Competitive LLM Assessment: Using Large Language Models for University Exam Evaluations through Iterative Pairwise Comparisons

Bachelor's Thesis of

Isik Baran Sandan

Artificial Intelligence for Language Technologies (AI4LT) Lab
Institut for Anthropomatics and Robotics (IAR)
KIT Department of Informatics

Reviewer:	Prof. Dr. Jan Niehues
Second reviewer:	Prof. Dr.-Ing. Tamim Asfour
Advisor:	M.Sc. Tu Anh Dinh

28. October 2024 – 28. February 2025

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

. . . . . .

(Isik Baran Sandan)

Abstract

Large Language Models (LLMs) have been shown to be good evaluators in the scientific domain. However, while LLMs perform overall well in grading university-level exams with a Pearson correlation of 0.948 on exam-level, question-level correlation remains lower around 0.6, indicating a lower alignment between expert grading and LLM grading on grading individual answers. One possible limiting factor for the low question level accuracy is the current LLM-as-a-Judge approaches relying on individual assessments, thus preventing the judge LLM from having a global ranking perspective of all possible responses.

In order to address this challenge, this study presents an LLM-as-a-Judge method called competitive assessment, in which a knockout tournament system with iterative pairwise comparisons is used. Instead of evaluating each answer in isolation, this methodology uses repeated pairwise comparisons by providing the LLM with two answers each time. The answers which are deemed stronger advance, and the final score of each response is derived from the average of all evaluations.

Our experiments conducted across three different LLMs on two different datasets provide strong evidence that competitive assessment improves grading accuracy, increasing Pearson correlation with expert evaluations by up to 0.09, thereby aligning LLM assessments more closely with human grading.

Zusammenfassung

Large Language Models (LLMs) haben sich im wissenschaftlichen Bereich als gute Evaluatoren erwiesen. Während LLMs bei der Bewertung von Universitätsprüfungen mit einer Pearson-Korrelation von 0,948 auf Prüfungsebene insgesamt gut abschneiden, bleibt die Korrelation auf Fragenebene mit etwa 0,6 niedriger, was auf eine geringere Übereinstimmung zwischen der Expertenbewertung und der LLM-Bewertung bei der Bewertung einzelner Antworten hindeutet. Aktuelle LLM-as-a-Judge-Ansätze verlassen sich jedoch normalerweise auf die individuelle Bewertung jeder Antwort, was möglicherweise die Bewertungsfähigkeit der LLMs einschränkt.

Um diese Herausforderung anzugehen, stellt diese Studie eine LLM-als-Judge-Methode namens Competitive Assessment vor, bei der ein KO-Turniersystem mit iterativen Paarvergleichen verwendet wird. Anstatt jede Antwort isoliert zu bewerten, verwendet diese Methode wiederholte Paarvergleiche, indem dem LLM jedes Mal zwei Antworten zur Verfügung gestellt werden. Die Antworten, die als stärker gelten, kommen weiter, und die endgültige Punktzahl jeder Antwort ergibt sich aus dem Durchschnitt aller Bewertungen.

Unsere Experimente, die wir mit drei verschiedenen LLMs anhand von zwei unterschiedlichen Datensätzen durchgeführt haben, liefern überzeugende Beweise dafür, dass wettbewerbsorientierte Beurteilungen die Genauigkeit der Notengebung verbessern, indem sie die Pearson-Korrelation mit Expertenbewertungen um bis zu 0,09 erhöhen und dadurch LLM-Beurteilungen stärker an die menschliche Notengebung anpassen.

Contents

List of Figures

List of Tables

1. Introduction

Across various domains, and especially for scientific research, accurate and consistent evaluations are very crucial for informed decision-making. However, the inherent scale and subjectivity make this task very challenging and time-consuming. In recent years, the methodology of "LLM-as-a-Judge" [13] has emerged to tackle this challenge, where instead of humans, Large Language Models (LLMs) take the role of the expert to evaluate complex tasks. Using LLMs as evaluators allows us to mimic the abilities of human experts, making evaluations cost-effective and scalable.

Although many approaches to LLM-as-a-Judge exist, the most common is individual assessment, in which the evaluation prompt consists of only the question and the corresponding answer, which is to be evaluated [1]. While this approach has already shown to yield good evaluation results next to providing scalability [1, 3], it does not consider the relative strength of answers in a set to a given question. The more recent approach of pairwise assessment tries to address this issue by providing two responses to the judge LLM each time, however, it still fails to account for a global ranking perspective, as pairwise comparisons do not analyze how all responses compare to each other in the broader sense.

This thesis presents an LLM-as-a-Judge method called **competitive assessment** to address this challenge, which can be seen as a variation of the tournament system used in the "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena" paper [13], differing only in how it makes use of multiple pairwise comparisons. Instead of isolating responses individually or in pairs for evaluation, competitive assessment focuses on an iterative process where responses are compared against one another multiple times in a tournament manner. In each round, stronger responses advance to compete against each other in later rounds, allowing us to refine the scores progressively throughout the tournament. This approach allows the judge LLMs to develop a global perspective on responses without requiring all replies to be included in a single prompt, which would otherwise result in an impractically long context length.

We conducted experiments on two distinct datasets to ensure that our new LLM-as-a-Judge method generalizes beyond a specific domain. The primary dataset we focused on is SciEx, a dataset consisting of university exams and answers, which were graded by human experts. Here, our experiments will analyze if competitive assessment in fact improves grading accuracy compared to individual assessment, and identify what factors influence the performance of the grading LLMs. The second dataset is MT-Metrics-Eval from WMT Metrics Shared Task [4], in which we aim to validate our findings from SciEx dataset and analyze the performance of competitive assessment on evaluating Machine Translations

(MTs).

To ensure that competitive assessment remains effective across different LLMs of different capabilities, we will conduct each experiment with three different open-source models of various sizes. Furthermore, we will use a simplified version of the debiasing methodology introduced in the "LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models" paper [6] for competitive assessment to account for the potential bias introduced by the order of the answers and analyze how much this potential bias influences the performance of our methodology. Thus, our experiments will aim to analyze if competitive assessment is a better LLM-as-a-Judge method than the already existing ones.

For this, chapter 2 will cover all preliminary knowledge needed and the state of related research. The following chapter 3 will present our hypothesis and focus on the proposed methodology of competitive assessment. Our experiments, their results and analysis thereof will be presented in chapter 4. The final chapter 5 will conclude our findings.

2. Background and Related Work

In this section, the fundamental concepts and methods are described which are essential to understand our experiments. First, we will provide basic knowledge about Large Language Models (LLMs) and Natural Language Generation (NLG). Afterwards, we will elaborate the relatively new method of LLM-as-a-Judge and the current approaches for this method.

2.1. Large Language Models

Large Language Models can be defined as language-based artificial intelligence systems that can process and generate text and generalize to different tasks [7]. These tasks include machine translation, summarization, information retrieval, and question-answering.

The general purpose of such a system is to predict the next possible tokens given a sequence of input tokens, where tokens can be characters, symbols, sub words or words that make up the language [11]. Such a model usually excels at generating text similar to what it has seen in training data, however it will have difficulty handling tasks it has not encountered before. This challenge is known as Zero-Shot-Learning, where an LLM must handle tasks it has never seen before.

Regular LLMs which solely focus on next word generation are thus not very good at Zero-Shot tasks. Instead, often so-called instruction-models are used to handle such tasks, which were fined tuned on data, structured with instructions and input-output-pairs [2]. This teaches the model to respond to the user prompt and not just follow it up with the most probable next word and therefore improves zero-shot generalization.

In our thesis, we use such instruction-tuned-models of different sizes, as we need the language model to be able to perform the user instructions (namely grading exam answers or evaluating machine translations), without the need for few-shot prompting. Avoiding few-shot prompting is crucial for our thesis, as it would increase the context-length by multiple factors and thus might result in a context-window overflow and cause drops in performance.

2.2. LLM-as-a-Judge

LLM-as-a-Judge is the process in which the evaluation of a response to a prompt or question is carried out by an LLM. This method was introduced in the "Judging LLM-as-a-Judge with MT-Bench" paper [13] as an alternative to human evaluation for the benefits of scalability and explainability. The scalability aspect arises from the fact that human evaluations take more time and cost, and the explainability aspect is a result of being able to ask for an explanation for each evaluation.

In the following subsections, the different existing methods for LLM-as-a-Judge will be explained, before presenting our alternative method for LLM-as-a-Judge in Chapter 3.

2.2.1. Individual Assessment

One approach to LLM-as-a-Judge is individual assessment, where the judge LLM is provided with a prompt or a question, the corresponding answer, and the scoring criteria and is asked to provide an evaluation such as a grade. There have been various studies so far which have used this method with state-of-the-art LLMs in the time of the study such as GPT4, in order to evaluate the quality of story generation [1] or score quality of different texts according to different criteria [10]. Both of these studies showed that LLMs can provide comparable judgment to human evaluations, however the tests were conducted on tasks such as summarization and story generation, which modern LLMs excel at, and not on difficult reasoning-based tasks such as university questions.

The recent SciEx paper of Dinh et al. [3] has addressed this issue, where they made use of individual assessment to grade university-level exam question-answer pairs, and showed how the judge LLM graded the answers similarly to human-experts in the domain, especially in the exam level. The study made use of Zero-Shot and Few-Shot prompting, but only used individual grading, thus leaving room for further research.

2.2.2. Pairwise Assessment

There have also been recent studies that have focused on another LLM-as-a-Judge approach, namely pairwise prompting, in which the LLM judge is provided with two responses to the prompt, compared to the single response in individual assessment. The judge LLM is then asked to evaluate both responses. This has shown to be an effective LLM-as-a-Judge method for ranking documents, as it gives the judge LLM direct comparison points while making evaluations [6]. However, as in most Zero-Shot LLM-as-a-Judge studies, these documents were based on tasks that LLMs are good at, and so far, there has not been a study where the LLM had to evaluate reasoning-based responses using pairwise prompting.

2.2.3. Chatbot Arena

To address the issue of LLMs not developing a global perspective over the data while evaluating, when individual assessments or single pairwise assessments are used, the paper "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena" [13] introduced a new LLM-as-a-Judge method. For a given set of answers to a question, all possible answer pairs are evaluated against one another and the score for an answer is assigned using an ELO system. This approach is thus able to make implicit use of a global view over the dataset while assigning scores. However, pairing all possible answers results in a computational time of $O(N^2)$, which is too expensive for our experiments. Thus, our thesis uses an alternative version of this approach with $O(NlogN)$ computation time instead, which is explained further in the following chapter 3.

2.2.4. Sorting Based Approaches

The paper "Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting" [8] introduces two further methods to enable a global ranking of all responses while using LLM-as-a-Judge, both of which address the $O(N^2)$ computation time of all-pairs comparisons used in Chatbot-Arena [13]. First is a Heapsort approach, which uses Heapsort with pairwise comparisons to sort out the possible answers, thus guaranteeing $O(NlogN)$ computation complexity. Second is a sliding window approach, making use of individual passes in the Bubble Sort algorithm for a constant number K times, thus having only $O(N)$ computation complexity.

3. Methodology

3.1. Hypothesis

We have seen from the SciEx paper from Dinh et al. [3] that even if current state-of-the-art Large Language Models (LLMs) do not perform well on solving university level exams, they perform well on grading them as judges. Especially on exam-level grading, the best LLMs were able to achieve a 0.948 Pearson correlation with expert grading. However, this exam-level correlation can mislead one to believe the LLM grades the answers very similarly to human experts, which is untrue, as shown by the lower question-level Pearson correlation around 0.6.

The suggested LLM-as-a-Judge methodology that is used in SciEx focuses on individual assessment, where the grading LLM is given a question and an answer to grade each time. We hypothesize that this leaves a massive gap for improvement by using what we call competitive assessment, an LLM-as-a-Judge method in which the system iteratively uses pairwise comparisons on pairs of answers to the same question in a tournament manner, until it reaches the best answer. This methodology is explained in detail in Section 3.2.

By iteratively comparing the better answers with each other and averaging the grades an answer got throughout the process, we argue that the grading LLM performs better, since this way it can evaluate high-quality responses in relation to other high-quality responses, rather than grading each one in isolation.

In the following we illustrate the design of our experiment in order to increase the grading performance via competitive assessment and motivate the reasoning behind our choices.

3.2. Competitive Assessment

In the following, we present our new LLM-as-a-Judge method called competitive assessment, that uses iterative pairwise comparisons, where each response to the same task competes to be the best response in a knockout tournament manner.

3.2.1. Question-Level Match

Instead of individual assessment, which was used in the original SciEx [3], multiple pairwise assessments are used for competitive assessment. In each pairwise assessment, one pairwise ranking prompt similar to the comparative prompt introduced in the LLM Comparative Assessment paper [6] is used. In each prompt, one question and two responses to that question are provided to the judge LLM, which is asked to evaluate each response. We call this a “question-level-match” and the exact prompts that we used while using

competitive assessment for the SciEx dataset, for both individual grading and a question level match, can be seen in Figure 3.1. The German version of each prompt, and the prompts which were used on the second dataset are in Appendix A.1.

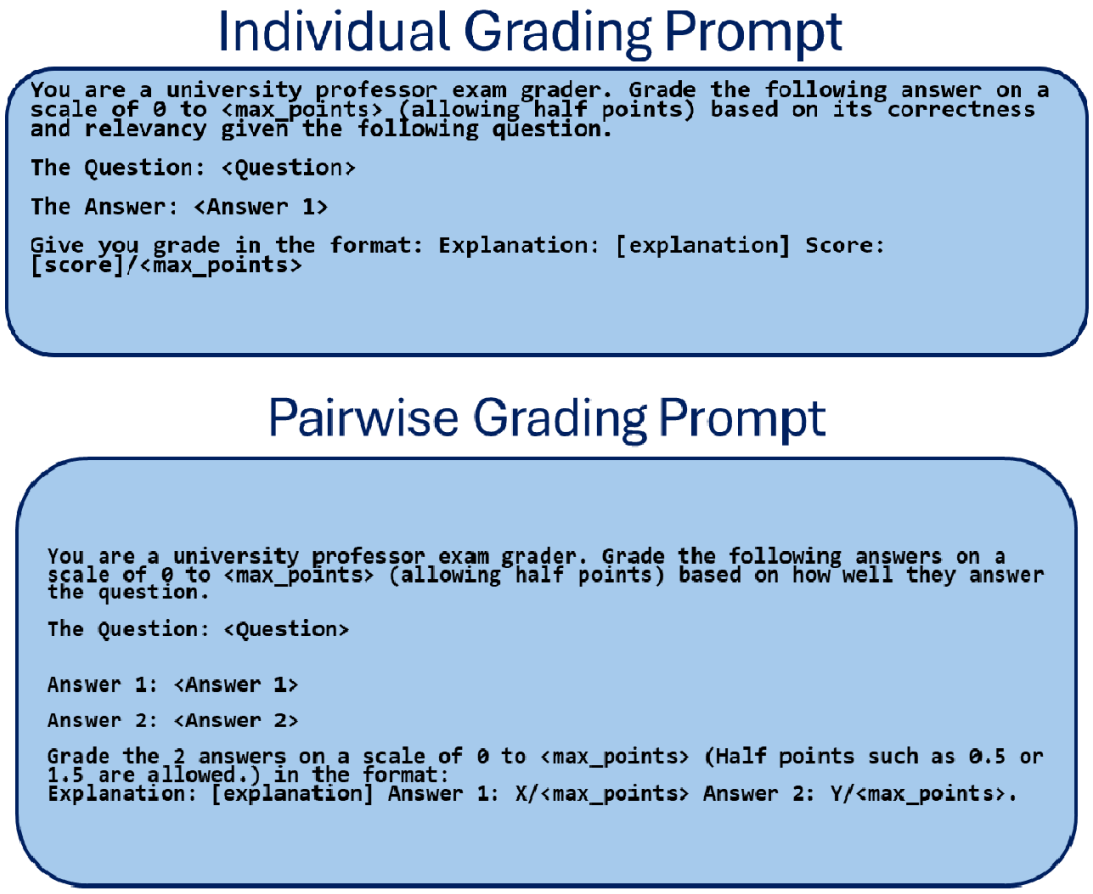

Figure 3.1.: LLM prompts for individual grading and pairwise grading

Furthermore, the judge LLM is asked for an explanation in the prompt before asking for the grades, since this is consistent with the original SciEx prompt and results in Chain-of-Thought prompting, which is shown to improve the grading LLM's ability to perform reasoning [12].

From the response generated by the judge LLM, the grade each individual answer got is parsed and saved to the list of grades for that answer, which keeps track of all the grades an answer got throughout all its question-level-matches. The answer which got the higher-grade advances to the next round to be matched up against another answer. This is depicted in Figure 3.2.

To observe the influence of the order of the competing answers, we sampled results both with and without debiasing in respect to the order of the answers in the prompt. Without debiasing, the process is exactly as shown in Figure 3.2. However, the order of texts in pairwise rankings has shown to be an influential factor in the LLMs decision making [9] , thus we also collected the results with using a debiasing methodology similar

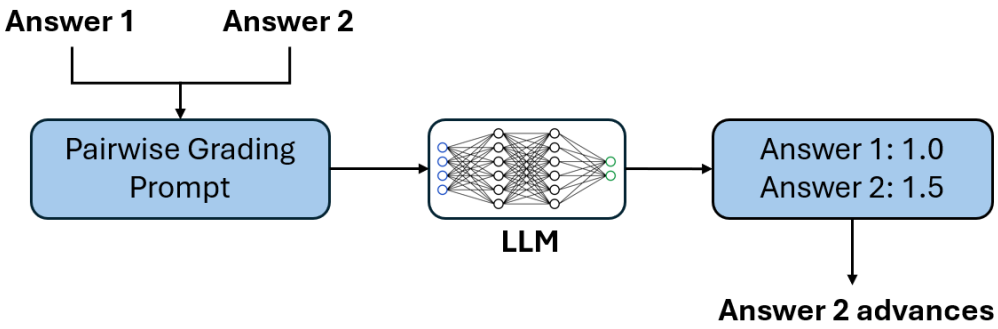

Figure 3.2.: A Question-Level-Match without debiasing

to the one introduced in the "LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models" paper [6], which worked as follows: For each answer pair, we ran the question-level-match first with one answer before the other, and a second time with the other one first. The final grade an answer got for that question-level-match is than the average of the two grades. Debiasing thus results in double the compute-time compared to a regular question-level-match, which only requires a single LLM response per question-level-match. A question-level-match with debiasing is illustrated in Figure 3.3.

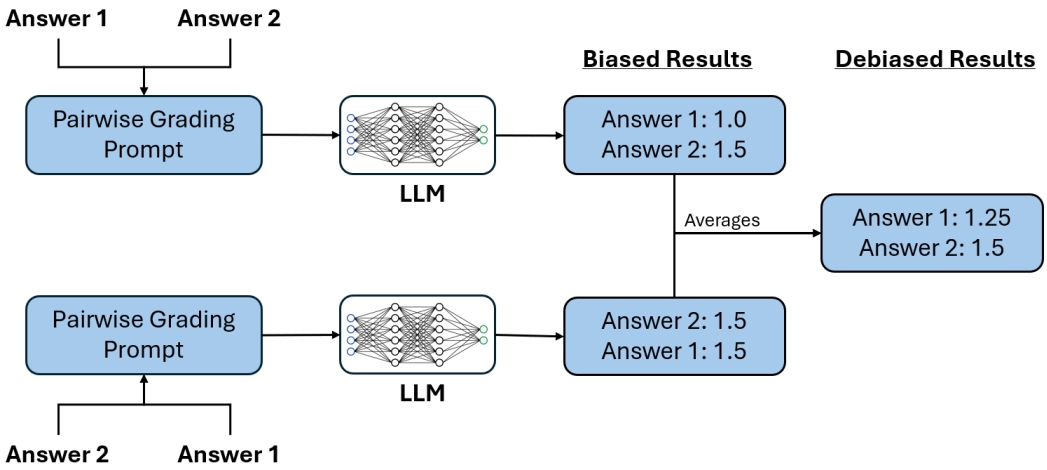

Figure 3.3.: A Question-Level-Match with debiasing

3.2.2. Knockout Tournament

Below is the algorithm of competitive assessment.

Algorithm 1 Competitive Assessment

Input: prompt P, Set of responses $\mathcal{R}$, grading function G (evaluates response quality)
Output: Final Champion Response, All Scores, Final Average Scores

while $|\mathcal{R}| > 1$ **do**
  Form **consecutive pairs** of responses from $\mathcal{R}$
  for $k = 1, 3, 5, \ldots, |\mathcal{R}| - 1$ **do** ▷ Match responses in consecutive pairs
    (Score1, Score2) ← `Question_Level_Match`(P, G, r_k, r_{k+1})
    Update `All Scores` with Score1 and Score2
    if Score1 > Score2 **then**
      Advance r_k to the next round
    else
      Advance r_{k+1} to the next round
    end if
  end for
  if $|\mathcal{R}|$ is odd **then**
    Advance the unmatched response to the next round
  end if
end while
Compute `Final Average Scores` for each response across rounds
Output: Final Champion r^* (last remaining response), `Final Average Scores`, `All Scores`

The main methodology behind competitive assessment is a knockout tournament system that iteratively uses the question-level-matches described in the previous chapter. It works as follows: For N available responses to a prompt P and a grading function G, N/2 question level matches are created by pairing responses consecutively from the dataset, where one response directly advances to the next round if N is odd. For each question-level-match, the response which got the better evaluation-score advances to the next round, which will have N/2 competitor answers and thus N/4 question-level-matches.

This process continues until we reach a tournament round where only a single answer remains, which is declared the winner of the tournament. Once the tournament ends, the final evaluation-scores for all responses can be computed. The final score that is assigned to a response is determined simply as the average of all the scores it received throughout all the rounds it competed in. An example tournament with N = 4 answers is depicted in Figure 3.4.

The tournament has the goal of improving the LLM gradings even more by giving more comparisons to the LLM by matching up the better answers with each other. This also allows us to determine which answer the LLM finds the best to a given question, without the need for prompting the LLM with all available answers, thus avoiding very long contexts and positional biases which might drop the LLM's performance. In addition, the knockout tournament system for grading has a computation time of $O(N \log N)$, where $N = 7$ for our experiments on the SciEx dataset, as the SciEx has 7 LLM-generated answers that were expert graded for each question in each exam, and $N > 3$ for our experiments

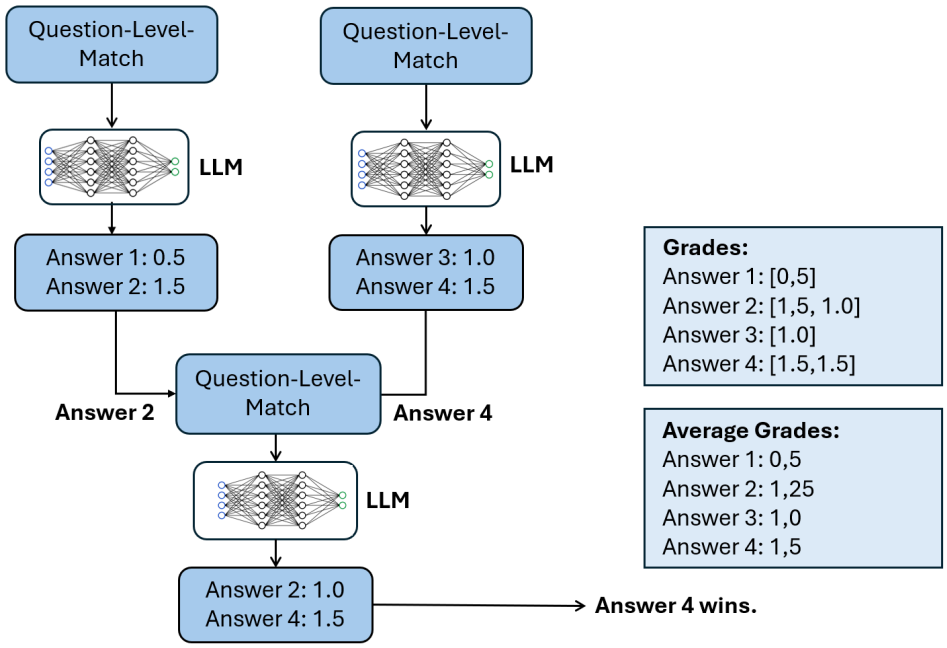

Figure 3.4.: An example Knockout-Tournament with 4 answers for a question

on the mt-metrics-eval dataset. More detailed explanations of each dataset can be found in section 4.1.1.

4. Experiments and Results

In this chapter we describe our experiments with competitive assessment and discuss their results.

4.1. Experimental Setup

In the following, we will describe our experimental setup to enable easy reproducibility of our experiments and results. Therefore, we first describe the datasets which were used to conduct our experiments on. Afterwards, we give information about the models we used to test our methods. Finally, we describe the hardware which was used to run the experiments.

4.1.1. Datasets

4.1.1.1. SciEx

The fundamental dataset curated for this thesis is the SciEx dataset from Tu Anh Dinh [3], which consists of four parts: Exams, LLM answers, human grades and LLM grades. Exams included 10 different exams from Karlsruhe Institute for Technology that are mostly related to computer science. LLM answers included answers generated by 7 different LLMs for each of those exams. Human grades were the expert gradings for these answers provided by the corresponding university professors.

LLM grades found in the original SciEx dataset were acquired through individual assessment, and these grades have been investigated in the original SciEx paper. Our research focuses instead on competitive assessment (see section 3.2) to grade the LLM answers. Our grading data consists of LLM gradings produced by multiple different judge LLMs, that were used with or without debiasing (for debiasing, see section 3.2.1). The SciEx dataset also included difficulty information for each question, allowing us to investigate the influence of difficulty while using competitive assessment, in addition to the influence of language and influence of different examinees.

The dataset has 1120 answers from 7 different LLMs across 10 different exams. Thus, there are 160 unique questions with 7 different answers each time. Out of the 160 questions, 52 are classified as easy, 72 as medium, and 36 as hard based on their difficulty. Detailed information about the exams and answering LLMs can be found in the original SciEx paper [3].

4.1.1.2. MT-Metrics-Eval

The second dataset which was used to conduct additional experiments using our methodology was a subset of the machine-translation-metrics-evaluation (mt-metrics-eval) dataset from WMT Metrics Shared Task [4]. The original dataset is a list of source sentences and translations, accompanied by a human evaluation on a scale of 0 to 100 and additional information such as the language pairing or the year the translation was acquired.

The subset we used for our thesis was created by filtering the original dataset by language, year, and number of translations per source sentence. The translations and source sentences in our subset are from the languages supported by the Llama models (English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai). In addition, we used translations from the years 2023 and 2024 and only included translations which can be grouped by the source sentence to groups of four or larger, so that competitive assessment could yield meaningful results.

Final adjustment to the subset was the removal of identical translations of the same source sentence, as sometimes the human evaluations differed for the same translation, causing inconsistencies. The resulting subset has 2087 unique translations to 211 unique source sentences. As a result, there are on average 9,89 competing translations per source sentence, with a minimum of 4 and a maximum of 14 translations.

4.1.2. Models

The models we used for our experiments are all open-source models. We used Meta's Llama 3.2 1B parameter model, Llama 3.2 3B parameter model and Llama 3.1 70B parameter model. All the model checkpoints for our experiments were obtained from the HuggingFace model hub. Llama 3.1 70B was selected as the successor of Llama 3 70B, which was used in the original SciEx paper [3]. We chose the other two smaller models since no experiments were conducted on smaller models on the SciEx dataset, allowing us to analyze how good they are at grading; thus, this thesis also is the first to research the capabilities of such smaller models on university-level question-answer pairs.

4.1.3. Hardware

For evaluation and usage of the named models we need high-performance GPUs. Therefore, we use the bwUniCluster 2.0 for our experiments. The experiments on the 70B parameter model were conducted on 4 NVIDIA Tesla V100 GPUs with 32GB VRAM each. The experiments for the smaller models were conducted on 1 NVIDIA Tesla V100 GPU with 32GB VRAM.

4.2. Evaluation Metrics

4.2.1. Pearson Correlation

We will use the Pearson Correlation Coefficient (PCC) as our primary evaluation metric. This is done to ensure compatibility with other research, such as the original SciEx paper [3], in addition to being appropriate regarding our specific focus.

Pearson correlation is a statistical similarity measure, where the corresponding coefficient PCC is a value between 1 and -1 which quantifies the linear correlation between two sets of data. This suits our case perfectly since the goal is to measure how well the grades the judge LLM gave to student answers align with the grades the human experts gave to the same answers. The formula for calculating PCC can be defined as:

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{4.1}$$

What is important to note is that a PCC value between 0 and 1 implies a positive covariance of the two datasets, and thus a positive correlation. The interpretation of such a positive correlation is that the change for a variable in one direction would result in change in the same direction for the other variable, where a value closer to 1 indicates a stronger correlation and thus a stronger similarity between the datasets.

4.2.2. Pairwise Ranking Accuracy

Additionally to Pearson Correlation, we also evaluate our methodology by pairwise ranking accuracy. This metric measures how well the choice of the grading LLM, when picking the better answer/exam out of two options, aligns with the choice of the human expert. We use the accuracy metric of Kocmi et al. from the "To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation" [5], which defines accuracy as follows:

$$\text{Accuracy} = \frac{|\text{sign}(\Delta_{\text{LLM}}) = \text{sign}(\Delta_{\text{human}})|}{|\text{all system pairs}|} \tag{4.2}$$

where:

$$\Delta = \text{score}(\text{System A}) - \text{score}(\text{System B}) \tag{4.3}$$

This metric is used only on the SciEx dataset and only on exam-level for our thesis, since on question level, for a given pair of answers to a question, human experts will often grade them equally. This equal-grading is however very improbable for competitive assessment because of the averaging of grades at the end for each answer (see section 3.2). The same problem occurs on the mt-metrics-eval dataset.

4.3. Results

Below, we introduce the results of our experiments and their analysis. First the overall results on the SciEx dataset are presented. Second, the possible influential factors that resulted in these overall results will be analyzed. Finally, we will present our results with the second dataset that covers machine translation.

For our analysis of the SciEx Dataset, we consider the exam-level and question-level grades. An exam-level grade is the sum of the grades of all questions in the exam. If the Pearson correlation is not explicitly stated to be on exam-level, it refers to question-level results.

Note that we assume the human expert grades as the golden grades, and human translation evaluations as golden evaluations. Thus, a higher correlation indicates a better performance by the judge LLM on scoring the answers/translations.

4.3.1. SciEx Baselines

In order to compare our methodology to other LLM-as-a-Judge methods, we need a baseline which we can reference. As the main goal is to compare the performance of competitive assessment to individual assessment, we replicated the results of the original SciEx paper, with the models we wanted to use for our thesis. We did not use the results of SciEx directly as an older LLM was used to gather the individual assessment results. Instead, the Llama 3.1 70B and the smaller Llama 3.2 1B and 3B parameter models were used, so that our results reflect the capabilities of current state-of-the-art LLMs.

Model	Pearson Correlation
LLaMA 3.2 1B	0.4001
LLaMA 3.2 3B	0.3655
LLaMA 3.1 70B	0.6151

Table 4.1.: LLM grader's performance (i.e., Pearson correlation to expert grading) using individual assessment.

Looking at table 4.1, we see that the bigger model performs better than the two smaller models, but still only has a correlation around 0.61, indicating that LLM grades and expert grades are not highly correlated on the question-level. However, using a better model has increased the performance, as the correlation of Llama 3 70B to expert grades was around 0.46 according to the original SciEx paper [3].

4.3.2. Impact of Competitive Assessment on Question Level

To measure the performance of our methodology, all the answers in the SciEx dataset were graded using competitive assessment with three different models. The grades were sampled twice for each model, once without debiasing and once with debiasing as described in section 3.2.1. Afterwards we calculated the Pearson correlation to expert grades for

each set of grades we gathered and compared these to the Pearson correlations between individual assessment grades and expert grades.

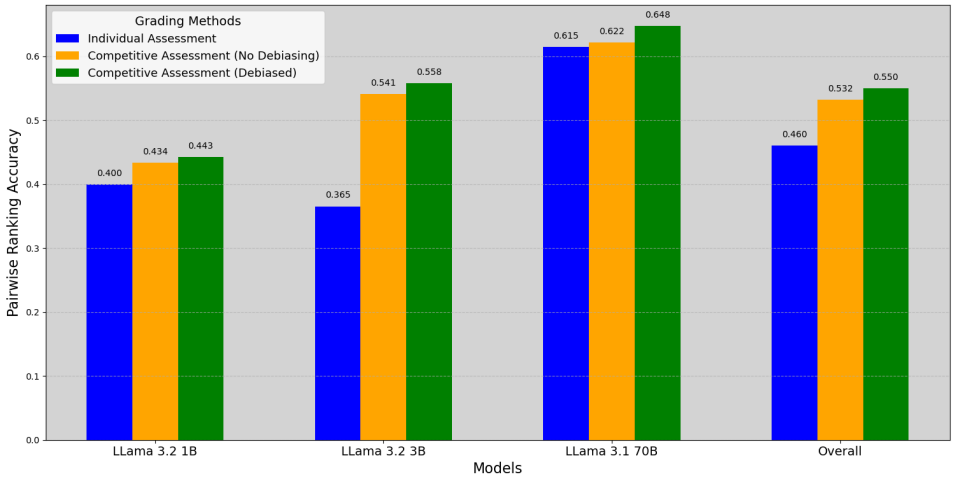

Figure 4.1.: Question Level Performance of Different Grading Methods Across Different Models

In figure 4.1 it can be seen that competitive assessment improved performance on all three models on question-level. Grades sampled using debiasing have improved performance even further on all three models from the original competitive assessment grades. The overall performance impact of each grading methodology can be seen in the final three columns, where we averaged the performance of each methodology across the three models.

4.3.3. Impact of Competitive Assessment on Exam Level

In addition to question level, we investigated the exam level Pearson correlations to expert grades for both competitive assessment and individual assessment. The results mirror the effect of competitive assessment on question level, as it can be seen in figure 4.2.

When every model is observed individually, the order of the methodologies regarding their performance is not as consistent as on question level, however the overall order of the methodologies is the same as on question level, with debiased competitive assessment yielding the best performance.

One explanation for the inconsistency is the smaller dataset size used for the exam level assessment (70 exam results) compared to the larger dataset size for question level assessment (1120 question results). The smaller dataset of 70 exam-level results may simply be not enough information for the models to make consistent decisions, which would result in high variance and explain the model-based differences. Additionally, the biggest inconsistency being for the bigger model Llama 3.1 70B (individual assessment performing better than competitive assessment) may be a result from bigger models potentially requiring bigger datasets to fully generalize and achieve consistency.

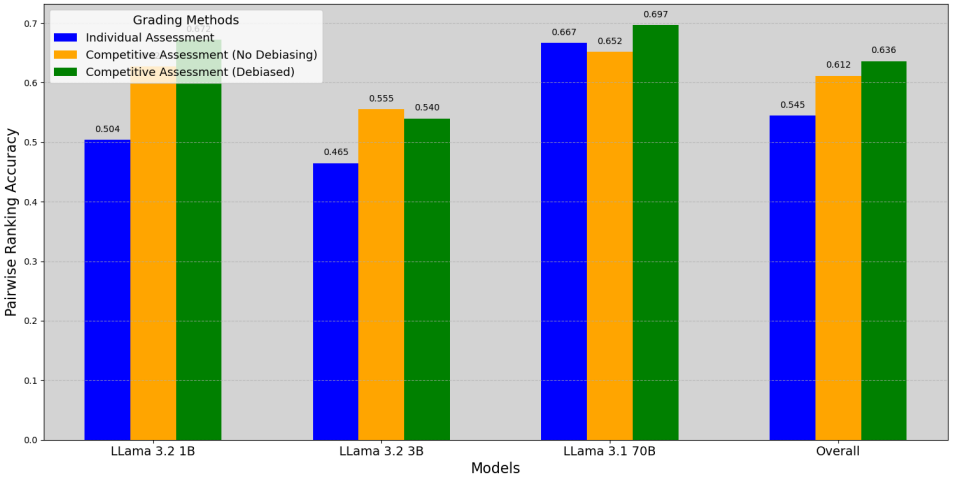

Figure 4.2.: Exam Level Performance of Different Grading Methods Across Different Models

4.3.4. Influential Factors

There are several possible factors of the SciEx dataset and the competitive assessment methodology that may influence the Pearson correlation. This section explores the key aspects of question difficulty, examinee based characteristics, language related factors and the elimination round of answers in competitive assessment.

4.3.4.1. Difficulty Based Results

Parallel to the SciEx paper [3] , one influential factor that we investigated for calculating the Pearson correlation was the difficulty of the questions. While observing the grades acquired through individual assessment, the smaller models performed better with harder questions, and the bigger model Llama 3.1 70B performed similarly for easy and hard questions, but worse for medium difficulty questions (see Figure 4.3).

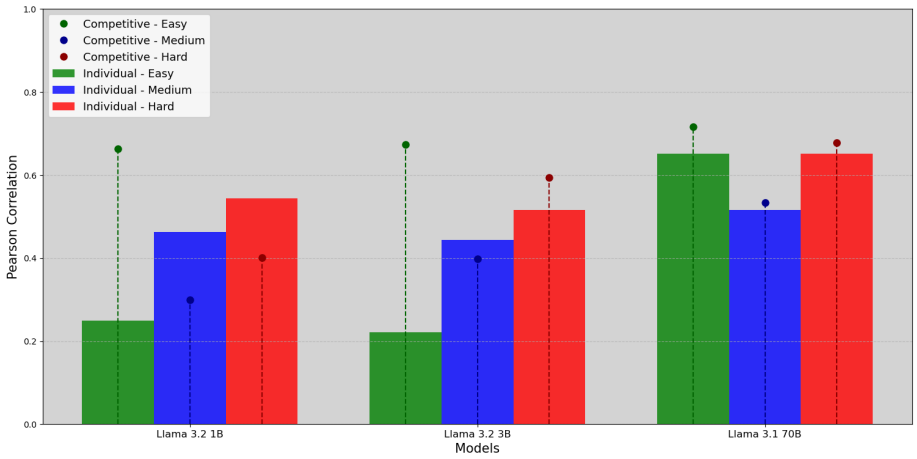

Figure 4.3.: Performance by Model and Difficulty Level: Competitive (debiased) vs. Individual Assessment

Using competitive scoring improved the Pearson correlation on easy questions on all three models, where smaller models saw a dramatic increase in performance with a difference to individual assessment correlation of 0.41 and 0.45. Furthermore, the best performing bigger model saw an increase in performance for all the difficulties, yet the smaller models saw little or no increase in performance for difficulties other than easy. As a result, the average performance difference to individual grading is very positive for easy questions and slightly negative for medium and hard questions across all the LLMs (see table 4.2) when using competitive assessment.

Difficulty	Difference in Performance when using competitive assessment			
	LLama 3.2 1B	LLama 3.2 3B	LLama 3.1 70B	**Overall**
easy	+0.41	+0.45	+0.06	**+0.31**
medium	-0.16	-0.05	+0.03	**-0.06**
hard	-0.14	+0.08	+0.03	**-0.01**

Table 4.2.: Difference in LLM grader's performance when using competitive assessment per difficulty level, subdivided by model.

One possible reason for this is that competitive assessment reduces the gap in average grades between LLMs and human experts most for easier questions, where the range of possible grades is narrower and thus each change is more pronounced. As can be seen in table 4.3, the LLMs consistently graded the answers higher than the experts. However, the difference to expert grades is lower for competitive assessment on all difficulties and models. This reduction in grading gap was most pronounced for easy questions, where the difference in grade discrepancy to expert grades between individual and competitive assessments was 71.7%. In comparison, the differences for medium and hard questions were 35.5% and 48.9%, respectively.

Difficulty	Expert	Individual			Competitive		
		1B	3B	70B	1B	3B	70B
easy	1.33	+1.79 (+135%)	+1.60 (+120%)	+1.32 (+99.2%)	+0.78 (+58.6%)	+0.44 (+33.1%)	+0.63 (+47.4%)
medium	1.80	1.70 (+99.4%)	1.57 (+87.2%)	1.92 (+107.7%)	1.21 (+67.2%)	1.06 (+58.9%)	1.11 (+61.7%)
hard	2.32	+3.56 (+153.4%)	+3.28 (+141.4%)	+3.67 (+158.2%)	+2.64 (+113.8%)	+2.14 (+92.2%)	+2.33 (+100.4%)

Table 4.3.: Difference to Expert Average Grades by Difficulty for each Grading Method.

The average grades each class of answers got from each model while using individual assessment or competitive assessment can be seen in table 4.4. The average grades that are shown in both tables for competitive grading are based on debiased competitive grading, since it was the best performing methodology.

Difficulty	Expert	Individual			Competitive		
		1B	3B	70B	1B	3B	70B
easy	1.33	3.12	2.93	2.65	2.11	1.77	1.96
medium	1.80	3.50	3.37	3.72	3.01	2.86	2.91
hard	2.32	5.88	5.60	5.99	4.96	4.46	4.65

Table 4.4.: Average Grades by Difficulty for each Grading Method.

4.3.4.2. Examinee Based Results

The average performance of competitive assessment based on individual examines can be seen on table 4.5.

	Llama 3.2 1B		Llama 3.2 3B		Llama 3.1 70B		**Overall**	
	Biased	Debiased	Biased	Debiased	Biased	Debiased	**Biased**	**Debiased**
Llava	0.166	0.1696	0.169	0.1845	0.2631	0.3289	**0.1994**	**0.2276**
Mistral	0.2915	0.3436	0.1706	0.2246	0.3216	0.3291	**0.2612**	**0.2992**
Mixtral	0.4674	0.4375	0.527	0.586	0.5443	0.568	**0.5129**	**0.5305**
Qwen	0.428	0.4449	0.5156	0.5548	0.5206	0.5626	**0.488**	**0.5216**
Claude	0.4511	0.5326	0.8306	0.8546	0.8521	0.8489	**0.7113**	**0.7453**
GPT-3.5	0.4075	0.436	0.5297	0.4727	0.601	0.583	**0.5127**	**0.4973**
GPT-4V	0.698	0.686	0.8004	0.8358	0.8992	0.8819	**0.7992**	**0.8026**

Table 4.5.: LLM graders performance (i.e., Pearson correlation to expert grading) on different examinees with and without debiasing, including overall performance.

Examinee	Difference in Performance compared to individual assessment			
	LLama 3.2 1B	LLama 3.2 3B	LLama 3.1 70B	**Overall**
Llava	+0.0150	-0.0040	+0.0724	**+0.0283**
Mistral	+0.2452	+0.1043	-0.0247	**+0.1082**
Mixtral	-0.1247	+0.0407	-0.0776	**-0.0539**
Qwen	-0.1263	+0.0184	-0.0106	**-0.0395**
Claude	-0.2975	+0.0760	-0.0018	**-0.0744**
GPT-3.5	-0.0310	+0.1943	+0.0228	**+0.0620**
GPT-4V	+0.1903	+0.3529	+0.0046	**+0.1826**

Table 4.6.: Difference in LLM grader's performance when using competitive assessment per difficulty level, subdivided by model.

In addition, table 4.6 shows that, across all model sizes, models have similar examinee-based performance using competitive assessment with individual assessment. Our results also parallel the findings of SciEx with a similar distribution of examinee-based performance [3]. The difference to individual assessment on an examinee-based level is therefore

also evenly distributed, with the average difference in performance overall being only +0.03 per examinee.

One thing to note is that the highest performance increase per examinee is for GPT-4V with 0.1826. This might suggest that competitive assessment helps more with grading better examinees, however, it would need to be researched further, as the best examinee out of the SciEx dataset, Claude, saw a performance decrease of 0.0744.

4.3.4.3. Language Based Results

Another influential factor we investigated was the difference in performance of the grading methodologies, when we focused on different languages. SciEx dataset consists of exams in two languages. Six of the ten exams in the dataset are in German and the remaining four exams are in English. With the larger model Llama 3.1 70B, there was an increase in performance when using debiased competitive assessment compared to individual assessment for both languages.

However, for the smaller models, most of the overall performance increase while using competitive assessment resulted from the performance increase for the English exams (see Table 4.7). The increase in performance when using debiased competitive assessment compared to individual assessment for English exams, is 0.5133 with Llama 3.2 3B and 0.3343 with Llama 3.2 1B. For German exams however, 3B parameter model barely gained any performance (0.018 increase) and 1B parameter model even lost performance (0.165 decrease).

Language	Individual			Competitive					
	1B	3B	70B	1B		3B		70B	
				Biased	Debiased	Biased	Debiased	Biased	Debiased
English	0.2348	0.176	0.6759	0.5695	0.5691	0.6365	0.6892	0.6429	0.6952
German	0.5474	0.5451	0.6263	0.3706	0.3824	0.5456	0.5628	0.6497	0.6790

Table 4.7.: Pearson correlations for LLM graders' performance across languages (English and German), for individual and competitive assessment.

One possible reason for the difference based on language is the disparity in language proficiency between the smaller LLMs and bigger ones. Assuming the 1B and 3B parameter models have weaker capabilities in German compared to English, this would make their responses already closer to an upper bound of quality when using individual assessment. This would also mean, comparatively, they have more variance in response quality in English, hence the big performance increase while using competitive assessment.

4.3.4.4. Elimination-Round Based Results

To check whether competitive assessment results in any additional performance increase compared to regular pairwise comparisons with only one round, Pearson correlations of the sets of grades the answers got, based on their round of elimination were investigated.

The answers which got eliminated on the first round of the knockout tournament (see chapter 3.2.2), got only one pairwise comparison, compared to the multiple pairwise comparisons the answers which advanced to later rounds got.

Elimination	Competitive Assessment							
	1B		3B		70B		**Overall**	
	Biased	Debiased	Biased	Debiased	Biased	Debiased	**Biased**	**Debiased**
First Round	0.3737	0.3223	0.5400	0.5400	0.5264	0.5801	**0.4800**	**0.4808**
Later Rounds	0.4816	0.4428	0.5393	0.5692	0.6602	0.6782	**0.5604**	**0.5634**
Difference	+0.1079	+ 0.1205	-0.0007	+0.0292	+0.1338	+0.0981	**+0.0804**	**+0.0826**

Table 4.8.: Comparison of LLM Grader's performance (on the SciEx dataset) on the answers which only graded once versus the answers which got graded multiple times.

As can be seen in Table 4.8, the answers which got eliminated in later rounds have an overall higher Pearson correlation in the grades they got, across the three models we used. This shows that more pairwise comparisons result in more accurate grades from the judge LLM.

4.3.5. Adding Reference Answers

As SciEx dataset also provides the reference answers for a large subset of the available questions, we investigated the effects of adding reference answers.

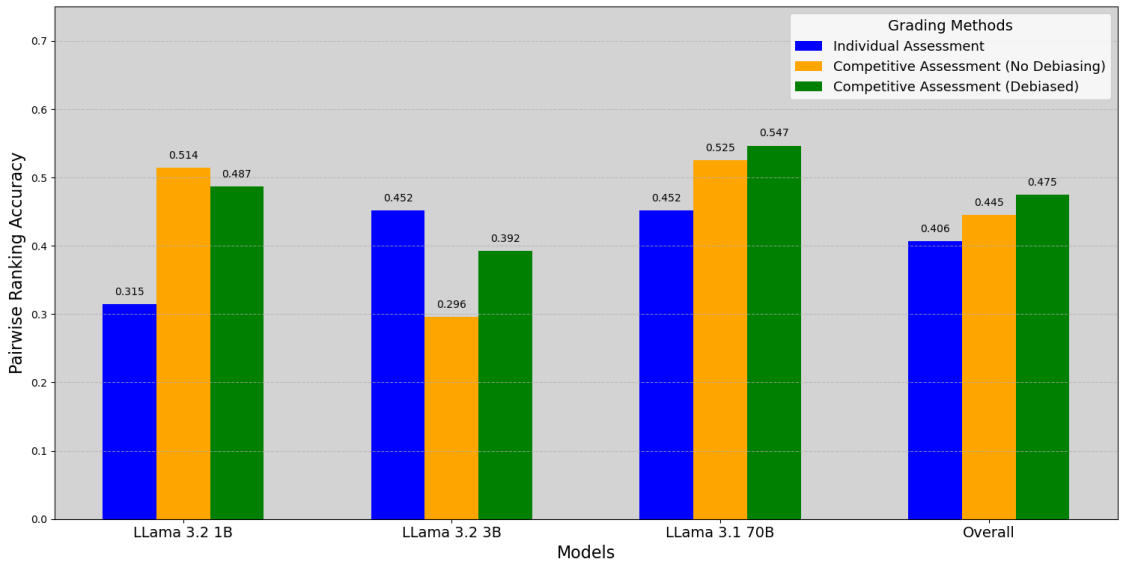

Figure 4.4.: Question Level Performance of Different Grading Methods Across Different Models with Reference Answer Provided

For this purpose, the judge LLM was also provided with the reference answer in the prompt, in addition to the question and the single answer (individual assessment) or the pair of answers (competitive assessment). Again, the grades were sampled using the three models, first with competitive assessment with and without debiasing, and second with individual assessment. The exact prompts we used for this experiment can be found in Appendix A.1. Although not as consistent across all three models, the overall order of the methodologies regarding performance is the same as without reference answers. This can be seen in figure 4.4.

One thing to note however is that providing the reference answers dropped the overall performance across the models independently from the grading methodology (see tables 4.9 and 4.10). This is contrary to the findings of the SciEx paper [3], which showed that providing reference answers increases performance.

	Individual Assessment			
	1B	3B	70B	**Overall**
Without reference answers	0.4001	0.3655	0.6161	**0.4606**
With reference answers	0.3149	0.4523	0.4518	**0.4063**
Performance change	-0.0852	+0.0868	-0.1643	**-0.0543**

Table 4.9.: Comparison of LLM Grader's performance with or without providing the reference answer, while using individual assessment.

	Competitive Assessment							
	1B		3B		70B		**Overall**	
	Biased	Debiased	Biased	Debiased	Biased	Debiased	**Biased**	**Debiased**
Without ref	0.4341	0.4429	0.5409	0.5585	0.6215	0.6478	**0.5322**	**0.5497**
With ref	0.5144	0.4867	0.2964	0.3923	0.5255	0.5467	**0.4454**	**0.4755**
change	+0.0803	+ 0.0438	-0.2445	-0.1662	-0.0960	-0.1011	**-0.0867**	**-0.0745**

Table 4.10.: Comparison of LLM Grader's performance with or without providing the reference answer, while using competitive assessment.

4.3.6. Pairwise Ranking Accuracy

For our other second metric -Pairwise Ranking Accuracy- we measured how often the judge LLM correctly identified the better performing exam in alignment with the human expert judgments. Again, the accuracy was calculated across the three models, comparing

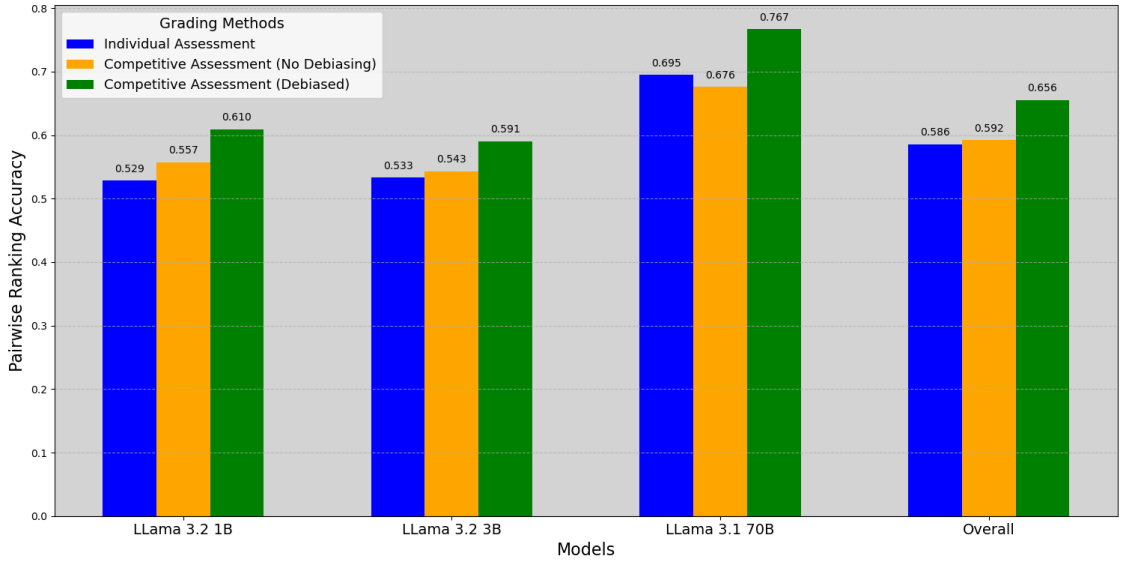

Figure 4.5.: Pairwise Ranking Accuracy of Different Grading Methods Across Different Models

grades obtained through individual assessment, competitive assessment without debiasing and competitive assessment with debiasing.

As can be seen in Figure 4.5, all methodologies were able to identify the better performing exam correctly more than half of the time. When comparing the LLM-as-a-Judge methodologies, we observed the same ranking pattern as the Pearson Correlation evaluation. Individual assessment had the lowest overall pairwise ranking accuracy with a score of 0.586, followed by the slightly better performing competitive assessment without debiasing at 0.592. Debiased competitive assessment performed best once again with an accuracy of 0.656 across three models. The trend was generally consistent across individual models as well, with debiased competitive assessment performing best for each model. The best individual performance was Llama 3.1 70B with debiased competitive assessment with an accuracy of 0.7667, indicating a strong ability of picking the better performing exams correctly.

4.3.7. Results on the MT-Metrics-Eval Dataset

To further generalize and also validate our findings from the SciEx dataset, competitive assessment and individual assessment methodologies were both used on mt-metrics-eval dataset (see section 4.1.1.2). By analyzing this second dataset, we assess if the performance increase in evaluation observed while using competitive assessment extends beyond the domain of scientific argumentation into the task of machine translation. The exact prompts we used to collect the grades for this experiment can be found in Appendix A.1. Similar to SciEx, the performance of each LLM-as-a-Judge method is measured by calculating the Pearson correlations to assessments by human experts. The correlations are once again calculated over the same three models of different sizes.

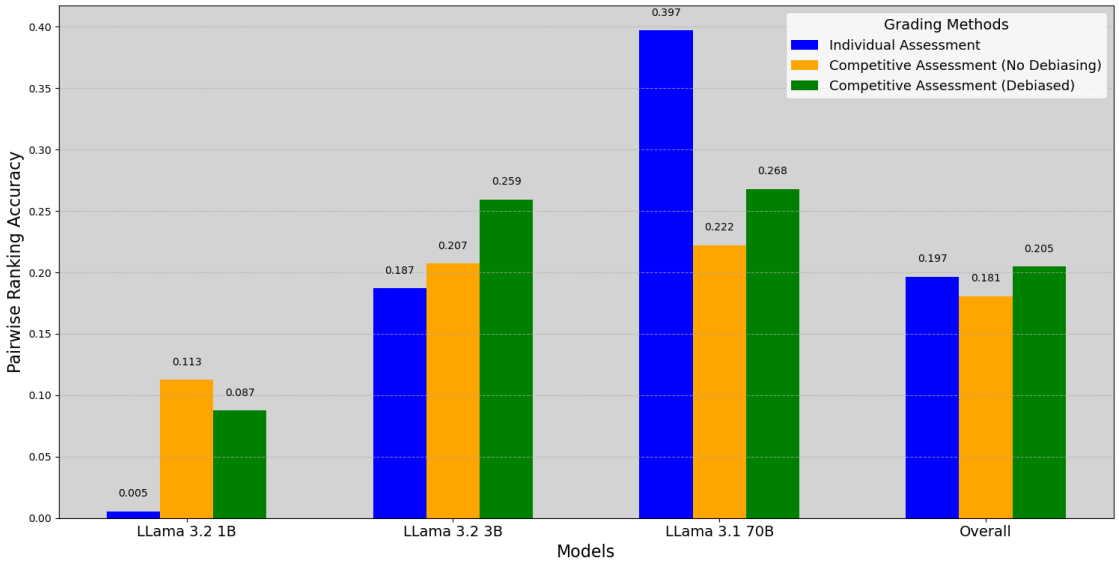

Figure 4.6.: Performance of Different Grading Methods Across Different Models on the MT-Metrics-Eval Subset

Overall, competitive assessment with debiasing yielded the best performance across the three models (see Figure 4.6). Second best performing method was individual assessment, with competitive assessment without debiasing having the least accurate results. The biggest difference to our results on the first dataset is that the best performing individual model-method pair on mt-metrics-eval was individual assessment with Llama 3.1 70B, whereas debiased competitive assessment with Llama 3.1 70B had the best performance on the other dataset. Thus, the overall increase in performance while using debiased competitive assessment on mt-metrics-eval results directly from the large performance increases for the smaller models.

One possible reason for this difference in the two datasets is the varying nature of each task. The SciEx dataset requires complex scientific reasoning while evaluating, which benefits from an iterative ranking and pairwise comparisons, as this might improve reasoning. Machine translation quality evaluations, however, might rely more on direct linguistic comparisons without complex reasoning, a task which larger LLMs handle well without the need for iterative ranking. Smaller models on the other hand, might benefit from the structured comparisons provided by competitive assessment, as it allows them to refine their evaluations progressively, explaining the increase in performance with competitive assessment.

This difference in each task would also explain the difference between SciEx and mt-metrics-eval datasets while analyzing elimination round based results. As can be seen in Table 4.11, the responses which advanced further in the tournament showed lower alignment with human experts in mt-metrics-eval, whereas for SciEx, the results were the opposite with later round eliminations showing better alignment (see section 4.3.4.4). This suggests that the iterative comparisons do not increase the grading accuracy for simple tasks that LLMs already excel at, but rather introduce inconsistencies by adding too many

Elimination	Competitive Assessment							
	1B		3B		70B		**Overall**	
	Biased	Debiased	Biased	Debiased	Biased	Debiased	**Biased**	**Debiased**
First Round	0.1245	0.0836	0.2222	0.2917	0.2688	0.3173	**0.2052**	**0.2309**
Later Rounds	0.0777	0.0875	0.1603	0.2013	0.1150	0.1581	**0.1177**	**0.1490**
Difference	-0.0468	+0.0039	-0.0619	-0.0904	-0.1538	-0.1592	**-0.0875**	**-0.0819**

Table 4.11.: Comparison of LLM Grader's performance (on the mt-metrics-eval dataset) on the answers which only graded once versus the answers which got graded multiple times.

comparisons. This is also supported by the fact that the effect is most prominent on the larger model, as it has the strongest evaluation capabilities out of the three models.

5. Conclusion

This study set out to address a key limitation of existing LLM-as-a-Judge methods, such as individual assessment or pairwise assessment: not having a global view over the responses while evaluating them. This prevents evaluations from taking the relative strength of each response into account, which is information the human experts inherently consider while evaluating, leading to a drop in the accuracy of LLM judgments.

To address this, we proposed an alternative LLM-as-a-Judge method called competitive assessment and tested it with three different LLMs on two different datasets. Our methodology uses iterative pairwise comparisons in a tournament manner, with better evaluated responses advancing through successive rounds until the best response is determined, thereby giving the judge LLM a more global view of all responses while evaluating. On both datasets, first, the evaluations for all the responses were collected using both individual assessment and competitive assessment, and later compared to human expert evaluations to determine their accuracy.

In our experiments, competitive assessment managed to improve both question level and exam level Pearson correlation by around 0.09 from individual assessment for complex university-exam evaluations. Furthermore, the responses which progressed further in the competitive assessment process had 0.08 better accuracy compared to the responses which got eliminated on the first round. This indicates that competitive assessment results in a performance increase from regular pairwise assessments as well.

However, the performance increase was not as significant in machine translation evaluation, especially for the larger LLM. This may be because translation tasks rely more on direct linguistic pattern matching rather than complex reasoning, reducing the benefits of iterative ranking. To further investigate the effects of competitive assessment on automatic evaluation, it can be tested on a broader range of tasks.

Bibliography

[1] Cheng-Han Chiang and Hung-yi Lee. "Can Large Language Models Be an Alternative to Human Evaluations?" In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 15607–15631. DOI: 10.18653/v1/2023.acl-long.870. URL: https://aclanthology.org/2023.acl-long.870/.

[2] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG]. URL: https://arxiv.org/abs/2210.11416.

[3] Tu Anh Dinh et al. *SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading*. 2024. arXiv: 2406.10421 [cs.CL]. URL: https://arxiv.org/abs/2406.10421.

[4] Markus Freitag et al. "Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn et al. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 578–628. DOI: 10.18653/v1/2023.wmt-1.51. URL: https://aclanthology.org/2023.wmt-1.51/.

[5] Tom Kocmi et al. "To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation". In: *Proceedings of the Sixth Conference on Machine Translation*. Ed. by Loic Barrault et al. Online: Association for Computational Linguistics, Nov. 2021, pp. 478–494. URL: https://aclanthology.org/2021.wmt-1.57/.

[6] Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. *LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models*. 2024. arXiv: 2307.07889 [cs.CL]. URL: https://arxiv.org/abs/2307.07889.

[7] Humza Naveed et al. *A Comprehensive Overview of Large Language Models*. 2024. arXiv: 2307.06435 [cs.CL]. URL: https://arxiv.org/abs/2307.06435.

[8] Zhen Qin et al. *Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting*. 2024. arXiv: 2306.17563 [cs.IR]. URL: https://arxiv.org/abs/2306.17563.

[9] Philip Resnik. *Large Language Models are Biased Because They Are Large Language Models*. 2024. arXiv: 2406.13138 [cs.CL]. URL: https://arxiv.org/abs/2406.13138.

[10] Jiaan Wang et al. *Is ChatGPT a Good NLG Evaluator? A Preliminary Study*. 2023. arXiv: 2303.04048 [cs.CL]. URL: https://arxiv.org/abs/2303.04048.

[11] Jonathan J. Webster and Chunyu Kit. "Tokenization as the initial phase in NLP". In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*. COLING '92. Nantes, France: Association for Computational Linguistics, 1992, pp. 1106–1110. DOI: 10.3115/992424.992434. URL: https://doi.org/10.3115/992424.992434.

[12] Jason Wei et al. "Chain of Thought Prompting Elicits Reasoning in Large Language Models". In: *CoRR* abs/2201.11903 (2022). arXiv: 2201.11903. URL: https://arxiv.org/abs/2201.11903.

[13] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL]. URL: https://arxiv.org/abs/2306.05685.

A. Appendix

A.1. Prompts

Individual Grading Prompt in German

```
Sie sind ein Universitatsprofessor. Bewerten Sie die folgende Antwort auf die unten
stehende Frage. Geben Sie eine Punktzahl von 0 bis {max_points} basierend auf
Korrektheit und Relevanz an.

Die Frage: <Question>

Die Antwort: <Answer 1>

Give you grade in the format: Explanation: [explanation] Score: [score]/<max_points>

Bewerten Sie die Antwort auf einer Skala von 0 bis {max_points} (halbe Punkte wie 0,5
oder 1,5 sind erlaubt) im Format:
Begründung: [begründung] Punktzahl: X/{max_points}
```

Pairwise Grading Prompt in German

```
Sie sind ein Universitätsprofessor und bewerten Prüfungsantworten. Bewerten
Sie die folgenden Antworten auf einer Skala von 0 bis {max_points} (halbe
Punkte sind erlaubt) basierend darauf, wie gut sie die Frage beantworten.\n\n

Die Frage: <Question>


Antwort 1: <Answer 1>

Antwort 2: <Answer 2>

Bewerten Sie die beiden Antworten auf einer Skala von 0 bis {max_points}
(halbe Punkte wie 0,5 oder 1,5 sind erlaubt) im Format:
Begründung: [begründung] Antwort 1: X/{max_points} Antwort 2: Y/{max_points}
```

Figure A.1.: LLM prompts for individual grading and pairwise grading in German for the SciEx dataset

Individual Grading Prompt in English with Reference

```
You are a university professor exam grader. Grade the following answer on a
scale of 0 to <max_points> (allowing half points) based on its correctness
and relevancy given the following question.The correct answer is provided as
reference.

The Question: <Question>
The Reference Answer: <Reference Answer>

The Answer: <Answer 1>

Give you grade in the format: Explanation: [explanation] Score:
[score]/<max_points>
```

Pairwise Grading Prompt in English with Reference

```
You are a university professor exam grader. Grade the following answers on a
scale of 0 to <max_points> (allowing half points) based on how well they answer
the question. A correct answer is provided as reference.

The Question: <Question>

The reference answer: <Reference Answer>


Answer 1: <Answer 1>

Answer 2: <Answer 2>

Grade the 2 answers on a scale of 0 to <max_points> (Half points such as 0.5 or
1.5 are allowed.) in the format:
Explanation: [explanation] Answer 1: X/<max_points> Answer 2: Y/<max_points>.
```

Figure A.2.: LLM prompts for individual grading and pairwise grading with reference in English for the SciEx dataset

Individual Grading Prompt in German with Reference

```
Sie sind ein Universitätsprofessor. Bewerten Sie die folgende Antwort auf die
unten stehende Frage. Geben Sie eine Punktzahl von 0 bis {max_points}
basierend auf Korrektheit und Relevanz an. Berücksichtigen Sie die
Referenzantwort für Ihre Bewertung.\n\n

Die Frage: <Question>

Referenzenantwort: <Reference Answer>

Antwort 1: <Answer 1>

Antwort 2: <Answer 2>

Bewerten Sie die beiden Antworten auf einer Skala von 0 bis {max_points}
(halbe Punkte wie 0,5 oder 1,5 sind erlaubt) im Format:
Begründung: [begründung] Antwort 1: X/{max_points} Antwort 2: Y/{max_points}
```

Pairwise Grading Prompt in German with Reference

```
Sie sind ein Universitätsprofessor und bewerten Prüfungsantworten. Bewerten
Sie die folgenden Antworten auf einer Skala von 0 bis {max_points} (halbe
Punkte sind erlaubt) basierend darauf, wie gut sie die Frage beantworten.
Berücksichtigen Sie die Referenzantwort für Ihre Bewertung.\n\n

Die Frage: <Question>

Referenzenantwort: <Reference Answer>

Antwort 1: <Answer 1>

Antwort 2: <Answer 2>

Bewerten Sie die beiden Antworten auf einer Skala von 0 bis {max_points}
(halbe Punkte wie 0,5 oder 1,5 sind erlaubt) im Format:
Begründung: [begründung] Antwort 1: X/{max_points} Antwort 2: Y/{max_points}
```

Figure A.3.: LLM prompts for individual grading and pairwise grading with reference in German for the SciEx dataset

Individual Grading Prompt for MT

```
You are a translation evaluator. Evaluate the quality of the translation provided.
Give a score from 0 to 100 based on clarity, accuracy, and grammar.

Source: {src}
Translation: {tgt}

Output only: Explanation: [explanation] Score: [Score]/100.
```

Pairwise Grading Prompt for MT

```
You are a translation evaluator. Your task is to evaluate the quality of two
translations for a given source sentence. You will provide the scores on a scale
from 0 to 100, based solely on the clarity, accuracy, and grammar of the
translations.

Source: {src}
Translation 1: {tgt1}
Translation 2: {tgt2}


Output only: Explanation: [explanation] Translation 1: X/100, Translation 2: Y/100.
```

Figure A.4.: LLM prompts for individual grading and pairwise grading for the mt-metrics-eval dataset