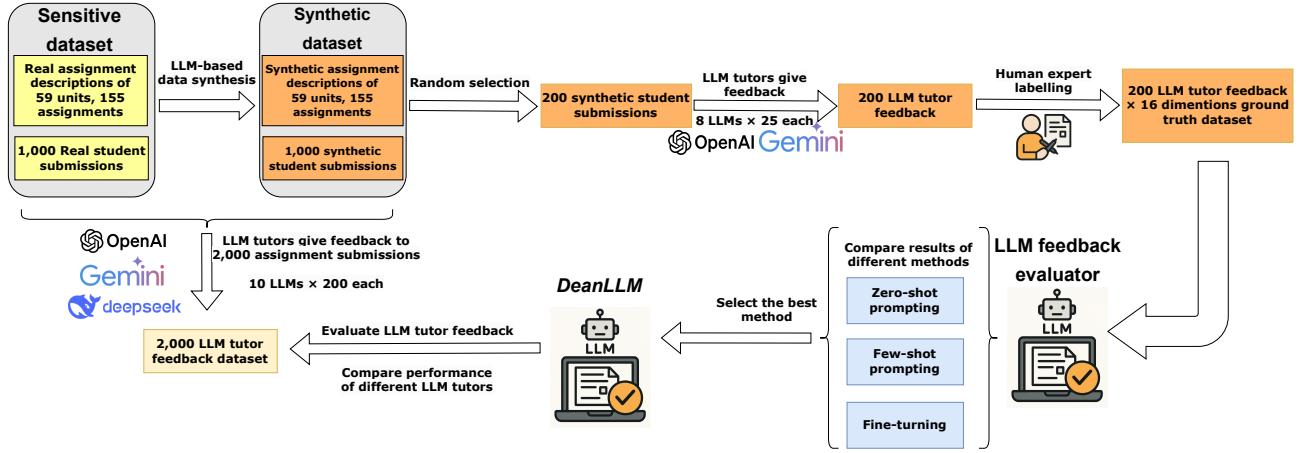


Figure 1: Overview of experiment process and data flow.

standing or completing the task (Cavalcanti et al. 2020).

- iii. **Feedback on self-regulation.** This refers to informative developmental guidance component in the feedback (Derham, Balloo, and Winstone 2022). Feedback for students can be focused on the level of self-regulation, including greater self-assessment or confidence skills, which can have major influences on self-efficacy, self-regulatory proficiency, and students’ personal beliefs as learners (Cavalcanti et al. 2020).
- iv. **Feedback on self.** This refers to the feedback component that is personal in the sense that it is directed to the self. It is often unrelated to task performance (Derham, Balloo, and Winstone 2022).

In the labelling, human coders labelled these dimensions as binary values (Dai et al. 2024b, 2025) indicating whether these feedback effectiveness components are included in the feedback or not.

- **Types of hallucinations.** This aspect detects hallucination issues in the feedback. These includes Input-conflicting Hallucinations, Context-conflicting Hallucinations, and Fact-conflicting Hallucinations (Zhang et al. 2023b; Rawte, Sheth, and Das 2023).
- 1). **Input-conflicting Hallucinations.** This type of hallucinations replace correct input information with errors, such as swapping a person’s name in a summary. These often stem from limited contextual understanding.
- 2). **Context-conflicting Hallucinations.** This type of hallucinations provide contradictory information within the same context, leading to confusion and misinformation.
- 3). **Fact-conflicting Hallucinations.** Fact-conflicting hallucinations produce text that contradicts known facts.

In the labelling, human coders labelled these dimensions

as binary values indicating whether the feedback contained these hallucinations or not.

Human-labelled Feedback Quality Dataset

Our experiment process is illustrated in Figure 1. We first retrieved the assignment dataset of university courses delivered during the first semester of 2021 from the computer science school in a [country blinded] university. The dataset consists of 1,000 student submissions and corresponding assignment descriptions (including rubric and assignment materials), spanning 155 assignments from 59 courses.

As the real assignment data is sensitive with student private information and copyright of the university, which prevent us from making the human-labelled feedback quality dataset publicly available, we deployed o4-mini to generate a synthetic assignment dataset from the real assignment dataset, which also contains 1,000 synthetic assignment submissions spanning 155 synthetic assignments. This is done through LLM-based one-to-one imitation: we first let the LLM imitate the theme and writing style of each real assignment description to generate a corresponding synthetic assignment description, and then let the LLM imitate the writing style, level of detail and correctness of a real assignment submission to generate a synthetic assignment submission according to the corresponding real and synthetic assignment descriptions, using step-by-step prompt instructions.

Next, we randomly selected 200 assignment submissions and corresponding assignment descriptions from the virtual assignment dataset, spanning 85 different assignments from 43 courses. We employed 8 common commercial LLMs from OpenAI (GPT-4.1 nano, GPT-4.1, o4-mini and o3) and Google (Gemini 2.0 Flate-Lite, Gemini 2.0 Flash, Gemini 2.5 Flash Preview and Gemini 2.5 Pro Preview) as LLM tutors to give feedback to synthetic assignment submissions according to corresponding assignment description, where each LLM generate 25 feedback instances. We chose these LLMs to represent a wide variety of current common LLMs, spanning from old models to the newest models, from weak models to state-of-the-art models, from non-reasoning mod-

els to reasoning models. All reasoning models were used with high reasoning effort hyper-parameter and temperature of all models were set to 0 to achieve the highest performance.

We recruited 3 researchers who had did research in LLM-generated educational feedback and feedback effectiveness to as coders of the human-labelled feedback quality dataset. Coders first discussed and concluded a detailed labelling rubric, then labelled 16 feedback instances, 2 for each of 8 LLM tutors, in the inter-rater reliability test. The coders passed the test with Fleiss' Kappa =0.77, and then labelled the rest of dataset in the way that feedback instances generated by each LLM tutors was evenly allocated to 3 coders. During the whole labelling process, coders used hallucination labels and explanation of label results generated by o3 and o3-pro as reference to increase labelling result accuracy.

Except for assigning feedback quality labels, we asked coders to provide explanations of their labelling results when they felt these results are difficult to judge, especially when labelling the hallucination dimensions. As a results, 45 labelled feedback instances are provided with explanations of labelling results. These explanatory labelling data can be useful in fine-tuning LLMs to label feedback quality dimensions, as explanatory data can provide the model with the basis for generating its final output, thereby improving its comprehension and transparency (Krishna et al. 2023; Kassner et al. 2023; Nirmal et al. 2024; Jiang et al. 2024).

Experiments

LLM Feedback Evaluators

We used the labelling rubric as part of the prompt inputs to make LLMs serve as feedback evaluators of LLM tutor feedback. We tested 4 LLMs as LLM tutor feedback evaluators: GPT-4.1, o4-mini, o3 and o3-pro, which were commonly-used advanced LLMs of OpenAI.

We explored zero-shot prompting, few-shot prompting and ways of fine-tuning to employ LLMs as feedback evaluators. We split the human-labelled feedback quality dataset into a train dataset for fine-tuning (where the explanatory data mentioned above are also included) and a test dataset for testing the performance of each LLM feedback evaluator. We ensured that feedback generated by different LLM tutors are evenly distributed into the train and test dataset.

For zero-shot prompting LLMs as feedback evaluators, we first provided the synthetic assignment description and corresponding synthetic student submission at the beginning of prompts, then included our labelling rubric, where examples of labelled feedback were excluded, and finally restricted the output format. The different between the few-shot prompts and the zero-shot prompts was that the few-shot prompts included examples of feedback for different labelling results of the labelled feedback dimensions, with explanations of the labelling results. We illustrated the labelling rubric prompt for the first feedback quality dimension, Alignment with goals, of zero-shot prompts and few-shot prompts, respectively, as follows:

```
1. Alignment with goals: Alignment with
goals evaluates to what extent most
comments are aligned to specific goals.
Score 0: No reference to assignment
goals in specific comments.
Score 1: References goals partly and
vaguely.
Score 2: References goals clearly in
most comments.
```

Listing 1: Zero-shot prompt for labelling Alignment with goal dimension.

```
1. Alignment with goal: Alignment with
goals evaluates to what extent most
comments are aligned to specific
learning goals.
Score 0: No reference to assignment
goals in specific comments.
Example Feedback:
"Nice work|keep going!"
(No reference to any assignment goals or
criteria.)
Score 1: References goals partly and
vaguely.
Example Feedback:
"Your introduction is okay, but make it
stronger to add depth."
(Mentions a goal vaguely but does not
tie it to a specific learning outcome or
stated goal.)
Score 2: References goals clearly in
most comments.
Example Feedback:
"Excellent work linking your data
analysis to the learning outcome of
critically evaluating statistical
models; each example demonstrates how
you tested and interpreted the
hypothesis."
"Please also review the FIT citation
style as in-text citation is missing in
your report."
(Most comments are explicitly framed in
objective-language and aligned to the
rubric or task requirements or other
goals for improvement.)
```

Listing 2: Few-shot prompt for labelling Alignment with goal dimension.

As for experiments of fine-tuned LLM feedback evaluators, we investigated three practices in the literature:

- **Fine-tuning with plain-labelled instances.** In this way, we formed the fine-tuning dataset with 100 human labelling instances in the train dataset that taking few-shot prompts as training inputs and using human labels of feedback quality as expected outputs. This is the com-

LLM Feedback Evaluator	Overall Accuracy	Overall F1	Content Accuracy	Content F1	Effectiveness Accuracy	Effectiveness F1	Hallucinations Accuracy	Hallucinations F1
Zero-shot Prompting								
GPT-4.1	0.734	0.738	0.607	0.360	0.811	0.756	0.807	0.519
o4-mini	0.738	0.742	0.652	0.391	0.780	0.726	0.813	0.536
o3	0.741	0.742	0.603	0.358	0.816	0.741	0.833	0.632
o3-pro	0.744	0.745	0.607	0.366	0.821	0.738	0.837	0.635
Average	0.739	0.742	0.617	0.369	0.808	0.740	0.822	0.580
Few-shot Prompting								
GPT-4.1	0.733	0.737	0.598	0.366	0.816	0.756	0.810	0.533
o4-mini	0.749	0.752	0.657	0.388	0.807	0.743	0.800	0.527
o3	0.735	0.736	0.595	0.356	0.811	0.729	0.837	0.664
o3-pro	0.743	0.744	0.590	0.350	0.824	0.745	0.860	0.703
Average	0.740	0.742	0.610	0.365	0.815	0.743	0.827	0.607
Fine-tuned GPT-4.1								
with plain-labelled data	0.798	0.794	0.737	0.532	0.843	0.725	0.813	0.679
with explanatory data	0.721	0.705	0.675	0.269	0.799	0.667	0.633	0.590
with both datasets	0.703	0.674	0.675	0.269	0.810	0.687	0.507	0.499
Human coders								
Coder 1	0.804	0.850	0.678	0.743	0.838	0.880	0.978	0.969
Coder 2	0.779	0.818	0.667	0.720	0.790	0.805	0.978	0.969
Coder 3	0.767	0.810	0.633	0.678	0.810	0.846	0.933	0.900
Average	0.783	0.826	0.659	0.714	0.813	0.844	0.963	0.946

Table 1: Summary of performance of LLM feedback evaluators based on human-labelled test dataset. Performance of human coders were concluded in the inter-rater test, as discussed in Method section.

mon practice to fine-tune LLMs for automated scoring tasks (Latif and Zhai 2024).

- **Fine-tuning with explanatory data instances.** We formed the fine-tuning dataset with 45 explanatory instances in the train dataset. We followed the practice from the literature to fine-tune LLM with explanatory data (Krishna et al. 2023; Bilal, Ebert, and Lin 2025). For training inputs, we added an additional instruction to the end of few-shot prompts that “Here you are free to briefly explain some labelling results that you think might be not straight-forward to others. Specially, I suggest that you explain your label of hallucinations here if and only if you labelled some hallucinations above.” This would allow the LLM to output explanations of difficult labelling decisions especially in hallucination detection, which was what human coders did in their explanatory instances. For expected training outputs, we used human labels of feedback quality followed by human explanations.
- **Fine-tuning with both plain labelled instances and explanatory data instances.** Another practice is to fine-tuning on a dataset with a mixture of a small subset of explanatory data and a larger amount of plain-labelled data using different instructions (Longpre et al. 2023; Liu et al. 2024; Zhang et al. 2025). And inspired by Liu et al. (2024), we used a separate training method to fine-tune the dataset. First, we randomly chose half of explanatory data instances to fine-tune the LLM, followed by fine-tuning with all plain-labelled instances, and finally fine-tuned in the other half of explanatory data instances.

We applied these fine-tuning methods in GPT-4.1, and set hyper-parameter $n_{epochs} = 2$.

All results of our LLM feedback evaluators are shown in

Table 1. The results show that o3-pro demonstrated the best performance in zero-shot labelling of feedback (Overall Accuracy 74.3%, F1-score 74.5%) while o4-mini demonstrated the best performance in few-shot labelling of feedback (Overall Accuracy 74.9%, F1-score 75.2%). In both prompting methods, o3-pro achieved the highest accuracy in grading feedback effectiveness (zero-shot Accuracy 82.1%, few-shot Accuracy 82.4%) and hallucinations dimensions (zero-shot Accuracy 83.7%, few-shot Accuracy 86.0%), while o3 achieved the second highest accuracy in both prompting methods (Accuracy – 83.3% zero-shot and 83.7% few-shot). On the other hand, the 4 tested LLMs shared the same level of performance in evaluating LLM tutor feedback quality, as the highest accuracy model only outperformed the lowest accuracy model by 1% with zero-shot prompting setting, and 1.6% with few-shot prompting setting.

The few-shot prompting method did not outperform the zero-shot method in our experiment, as the average overall F1 of 4 LLMs using few-shot prompting (74.2%) is the same as the average overall F1 using zero-shot prompting (74.2%), and there was no significant improvements in average Accuracy and F1 of all three feedback quality aspects. However, o3-pro benefited from a notable performance improvement switching from zero-shot prompting to few-shot prompting in detection of hallucinations, with an increase of 2.3% in Accuracy and 6.8% in F1, while o3-pro is already the best-performance model with zero-shot prompting.

Moreover, GPT-4.1 was largely improved by fine-tuning with plain-labelled instances, with human expert level performance after fine-tuning (fine-tuned GPT-4.1 Overall Accuracy 79.8%, F1-score 79.4%; human average Accuracy 78.3%, F1-score 82.6%). However, the fine-tuned model that was trained with explanatory data (Overall Accuracy 72.1%, F1 70.5%) showed even worse performance than the origi-

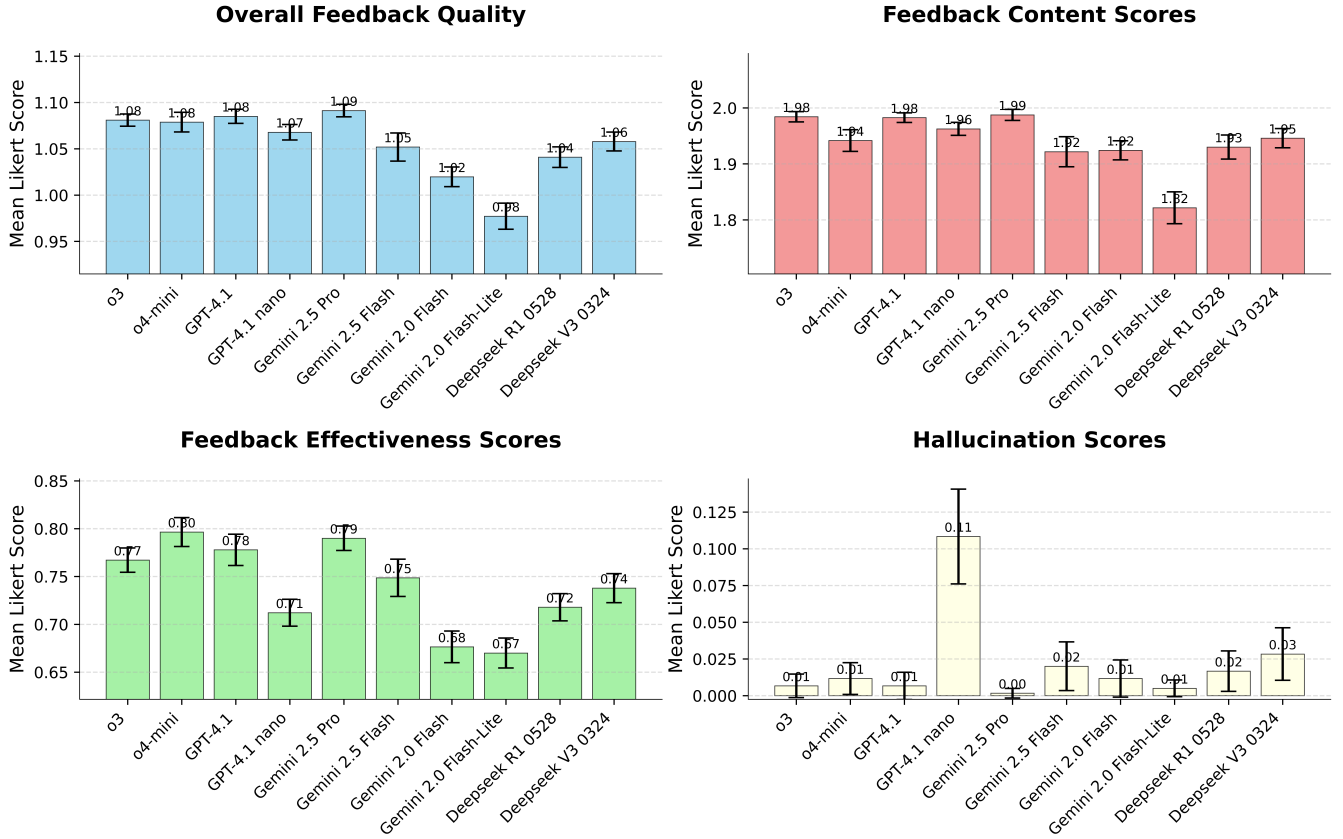


Figure 2: Feedback quality scores of different LLM Tutors. Overall feedback quality stands for mean Likert score of all feedback quality dimensions (except hallucinations) in our feedback evaluation framework. Other scores are the mean Likert scores of feedback quality dimensions within a specific aspect shown in the titles.

nal GPT-4.1 model (Accuracy 73.4%, F1 73.8%). Furthermore, the fine-tuned model trained with both plain-labelled data and explanatory data (Accuracy 70.3%, F1 67.4%) performed even worse than simply training on explanatory data. The accuracy and F1 of these two fine-tuned models trained on explanatory data decreased the most in detection of Hallucinations dimensions, where their accuracy was decreased by 17.7% and 30.3% after fine-tuning with explanatory data and both datasets, respectively.

Comparing Feedback across LLMs

We used the best-performance feedback quality evaluation model found in the last experiment, which is the model fine-tuned with plain-labelled data, to serve as a *Dean* of LLM tutors (in other words, a *DeanLLM*) to compare quality of feedback generated by different LLMs as LLM tutors.

To do this, we employed 10 common commercial LLMs – OpenAI (GPT-4.1 nano, GPT-4.1, o4-mini-high and o3-high), Google (Gemini 2.0 Flash-Lite, Gemini 2.0 Flash, Gemini 2.5 Flash and Gemini 2.5 Pro) and Deepseek (Deepseek R1 0528 and Deepseek V3 0324). Each model generated 100 feedback instances for the synthetic assignment submissions, and also generated 100 feedback instances for their corresponding original real assignments,

resulting in 2,000 feedback samples - i.e, 1,000 for synthetic assignment submissions and 1,000 for the real ones. We chose these LLMs to represent a wide variety of current common LLMs. We then used our DeanLLM to evaluate these feedback. The results are illustrated in Figure 2.

From Figure 2, Gemini 2.5 Pro achieved the highest overall feedback quality in our feedback evaluation framework, but the difference between Gemini 2.5 Pro (mean Likert score 1.09) and GPT-4.1 (mean Likert score 1.08) is not significant. Almost all tested model were at the same level of performance in overall feedback quality, except Gemini 2.0 Flash and Gemini 2.0 Flash-Lite: Gemini 2.0 Flash was demonstrated to be significantly worse than other 8 LLM tutors while Gemini 2.0 Flash-Lite was significantly worse than other 9 LLM tutors including Gemini 2.0 Flash in overall feedback quality, according to 95 Confidence Interval error bars shown in the figure. These two LLMs are also significantly worse than others in feedback effectiveness scores, and Gemini 2.0 Flash-Lite is also significantly worse than other LLM tutors in the feedback content scores.

Hallucination scores with higher values indicating more hallucination issues, GPT-4.1 nano resulted in significantly more hallucinations than any other LLM when acting as an LLM tutor, with a hallucination detection rate of 11% on

average of detection rates of 3 hallucination types. In contrast, the hallucination detection rate for other 9 LLMs was relatively small ($\leq 3\%$ for each of 3 hallucination type), and Gemini 2.5 Pro showed no hallucination issues in its 200 feedback instances, significantly better than any other LLMs.

Discussion

As o3-pro is known as the most powerful reasoning model among the tested 4 models, and o3 is the second most powerful reasoning model, the result in Table 1 that o3-pro demonstrates the best performance in detection of different types hallucinations in LLM tutor feedback and o3 wins the second place on that aligns with previous literature that chain-of-thought reasoning improves hallucination detection (Akbar et al. 2024; Wang et al. 2025) and automated scoring (Lee et al. 2024). Since avoiding hallucinations in feedback to students is crucial for meeting the reliability, safety, and trustworthiness requirements of educational feedback (Alfredo et al. 2024), we recommend to choose powerful reasoning models to serve as the DeanLLMs, and expect the rapid improvement of reasoning models would bring us better LLM feedback evaluators under our comprehensive LLM tutor feedback evaluation framework. Moreover, we note that we did not investigate use of step-by-step thinking prompt design or chain-of-thought prompting in the current work, future explorations of these prompt engineering methods are needed, which may improve especially in hallucination detection dimensions according to our results.

The result that few-shot prompting did not outperform the zero-shot method in our experiment aligns with literature that providing examples does not always lead to improvements and zero-shot prompts can actually outperform few-shot prompts in some tasks (Li 2023). As illustrated in listing 1 and Listing 2, the length of few-shot rubric was around 3 times the length of zero-shot rubric, which added complexity for LLMs to understand the rubric. Moreover, few-shot prompting can lead to over-fitting the specific cases provided in the prompt (Yang et al. 2022). However, the contradiction result that few-shot prompt significantly facilitated o3-pro’s ability in hallucination detection even though o3-pro already performed the best in zero-shot setting reveals that the impact of same prompt for different models are different. As the model with the highest reasoning effort we included in our experiments, o3-pro may benefit from a more complex, long instructions in the prompt while other weaker models may not. This demonstrates the need for more experiments in prompt designs and involvement of other state-of-the-art reasoning models like Gemini 2.5 Pro in future explorations.

The reason why fine-tuning with explanatory instances led to sharp drop in evaluation accuracy also remains an unsolved question for future investigations, as we learnt from the literature that fine-tuning with explanatory data can improve LLM performance (Longpre et al. 2023; Liu et al. 2024; Zhang et al. 2025) and fine-tuning with mixed prompt settings improves LLM’s reasoning abilities (Chung et al. 2024; Wei et al. 2021). One possible explanation is that we fine-tuned GPT-4.1 on a small dataset of only 45 samples, for a separate prompt that only added one instruction to the

end of our long few-shot prompt but desired a quite different expected output, this confused the fine-tuned LLM regarding the meaning of the few-shot prompt instructions and led to decrease in evaluation accuracy when testing with the few-shot prompt. Future experiments involving other prompting methods and better models in reasoning (e.g., o3-pro) can be conducted to verify this explanation.

According to evaluation results of feedback generated by different LLMs, Gemini 2.5 Pro seems to be the best choice for its highest feedback quality and 0 detection rates of hallucination issues in feedback. In contrast, the smallest and cheapest LLMs involved in the experiment, GPT-4.1 nano and Gemini 2.0 Flash-Lite, are the least considerable choices considering feedback quality and hallucination rates, as GPT-4.1 nano is demonstrated to have the highest hallucination detection rate and the overall feedback quality score of Gemini 2.0 Flash-Lite is significantly lower than other 9 LLMs.

Conclusion and Future Work

In this paper, we addressed the critical challenge of ensuring the quality and safety of feedback from LLM tutors by introducing the ‘Dean of LLM Tutors’ (*DeanLLM*) framework, an automated evaluation system that uses an LLM to screen feedback before it reaches students. We established a comprehensive 16-dimension evaluation framework covering content, effectiveness, and hallucinations, and created a corresponding synthetic dataset to facilitate research. Our key finding is that LLMs can serve as effective evaluators; specifically, a fine-tuned GPT-4.1 model achieved human-expert-level performance in this role. By applying this top-performing *DeanLLM*, we benchmarked ten common commercial LLM tutors and identified Gemini 2.5 Pro as providing feedback with the highest quality and zero detected hallucinations. This work validates a practical and scalable approach to systematically ensure that AI-generated educational feedback is both reliable and of high pedagogical value.

Looking ahead, several promising avenues for future research emerge from this work. An essential direction is to enhance the *DeanLLM*’s capabilities by exploring advanced prompt engineering techniques like chain-of-thought to improve grading accuracy of llm feedback quality and hallucination detection. Another critical area for investigation is the unexpected performance decrease observed when fine-tuning with explanatory data to optimise our training methods. Moreover, a key priority is to evolve the current system into a closed-loop model where the *DeanLLM*’s evaluation automatically directs the LLM tutor to revise and improve its output. Finally, a crucial next step is to deploy and test this framework in a live educational setting, allowing us to conduct student studies that measure the real-world impact of quality-controlled AI feedback on learning outcomes, student engagement, and trust in AI-powered educational tools.

References

- Akbar, S. A.; Hossain, M. M.; Wood, T.; Chin, S.-C.; Salinas, E. M.; Alvarez, V.; and Cornejo, E. 2024. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15020–15037.
- Alfredo, R.; Echeverria, V.; Jin, Y.; Yan, L.; Swiecki, Z.; Gašević, D.; and Martinez-Maldonado, R. 2024. Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 100215.
- Balse, R.; Valaboju, B.; Singhal, S.; Warriem, J. M.; and Prasad, P. 2023. Investigating the potential of gpt-3 in providing feedback for programming assessments. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 292–298.
- Banerjee, S.; Agarwal, A.; and Singla, S. 2024. LLMs Will Always Hallucinate, and We Need to Live With This. *arXiv preprint arXiv:2409.05746*.
- Bennett, N.; and Kell, J. 1989. *A good start?: four year olds in infant schools*. Basil Blackwell.
- Bilal, A.; Ebert, D.; and Lin, B. 2025. LLMs for explainable ai: A comprehensive survey. *arXiv preprint arXiv:2504.00125*.
- Brookhart, S. M. 2017. *How to give effective feedback to your students*. Ascd.
- Cavalcanti, A. P.; Diego, A.; Mello, R. F.; Mangaroska, K.; Nascimento, A.; Freitas, F.; and Gašević, D. 2020. How good is my feedback? a content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge*, 428–437.
- Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.
- Dai, W.; Cheng, Y.; Aldino, A. A.; Tsai, Y.-S.; Gašević, D.; and Chen, G. 2025. Evaluating the Capability of Large Language Models in Characterising Relational Feedback: A Comparative Analysis of Prompting Strategies. *Computers and Education: Artificial Intelligence*, 100427.
- Dai, W.; Tsai, Y.-S.; Lin, J.; Aldino, A.; Jin, H.; Li, T.; Gašević, D.; and Chen, G. 2024a. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7: 100299.
- Dai, W.; Tsai, Y.-S.; Lin, J.; Aldino, A.; Jin, H.; Li, T.; Gašević, D.; and Chen, G. 2024b. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7: 100299.
- Denny, P.; Prather, J.; Becker, B. A.; Finnie-Ansley, J.; Hellas, A.; Leinonen, J.; Luxton-Reilly, A.; Reeves, B. N.; Santos, E. A.; and Sarsa, S. 2024. Computing education in the era of generative AI. *Communications of the ACM*, 67(2): 56–67.
- Derham, C.; Balloo, K.; and Winstone, N. 2022. The focus, function and framing of feedback information: Linguistic and content analysis of in-text feedback comments. *Assessment & Evaluation in Higher Education*, 47(6): 896–909.
- Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.
- Fang, J.; Fleck, M.; Green, A.; McVilly, K.; Hao, Y.; Tan, W.; Fu, R.; and Power, M. 2011. The response scale for the intellectual disability module of the WHOQOL: 5-point or 3-point? *Journal of Intellectual Disability Research*, 55(6): 537–549.
- Goodman, J. S.; Wood, R. E.; and Hendrickx, M. 2004. Feedback specificity, exploration, and learning. *Journal of applied psychology*, 89(2): 248.
- Hattie, J.; and Timperley, H. 2007. The power of feedback. *Review of educational research*, 77(1): 81–112.
- Hellas, A.; Leinonen, J.; Sarsa, S.; Koutchme, C.; Kujanpää, L.; and Sorva, J. 2023. Exploring the responses of large language models to beginner programmers’ help requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, 93–105.
- Jiang, D.; Wang, G.; Lu, Y.; Wang, A.; Zhang, J.; Liu, C.; Van Durme, B.; and Khashabi, D. 2024. Rationalyst: Pre-training process-supervision for improving reasoning. *arXiv preprint arXiv:2410.01044*.
- Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; and McHardy, R. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.
- Kassner, N.; Tafjord, O.; Sabharwal, A.; Richardson, K.; Schuetze, H.; and Clark, P. 2023. Language models with rationality. *arXiv preprint arXiv:2305.14250*.
- Kim, H.; Hwang, J.; Kim, T.; Choi, M.; Lee, D.; and Ko, J. 2025. Impact of Generative Artificial Intelligence on Learning: Scaffolding Strategies and Self-Directed Learning Perspectives. *International Journal of Human-Computer Interaction*, 1–23.
- Krishna, S.; Ma, J.; Slack, D.; Ghandeharioun, A.; Singh, S.; and Lakkaraju, H. 2023. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 36: 65468–65483.
- Lakera AI. 2024. Guide to Hallucinations in Large Language Models. Accessed: 2024-06-04.
- Latif, E.; and Zhai, X. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6: 100210.
- Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; and Zhai, X. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6: 100213.
- Leinonen, J.; Hellas, A.; Sarsa, S.; Reeves, B.; Denny, P.; Prather, J.; and Becker, B. A. 2023. Using large language models to enhance programming error messages. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 563–569.

- Li, Y. 2023. A practical survey on zero-shot prompt design for in-context learning. *arXiv preprint arXiv:2309.13205*.
- Liffiton, M.; Sheese, B. E.; Savelka, J.; and Denny, P. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, 1–11.
- Lin, J.; Chen, E.; Han, Z.; Gurung, A.; Thomas, D. R.; Tan, W.; Nguyen, N. D.; and Koedinger, K. R. 2024. How can i improve? using gpt to highlight the desired and undesired parts of open-ended responses. *arXiv preprint arXiv:2405.00291*.
- Liu, J.; Huang, Z.; Xiao, T.; Sha, J.; Wu, J.; Liu, Q.; Wang, S.; and Chen, E. 2024. SocraticLM: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37: 85693–85721.
- Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 22631–22648. PMLR.
- Middleton, T.; Ahmed Shafi, A.; Millican, R.; and Templeton, S. 2023. Developing effective assessment feedback: Academic buoyancy and the relational dimensions of feedback. *Teaching in Higher Education*, 28(1): 118–135.
- Nguyen, A.; Piech, C.; Huang, J.; and Guibas, L. 2014. Codewebs: scalable homework search for massive open online programming courses. In *Proceedings of the 23rd international conference on World wide web*, 491–502.
- Nicol, D. J.; and Macfarlane-Dick, D. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2): 199–218.
- Nirmal, A.; Bhattacharjee, A.; Sheth, P.; and Liu, H. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. *arXiv preprint arXiv:2403.12403*.
- of Sciences, N. A.; Medicine; on Engineering, D.; Sciences, P.; Science, C.; Board, T.; Affairs, G.; on Higher Education, B.; and on the Growth of Computer Science Undergraduate Enrollments, C. 2018. *Assessing and responding to the growth of computer science undergraduate enrollments*. National Academies Press.
- Osakwe, I.; Chen, G.; Whitelock-Wainwright, A.; Gašević, D.; Cavalcanti, A. P.; and Mello, R. F. 2022. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, 3: 100059.
- Pankiewicz, M.; and Baker, R. S. 2023. Large Language Models (GPT) for automating feedback on programming assignments. *arXiv preprint arXiv:2307.00150*.
- Phung, T.; Cambronero, J.; Gulwani, S.; Kohn, T.; Majumdar, R.; Singla, A.; and Soares, G. 2023a. Generating high-precision feedback for programming syntax errors using large language models. *arXiv preprint arXiv:2302.04662*.
- Phung, T.; Pădurean, V.-A.; Cambronero, J.; Gulwani, S.; Kohn, T.; Majumdar, R.; Singla, A.; and Soares, G. 2023b. Generative AI for programming education: benchmarking ChatGPT, GPT-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*, 41–42.
- Prather, J.; Denny, P.; Leinonen, J.; Becker, B. A.; Albluwi, I.; Craig, M.; Keuning, H.; Kiesler, N.; Kohn, T.; Luxton-Reilly, A.; et al. 2023. The robots are here: Navigating the generative ai revolution in computing education. In *Proceedings of the 2023 working group reports on innovation and technology in computer science education*, 108–159. ACM.
- Rawte, V.; Sheth, A.; and Das, A. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.
- Rudolph, J.; Tan, S.; and Tan, S. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching*, 6(1): 342–363.
- Ryan, T.; Henderson, M.; Ryan, K.; and Kennedy, G. 2023. Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, 28(7): 1565–1582.
- Sangaré, M. J.-P. 2023. ChatGPT is Doomed. Accessed: 2024-06-12.
- Scarlatos, A.; Smith, D.; Woodhead, S.; and Lan, A. 2024. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, 280–294. Springer.
- Shute, V. J. 2008. Focus on formative feedback. *Review of educational research*, 78(1): 153–189.
- Wang, C.; Su, W.; Ai, Q.; and Liu, Y. 2025. Joint Evaluation of Answer and Reasoning Consistency for Hallucination Detection in Large Reasoning Models. *arXiv preprint arXiv:2506.04832*.
- Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Yan, L.; Greiff, S.; Teuber, Z.; and Gašević, D. 2024. Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8(10): 1839–1850.
- Yang, J.; Jiang, H.; Yin, Q.; Zhang, D.; Yin, B.; and Yang, D. 2022. SEQZERO: Few-shot compositional semantic parsing with sequential prompts and zero-shot models. *arXiv preprint arXiv:2205.07381*.
- Zhang, M.; Press, O.; Merrill, W.; Liu, A.; and Smith, N. A. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.
- Zhang, X.; Chen, Z. Z.; Ye, X.; Yang, X.; Chen, L.; Wang, W. Y.; and Petzold, L. R. 2025. Unveiling the impact of

coding data instruction fine-tuning on large language models reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25949–25957.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023b. Siren’s song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.