Overview of the PISA Quiz Item Scoring (Sensemaking) Task of the ELOQUENT 2026 Lab for Evaluating Generative Language Model Quality

Pavel Šindelář[1,*], Ondřej Bojar[1,*], Mario Piacentini[2,‡], Katherina Thomas[2,‡] and Said Ettejjari[2,‡]

[1]*Charles University, Prague, Czech Republic*
[2]*OECD*

Abstract

ELOQUENT is a set of shared tasks that aims to create easily testable high-level criteria for evaluating generative language models. Sensemaking is one such shared task, and so ELOQUENT is hosting Sensemaking's second instance in 2026. In Sensemaking, we try to assess how well generative models "make sense out of a given text". In the previous year, this encompassed producing questions given an input material, answering the questions, and evaluating the correctness of the answers, adhering rather strictly to the input materials. In 2026, we focus on the evaluation area; a sister task considers the construction of questions. This work presents the assignment, participating systems, our baselines, and empirical results for the evaluation, considering two variants: simple evaluation (only the material, question, and answer are provided) and rubric-based evaluation with detailed rubrics with instructions for the evaluator. Our test set spans 13 languages and 8 topical domains, of which one comes from the PISA tests and includes genuine student responses from across Europe.

Keywords

evaluation, LLM, automated short answer grading, PISA tests

1. Introduction

The shared task experiment described in this report is part of the ELOQUENT lab [1] at CLEF 2026.

This task traces its roots to previous tasks on quiz handling in 2024 [2] and in 2025 [3] when we adopted the name "Sensemaking". This name change signifies that we are interested in how well a model can make sense of given text for various quiz-related tasks, and not just topical competence in the context of world knowledge.

The Sensemaking task in 2026 focuses exclusively on automated short answer grading (ASAG), but it is also closely related to its sister task, which handles quiz generation [4] and was happening concurrently.

Building upon the experience of Šindelář and Bojar [3], who demonstrated that evaluation is still particularly problematic for current LLMs, we decided to focus only on the Evaluator subtask. This task could be described as context-aware automated short-answer grading. We use multiple metrics to measure alignment with human graders and hypothesise that especially smaller LLMs will fail to achieve sufficient agreement.

We wish to examine how the model responses change when they are provided with a grading rubric to ground their evaluations. To this end, we created one track where a rubric is provided to constrain the grade and one where it is not. We hypothesise that there will be a significant difference in the performance of systems showing a lack of context understanding and alignment with human grading criteria.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

‡The opinion expressed and arguments employed in this article are solely those of the authors and do not necessarily reflect the official views of the OECD or of its member states.

sindelar@ufal.mff.cuni.cz (P. Šindelář); bojar@ufal.mff.cuni.cz (O. Bojar)
0000-0002-0877-7063 (P. Šindelář); 0000-0002-0606-0050 (O. Bojar)

Due to recent findings such as the one presented in Fu and Liu [5], we decided to test how reliably these models rate student responses in different languages. We collected medium and low-resource European languages and hypothesise that the model performance will suffer in this non-English subset.

In this work, when discussing neural network models, we often refer to the Hugging Face website[1] and the associated transformers [6] and sentence-transformers [7] libraries. When mentioning each model we used for the first time, we will provide the Hugging Face commit hash so that the exact version of the model used can be found to ease reproducibility.

2. Task variants and evaluation

Sensemaking runs in two variants this year: with instructions specifying the rating scale and expectations for each level of each question, called the *Rubric-based track*, and without them, called the *Simple track*.

2.1. The simple rating track

In the simple rating track, we tested how well the models can grade student responses without relying on a rubric. That is, they must only implicitly infer a loose rating based on the context, question, and answer. Participants were required to estimate the desired scores for each answer independently. They were presented with a short description of the levels, and while scoring, they were allowed to use examples from the training and development sets to determine what the levels should be. Ultimately, the systems were required to return the score of 0–4 for each scored answer based on the input materials (called *context*) and the question.

We used a simple, general description for each of the levels:

0 No relevant effort or fundamental misunderstanding of the task
1 Minimal attempt to identify or apply relevant reasoning
2 Valid reasoning for core parts, but flawed overall execution
3 Correct methodology with only minor slips or omissions
4 Clearly correct execution fully answering the question

2.2. The rubric-based rating track

The rubric-based track tests how reliably the models apply rather detailed instructions ("rubric") provided with each individual question.

Here, the model has to match the answer to one of the three levels using the context, the question, and the rubric descriptor of the level. The considered levels are:

NC when the answer fits best the *no credit* (label 0) descriptor.
PC when the answer fits best the *partial credit* (label 1) descriptor.
FC when the answer fits best the *full credit* (label 2) descriptor.

Often, the rubric descriptors contain examples of the answers that fit the level. In an adversarial variant, we also deleted these examples to test the ability of the models to work without examples.

2.3. The relationship between the two scales

Depending on the source (see Section 3 below) of the texts and the questions, ground truth ratings were available in different scales.

The conversion from NC/PC/FC to 0–4 was carried out deterministically for some domains, with NC getting 0, PC getting 2, and FC getting 4. For PISA, the grade was determined on a question-by-question basis, and for generated items, it was decided by the generating LLM. The conversion from 0–4 to NC/PC/FC is only done for comparison, and the conversion is done by mapping 0 to NC, 4 to FC, and the rest to PC.

[1] www.huggingface.co

For sources that did not have a ground truth rubric, we generated the rubrics automatically (see below) and also generated the corresponding ratings in both scales at the same time to ensure the ratings matched. For full credit, the vast majority of the answers were rated as 4, with a few rated as 3. For no credit, there was more variety, and the answers were rated as either 0 or 1. For partial credit, most answers were rated as 2 or 3, with a few rated as 1.

2.4. Evaluation metrics

The main evaluation metric we use is quadratic weighted Cohen kappa (QWK) [8]. QWK has the benefit of considering both the imbalance in the number of classes and the ordinality of the ratings (NC<PC<FC). In each track, the participants were allowed to return invalid ratings, but such ratings were simply mapped to the closest valid rating (i.e., we interpret -1 as 0) and were not evaluated separately.

In addition to this metric, we also report accuracy. We find that although it is not as generalisable or as good at comparing systems, it is much more intuitive. It also allows us to better understand the system results when paired with QWK, without the need to examine more than two measures (as would be the case, e.g., with QWK, precision, recall, and F1) or the entire confusion matrix.

In some circumstances, such as when examining only adversarial inputs, QWK cannot be used at all because it requires there to be more than one class in the set of golden labels. This could be, of course, fixed by simulating predictions or adding a random subset from a more balanced set, but it would introduce further noise to the evaluation.

3. Dataset

The domains tested this year come from three main sources: the sources we used last year, three new domains from our sources, and one new domain created in partnership with the OECD (Organisation for Economic Co-operation and Development), reusing items from their PISA (Programme for International Student Assessment) datasets. Specifically, the texts we used come from the website-like PISA participation interface.

Because we wanted to utilise the PISA data to the highest possible extent, we decided to use only PISA data in the test set. This meant we had to collect the development set and training set data points separately. The only PISA quizzes collected for the development set were from a very small selection of publicly available PISA quizzes, and the responses provided were designed by us.

Because the test set created from the PISA data (this subsection is referred to as **PISA** test set) had a relatively small variety of questions, we decided to collect additional data for the test set (this subsection is referred to as **Additional** test set).

For **Additional** test set, we collected a number of responses from students and annotators (such responses are **Native**) and generated other data by LLM (such responses are **Generated**). The data creation procedure will be described in more detail in other publications. A publicly available example data point can be seen in Figure 2. All datapoints in **PISA** test set are marked as **Native** to distinguish them from such **Generated** items.

In the end, we used the same source dataset for both tracks. In principle, the participants could have cheated by also using the rubrics in the simple track, but we do not think this has happened. The only difference between the tracks is that some data points differ only in their rubrics; such duplicate data points were not included in the simple rating track dataset.

3.1. Level of human agreement

Each year, PISA records the statistics of inter-rater agreement. These are not broken down by item, so there is no way for one to get the exact agreement one should expect for our subset. For a rough comparison per domain across languages, one can look at Table 15.A.7. in OECD [9]. We can see that

all items not in the creative thinking category were rated with an agreement above 92.7%. We do not have any agreement estimates for our data sources.

3.2. Data domains and languages from last year

In the UKRBIO domain, we use a subset of the data points from the UKR-BIOLOGY test set domain contexts of last year. The UKR-BIOLOGY domain was based on sections of freely accessible Ukrainian textbooks. The questions were selected from those provided by the textbook authors.

In the GERSPEECH domain (we reuse a subset of the test set part, which was called POPULAR last year), we examine popular science videos and European Parliament speech recordings. The questions were provided by the authors of Javorský et al. [10] who were using these speeches to evaluate automated captioning systems.

In the WORLDHIST domain (a subset of last year's test set part of WORLD-HISTORY), we extracted chapters from a Creative Commons-licensed English world history textbook. The questions were selected from those provided by the authors.

In the DEMAGOG domain (similar but different data points from last year's DEMAGOG), we examine data points from the Demagog.cz project.[2] The project concerns the fact-checking of Czech politicians' public statements. The provided database contained a short excerpt of a claim followed by a fact-check article that either confirmed it, refuted it, or stated that the factuality cannot be verified. For each selected article, we converted the HTML of the fact-check article into plaintext. We then prepended the name of the author of the statement and the statement before this text. This was done because the article refers to the claim and its presenter, so this information is necessary to create a coherent context. Finally, we used a combination of an LLM, regular expressions, and manual parsing to remove the final verdict about the truthfulness. This was done because it serves as a summary of the article and makes it artificially easier. The questions were created manually by us.

3.3. New data domains

For the **PISA** test set domain, we take the provided contexts, test questions, and annotator rubrics from **PISA** test set tests created by OECD experts and granted to us by the OECD. Each had a set of authentic student responses collected during their administration. These responses were deidentified prior to being provided to us by removing any information about the student. A maximum of 500 responses was sampled for each language (labelled with the **NotTranslated** property). Then, more data points were created to ensure that each language was represented by 1000 data points. These additional data points were created by machine translating responses from the original response language into English and then from English into the target language (labelled with the **Translated** property). As mentioned previously in this section all datapoints in **PISA** test set are marked as **Native**.

For the COUNSELLING domain, the questions were taken from the Czech Law Advisory answers to the questions publicly submitted by members of the public to https://frankbold.org/poradna. Some answers were automatically extracted from the summary response provided at the beginning of the response as the lead paragraph. This summary was then removed from the text.

For the LAWF domain, the questions and answers were created by students of Charles University Faculty of Law using legal documents from the constitutional court as context. These were selected on the basis of the student's own areas of interest. The students equipped each text with one or more questions and also one or more good and one or more bad answers. The questions usually concern concepts from legal theory, constitutional court decisions, or the recall of factual information mentioned in the context that was important to the case.

For the POPSCI domain, we collected popular science articles from several websites and created questions for them manually.

[2] www.demagog.cz

1. Germanic
    a) English (5)
    b) German (5)
    c) Swedish (4)
    d) Danish (3)
2. Finno-Ugric
    a) Hungarian (4)
    b) Finnish (4)
3. Romance
    a) Portuguese (4)
    b) Romanian (3)
4. Slavic
    a) Czech (4)
    b) Serbian (4)
    c) Ukrainian (3)
5. Hellenic
    a) Greek (3)
6. Celtic
    a) Irish (2)

Figure 1: The languages selected for the experiment. Language classes [11] provided in parentheses.

3.4. Languages

For the experiments, we selected a mixture of high resource, medium resource, and occasionally low resource languages covering the geographic and linguistic landscape of Europe (as can be seen in Figure 1). For classifying languages by the amount of resources available, we use the classification in Joshi et al. [11].[3] . This categorisation correctly notes that languages with few official resources but a large amount of online content tend to be easier to work with when using language models pretrained on content from the Internet. Most of the languages collected are classified as 4 (a large amount of dedicated unlabelled data and some labelled data) or 3 (strong Internet presence, insufficient labelled data). English and German are categorised as 5[4] (massive amounts of resources), and Irish is categorised as 2 (weak Internet presence, small set of labelled datasets). We hypothesise that performance will be lower for languages categorised as having fewer resources available.

The **NotTranslated** PISA data points originate in these languages authentically, except for Irish, which was machine-translated from English. Across the other domains (the **Additional** test set portion of the dataset), we synthesise part of the data points using an LLM. These items are marked as LLM-Generated and were originally produced in English or Czech (this is the **NotTranslated** section) and subsequently translated into the languages used in the experiment through English (this is the **Translated** section).

Most of the languages selected, other than English and German, have more than 9 million native speakers, are official languages of the European Union, or have language resources available for them or a closely related language. For these reasons, almost all of them, except for Irish, fall into categories 3 and 4 [11], both of which have a significant presence in semi-supervised pretraining datasets due to their significant presence on the Internet.

Unfortunately, the responses for the Irish language could not be provided by the OECD. It is the only language classified as category 2 [11]. It has a low number of resources, a low number of native speakers, and no large related languages (as discussed in Section 3.4). The same is true for Romanian. We also could only get a subset of answers for Ukrainian, which has fewer resources than most other languages, but has a high-resource related language.

3.5. Adversarial samples

We extended the dataset by creating artificially hard (i.e., adversarial) samples in two ways.

The first way consisted of two methods for creating adversarial answers. The PISA data already

[3]Available at https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

[4]Portuguese is also considered a high-resource language in some categorizations, but we choose to follow the categorization in Joshi et al. [11].

Context: Eva Decroix: Here we have the case of Mr Novotný, in which the minister decided not to appoint him as a judge for reasons that he published.
The reasons why Mr Novotný was not appointed … *[text omitted]* … had previously objected because "linking the specific decisions … to whether he will be appointed or not" is considered unfortunate.

Question: How did the president of the Judges' Union's perspective on Minister Blažek's decision change after learning about the earlier reprimand, and how does this relate to the original reasons for not appointing Judge Novotný?

Rubrics:

- **FC:** The answer correctly explains that the president of the Judges' Union, Libor Vávra, indicated that had they known about the reprimand earlier, they might not have criticized Minister Blažek's decision. This relates to the original reasons for not appointing Judge Novotný, which were initially based on alleged illegal decisions causing financial damage, but later included the reprimand as a justification.
- **PC:** The answer partially explains the change in perspective of the president of the Judges' Union, Libor Vávra, regarding Minister Blažek's decision after learning about the reprimand. It mentions that this knowledge could have influenced their criticism, but lacks detail on how it directly relates to the original reasons for not appointing Judge Novotný.
- **NC:** The answer fails to explain how the president of the Judges' Union's perspective changed after learning about the reprimand and does not relate this to the original reasons for not appointing Judge Novotný.

Answer: Vávra said that if the Judges' Union had known about the 2020 reprimand before the debate, they probably wouldn't have objected to Blažek's refusal – the reprimand gave the minister an extra, concrete reason beyond the vague claim of illegal decisions that caused state losses. **Label:** 2

Figure 2: A sample instance from the DEMAGOG domain showing context, question, grading rubrics, and labeled answer. The dataset entry originally comes from Czech; this context utilizes our machine translation into English.

contained a wide variety of responses for each question. Therefore, to avoid unnecessarily increasing the dataset size, both methods were applied only to the English data in the **Additional** test set section of the dataset.

In the first way, we focused on adding answers to the questions in the **Additional** test set.

- The **Persona** method worked by generating a persona illustrating typical student failures based on a few selected traits known from the PISA literature. We uniformly sample three disadvantage traits for each answer from those in Figure 3 and use an LLM to create an answer given the persona description. After that, we use an LLM to explicitly clean any unnatural persona manifestations from the answer.
- The **Similarity** method was designed to create very hard adversarial examples by starting from reliably bad answers and selecting the ones with the highest semantic similarity to the full credit answers for a given question. The initial set of bad answers was, in part, made up of full credit answers belonging to items with different questions but with the same context. In part, it was made up of sequences of context sentences. Such sequences are copied verbatim from the text without providing an additional explanation as requested by all questions and are therefore considered No credit. A sequence with the best semantic similarity was selected from each of the approaches. The semantic similarity model used was `all-MiniLM-L6-v2` (version `c9745ed1d9f207416be6d2e6f8de32d1f16199bf`). We also considered applying this method to non-English data, but concluded that it is not worth the increase in the size of the **Additional** test set.

The second way to create adversarial samples focused on the rubrics rather than on the actual questions and answers. We utilized a method of manually altering the rubrics to create up to four

material_resource_deficit: Lack of quiet study space and modern hardware. – Differs from norm by forcing shared-space multitasking and tech-latency hurdles.

financial_precarity: Active concern regarding food security and household bills. – Differs from norm by redirecting cognitive bandwidth from learning to survival anxiety.

educational_capital_gap: Parents lack experience with higher education systems (ISCED < 3). – Differs from norm by reducing available at-home academic scaffolding and systemic 'know-how'.

linguistic_dissonance: Home language differs significantly from the test/instructional language. – Differs from norm by imposing a 'translation tax' on every cognitive task.

acculturation_friction: 1st or 2nd generation migrant status with evolving social networks. – Differs from norm through the added labor of navigating dual cultural expectations.

persistence_fragility: Tendency to disengage quickly when faced with non-obvious solutions. – Differs from norm by lacking the 'academic safety net' of previous successful struggle.

cognitive_ceiling_bias: Belief that ability is fixed and effort is a sign of low intelligence. – Differs from norm by interpreting errors as permanent failures rather than data points.

numerical_affective_load: High physiological stress response to mathematics and abstract logic. – Differs from norm by triggering an 'avoidance' reflex in place of analytical curiosity.

perceived_agency_deficit: Deep-seated doubt in one's ability to solve complex, novel problems. – Differs from norm by creating a reliance on external validation rather than internal logic.

metacognitive_burden: Compulsion to over-explain simple steps to ensure comprehension. – Differs from norm by wasting time on 'obvious' proofs, often due to a history of being misunderstood.

Figure 3: List of the used persona disadvantage traits.

distinct versions. Due to its labor-intensive nature, it was only viable with a few questions, so we chose those provided in the PISA data.

- As provided or **RubrOrig** – The PISA rubrics as provided do not expect the task to be formulated in plain natural language. Instead, they mention HTML elements like select boxes or text input fields based on their original setting in the web-like PISA test interface.
- Cleaned or **RubrClean** – In the cleaned version, make the rubrics "plaintext-savvy" by removing all the incompatible references to HTML elements and move information around so that the separate rubric parts do not reference each other.
- Removed examples or **RubrNoEx** – In this version, we take the **RubrClean** version and remove all the example answers from all the rubrics.
- Inexplicit with one example or **RubrAbstract** – This version is also based on the **RubrClean** version. Here, we remove all explicit mention of the correct solution and remove all but one example answer in each rubric.

Some of these modifications could not always be applied, e.g., when there is only one example in each rubric. For example, when the rubric did not contain any mentions of HTML elements, and the rubric parts were sufficiently standalone. In such cases, the initial rubric was already categorized as **RubrClean**, and as many of the remaining versions as possible are created.

3.6. PISA quiz item properties

In addition to the quizzes, the OECD data can be divided for analysis purposes into a number of properties collected experimentally or assigned by the quiz designers.

1. Some rubrics contained partial credit as an option (**FCPCNC**). Some did not (**FCNC**).
2. Most of the answers were strings of valid written language (**AnswerVerifiable**). However, some answers, despite being technically present, were simply invalid strings (**AnswerMissing**).
3. Some items required the student to compare multiple sections of text (**MultiContext**). Some sections concerned only one piece of text (**SingleContext**).
4. Some items were designated **Easy**, some **Difficult**, based on overall response accuracy.

4. Related works

4.1. Other methods for PISA item ASAG

Articles such as Zehner et al. [12], Yamamoto et al. [13], and Andersen et al. [14] already compare some methods of ASAG or human-assisted ASAG of PISA items. However, they do so only for a subset of languages selected by us or for a different set.

Our task tries to compare the performance of systems in a wider variety of domains and languages. Our task's baselines and contestant submissions also work with models that were not available at the time these previous articles were written.

4.2. Other ASAG datasets

There are a few existing ASAG datasets. However, they often do not provide course material as context and instead give a list of reference answers. For example, both Mohler et al. [15] and Nielsen et al. [16] provide only a single reference answer. This causes the questions examined in existing datasets to often be very simple and to focus on testing factual recall over complex understanding.

It is difficult for a teacher to be able to provide reference answers that completely cover the material without leaving out nuance, especially for answers that require providing reasoning and explanations using the course material.

The number of distinct questions in these datasets is relatively small, as is the diversity of domains they cover. Most focus on a single domain and do not evaluate how domain shift affects the approaches created using their train split. Also, many of these datasets are older than the dataset etiquette of the Internet scraping age and could have leaked into the training sets of many modern language models pre-trained on data from the Internet.

With the rise of generative models, it is also possible to create datasets to evaluate how well a model can explain what errors a pupil has made. The article Filighera et al. [17] is the first large-scale example of such a dataset.

4.3. Common ASAG methods

ASAG datasets, including those mentioned in the previous section, have led to many methods being developed. Some of these methods were evaluated in shared tasks or small-scale experiments that compared them, mainly using QWK, on a closed test set. The methods fall into a few distinct categories.

4.3.1. Feature based methods

Older feature-based methods often required the models to be trained or calibrated on a question-by-question basis, which limits their use to large-scale standardized student evaluation.

A parse tree based scoring system can be seen in Mohler et al. [15]. The article examined a perceptron trained on parse subtree matching features and a support vector machine trained on semantic similarity using methods such as LSA (latent semantic analysis). It also compared how well the features could be used for ranking compared to classification. The models achieved an RMSE of 0.978, which is substantially closer to the random baseline RMSE of 1.097 than to the inter-annotator RMSE of 0.659.

Another attempt at automated essay scoring, which has been applied in practice, is examined in Foltz et al. [18]. They trained a classifier using unspecified machine learning methods. The features used were a combination of simple grammatical and stylistic features and more complex features, such as LSA semantic similarity with reference answers.

This method was primarily developed for essay grading and later modified for ASAG. It was used as the second evaluator to complement a manual evaluation. Even then, only some kinds of questions could be partially scored automatically, and others required two human evaluators.

Foltz et al. [18] mentions that, in the experience of the auhtors, about half to two-thirds of the questions can be partially scored automatically, even though their experimental results in Table 5.3 of their work seem to show near human agreement for all the items chosen.

4.3.2. Sequence classification methods

The work Condor et al. [19] is very similar to our task setting. It examines the use of context sequences for the generalization of grading systems to unseen questions on a 0–4 scale.

The set of context sequences contains one or more of the following.

1. bundle one-hot (domain indication)
2. question
3. question context (any additional text such as figure descriptions)
4. scoring rubric (rubric giving about a sentence worth of description and a single example for each of the scores)

The approach used sentence embedding models (Sentence-BERT, Word2Vec, Bag-of-words) to create embeddings of the set of context sequences, and concatenated these embeddings and the embedding of the answer together into one vector. Then it trained a small classifier model on top of these embeddings. It compared a multinomial logistic regression (MLR) model with a multi-layer perceptron (MLP) model.

4.3.3. Methods based on sequence embedding similarity

All sequence embedding methods require a set of reference answers (sometimes referred to as a model or model answers in the literature), whose embeddings are compared to the given answer.

They can, for example, use a model pretrained with weakly supervised contrastive pretraining (applying contrastive loss to pairs that are not certain to be related) and finetuned on retrieval re-ranking, natural language inference (NLI), and other datasets that can be framed as sentence similarity problems. This is the approach used in Wang et al. [20]. Alternatively, any model finetuned using contrastive loss can be used, including a decoder-only LLM.

4.3.4. LLM based methods

In recent years, LLM as a judge has become a common way to benchmark question answering models. Several articles have been written on the topic of their reliability as reference-free benchmarks, especially in the multilingual setting [21].

There have also been several works on the use of LLMs for ASAG of student responses. For example, Chang and Ginter [22] examine the performance of the GPT-3.5 and GPT-4 models in ASAG of Finnish responses. The newer model, GPT-4, achieved a QWK score of more than 0.6 but its performance degraded on longer answers.

Team	System	Simple	Rubric	PISA	Additional	Paper
ASTeam	`Qwen3-32B few-shot`	✓	✓	–	✓	Jing Zhang and Dai [27]
Deep Thinkers	`HPLT`	✓	–	✓	✓	–
DIPF-TBA	`Tolegraa`	–	✓	✓	✓	–
FiAnSo	`Qwen3-Embedding-8B`	–	✓	✓	✓	Prášil and Otmar [28]
Pythoneers	`ModernBERT+bge-m3`	–	✓	✓	✓	Ahmetović and Olovčić [29]
Writers logic	`PoE Combination DeBERTa+other`	✓	✓	✓	✓	Condrey [30]
WSE-Research	`Qwen3.5`	✓	✓	✓	✓	Gwozdz and Both [31]

Table 1
The participating teams and their systems, indicating the tracks they took part in (*Simple* vs. *Rubric-based*) and which parts of the test set they provided outputs for (the **PISA** test set and/or the **Additional** test set). The last column gives the reference to the participant's system description paper, or "–" if no paper was provided

4.3.5. Other methods used for evaluating question answering

Historically, text generation models have been scored using metrics such as ROUGE [23] or BLEU [24].

Nowadays, there has been a shift towards trained metrics. These are either trained as sequence similarity models (e.g., BERTScore [25]), but can also be trained as sequence-to-sequence models (e.g., BARTScore [26]). Depending on the task that these are used to evaluate, they may be directly used for ASAG. For example, in the case where these are used to evaluate question answering models with reference to some golden answer. They may also at least be used to check if the answer is paraphrasing some other correct answer.

5. Participation

Of the 67 teams registered for the Topical Quiz task (of which Sensemaking is a subtask), 24 teams completed the additional registration step required for our Sensemaking task.

By the deadline, seven teams participated. Three teams participated in both tracks, three in the rubric-based track only, and one only in the simple track.

In addition to the primary submissions, there were 5 submitted experiments using different models for parts of the test set. Of these, eight were in the simple track, and seven were in the rubric-based track. In this overview, we focus mainly on the primary submissions.

6. Submissions

This section briefly describes our baselines and all participating systems. The submissions are summarized in Table 1.

6.1. LLM baselines

Our LLM baseline uses two models, `gemma-4-E4B-it` (version `d6436b3d62967e1af08bbb046c630-0b2a9ae8e85`) and `gemma-4-E2B-it` (version `905e84b50c4d2a365ebde34e685027578e6728db`). The architecture used by `gemma-4-E4B-it` and `gemma-4-E2B-it` allowed the effective size of the model for its effect on the number of operations performed to be around half that it would be if a normal transformer architecture with the same number of parameters was used. Both models still require roughly the same amount of memory as is common for other 8B and 5.1B models, for `gemma-4-E4B-it` and `gemma-4-E2B-it`, respectively. However, `gemma-4-E4B-it` requires roughly as many operations as a 4.5B model would require, and `gemma-4-E2B-it` requires as many operations as a 2.3B model would require.

Please rate the given answer on a scale from 0 to 2 based on the rubric provided with respect to the question and the information in the context. Think for a bit in English before giving your final rating. Only put the argumentation before the rating number, not after.

Figure 4: `gemma-4-E4B-it` prompt used for the rubric rating track

You are an expert academic evaluator known for meticulous attention to detail and strict adherence to scoring rubrics. Please rate the given answer as NC PC or FC based on the rubric provided with respect to the question and the information in the context. Think for a bit in English before giving your final rating as [[## 0 ##]], [[## 1 ##]] or [[## 2 ##]]

Figure 5: `gemma-4-E2B-it` prompt used for the rubric rating track

This decrease in the number of operations is achieved by dedicating part of the parameters to per-layer embeddings calculated by indexing an embedding matrix using the input token IDs. This indexing operation only requires a linear number of operations per sequence with respect to the length of the input sequence, as opposed to the quadratic number of operations full matrix multiplication takes. Such embeddings are then integrated into the hidden state in each layer after the attention and feed-forward operations. These details are drawn from the implementation here https://github.com/huggingface/transformers/blob/main/src/transformers/models/gemma4/modeling_gemma4.py.

The larger model `gemma-4-E4B-it` was used with a simple natural prompt (Figure 4), prompting the model to return its output after providing some reasoning. We did not tune the prompt much and included a few errors, which should not affect performance [32]. The final model answer was extracted from the tail part of its output. We either take the last character or, while preparing the system on the development set, we noticed it sometimes outputted the full word grade, so we check if it does not end with it. Otherwise, we just take the last number that occurs in the response. If there was no number, we output -1 (which is then mapped to 0 as per Section 2.4).

The inference for all LLM baselines was carried out using the vLLM engine with AWQ quantization and the temperature of 0.

The smaller model, `gemma-4-E2B-it`, had a significant failure rate when used with this approach, so instead we opted for prompting around an explicit format, see Figure 5. Again, the prompt was not tuned and contained a few errors.

The inference for all LLM baselines was carried out using the vLLM engine with the AWQ quantization and the temperature of 0. Contexts above 9000 tokens have been truncated to 9000 tokens so that the question, rubrics, answer, and response can fit into a 10000 overall limit.

For the next year, we are considering adding such extremely long contexts as part of a separate track or sub-track.

6.2. Encoder-only baselines

We created a number of models based on `EuroBERT`. After hyperparameter selection on the development set, we decided to use natural language inference (NLI) to compare how well different grade descriptions fit the answer.

We report the results of two models. The first one utilized context as the premise and the question, rubric, and answer sections together as a hypothesis. The other one did not use context at all and instead utilized the question and one of the rubrics as the premise and the answer as the hypothesis. The most fitting grade was then selected by taking the rubric with the highest entailment logit, see Figure 6. Both models were finetuned on a mix of a dataset similar to XNLI, named multilingual-NLI-26lang-2mil7[5] [33], the training set, and the development set.

We decided on using two variations of `EuroBERT-210m` (as provided in huggingface commit `39b51e15dd1f1a06f58b5cbf6a8a188cec60bd0e`). One variation was trained to take context into

[5] https://huggingface.co/datasets/MoritzLaurer/multilingual-NLI-26lang-2mil7

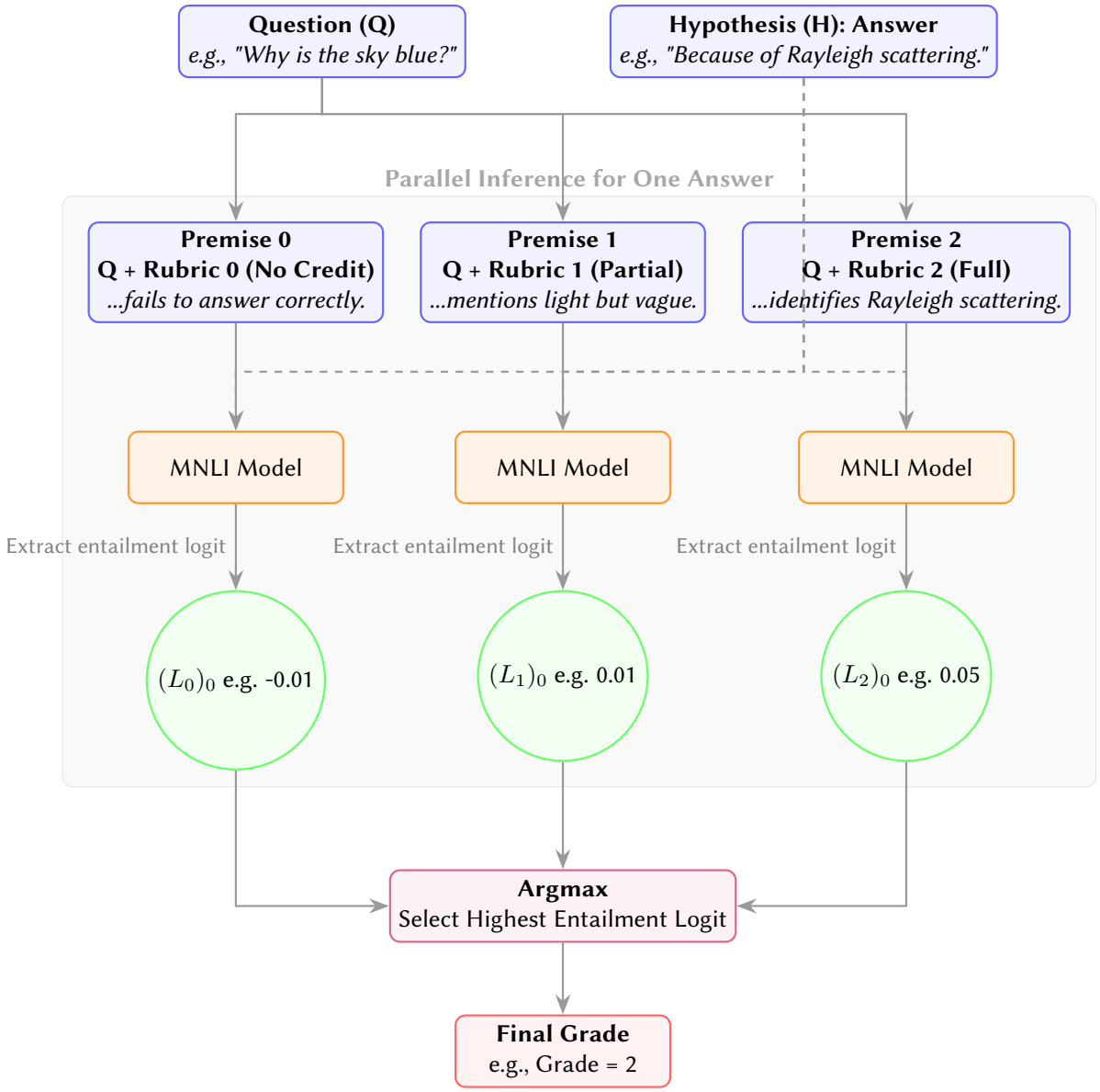

Figure 6: A visual summary of how our inference-based baseline classification model builds upon a pretrained model for NLI.

account when grading, and the other was trained to only rely on the rubrics.

We also experimented with a translation-based baseline based on `ModernBERT`, but because of its low score, we do not report its results in some of the charts.

The inference for all classifier encoder-only baselines was conducted using the transformers library with the weights cast to 16-bit floating point.

6.3. The baseline using embedding similarity

Our last baseline was based on embedding similarity between the answer and each of the rubrics. For the embedding we use the `intfloat/multilingual-e5-large` (version `0dc5580a448e4284468b-8909bae50fa925907bc5`) model [20]. Then we train a random forest classifier as provided by the scipy library [34] on 40 selected examples from the development set and select the number of trees using the entire development set using five-fold cross-validation.

The inference for the embedding similarity baseline was conducted using the sentence-transformers library without any quantization.

6.4. WSE-Research submission

The WSE-Research submission used `Qwen3.5-27B` with the temperature of 0 and reasoning disabled. They then analyze the continuation probabilities of the string representations of the grade numbers. Since `Qwen3.5-27B` assigns each number between 0 and 9 a single token, this is mathematically equivalent to analyzing the first token probabilities.

To improve their method, they use the few-shot retrieval method. Details can be found in Gwozdz and Both [31].

6.5. FiAnSo submission

The FiAnSo team relied on the pretrained `Qwen3-Embedding-8B` embedding model and used logistic regression on top of a set of similarities between pairs of input fields, except for the context. This team's system does not use the context provided at all.

More details are available in Prášil and Otmar [28].

6.6. Writers logic submission

The Writers logic team used a combination of systems using a kernel density estimation-calibrated Product of Experts [35]. The systems used ended up with relatively balanced weights.

The first system trained the CORAL [36] ordinal regression head on top of `mDeBERTa-v3-base` and `xlm-roberta-large`, with the second model having a weight of 70%.

The second system trained a gradient-boosted decision tree on top of the `SBERT` [37] model.

The third system retrieved similar items from the training dataset using `SBERT` nearest-neighbor retrieval with KDE-smoothed class posteriors. When no sufficiently close training dataset item existed, the scorer returned a fixed prior of $[0.28, 0.28, 0.44]$.

More details can be found in Condrey [30].

6.7. Deep Thinkers submission

The Deep Thinkers team only participated in the simple track. They submitted one primary submission for both the **Additional** test set and the **PISA** test set.

For the primary submission, the team used ModernBERT to get an entailment prediction for each pair of context/question/answer and separate rubric descriptors. They choose the descriptor with the highest overall score.

For HPLT models, the team constructs a prompt by explaining the rules, giving context, a question, an answer, and a one-shot example. From there, the procedure differs depending on the type of model.

For HPLT v2 and HPLT v3 GPT, they took whatever the model generated in an auto-regressive fashion and tried to extract any digit from its output. If there was no suitable digit in the output, an invalid value of -1 was returned.

For HPLT v3 BERT, the team includes a masked token as the very last token in the prompt, then they took the largest logit on that position and decoded it using the tokenizer. The decoded output was then searched for one of the labels.

The team also experimented with finetuning HPLT v3 BERT for classification and regression, as it showed the best results in the base experiments (Masked LM). When training for classification, the model tries to predict one of the labels. The model trained for regression outputs a single number, which was then clamped and rounded to the nearest result.

6.8. Pythoneers submission

The Pythoneers team trained a sequence similarity head on top of the `ModernBERT-base` model (for English responses) and the `BAAI/bge-m3` model (for non-English responses). The first sequence

compared was a string containing the question and the grade descriptor, the second was a string containing the answer. This system did not use the context in any way.

A more detailed description of the procedure can be found in Ahmetović and Olovčić [29].

6.9. DIPF-TBA submission

The DIPF-TBA team used the Tolegraa [38] rubric-pointer model trained on the rubric-based training set. The model jointly encodes the three rubric criteria, the student answer, and the question with a multilingual GTE backbone.

Then it assigns each label by computing alignment scores between the answer span and the corresponding rubric span. These alignment scores are treated as logits over the three rubric labels.

The maximum softmax probability was used as a confidence estimate for the Tolegraa prediction. To improve robustness, the team combines this model with an LLM-based rubric-prompting system. Predictions with Tolegraa confidence above 0.5 are retained, whereas lower-confidence cases are replaced by an LLM prediction.

6.10. ASTeam submission

The ASTeam submission only contains results for the **Additional** test set because they were not allowed to access the PISA data due to administrative reasons.

They used the model Qwen3-32B with the temperature of 0 and reasoning enabled, asking the model to first provide chain-of-thought reasoning and outputting the label last. For non-English answers, they also provide the English translation in addition to the answer. They provide a single example per class label. When extracting the final rating assigned by the model, they simply look at the last number output. More details can be found in Jing Zhang and Dai [27].

7. Overview of the simple rating track results

As mentioned before, to see how much the use of a rubric affects performance, we ran the simple rating track, which was rubric-less.

An overview of the results on most sections of the dataset can be seen in Figures 7 and 8. We can see that QWK here hardly ever surpasses 60% in some domains (e.g. LAWF or WORLDHIST) with some systems failing almost completely.

When required to grade using the full 5-point scale, the overall QWK is somewhat lower than when we map the scales into three classes, as can be seen in Tables 2 and 3, e.g. Overall QWK for wse_research of 46.4% vs. 43.4%. This is most likely caused by the change in the penalization QWK applies.

Overall, we see a big range of obtained QWK scores, from 0 to 46%, or up to 52% on the **PISA** test set.

According to QWK, the best performing Qwen model from WSE-Research and the 4B baseline Gemma are clearly more adept in the **PISA** test set since they reach higher scores than in the **Additional** test set.

A mixed pattern is seen when comparing all vs. **Native Additional** test set items. **Native** items are those with non-synthetic questions and answers: the participating models including the best WSE-Research benefit from synthetic questions (column "**Addi test set**" in Tables 2 and 3) while the baselines reach better scores only on the authentic data (column "**Addi test set Native**").

Tables 4 and 5 provide the accuracies. We can see that most models surpass the random baseline of 0.20 for 5-way classification or $0.\overline{3}$ for 3-way classification but even the best results of 0.397 in 5-way or of 0.610 in 3-way classification, which were obtained on **PISA** test set, still call for a serious improvement. In the simple rubric-less rating, participating models are not sufficiently reliable at scoring answers.

Compared to the QWK scores, the accuracies cover a smaller range of outcomes, ranging from 15% to 36% in the overall 5-way scores. The ranking of participating systems also somewhat changes: the 4B baseline Gemma fares substantially better on PISA than WSE-Research. Gemma 2B is more similar

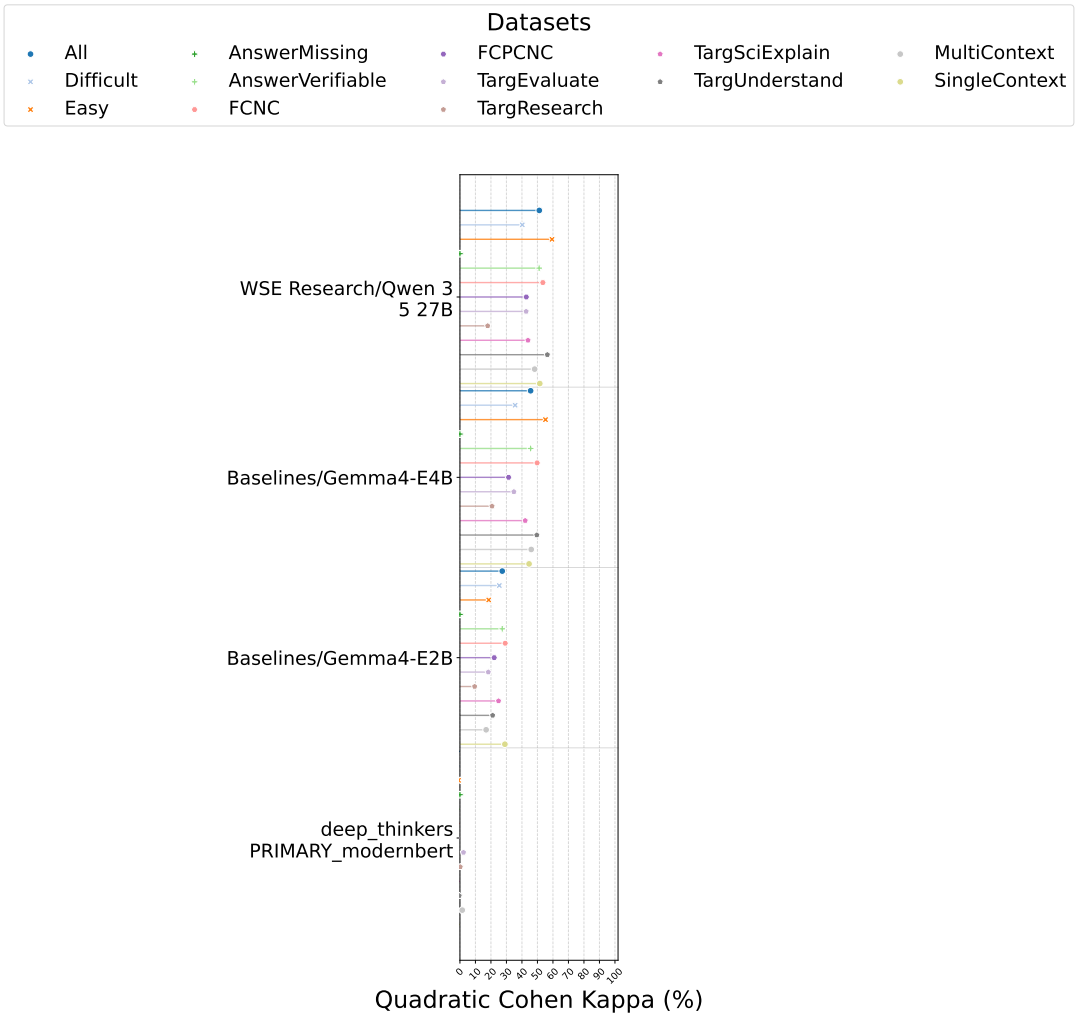

Figure 7: Comparison of the simple track model QWK scores across different subsets of the PISA dataset. Best viewed in color. Only systems that provided outputs for the simple track are included.

to WSE-Research than to the 4B version in that it reaches higher accuracies on **Additional** test set than on **PISA** test set. If our primary concern was the overall best performance, the weighting scheme would have been obviously critical.

When we focus on the **NotTranslated Native** items in Table 6, which compares the accuracies of **Additional** test set and **PISA** test set sources, we see that, after mapping to 3-way rating, the accuracy on both **Additional** test set and **PISA** test set is lower than the Overall ones: The WSE research submission achieved an accuracy of 0.314 on the **NotTranslated Native Additional** test set. This is a decrease from the QWK of 0.482 on the subset including **Native** responses in both **NotTranslated** and **Translated** varieties. It also achieved a lower QWK on the **PISA** test set, going from 0.581 to 0.297 (as mentioned in Section 3, **PISA** test set only contains **Native** data.). One possible explanation for the higher **Translated** scores is that the models are confused by shallow errors (e.g. typos and grammar) for grading correct answers and these errors tend to disappear with automatic translation. Furthermore, these shallow errors are present in **PISA** test set items and not so much in the **Additional** test set items, where the answers were created by adults, some of them experts in the given fields. This difference can also in part explain the WSE research increase in accuracy from 0.297 in the **PISA** test set up to 0.314 accuracy in the **Additional** test set for **NotTranslated Native** items (Table 6).

When one compares Table 3 and Table 9 (discussed in the following section), one can see the QWK increases substantially when we add a rubric to ground the grade, for example, the wse_research submission goes from 0.434 overall QWK in simple rating to 0.648 in rubric-based rating. QWK in

Figure 8: Comparison of the simple track model QWK across different domains of the **Additional** test set. Only systems that provided outputs for the simple track are included.

rubric-based rating.

Model Name	Overall QWK	Addi test set	PISA test set	Addi test set Native	devel set
WSE Research ↳Qwen 3.5 27B	**0.464**	**0.416**±008	**0.512**±020	**0.378**±009	**0.714**±018
Baselines ↳Gemma4-E4B	**0.378**	0.299±008	**0.456**±023	0.344±008	**0.662**±018
Baselines ↳Gemma4-E2B	**0.265**	0.256±008	**0.273**±022	0.304±007	0.465±022
WritersLogic ↳mDeBERTa-v3 PoE	NA	**0.423**±008	NA	**0.361**±009	0.536±023
ASTeam ↳qwen3_fewshot	NA	**0.416**±007	NA	**0.366**±008	**0.545**±023
deep_thinkers ↳modernbert	-0.030	-0.018±008	-0.043±025	-0.011±010	0.007±028
deep_thinkers ↳hpltv3_bert_base	NA	NA	-0.018±018	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_classification	NA	NA	-0.022±013	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_regression	NA	NA	-0.029±009	NA	NA

Table 2
The simple track QWK. Here, we compare the results for predicting the full 5-class scale. The value after ± signifies the bootstrap confidence interval [39] ($p < 0.05$) based on 999 resamples. Colors are scaled from best (dark red) to worst (dark blue) for each column separately. The WritersLogic team provided their simple-rating grades only for the **Additional** test set, not the **PISA** test set. Similarly, Deep Thinkers did not provide **Additional** test set grades for this track. ASTeam did not provide **PISA** test set grades in any track.

Model Name	Overall QWK	Addi test set	PISA test set	Addi test set Native	devel set
WSE Research ↳Qwen 3.5 27B	**0.434**	**0.349**±007	**0.520**±022	**0.295**±010	**0.719**±018
Baselines ↳Gemma4-E4B	**0.351**	0.262±009	**0.439**±025	0.283±008	**0.639**±020
Baselines ↳Gemma4-E2B	**0.247**	0.221±008	**0.274**±025	0.252±007	0.391±023
WritersLogic ↳mDeBERTa-v3 PoE	NA	**0.355**±008	NA	**0.292**±010	**0.562**±022
ASTeam ↳qwen3_fewshot	NA	**0.363**±008	NA	**0.311**±008	0.536±024
deep_thinkers ↳modernbert	-0.018	-0.016±009	-0.019±028	-0.012±011	0.016±030
deep_thinkers ↳hpltv3_bert_base	NA	NA	-0.024±020	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_classification	NA	NA	-0.022±012	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_regression	NA	NA	-0.035±011	NA	NA

Table 3
The simple track QWK. The predictions are mapped to 3 different ratings to allow for comparison with the rubric-based track, even though the model predicted one of the 5 different ratings. The value after ± signifies the bootstrap confidence interval [39] ($p < 0.05$) based on 999 resamples. Colors are scaled from best (dark red) to worst (dark blue) for each column separately.

Model Name	Overall Accuracy	Addi test set	PISA test set	Addi test set Native	devel set
WSE Research ↳Qwen 3.5 27B	**0.300**	**0.311**	0.290	**0.307**	**0.517**
Baselines ↳Gemma4-E4B	**0.364**	**0.331**	**0.397**	**0.320**	**0.480**
Baselines ↳Gemma4-E2B	**0.259**	0.297	0.220	0.280	**0.450**
WritersLogic ↳mDeBERTa-v3 PoE	NA	0.306	NA	0.290	0.415
ASTeam ↳qwen3_fewshot	NA	**0.328**	NA	**0.307**	0.444
deep_thinkers ↳modernbert	0.154	0.188	0.121	0.188	0.209
deep_thinkers ↳hpltv3_bert_base	NA	NA	**0.346**	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_classification	NA	NA	**0.362**	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_regression	NA	NA	0.023	NA	NA

Table 4
The simple track accuracy. Here, we compare the unmapped results given when predicting one of the 5 classes. The random baseline would thus be 0.2.

Model Name	Overall Accuracy	Addi test set	PISA test set	Addi test set Native	devel set
WSE Research ↳Qwen 3.5 27B	**0.537**	**0.494**	**0.581**	**0.482**	**0.712**
Baselines ↳Gemma4-E4B	**0.550**	0.490	**0.610**	**0.473**	**0.657**
Baselines ↳Gemma4-E2B	**0.465**	0.431	**0.498**	0.410	0.548
WritersLogic ↳mDeBERTa-v3 PoE	NA	**0.498**	NA	0.466	**0.627**
ASTeam ↳qwen3_fewshot	NA	**0.514**	NA	**0.481**	0.613
deep_thinkers ↳modernbert	0.373	0.358	0.389	0.358	0.360
deep_thinkers ↳hpltv3_bert_base	NA	NA	0.420	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_classification	NA	NA	0.397	NA	NA
deep_thinkers ↳hpltv3_bert_finetuned_regression	NA	NA	0.308	NA	NA

Table 5
The simple track accuracy with outputs mapped to 3-way ratings to allow for comparison with the rubric-based track, even though the simple-rating systems originally return 5-way ratings. The random baseline in this interpretation would be $\overline{0.3}$.

Model Name	Addi manual non translated	PISA manual non translated
WSE Research		
↳Qwen 3.5 27B	**0.314**	0.297
Baselines		
↳Gemma4-E4B	**0.325**	**0.412**
Baselines		
↳Gemma4-E2B	0.292	0.207
WritersLogic		
↳mDeBERTa-v3 PoE	0.299	NA
ASTeam		
↳qwen3_fewshot	**0.319**	NA
deep_thinkers		
↳modernbert	0.191	0.116
deep_thinkers		
↳hpltv3_bert_base	NA	**0.335**
deep_thinkers		
↳hpltv3_bert_finetuned_classification	NA	**0.344**
deep_thinkers		
↳hpltv3_bert_finetuned_regression	NA	0.024

Table 6
Comparison of accuracy for the **NotTranslated Native** subset in the simple track in the 3-way simplification of the original 5-way rating. Colors are scaled from best (dark red) to worst (dark blue) for each column separately.

8. Overview of the rubric-based rating track results

As stated before, the rubric rating-based track used a smaller set of labels (NC/PC/FC) and gave the models access to a rubric to ground the grading in.

To get a broad overview of the results in different sections of **PISA** test set consult Figure 9. To get a broad overview of the rubric-based rating track results in different **Additional** test set domains, consult Figure 10.

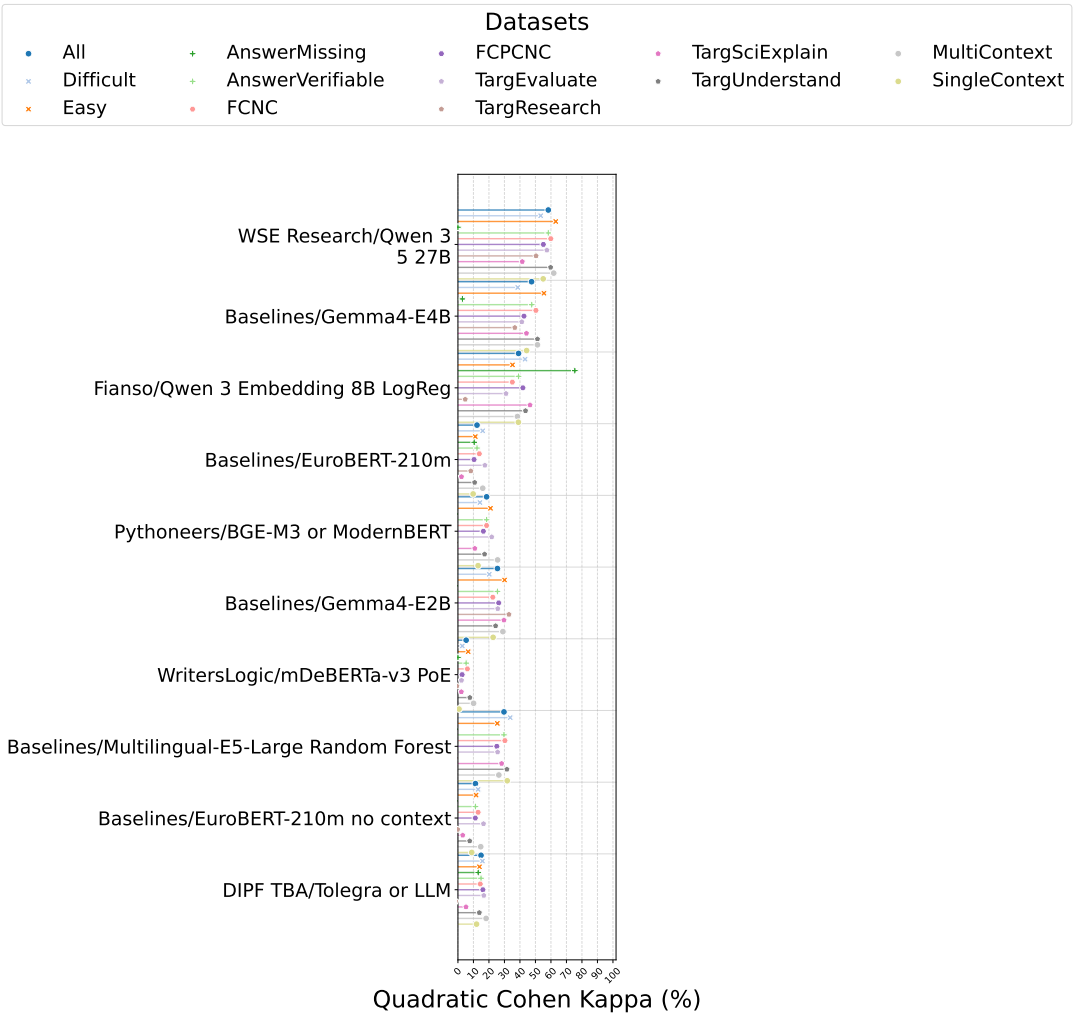

Figure 9: Comparison of the rubric-based track model QWK across different parts of the PISA dataset.

8.1. Overall model performance comparison

As can be seen in Table 9 below, the larger LLM models had the best results. The best overall results are achieved by WSE-Research's `Qwen3.5-27B` with 0.648 Overall and 0.583 QWK in the **PISA** test set, and by ASTeam's `qwen35_fewshot` with 0.718 QWK in the **Additional** test set. The encoder-only and embedding based models (where the best one is based on a decoder only model) follow relatively far behind, with the FiAnSo `Qwen3-Embedding-8B` based model achieving 0.391 QWK in the **PISA** test set and EuroBERT-210m baseline reaching 0.539 QWK in the **Additional** test set.

The test set had three sections, which used different generation methods. All sections except for the **PISA** test set items had their rubrics automatically generated. Also annotator-created (called **Native**) answers in the original language were, except for **PISA** test set, provided for only a few languages and only a few domains.

8.2. Performance on the Native NotTranslated subset

As can be seen in Table 7, the best model achieves a solid QWK of 0.675 and the second best 0.606 on the "most genuine" student answers, i.e. **Native NonTranslated** subset of items. However, in Table 8 we can see that for both models a high QWK in this section of the test set does not result in a particularly high accuracy: the best model achieves an accuracy of 0.638 and the second best an accuracy of 0.631. In simple terms, about 4 out of 10 student responses would be rated badly if we relied on these models.

Model Name	Addi manual non translated	PISA manual non translated
WSE Research		
↳Qwen 3.5 27B	**0.675**±027	**0.593**±011
Baselines		
↳Gemma4-E4B	**0.606**±034	**0.511**±012
Fianso		
↳Qwen 3 Embedding 8B LogReg	0.479±037	**0.430**±012
Baselines		
↳EuroBERT-210m	0.450±040	0.118±014
Pythoneers		
↳BGE-M3 or ModernBERT	0.368±043	0.213±012
Baselines		
↳Gemma4-E2B	0.304±041	0.295±011
WritersLogic		
↳mDeBERTa-v3 PoE	0.351±035	0.052±005
Baselines		
↳Multilingual-E5-Large Random Forest	0.182±040	0.327±014
Baselines		
↳EuroBERT-210m no context	0.391±042	0.125±015
DIPF TBA		
↳Tolegra or LLM	0.136±037	0.151±011
ASTeam		
↳qwen3_fewshot	**0.651**±029	NA

Table 7
Comparison of QWK for the **NotTranslated Native** answers in the rubric track. The value after ± signifies the bootstrap confidence interval [39] ($p < 0.05$) based on 999 resamples. Colors are scaled from best (dark red) to worst (dark blue) for each column separately.

Model Name	Addi manual non translated	PISA manual non translated
WSE Research		
↳Qwen 3.5 27B	**0.638**	**0.712**
Baselines		
↳Gemma4-E4B	**0.631**	**0.652**
Fianso		
↳Qwen 3 Embedding 8B LogReg	0.542	0.494
Baselines		
↳EuroBERT-210m	0.606	0.449
Pythoneers		
↳BGE-M3 or ModernBERT	0.536	0.357
Baselines		
↳Gemma4-E2B	0.513	**0.537**
WritersLogic		
↳mDeBERTa-v3 PoE	0.444	0.179
Baselines		
↳Multilingual-E5-Large Random Forest	0.375	0.535
Baselines		
↳EuroBERT-210m no context	0.522	0.444
DIPF TBA		
↳Tolegra or LLM	0.363	0.492
ASTeam		
↳qwen3_fewshot	**0.624**	NA

Table 8
Comparison of accuracy for the **NotTranslated Native** answers in the rubric track. Colors are scaled from best (dark red) to worst (dark blue) for each column separately.

8.3. Comparison of the results in the development and test sets

The development set items were available to the contestants for the sake of developing their systems. Many contestants either trained on some large portion of this development set or did substantial model selection.

Due to this, as can be seen in Table 9, the results on the development set are substantially inflated for some systems. An additional factor could be that the development set used fewer domains than the full test set, and so it was generally easier. This is further supported by the observation that higher devset scores were achieved even by systems such as the `gemma-4-E2B-it` baseline, where we only used the development set to ensure basic functionality.

8.4. Comparison of QWK and Accuracies

When we compare Tables 9 and 10, we can see that high QWKs do not correspond to particularly high accuracies. This means that the models often predict a wrong rating which is nevertheless close to the correct one (e.g., they grade a No Credit answer as Partial Credit); as we know QWK is more forgiving for close misses than Accuracy. In the **Additional** test set best model, an accuracy of only 0.656 is obtained despite a QWK of 0.713. However, the pattern seems to be the opposite for **PISA** test set where the best model gets 0.583 QWK and an accuracy of 0.697. This is likely due to the smaller number of responses in that dataset graded as Partial Credit.

8.5. Granular comparison of the results

For the sake of comparison, we split the previously described data collection methods into three main categories. We compare how much the results drop from the development set results in each of them.

Model Name	Overall QWK	Addi test set	PISA test set	Addi test set Native	devel set
WSE Research					
↳Qwen 3.5 27B	**0.648**	**0.713**±005	**0.583**±008	**0.723**±007	**0.863**±010
Baselines					
↳Gemma4-E4B	**0.506**	0.536±008	**0.475**±008	**0.634**±009	0.764±017
Fianso					
↳Qwen 3 Embedding 8B LogReg	**0.433**	0.476±008	**0.391**±009	0.462±010	0.517±025
Baselines					
↳EuroBERT-210m	0.331	**0.539**±008	0.123±009	0.564±009	0.548±024
Pythoneers					
↳BGE-M3 or ModernBERT	0.317	0.449±009	0.185±010	0.458±011	**0.825**±018
Baselines					
↳Gemma4-E2B	0.286	0.317±011	0.255±008	0.363±013	0.527±026
WritersLogic					
↳mDeBERTa-v3 PoE	0.243	0.433±008	0.053±004	0.410±009	**1.000**±000
Baselines					
↳Multilingual-E5-Large Random Forest	0.243	0.189±009	0.297±010	0.165±010	0.415±025
Baselines					
↳EuroBERT-210m no context	0.238	0.363±009	0.113±010	0.375±012	0.528±027
DIPF TBA					
↳Tolegra or LLM	0.188	0.226±007	0.149±008	0.154±008	0.368±024
ASTeam					
↳qwen3_fewshot	NA	**0.718**±005	NA	**0.705**±007	0.816±013

Table 9
Comparison of QWK across sub-datasets. Colors are scaled from best (dark red) to worst (dark blue) for each column separately. The value after ± shows the confidence interval ($p < 0.05$) based on 999 bootstrap samples.

8.5.1. The Generated Additional test set items

These items had answers and "golden-truth" ratings automatically generated based on the context and question. We used the same model as we used to generate the development set. They dominate the overall **Additional** test set, and to examine them, we need only look at the overall results. The drop in accuracy and QWK compared to the results on the development set was rather small, as can be seen in Table 9. The best model drops from the development set QWK of 0.863 to 0.713 in overall QWK. This is a large drop in absolute terms, but considering that the model was prepared on the development set, and compared to the other subsections, it is relatively small.

8.5.2. The Additional test set Native items

These items had only their rubric automatically generated and were partially automatically translated; the drop here compared to the development set is similar to the drop in the fully automatically generated data (discussed in the previous paragraph). For the best model, the QWK of 0.723 in this section is actually slightly higher than the **Additional** test set QWK of 0.713 so the drop from 0.863 is less severe but for example for the `Qwen3-Embedding-8B`-based model, the drop from the development set is higher. It goes from 0.517 to 0.462 vs 0.476 on the **Additional** test set.

8.5.3. The PISA items

These items were the most natural ones and had human-authored rubrics. The drop in this part of the test set was the most pronounced, e.g. from development QWK of 0.764 to 0.475 for `gemma-4-E4B-it`. LLMs maintained their metric scores better than the pretrained models finetuned on the training data.

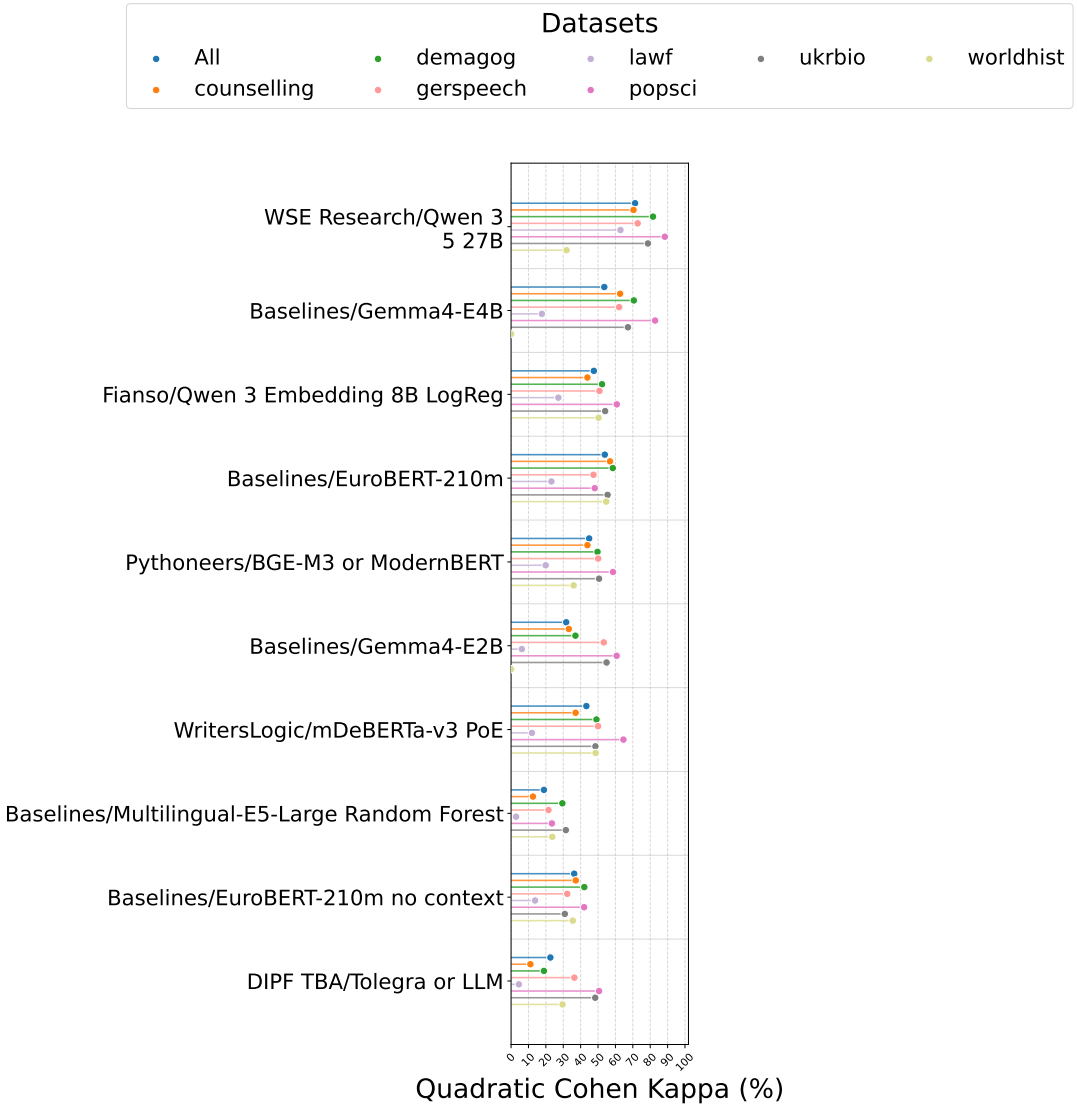

Figure 10: Comparison of the rubric-based track model QWK across different domains of the **Native** dataset.

Those models suffered significant drops in both metrics (e.g. drop from development QWK of 0.548 to 0.123 `EuroBERT-210m`).

Model Name	Overall Accuracy	Addi test set	PISA test set	Addi test set Native	devel set
WSE Research					
↳Qwen 3.5 27B	**0.676**	**0.656**	**0.697**	**0.666**	0.809
Baselines					
↳Gemma4-E4B	**0.609**	0.594	**0.624**	**0.629**	0.756
Fianso					
↳Qwen 3 Embedding 8B LogReg	0.502	0.521	0.482	0.511	0.588
Baselines					
↳EuroBERT-210m	**0.528**	**0.606**	0.451	0.618	0.657
Pythoneers					
↳BGE-M3 or ModernBERT	0.447	0.535	0.359	0.538	**0.882**
Baselines					
↳Gemma4-E2B	0.526	0.535	0.516	0.556	0.699
WritersLogic					
↳mDeBERTa-v3 PoE	0.338	0.487	0.189	0.474	**1.000**
Baselines					
↳Multilingual-E5-Large Random Forest	0.443	0.366	**0.519**	0.356	0.526
Baselines					
↳EuroBERT-210m no context	0.464	0.493	0.434	0.501	0.635
DIPF TBA					
↳Tolegra or LLM	0.431	0.379	0.484	0.337	0.486
ASTeam					
↳qwen3_fewshot	NA	**0.671**	NA	**0.663**	0.773

Table 10
Comparison of accuracy across sub-datasets. Colors are scaled from best (dark red) to worst (dark blue) for each column separately.

9. Performance across different domains and task varieties

In this section, we continue with detailed analyses of the rubric-based rating task. As could be seen in Section 8, the performance of the models varies widely across different sections of the data. The resulting performance drops are also not unique to comparing human-annotated and machine-generated data, but also happen when comparing across different domains.

9.1. Performance across different domains

The test set data contained domains where contexts were of different lengths and levels of formatting than in other domains. For example, in the history and law domains, the texts were often very long, and the questions dealt with a nuanced understanding of different sections of the text. It is likely due to this that most models have substantially lower QWK in these domains. For instance, the best model in Table 11 suffers a drop from the overall 0.713 to 0.319 when dealing with history and a drop from 0.704 to 0.629 when comparing the performance in constitutional court documents with the similar domain of legal counselling.

9.2. Performance across different rubric varieties

As was mentioned previously, we created up to four versions of each rubric in the PISA rubric-based track. As can be seen in Table 12, the two rubric variants between which we see a significant decrease in performance are **RubrClean** and **RubrAbstract**. Here, all LLMs suffer a significant decrease in QWK. The best model goes from the QWK of 0.649 for **RubrClean** to QWK of 0.612 for **RubrAbstract** on the very same set of underlying questions and answers. The only difference between these sets of inputs for these ratings is that we have removed the explicit mention of the correct answer and all but one

Model Name	Overall	Legal Counselling	History	Constitutional Law
WSE Research				
↳Qwen 3.5 27B	**0.713**±005	**0.704**±009	0.319±024	**0.629**±030
Baselines				
↳Gemma4-E4B	0.536±008	**0.627**±010	0.000±000	0.177±039
Fianso				
↳Qwen 3 Embedding 8B LogReg	0.476±008	0.439±012	**0.503**±026	**0.272**±045
Baselines				
↳EuroBERT-210m	**0.539**±008	0.569±011	**0.547**±024	0.232±047
Pythoneers				
↳BGE-M3 or ModernBERT	0.449±009	0.439±013	0.360±030	0.199±050
Baselines				
↳Gemma4-E2B	0.317±011	0.332±015	0.000±000	0.062±049
WritersLogic				
↳mDeBERTa-v3 PoE	0.433±008	0.371±011	0.486±023	0.120±039
Baselines				
↳Multilingual-E5-Large Random Forest	0.189±009	0.126±012	0.237±026	0.029±035
Baselines				
↳EuroBERT-210m no context	0.363±009	0.372±014	0.355±032	0.138±046
DIPF TBA				
↳Tolegra or LLM	0.226±007	0.111±008	0.296±024	0.045±025
ASTeam				
↳qwen3_fewshot	**0.718**±005	**0.688**±008	**0.709**±017	**0.640**±033

Table 11
The comparison of the QWK across different domains of the **Additional** test set. The value after ± signifies the bootstrap confidence interval [39] ($p < 0.05$) based on 999 resamples. Colors are scaled from best (dark red) to worst (dark blue) for each column separately.

example answer. This means that the models are not able to make sense of the context to reconstruct these removed parts.

Additionally, as we can see in Table 12, most models suffer a drop in QWK when they are given the rubrics uncleaned or without examples. An especially large drop can be seen for LLMs when the rubric does contain examples (**RubrNoEx**). The QWK of the relatively low-performing `gemma-4-E2B-it` goes from 0.338 to 0.228. This means that it suffers a larger drop than `multilingual-E5`, which fell from 0.374 to 0.293. This is interesting because `multilingual-E5` does not use the context at all, while `gemma-4-E2B-it` has access to the context and so should be able to deduce the correct answers. All the other LLMs also suffer a drop in QWK. This suggests that at least some LLMs are not able to make sense of the context and the rubric enough to not require examples.

The only models that do not substantially suffer in the **RubrNoEx** split are the models where the **RubrClean** best case QWK was already quite low.

9.3. False positive rate across different rubrics

An interesting outcome we observed was that some systems substantially over-predicted the partial credit grade, see Table 13. We can see that the Writers logic submission has an extremely high false positive rate of 0.968 which goes up to 0.989 when presented with the rubrics as provided (i.e., **RubrOrig**). The Pythoneers' submission has a false positive rate of 0.396 on the cleaned rubrics. When given the rubrics as provided, i.e. with HTML-like markup and referring to non-text information, Pythoneers predict PC much more often and the false positive rate jumps to 0.605.

Model Name	Rubr Clean where Rubr Abstract exists	Rubr Abstract	Rubr Clean where Rubr Orig exists	Rubr Orig	Rubr Clean where Rubr NoEx exists	Rubr NoEx
WSE Research						
↳Qwen 3.5 27B	**0.649***	**0.612**	**0.592***	**0.579**	**0.677***	**0.558**
Baselines						
↳Gemma4-E4B	**0.562***	**0.498**	**0.500***	**0.474**	**0.591***	**0.469**
Fianso						
↳Qwen 3 Embedding 8B LogReg	**0.408**	**0.397**	0.331	**0.353***	**0.449**	**0.437**
Baselines						
↳EuroBERT-210m	0.127	0.146	0.114	0.104	0.129	0.152
Pythoneers						
↳BGE-M3 or ModernBERT	0.187	0.247*	0.106*	0.035	0.234	0.288*
Baselines						
↳Gemma4-E2B	0.339*	0.278	**0.334**	0.313	0.338*	0.228
WritersLogic						
↳mDeBERTa-v3 PoE	0.047	0.068*	0.026	0.021	0.041	0.079*
Baselines						
↳Multilingual-E5-Large Random Forest	0.357	0.343	0.275*	0.242	0.374*	0.293
Baselines						
↳EuroBERT-210m no context	0.094	0.078	0.108	0.168*	0.082	0.168*
DIPF TBA						
↳Tolegra or LLM	0.075	0.244*	0.040	0.079*	0.069	0.190*

Table 12
Comparison of QWK with the "golden-truth" label in the subset of items from **PISA** test set with adversarial rubrics. The QWK is compared on items where the modification is applicable. Colors are scaled from best (dark red) to worst (dark blue). The mark * indicates that a value is significantly ($p < 0.05$) larger than the other rubric variant evaluated in the same pair (e.g. "RubrClean where a RubrAbstract exists", compared to the "RubrAbstract" version) based on a permutation test [39] with 999 resamples.

Model Name	Rubr Abstract	Rubr Clean	Rubr Orig	Rubr NoEx
WSE Research				
↳Qwen 3.5 27B	**0.096**	**0.079**	**0.101**	**0.089**
Baselines				
↳Gemma4-E4B	**0.139**	**0.105**	0.131	**0.100**
Fianso				
↳Qwen 3 Embedding 8B LogReg	0.377	0.274	0.336	0.247
Baselines				
↳EuroBERT-210m	0.143	0.133	**0.129**	**0.153**
Pythoneers				
↳BGE-M3 or ModernBERT	0.452	0.396	0.605	0.428
Baselines				
↳Gemma4-E2B	0.190	0.169	0.151	0.157
WritersLogic				
↳mDeBERTa-v3 PoE	0.954	0.968	0.989	0.955
Baselines				
↳Multilingual-E5-Large Random Forest	0.301	0.306	0.193	0.291
Baselines				
↳EuroBERT-210m no context	0.200	0.140	0.169	0.166
DIPF TBA				
↳Tolegra or LLM	**0.141**	**0.023**	**0.020**	0.186

Table 13
Partial Credit false positive rates in **PISA** test set for different rubric kinds. PC FP rate reflects how often the model wrongly predicted PC, i.e. it predicted PC but the "golden truth" answer was a different one. Colors are scaled from best (dark red) to worst (dark blue).

10. Impact of the language

Our final analysis of the rubric-based rating results concerns multilinguality, i.e. the impact of the underlying language on the systems' performance.

10.1. Performance on lower- vs. higher-resourced languages

One hypothesis we had was that the performance of many models would be substantially worse in lower-resource languages as classified by Joshi et al. [11]. For this purpose, we include two high-resource languages, English and German, and a number of lower-resource languages.[6]

Most of the languages selected, other than English and German, have more than 9 million native speakers, are official languages of the European Union, or have language resources available for them or a closely related language. For these reasons, almost all of them, except for Irish, fall into categories 3 and 4 [11], both of which have a substantial coverage in semi-supervised pretraining datasets due to their significant presence on the Internet.

Unfortunately, the responses for the Irish, the only language classified as category 2 [11], could not be provided by the OECD. The same is true for Romanian. We could also only get a subset of answers for Ukrainian. In the end, our analysis across languages compares English (class 5) with Czech (class 4) and Danish and Greek (class 3).

The comparison of performance in subsets of different languages can be seen in Table 14. The comparison shows that while some models perform worse in data from some languages, the matter of which languages are problematic varies heavily from model to model. Across the examined set, the decreases in QWK are not substantial, especially for LLMs. Only the `gemma-4-E4B-it` model actually has a substantial drop in QWK across languages: 0.552 in English vs 0.511 overall. For other models, the performance on English is better than the overall performance but not substantially so. For example the best model's performance only fell from 0.593 to 0.613.

Ultimately, these results show either that the answers in the other languages are easier to rate (which is conceivable but unlikely because we received a balanced set of student answers from OECD) or that the models are indeed proficient in these languages.

10.2. Model consistency across languages

As mentioned in Section 3.3, we machine-translate some responses into all the selected languages to see how fair the tested models are across different languages (see Section 3.4). The contexts and questions used for these items were the professional context and question localizations used for that language, which sometimes change small details from the original question to better fit the language's cultural background, and so it is possible that the machine translation of the response does not adequately address the localizations. Finally, we measure how well the models agree with their rating of the original item when the item is presented in the language it has been translated into.

Overall, as can be seen in Table 15, the average agreement between the model's original prediction and its prediction for the translated answer across languages is relatively strong for large language models, embedding-based models, and substantially smaller for the encoder-only models trained on the training set. One issue is that the model with the best agreement also belongs to the same family of models used to perform the translation, which likely skews the results. Another factor that could be important here is that the non-English translations are sometimes not as fluent as the original answer and inherit some properties of the English original, and so could be harder for the systems to understand. We also do not check all translations for correctness; we acknowledge that translation errors could be a substantial factor; unfortunately we are forced to leave a more thorough analysis of the translations for future work.

[6] Portuguese is also considered a high-resource language in some categorizations, but we choose to follow the categorization in Joshi et al. [11].

Table 16 shows us that when we compare the QWK between the original and translated data on the same subsets, the significant differences mostly appear in the models where the results were poor to begin with. It is particularly interesting to see that the QWK is sometimes significantly higher on the translated data. This seems to happen most often with the encoder-only based systems. The highest QWK for which this occurs is the 0.250 QWK on the Greek translation as opposed to the 0.210 QWK on the **NotTranslated** set. This is especially strange given that the **NotTranslated** set contains answers in higher-resource languages.

Model Name	All langs	En (5)	Cs (4)	Da (3)	El (3)
WSE Research					
↳Qwen 3.5 27B	**0.593**±011	**0.613**±038	**0.578**±034	**0.590**±038	**0.710**±030
Baselines					
↳Gemma4-E4B	**0.511**±012	**0.552**±037	**0.531**±040	**0.560**±040	**0.540**±037
Fianso					
↳Qwen 3 Embedding 8B LogReg	**0.430**±012	**0.477**±037	**0.418**±035	**0.448**±041	**0.372**±041
Baselines					
↳EuroBERT-210m	0.118±014	0.187±047	0.075±039	0.147±052	0.025±041
Pythoneers					
↳BGE-M3 or ModernBERT	0.213±012	0.250±045	0.179±040	0.289±039	0.155±043
Baselines					
↳Gemma4-E2B	0.295±011	0.328±039	0.316±036	0.344±039	0.266±037
WritersLogic					
↳mDeBERTa-v3 PoE	0.052±005	0.094±023	0.045±015	0.038±014	0.059±019
Baselines					
↳Multilingual-E5-Large Random Forest	0.327±014	0.366±042	0.337±041	0.379±045	0.280±043
Baselines					
↳EuroBERT-210m no context	0.125±015	0.144±047	0.133±044	0.159±052	0.054±044
DIPF TBA					
↳Tolegra or LLM	0.151±011	0.229±043	0.095±033	0.210±045	0.109±038

Table 14
Comparison of QWK of **NotTranslated PISA** test set answers across languages. Colors are scaled from best (dark red) to worst (dark blue). The value after ± signifies the bootstrap confidence interval [39]($p < 0.05$) based on 999 resamples. The value in the parentheses after the language signifies its category (see Section 3.4). Colors are scaled from best (dark red) to worst (dark blue).

Model Name	Avg Agreement	Min Agreement	En (5) Agreement	Pt (4) Agreement	Da (3) Agreement	El (3) Agreement
WSE Research						
↳Qwen 3.5 27B	**0.870**	**0.852**	**0.865**	**0.885**	**0.878**	**0.852**
Baselines						
↳Gemma4-E4B	**0.719**	**0.693**	**0.715**	**0.725**	**0.742**	0.693
Fianso						
↳Qwen 3 Embedding 8B LogReg	**0.793**	**0.749**	**0.795**	**0.805**	**0.824**	**0.749**
Baselines						
↳EuroBERT-210m	0.346	0.285	0.285	0.370	0.381	0.346
Pythoneers						
↳BGE-M3 or ModernBERT	0.505	0.235	0.235	0.602	0.593	0.590
Baselines						
↳Gemma4-E2B	0.444	0.429	0.429	0.439	0.450	0.456
WritersLogic						
↳mDeBERTa-v3 PoE	0.656	0.569	0.569	0.596	0.736	**0.723**
Baselines						
↳Multilingual-E5-Large Random Forest	0.583	0.459	0.628	0.459	0.667	0.579
Baselines						
↳EuroBERT-210m no context	0.292	0.219	0.346	0.321	0.280	0.219
DIPF TBA						
↳Tolegra or LLM	0.699	0.657	0.657	0.719	0.704	0.714

Table 15
The comparison between the average and minimum QWK agreement of the predictions for the translation and the original over all languages. This QWK agreement is calculated similar to the QWK with the "golden-truth" grade except here we use the model's original response as the "golden truth" and see if its prediction for the translated response differs. Since all models gave at least one instance of each grade level (NC/PC/FC) the QWK is by its nature invariant to the order of comparison. The value in the parentheses after the language signifies its category (see Section 3.4). Colors are scaled from best (dark red) to worst (dark blue).

Model Name	Pt (4) Reference	Pt (4) Final	Da (3) Reference	Da (3) Final	El (3) Reference	El (3) Final
WSE Research ↳Qwen 3.5 27B	**0.558**	**0.566**	**0.564**	**0.567**	**0.527**	**0.531**
Baselines ↳Gemma4-E4B	0.444	0.452	0.460	0.479	0.456*	0.414
Fianso ↳Qwen 3 Embedding 8B LogReg	0.379	0.393	**0.363**	**0.346**	0.364	0.356
Baselines ↳EuroBERT-210m	0.103	0.165*	0.120	0.117	0.109	0.168*
Pythoneers ↳BGE-M3 or ModernBERT	0.092	0.186*	0.111	0.182*	0.118	0.114
Baselines ↳Gemma4-E2B	0.236	0.249	0.230*	0.190	0.231	0.203
WritersLogic ↳mDeBERTa-v3 PoE	0.046	0.061	0.050	0.058	0.048	0.060
Baselines ↳Multilingual-E5-Large Random Forest	0.261	0.239	0.239	0.254	0.210	0.250*
Baselines ↳EuroBERT-210m no context	0.115	0.188*	0.123*	0.063	0.130	0.093
DIPF TBA ↳Tolegra or LLM	0.124	0.163*	0.131	0.173*	0.148	0.142

Table 16
Comparison of **Reference** (items which were used as reference for translating into different languages) QWK with "golden-truth" labels and the QWK of the same answers in the final language in the **PISA** test set. Colors are scaled from best (dark red) to worst (dark blue). The mark * indicates that a value is significantly ($p < 0.05$) larger than the other variant evaluated on the same subset based on a permutation test [39] of 999 QWK resamples. The value in the parentheses after the language signifies its category (see Section 3.4). Colors are scaled from best (dark red) to worst (dark blue).

11. Lessons Learnt and Conclusion

Overall, as can be seen in Table 8 and discussed in Section 8.2, the accuracy of the models in predicting the grades of authentic human answers in the rubric based track answers in the **PISA** test set dataset is not bad, surpassing 60% with random baseline at 33%. However, the observed accuracies are quite below the minimum human agreement of 92% as discussed in Section 3.1. The WSE-Research submission with the highest accuracy is a whole 20% below it. Of course, it is possible that our setup causes the models to behave worse than they would under ideal conditions. Due to limited resources, we do not collect human agreement estimates for the non-PISA domains and the simple track.

As was discussed in Section 10.1, the hypothesis about the performance in lower-resource languages was not correct. Most models maintained their QWK across language resourcedness categories of 5 (very well covered), 4, 3 (moderately covered). For example, best model QWK falls only by 0.02, that is, from 0.613 on **NotTranslated** English data to 0.593 on all **NotTranslated** data. Next year we must ensure that our dataset contains also languages of lower categories.

The main Sensemaking suspicion about the models not being able to make sense of background contexts, especially long contexts, was correct. As discussed in Section 9.2 when the models are not provided a rubric with many examples and explicit descriptions of corresponding answers, their performance significantly degrades.

For the rubric variants without explicit description of what how the correct answers should exactly look like the QWK for the best system goes from 0.649 to 0.612 and drops also for all other systems that are also LLM based. This seems to indicate that the models are not able to make sense of the contexts to reconstruct the rubric. For the rubric variants without examples, the QWK of the best system also drops, this time from 0.677 to 0.558 and again drops also for all other systems that are also LLM based. This seems to additionally indicate the models are sometimes not even able to make sense of the rubric, which is substantially shorter than the context.

Overall, as can be seen in Table 9, the datasets we prepared in **Additional** test set had a different distribution of results compared to the **PISA** test set and were generally easier. There were some exceptions, such as inconsistent ratings that caused a decrease in accuracy, which we saw in Table 8. It seems that the inclusion of automatically generated responses did not contribute substantially to our analysis, and their quality needs to be improved next year. Overall, this observation highlights the need to collect and focus on genuine, rather than synthetic, items when assessing the reliability of LLMs as judges.

12. Plans for Next Year

We plan to run the task in the next edition of CLEF again. There is an ongoing discussion about what data can be provided by the OECD for next year.

We plan on remaking our parts of the test set (**Additional** test set) to match the **PISA** test set more closely and examine in a closer detail how models behave at different context lengths. We plan on changing the proportion of items and item fields that are automatically generated. We are discussing different modifications to the task structure. The number of tracks will likely increase next year. Another point being discussed is what other parts of the pipeline introduced in Sensemaking 2025 (Teacher and Student) to use.

Acknowledgments

The ELOQUENT lab is partially supported by the European Commission through the OpenEuroLLM (101195233-DIGITAL-2024-AI-06) and the DeployAI (101146490-DIGITAL-2022-CLOUD-AI-B-03) projects as part of their activities on building, evaluating, and disseminating generative language models. Additionally this lab was supported by the OECD.

Ondřej Bojar and Pavel Šindelář would also like to acknowledge the support from the Project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named “Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím”, and the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO, resp.

References

[1] J. Karlgren, M. Barrett, O. Bojar, M. I. Engels, D. Fabre, S. Ettejjari, L. Goeuriot, R. R. Gunti, J. Mothe, P. Mulhem, M. Piacentini, L. F. V. Madriz, D. Schwab, P. Šindelář, G. Stampoulidis, K. Thomas, M. Vartampetian, Overview of ELOQUENT 2026: shared tasks for evaluating generative language model quality, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[2] J. Karlgren, A. Talman, ELOQUENT 2024 — Topical Quiz Task, in: G. Faggioli, N. Ferro, M. Vlachos, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740, CEUR-WS.org, 2024.

[3] P. Šindelář, O. Bojar, Overview of the Sensemaking Task at the ELOQUENT 2025 lab: LLMs as Teachers, Students and Evaluators, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR-WS, 2025.

[4] Markarit, D. Fabre, J. Karlgren, M. O. Probably, Overview of the pisa quiz generation Task at the ELOQUENT 2026 lab for evaluating generative language model quality, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.

[5] X. Fu, W. Liu, How reliable is multilingual LLM-as-a-judge?, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 11040–11053. URL: https://aclanthology.org/2025.findings-emnlp.587/. doi:10.18653/v1/2025.findings-emnlp.587.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[8] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit., Psychological bulletin 70 (1968) 213.

[9] OECD, PISA 2022 Technical Report, PISA, OECD Publishing, Paris, 2024. URL: https://doi.org/10.1787/01820d6d-en. doi:10.1787/01820d6d-en.

[10] D. Javorský, D. Macháček, O. Bojar, Continuous rating as reliable human evaluation of simultaneous speech translation, in: P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, M. Zampieri (Eds.), Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 154–164. URL: https://aclanthology.org/2022.wmt-1.9/.

[11] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of

the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6282–6293. URL: https://aclanthology.org/2020.acl-main.560/. doi:10.18653/v1/2020.acl-main.560.

[12] F. Zehner, H. J. Shin, E. Kerzabi, A. Horbach, S. Gombert, F. Goldhammer, T. Zesch, N. Andersen, Down the cascades of omethi: Hierarchical automatic scoring in large-scale assessments (????).

[13] K. Yamamoto, Q. He, H. J. Shin, M. von Davier, Development and implementation of a machine-supported coding system for constructed-response items in pisa, Psychological Test (2018).

[14] N. Andersen, F. Zehner, F. Goldhammer, Semi-automatic coding of open-ended text responses in large-scale assessments, Journal of Computer Assisted Learning 39 (2023) 841–854.

[15] M. Mohler, R. Bunescu, R. Mihalcea, Learning to grade short answer questions using semantic similarity measures and dependency graph alignments, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 752–762. URL: https://aclanthology.org/P11-1076/.

[16] R. D. Nielsen, W. Ward, J. Martin, M. Palmer, Annotating students' understanding of science concepts, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: https://aclanthology.org/L08-1166/.

[17] A. Filighera, S. Parihar, T. Steuer, T. Meuser, S. Ochs, Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8577–8591.

[18] P. W. Foltz, L. A. Streeter, K. E. Lochbaum, T. K. Landauer, Implementation and applications of the intelligent essay assessor, in: Handbook of automated essay evaluation, Routledge, 2013, pp. 68–88.

[19] A. Condor, M. Litster, Z. Pardos, Automatic short answer grading with sbert on out-of-sample questions., International Educational Data Mining Society (2021).

[20] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv, 2022. URL: https://www.microsoft.com/en-us/research/publication/text-embeddings-by-weakly-supervised-contrastive-pre-training/.

[21] R. Hada, V. Gumma, A. de Wynter, H. Diddee, M. Ahmed, M. Choudhury, K. Bali, S. Sitaram, Are large language model-based evaluators the solution to scaling up multilingual evaluation?, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1051–1070. URL: https://aclanthology.org/2024.findings-eacl.71/. doi:10.18653/v1/2024.findings-eacl.71.

[22] L.-H. Chang, F. Ginter, Automatic short answer grading for finnish with chatgpt, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 23173–23181.

[23] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[26] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, Advances in neural information processing systems 34 (2021) 27263–27277.

[27] J. L. Jing Zhang, D. Dai, Few-shot llm-as-judge with language-aware translation for multilingual answer scoring, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS,

2026.
[28] F. Prášil, A. Otmar, Lightweight interpretable pipeline for automated short answer grading, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.
[29] Z. Ahmetović, M. Olovčić, Metric learning for automated rubric-based answer grading: A bi-encoder approach, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.
[30] D. Condrey, The pc attractor: Tracing a generalization failure through a coral + kde ordinal regression pipeline, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.
[31] J. Gwozdz, A. Both, Wse-research at clef 2026 eloquent sensemaking: Llm-based scoring of student answers to pisa reading comprehension items, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.
[32] P. Schmidtová, N. Bafna, S. Aycock, G. Vico, W. Kamzela, K. Hämmerl, V. Zouhar, How important is 'perfect' English for machine translation prompts?, in: V. Demberg, K. Inui, L. Marquez (Eds.), Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 760–777. URL: https://aclanthology.org/2026.findings-eacl.38/. doi:10.18653/v1/2026.findings-eacl.38.
[33] M. Laurer, W. v. Atteveldt, A. S. Casas, K. Welbers, Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI, Preprint (2022). URL: https://osf.io/74b8k, publisher: Open Science Framework.
[34] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272. doi:10.1038/s41592-019-0686-2.
[35] G. E. Hinton, Training products of experts by minimizing contrastive divergence, Neural computation 14 (2002) 1771–1800.
[36] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, Pattern Recognition Letters 140 (2020) 325–331.
[37] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.
[38] S. Gombert, Z. Sun, F. Zehner, J. Lossjew, T. Wyrwich, B. K. Czinczel, D. Bednorz, M. Kubsch, D. Di Mitri, K. Neumann, H. Drachsler, Are rubrics all you need? towards rubric-based automatic short answer scoring via guided rubric-answer alignment, in: Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference, LAK '26, Association for Computing Machinery, New York, NY, USA, 2026, p. 272–282. URL: https://doi.org/10.1145/3785022.3785064. doi:10.1145/3785022.3785064.
[39] P. Good, Permutation, parametric and bootstrap tests of hypotheses, Springer, 2005.