Overview of the PISA Quiz Item Generation Task of the ELOQUENT 2026 Lab for Evaluating Generative Language Model Quality

Notebook for the ELOQUENT Lab at CLEF 2026

Markarit Vartampetian[1], Sarah Bouaraba[1], Diandra Fabre[1], Said Ettejjari[2], Lorraine Goeuriot[1], Jussi Karlgren[3,4], Luis Francisco Vargas Madriz[2], Philippe Mulhem[1], Mario Piacentini[2], Didier Schwab[1] and Katherina Thomas[2]

[1]*Université Grenoble Alpes*
[2]*OECD*
[3]*AMD Silo AI, Sweden*
[4]*University of Helsinki, Finland*

Abstract

The ELOQUENT lab for evaluation of generative language model quality and usefulness addresses high-level quality criteria for generative language models through a set of open-ended shared tasks implemented, where possible, to minimise human effort in assessment, and with an objective to study how much the languages that the foundation model has been trained on make a difference in its responses.

In this third ELOQUENT edition, the Topical Quiz task introduced in the first edition is organised together with OECD, which defines and administers the Programme for International Student Assessment (PISA). One of the challenges investigated is to generate reading skill assessment test items, and the PISA Quiz Item generation task has experimented with methods to automatically generate useful such items.

Evaluating Generative Language Model Quality

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

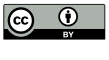
Previous Years

This task traces its roots to tasks on quiz handling given in earlier editions of ELOQUENT [2, 3] but in this third edition is anchored in the real world practice of assessing reading proficiency in primary education. Together with the task on quiz scoring [4], it this task is organised together with OECD, which defines and administers the Programme for International Student Assessment (PISA). PISA is an international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students nearing the end of their compulsory education. PISA assesses how well students can apply what they learn in school to real-life situations. So far, over 90 countries or economies have participated in the assessment since 2000 [5].

Question generation

Automatic question processing, both for general purpose question answering and for question generation has a well-established history as a subfield of language technology. Question generation has applications for both testing purposes (as for PISA), for self-study, and for powering discourse with chatbots. A relatively recent survey paper gives a useful overview [6]. This overview focuses mostly on reading comprehension tasks, where the task stimulus is given in language form, typically text, and is quite comprehensive. The questions and responses are expected to be given in text form.

CLEF 2026 Working Notes, September 2026, Jena



Types of questions

Types of questions are foundational to the generation effort. Questions are central to human language use and there is a wealth of linguistic study on question types. On a more technical level, Lehnert formulated a finegrained set of thirteen conceptual types of questions in 1977 [7] which have been used as a basis for much of the language technology and cognitive science work since. Her suggested set of categories — causal antecedent, goal orientation, enablement, causal consequent, verification, disjunctive, instrumental, concept completion, expectational, judgmental, quantification, feature specification and request — are based on traditional linguistic studies and are designed to aid her primary goal of building automated question answering systems. Similarly, in the educational field, Bloom's elaborate taxonomy of learning objectives [8] have been used in informing the formulation of test questions. In practice most efforts for question generation use a smaller set of more structural rather than conceptual question types to guide implementational work and an example is the one used in the survey cited above:

- Wh-questions which can be answered either extractively from a single location in the target text or through collating information from several locations in the text
- Yes-no questions
- Deep understanding questions

Recent experiments have guided the generation of test items using a number of criteria (such as difficulty level, Bloom's taxonomy) included in the prompts to the generative language model, to successfully achieve a fine-grained palette of questions [9, 10, 11]. Lehnert's more conceptually elaborate scheme has great potential to serve as a model for future fine-grained automatically generated question sets, and might be worth resurrecting for reading comprehension tests: the experiments referenced here have used standards language models, unmodified, and with even a small amount of specific fine tuning, adding more sophistication to the test item typology.

Participation

Two teams submitted experiments: AMD Silo AI and Université Grenoble Alpes (UGA). The experimental setups were loosely coordinated, with the team from AMD Silo AI focusing on commercially available systems (Claude Opus 4.6) and UGA on open-weight ones (Qwen3-32B[1] [12] and Mistral-Small-24B-Instruct [2]). In total, this resulted in 12 experimental submissions.

Data

This edition of the task evaluates text-to-text item generation for PISA reading literacy. PISA spans three domains, namely reading, mathematical, and scientific literacy, of which we target reading literacy exclusively. PISA texts may be continuous or non-continuous and appear in either fixed or dynamic formats. Given the text-to-text setting, we retained only text-based reading items and excluded any question requiring the interpretation of images, graphics, or other non-textual elements. All retained units and items are in English. The released PISA items are available online in pdf format (image-based scans for the 2000 cycle); we used Claude Opus 4.5 to extract their text and metadata into a serialized jsonl format, drawing on the freely available reading units from the 2000 [13, 14], 2015 [15], and 2018 [16] cycles. The dataset is organised into PISA units containing open-ended and multiple-choice QA items, for a total of 16 units and 77 QA items. Where available, we also retain official metadata such as cognitive process, response format, difficulty score, proficiency level, source requirement (single vs. multiple texts), scoring rubrics (including partial-credit and incorrect-response criteria), and item-level pedagogical rationale. An example from a previous test, together with the original PISA test questions,

[1] https://huggingface.co/Qwen/Qwen3-32B
[2] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

is shown in Figure 1. The dataset consists of 77 items, summarised in Table 1, with further statistics reported in Section 4. It is available through the task's GitHub page [3].

Stimulus


```
{
  "unit_title": "Cow's Milk", "year": 2018,
  "stimulus_text": "[Farm to Market] FARM TO MARKET DAIRY. The Nutritional Value of Milk: Countless
Benefits! [. . . ] milk products contain key nutrients: calcium, protein, vitamin D, [. . . ] The
IDFA suggests that many health professionals and groups would also agree.[. . . ] [Just Say No]
JUST SAY 'NO' TO COW'S MILK! (Dr. R. Garza) [. . . ] a study of 100 000 people in Sweden found
female milk drinkers suffered more bone fractures, and milk drinkers were more likely to suffer
heart disease and cancer. [. . . ]"
}
```


Multiple-choice


```
{
  "item_id": "CR557Q03",
  "question": "According to the IDFA, with which statement do leading health professionals and
organizations agree?",
  "options": {
     "A": "Consuming milk and milk products leads to obesity",
     "B": "Milk is a good source of essential vitamins and minerals",
     "C": "Milk contains more vitamins than minerals",
     "D": "Drinking milk is a leading cause of osteoporosis" },
  "answer": {"type": "single", "correct": "B"},
  "cognitive_process": "Represent literal meaning",
  "response_format": "Simple Multiple Choice - Computer Scored",
  "difficulty_score": 323, "difficulty_level": "1b", "source_requirement": "Single"
}
```


Open-ended


```
{
  "item_id": "CR557Q10", "source_tab": "Just Say No",
  "question": "Dr. Garza presents a few research results which may 'surprise' readers. State one
of them.",
  "answer": { "type": "open", "full_credit": {
    "criteria": "Quotes or paraphrases one of the research results stated in the text",
    "acceptable_answers": [
      { "answer": "Female milk drinkers suffered more bone fractures",
        "paraphrases": ["Women who drank milk had more broken bones"] },
      { "answer": "Both male and female milk drinkers were more likely to suffer from heart disease
and cancer",
        "paraphrases": ["People who drink milk had more heart disease and cancer"] }
    ] }
  },
  "cognitive_process": "Represent literal meaning",
  "response_format": "Open Response - Human Coded",
  "difficulty_score": 398, "difficulty_level": "3", "source_requirement": "Single"
}
```


Figure 1: Examples of items from the PISA unit *Cow's Milk* (2018): the stimulus text together with a multiple-choice item (CR557Q03) and an open-ended item (CR557Q10) derived from it. For simplicity, the stimulus is shown once since it is common to both items; in the released data it is repeated once per item.

Dataset statistics

The data cover three PISA cycles (2000, 2015, 2018), totalling 16 reading units and 77 QA items, of which 43 are multiple-choice and 34 are open-ended (Table 1). Only text-answerable items are included, and each unit contributes between 2 and 8 items. Stimulus text length ranges from 62 to 1,751 words (median 465), and question length from 2 to 86 words (median 17). For open-ended items, a scoring rubric is provided, specifying full-credit, partial-credit, and incorrect-response criteria. Every item carries

[3] https://github.com/eloquent-lab/eloquent-lab.github.io/tree/main/task-pisa-generate-items/data

```
{
  "item_id": "CR551Q11", "year": 2018, "unit_title": "Rapa Nui",
  "source_tab": ["Blog", "Book Review", "Science News"],
  "stimulus_text": "[Blog] The Professor's Blog [. . . ] the people of Rapa Nui used large trees
to move the moai; what happened to those trees is a mystery. [. . . ] [Book Review] Review of
Collapse (J. Diamond) [. . . ] the people cleared the land and cut down the trees. [. . . ] [Science
News] [. . . ] Lipo and Hunt argue the Polynesian rat ate the seeds, preventing new trees from
growing. [. . . ]",
  "question": "After reading the three sources, what do you think caused the disappearance of the
large trees on Rapa Nui? Provide specific information from the sources to support your answer.",
  "answer": { "type": "open", "full_credit": {
    "criteria": "Includes one or more of: (1) people cut down / used the trees to move the moai
or cleared land; (2) rats ate the seeds so new trees could not grow; (3) cannot say until further
research",
    "acceptable_answers": [
      { "answer": "People cut down / used the trees to move the moai or cleared land", "paraphrases":
["People cut too many down"] },
      { "answer": "The rats ate the seeds so new trees could not grow", "paraphrases": ["The rats
ate the seeds"] },
      { "answer": "Cannot say until further research", "paraphrases": ["We have to wait for more
information"] }
    ] }
  },
  "cognitive_process": "Detect and handle conflict",
  "response_format": "Open Response - Human Coded",
  "difficulty_score": 588, "difficulty_level": "4", "source_requirement": "Multiple",
  "item_explanation": "The student integrates the three sources, recognises that the theories
conflict, and supports a chosen theory (or none, citing the need for more research) with textual
evidence."
}
```


Figure 2: Open-ended item from the PISA unit *Rapa Nui* (2018), with full-credit criteria, acceptable answers and paraphrases, and an explanation of full-credit responses (`item_explanation`).

a cognitive-process label, while difficulty, source requirement, and item-level rationale are available mainly for the 2018 cycle, following the official OECD PISA frameworks.

Cycle	Units	Items	MC / Open
2000	11	46	22/24
2015	1	3	2/1
2018	4	28	19/9
Total	16	77	43/34

Table 1. Released PISA reading items by cycle.

Task Definition

The proposed setting focuses on generating stimulus material (a brief written passage) suitable for a PISA reading-literacy assessment and its corresponding QA items. Stimulus generation is considered under two configurations: *Topic-seeded* and *Self-selected topic*. In the first configuration, the LLM is given a topic drawn from the list used in the Topical Competence task of the 2024 ELOQUENT lab.[4] For the submitted runs, four topics were used in this setting: `006` (stenography), `019` (tea), `020` (startup life), and `021` (detecting conspiracy theories). See Appendix C for the prompt. In the second configuration, no predefined topic is provided. Instead, the LLM is prompted to formulate its own topic and generate a stimulus text appropriate for 15-year-old students and suitable for the PISA reading-literacy assessment. The prompt is provided in Appendix D.

[4] https://huggingface.co/datasets/Eloquent/TopicalQuiz

For the QA generation, the LLM is prompted in both configurations to generate 3 to 5 multiple-choice questions per stimulus text, with four answer options and a single correct answer. The generated output further includes the following item-level metadata: (i) the targeted cognitive process (e.g., Access and retrieve); (ii) the response format (e.g., Simple Multiple Choice); (iii) a difficulty score on a 0–1000 scale; (iv) the difficulty level (e.g., Easy, Medium, or Hard); and (v) an explanation justifying the correct response.

Submissions

The shared task drew two participating teams: AMD Silo AI (focussing on testing proprietary systems) and UGA (focussing on open-weight systems). The teams coordinated a shared design in which runs varied along three controlled dimensions; the model, the topic configuration, and the prompt variant, yielding 12 experiments in total (cf. Table 2). The two prompt variants are both elaborate instructions that specify the assessor's role, require a structured analysis of the source text, and define the expected response format and item-level metadata, i.d. cognitive process, item explanation and self estimated difficulty score and level. The baseline variant is intended to capture the professional aims of a PISA test author, framing the task as generating thoughtful and engaging PISA-style items that test comprehension beyond surface recall and support nuanced understanding. The strict variant reframes the same task around rigour and excellence, pushing the model toward more challenging items, eliminating the baseline's constraint on trickier questions, and calling for distractors designed to expose naive assumptions or partial understandings.

Configuration	Options
Model	Claude Opus 4.6; Qwen3-32B; Mistral-Small-24B-Instruct
Prompt variant	Baseline; Strict
Topic configuration	Topic-seeded; Self-selected topic

Table 2. Each experiment selects one value per row; experiments cover all combinations of model, prompt variant, and topic configuration.

Assessment procedure

The assessment is conducted using three complementary evaluation procedures: automatic metrics across four dimensions (diversity, readability, estimated difficulty, and guessability), described in the next subsection; a human evaluation conducted by OECD experts, detailed below; and an LLM-as-a-judge evaluation, described in the last subsection of the present section. The latter two procedures are based on a shared set of nine criteria: stimulus quality, question quality, question-stimulus alignment, answerability, non-guessability, distractor quality, correctness of the cognitive-process label, correctness of the difficulty label, and diversity of items.

Automatic assessment procedure

Automatic assessment is conducted through a set of metrics and metadata-based analyses. For diversity, we use Distinct-n [17] and Self-BLEU [18] at the lexical level, and mean pairwise cosine similarity between text embeddings obtained with a sentence transformer model at the semantic level. Distinct-n is applied only to intra-model item diversity, whereas Self-BLEU and cosine similarity are used for both intra-model and inter-model comparisons. The assessment considers generated items within a single model configuration, as well as generated stimuli and items between model/prompt configurations, in the topic-seeded setting (cf. Table 2). For readability, we use Flesch Reading Ease [19], Flesch-Kincaid Grade Level [20], the Gunning Fog Index [21], and Dale-Chall Readability [22]. All four combine

average sentence length with a measure of word difficulty, differing in how the latter is captured: Flesch Reading Ease and Flesch-Kincaid Grade Level use syllables per word, the Gunning Fog Index uses the proportion of polysyllabic words, and Dale-Chall Readability uses the proportion of words not included in a list of 3,000 words commonly understood by fourth-grade students. Readability is computed for the generated stimuli across all configurations. Estimated difficulty corresponds to the difficulty level each model assigns to its own generated items, as reported in the item metadata. We examine how this self-estimated difficulty varies across the experimental configurations in Table 2, i.e., according to model choice, prompt variant, and topic setting. This allows us to assess whether the attributed difficulty shifts depending on whether the topic is provided or self-selected, and, within the topic-seeded setting, whether some seeded topics lead models to assign higher or lower difficulty to their generated items. Finally, we estimate item guessability using artifact-based solvers that exploit answer-position, answer-length, and lexical-overlap cues. The following subsections detail each metric and its computation.

Dimension	Measure	Description
Diversity	Distinct-n [17]	Proportion of distinct n-grams; higher values indicate greater lexical diversity.
	Self-BLEU [18]	Average n-gram overlap between generated items; lower values indicate greater lexical diversity.
	Mean pairwise cosine similarity	Average cosine similarity between sentence embeddings; lower values indicate greater semantic diversity.
Readability	Flesch Reading Ease [19]	Readability score based on sentence length and syllables per word.
	Flesch-Kincaid Grade Level [20]	Estimated grade level based on sentence length and syllables per word.
	Gunning Fog Index [21]	Estimated reading difficulty based on sentence length and the proportion of polysyllabic words.
	Dale-Chall Readability [22]	Readability score based on sentence length and the proportion of words not included in a list of familiar words.
Estimated difficulty	Self-estimated difficulty labels	Distribution of difficulty levels assigned by the model to its own generated items in the item metadata.
Guessability	Artifact-based guessability score	Computed from artifact-solver accuracy as the best-performing solver's accuracy above the $1/k$ chance level, over the position, length, and lexical-overlap solvers; lower values indicate lower guessability.

Table 3. Automatic assessment dimensions and measures.

Diversity: We measure diversity lexically, using Distinct-n and Self-BLEU, and semantically, using mean pairwise cosine similarity between sentence embeddings. We consider three comparison cases: intra-model item diversity within a configuration, inter-model stimulus diversity for the same seeded topic, and inter-model item diversity for the same seeded topic. Distinct-n is used only in the intra-model setting, while Self-BLEU and cosine similarity are used in both settings. Higher Distinct-n indicates greater diversity, whereas lower Self-BLEU and cosine similarity indicate greater diversity.

Let $\mathcal{S}$ denote a set of generated stimulus texts and $\mathcal{I}$ a set of generated items. For inter-model comparisons, $\mathcal{S}^{(a)}$ and $\mathcal{S}^{(b)}$ denote two stimulus sets generated for the same seeded topic by two different configurations, while $\mathcal{I}^{(a)}$ and $\mathcal{I}^{(b)}$ denote the corresponding item sets. The embedding of an element x, here obtained with `gte-multilingual-base`, is denoted by $\mathbf{e}(x)$.

Distinct-n: Distinct-n [17] measures intra-model lexical diversity as the proportion of unique n-grams in the generated items. We compute it at the corpus level by pooling all items in $\mathcal{I}$ and dividing the number of distinct n-gram types by the total number of n-gram occurrences:

$$\text{Distinct-}n(\mathcal{I}) = \frac{\big|\,\text{unique}(\mathcal{G}_n(\mathcal{I}))\,\big|}{\big|\,\mathcal{G}_n(\mathcal{I})\,\big|}, \tag{1}$$

where $\mathcal{G}_n(\mathcal{I})$ is the multiset of all n-grams in $\mathcal{I}$. Normalising by the total number of n-grams avoids bias toward longer generated items and prevents the diversity score from being artificially inflated by item length.

Self-BLEU: Self-BLEU [18] adapts the BLEU score, traditionally used for machine translation, to measure diversity through lexical overlap within a set of generated stimulus texts or items. Unlike BLEU, which compares a system output to external reference translations, Self-BLEU compares generated elements with one another, without requiring external references. In the intra-model setting, each generated item is scored against the remaining items in the same set (Eq. 2); in the inter-model setting, stimulus texts or items from two configurations are compared symmetrically (Eq. 3). Lower Self-BLEU scores indicate higher lexical diversity. We report Self-BLEU-n for $n = 3$ and $n = 4$.

$$\text{Self-BLEU}_n(\mathcal{I}) = \frac{1}{|\mathcal{I}|}\sum_{x\in\mathcal{I}} \text{BLEU}_n\big(x, \mathcal{I}\setminus\{x\}\big). \tag{2}$$

$$\begin{aligned}\text{Self-BLEU}_n(\mathcal{X}^{(a)}, \mathcal{X}^{(b)}) = \frac{1}{2}\Bigg[&\frac{1}{|\mathcal{X}^{(a)}|}\sum_{x\in\mathcal{X}^{(a)}} \text{BLEU}_n\big(x, \mathcal{X}^{(b)}\big)\\ &+\frac{1}{|\mathcal{X}^{(b)}|}\sum_{x\in\mathcal{X}^{(b)}} \text{BLEU}_n\big(x, \mathcal{X}^{(a)}\big)\Bigg], \quad \mathcal{X}\in\{\mathcal{S},\mathcal{I}\}.\end{aligned} \tag{3}$$

Mean pairwise cosine similarity: At the semantic level, we measure intra-model diversity among generated items, and inter-model diversity among both generated stimuli and generated items in the topic-seeded configuration. First, for each element $x \in \mathcal{X}$ under evaluation, where $\mathcal{X} = \mathcal{I}$ for intra-model item comparisons and $\mathcal{X} \in \{\mathcal{S}, \mathcal{I}\}$ for inter-model comparisons, we compute sentence embeddings $\mathbf{e}(x)$ using *gte-multilingual-base*[5]. Semantic similarity is then measured as the average pairwise cosine similarity between embeddings. Lower values indicate lower semantic similarity and therefore higher semantic diversity.

$$\text{MeanCos}(\mathcal{I}) = \frac{1}{|\mathcal{I}|(|\mathcal{I}|-1)}\sum_{x\in\mathcal{I}}\sum_{\substack{x'\in\mathcal{I}\\ x'\neq x}} \cos\big(\mathbf{e}(x), \mathbf{e}(x')\big). \tag{4}$$

$$\text{MeanCos}(\mathcal{X}^{(a)}, \mathcal{X}^{(b)}) = \frac{1}{|\mathcal{X}^{(a)}|\,|\mathcal{X}^{(b)}|}\sum_{x\in\mathcal{X}^{(a)}}\sum_{x'\in\mathcal{X}^{(b)}} \cos\big(\mathbf{e}(x), \mathbf{e}(x')\big), \quad \mathcal{X}\in\{\mathcal{S},\mathcal{I}\}. \tag{5}$$

Readability: For readability, we report standard automatic metrics computed on the stimulus text only: Flesch Reading Ease (FRE) [19], Flesch-Kincaid Grade Level (FKGL) [20], Gunning Fog Index (GFI) [21], and Dale-Chall Readability (DC) [22]. Each metric is computed separately for each stimulus and then averaged within each evaluation set. Let W, S, and Y denote the number of words, sentences, and syllables in the stimulus, respectively. We further denote by C the number of complex words, defined as words with at least three syllables, and by D the number of difficult words, defined as words not included in the Dale-Chall list of words. Following the Dale-Chall formulation, let $\text{PDW} = 100\,D/W$ denote the percentage of difficult words. The metrics are computed as:

$$\text{FRE} = 206.835 - 1.015\,\frac{W}{S} - 84.6\,\frac{Y}{W} \tag{6}$$

[5] https://huggingface.co/Alibaba-NLP/gte-multilingual-base

$$\mathrm{FKGL} = 0.39\,\frac{W}{S} + 11.8\,\frac{Y}{W} - 15.59 \tag{7}$$

$$\mathrm{GFI} = 0.4\left(\frac{W}{S} + 100\,\frac{C}{W}\right) \tag{8}$$

$$\mathrm{DC} = 0.1579\,\mathrm{PDW} + 0.0496\,\frac{W}{S} + 3.6365 \cdot \mathbb{1}[\mathrm{PDW} > 5] \tag{9}$$

FRE conventionally ranges over $[0, 100]$, where higher scores indicate easier text. FKGL and GFI are interpreted as approximate grade-level scores, while DC is interpreted through grade-level bands; for these three metrics, lower scores indicate easier readability and higher scores indicate greater textual difficulty.

Estimated difficulty: During item generation, models are instructed to self-estimate the difficulty of each item by assigning both a numerical `difficulty_score` in $[0, 1000]$ and a categorical difficulty level in $\{\textit{Easy}, \textit{Medium}, \textit{Hard}\}$. We use these self-judgements to analyse how difficulty varies across experimental configurations, including model choice, prompt variant, and topic setting. In particular, we examine whether the strict prompt variant yields systematically higher self-estimated difficulty than the baseline, or whether certain models or topics tend toward easier or harder items.

Guessability: Generated items should not be answerable from surface cues alone. Thus, for each item, we run three artifact solvers that never read the stimulus: (i) majority-position solver, which always selects the globally most frequent correct option position; (ii) length solver, which picks the longest answer option; (iii) lexical-overlap solver, which chooses the option with the greatest token overlap with the question stem. For each solver s, we compute its accuracy A_s over the evaluation set and compare it with the chance baseline $1/k$, where $k = 4$ is the number of answer options in the generated items. We define guessability as the excess accuracy of the best artifact solver over chance:

$$A_s = \frac{1}{|\mathcal{I}_{\mathrm{MC}}|} \sum_{i \in \mathcal{I}_{\mathrm{MC}}} \mathbb{1}\left[\hat{y}_i^{(s)} = y_i\right], \qquad g = \max_{s \in \{\mathrm{maj},\,\mathrm{len},\,\mathrm{lex}\}} A_s - \frac{1}{k}, \tag{10}$$

where $\mathcal{I}_{\mathrm{MC}} \subseteq \mathcal{I}$ is the set of generated multiple-choice items, y_i is the correct answer for item i, and $\hat{y}_i^{(s)}$ is the option selected by solver s without access to the stimulus. The lexical-overlap solver compares normalized (lowercase) token sets, and all solvers break ties by first-option order. As the maximum over the three solvers, g is a conservative (upper-bound) estimate of guessability, with lower g indicating less exploitable items.

Instructions to assessors

The PISA reading assessment measures the reading literacy of 15-year-old students based on the OECD's reading framework. The human assessment was conducted following the OECD's 2018 theoretical framework [16]. Reading literacy encompasses a range of cognitive processes, including searching for and accessing information, understanding and interpreting texts, and evaluating and reflecting on sources and text content. Varying levels of difficulty arise from task characteristics (e.g., question complexity, response mode), text characteristics (e.g., length of the text, number of sources), and the interaction between text and task characteristics (e.g., alignment between question and text, location of target information). Following this, submissions are evaluated at two levels: the individual item and the full test form. A test form is a complete set of stimuli and items, corresponding to about one hour of testing time for a student. At the item level, experts assess the phrasing, content, and alignment between the text and the question. They also verify whether the declared cognitive process and difficulty level are appropriate. At the test form level, submissions are judged by how well the set of items taken together

reflects the distribution of cognitive processes in PISA, summarized in Table 5. The items should also show a reasonable variation in difficulty, which arises from the combination of item characteristics and cognitive processes.

Human evaluation was carried out by one OECD expert using the criteria in Table 4, on a subset of 60 QA items drawn entirely from the *Self-selected topic* configuration. To limit bias, submissions were anonymized during evaluation: each generating model was replaced by a system identifier, namely System 1, System 2, or System 3, and all metadata that could reveal the generator or the prompt variant was hidden. The three systems correspond to Claude Opus 4, Mistral-Small-24B-Instruct-2501, and Qwen3-32B. Each system contributed two stimuli generated under the *Self-selected topic* configuration, with ten items per stimulus: five standard and five strict. This gave 20 QA items per system and 60 QA items in total.

Group	Criteria	For each…	Explanation	Max
Text quality	Stimulus quality	Stimulus	Is the stimulus natural, realistic, matching the reading situation and free from biased, inappropriate or culturally-specific content?	2
	Question quality	Question	Is the question clear, concise and its requirements unambiguous?	5
Item function	Question-stimulus alignment	Question	Is the question aligned with the stimulus?	3
	Answerable	Question	Can the question be answered using *only* the stimulus and no prior knowledge?	3
	Non-Guessable	Question	Can the question be answered without reading the stimulus? (0 = guessable)	3
	Distractor quality	Question	Are distractors plausible, clearly incorrect, and not easily eliminated?	3
Framework alignment	Correct cog. process	Question	Does the cognitive process label match the processes required? Please indicate which process you would label to the exercise, see processes defined below (we will match with AI later)	1
	Correct difficulty	Question	Are the questions & answers covering a diversity of difficulty?	2
Testset	Diversity of items	Submission	Do the items cover diverse cognitive processes, difficulty levels, and reading situations?	2

Table 4. OECD evaluation criteria and scoring rubric.

Participant Results and Discussion

In this section, we report the automatic metrics (diversity, readability, estimated difficulty, and guessability) followed by the human assessment; the LLM-as-a-judge results are discussed separately in the next section. Across all dimensions we contrast the three models, the two prompt variants (baseline vs. strict), and the two topic configurations (topic-seeded vs. self-selected).

Diversity: Table 6 reports intra-model item diversity. In the topic-seeded configuration, the AMD Silo AI submission Claude Opus 4.6 is the most lexically diverse system, with the highest Distinct-2 and Distinct-3 values and the lowest Self-BLEU scores. The strict prompt further increases this lexical and surface-form diversity for Claude, although its mean pairwise cosine similarity changes only marginally and in the opposite direction. For the UGA submission (Mistral-Small-24B and Qwen3-32B), the strict prompt has the opposite effect, consistently reducing topic-seeded diversity: Distinct-n decreases, indicating fewer unique n-grams, while Self-BLEU and mean pairwise cosine similarity increase, indicating greater overlap between generated items. The strict prompt therefore does not act uniformly across models: it diversifies Claude's outputs on lexical metrics, but makes the open-weight

Area	Short title	Description	Share
Locate information	Access and retrieve	Access and retrieve information within a text - scanning a single text in order to retrieve target information consisting of a few words, phrases or numerical values.	15%
	Search and select	Search for and select relevant text - searching for information among several texts to select the most relevant text given the demands of the item/task.	10%
Understand	Represent literal information	Represent literal information - comprehending the literal meaning of sentences or short passages, typically matching a direct or close paraphrasing of information in the question with information in a passage.	15%
	Integrate and generate inferences in one item	Integrate and generate inferences - going beyond the literal meaning of information in a text by integrating information across sentences or even an entire passage. Tasks that require the student to create a main idea or to produce a summary or a title for a passage are classified as "integrate and generate inference" items.	15%
	Integrate and generate inferences across multiple sources	Integrate and generate inferences across multiple sources - integrating pieces of information that are located within two or more texts.	15%
Evaluate and Reflect	Assess quality and credibility	Assess quality and credibility - evaluating whether the information in a text is valid, current, accurate, unbiased, reliable, etc. Readers must identify and consider the source of the information and consider the content and form of the text or in other words, how the author is presenting the information.	20%
	Reflect and evaluate	Reflect on content and form - evaluating the form of the writing to determine how the author is expressing their purpose and/or point of view. These items often require the student to reflect on their own experience and knowledge to compare, contrast or hypothesize different perspectives or viewpoints.	—
	Detect and handle conflict	Detect and handle conflict - determining whether multiple texts corroborate or contradict each other and when they conflict, deciding how to handle that conflict. For example, items classified as "detect and handle conflict" may ask students to identify whether two authors agree on the stance of an issue or to identify each author's stance. In other cases, these items may require students to consider the credibility of the sources and demonstrate that they accept the claims from the more reliable source over the claims from the less reliable source.	10%

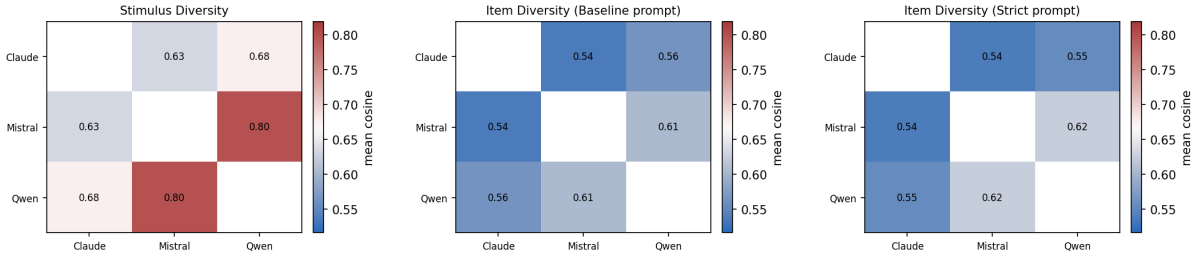
Table 5. Cognitive processes and item share in the PISA 2018 reading framework.

models produce more uniform and similarly formulated questions. The Self-selected topic configuration is more mixed, with no single model dominating. Figure 3 shows inter-model diversity as the pairwise cosine similarity between models, averaged over the four seeded topics, for (a) stimuli, (b) baseline items, and (c) strict items. In all three panels Mistral-Small-24B and Qwen3-32B are the most similar pair (their stimuli reach a cosine of 0.80), while Claude Opus diverges more clearly from both. This pattern indicates that the two open-weight models generate more semantically similar outputs under the same seeded topics. We observe a similar tendency across cosine similarity and Self-BLEU, and it is stable across all four topics. The two metrics diverge, however, on the stimulus-versus-item comparison: cosine similarity is higher for stimuli (0.63–0.80) than for items (0.54–0.62), which likely reflects the fact that stimuli are long, topically dense passages ($\sim$450–650 words), whereas items are short questions ($\sim$15–25 words) with little shared vocabulary. Self-BLEU, which is less sensitive to length, shows the reverse: items share more phrasing across models than stimuli do (Self-BLEU-3 up to 0.30 for Mistral–Qwen strict items versus ≤ 0.07 for stimuli). The two models reuse the same question formulations, recurring stems such as "Which of the following best describes…", even though their

generated stimuli share almost no common phrasing.

Model	Prompt	n	Distinct-2 ↑	Distinct-3 ↑	Self-BLEU-3 ↓	Self-BLEU-4 ↓	Mean Pairwise Cosine Similarity ↓
Topic-seeded							
Claude Opus 4.6	baseline	20	0.91	0.98	0.12	0.06	0.55
	strict	20	**0.92**	**0.99**	**0.09**	**0.04**	0.56
Mistral-Small-24B	baseline	20	0.81	0.92	0.22	0.15	**0.46**
	strict	20	0.64	0.79	0.46	0.36	0.49
Qwen3-32B	baseline	20	0.71	0.83	0.40	0.30	0.54
	strict	20	0.58	0.70	0.52	0.45	0.57
Self-selected topic							
Claude Opus 4.6	baseline	10	**0.91**	**0.96**	**0.20**	0.13	0.67
	strict	10	0.87	0.95	**0.20**	**0.12**	0.61
Mistral-Small-24B	baseline	10	0.79	0.90	0.28	0.17	0.57
	strict	10	0.84	0.91	0.26	0.18	**0.56**
Qwen3-32B	baseline	5	0.82	0.94	**0.20**	**0.12**	0.62
	strict	5	0.71	0.80	0.42	0.34	0.80

Table 6. Intra-model item diversity per configuration (n = number of items). Higher Distinct-n (↑) and lower Self-BLEU / cosine similarity (↓) indicate greater diversity. **Bold**: best value per metric within each configuration; underline: the more diverse of the two prompt variants for each model.



(a) (b) (c)

Figure 3: Inter-model diversity (topic-seeded): pairwise mean cosine similarity between models, averaged over topics: (a) stimulus diversity, (b) item diversity (baseline prompt), (c) item diversity (strict prompt).

Readability: Table 7 reports stimulus readability.[6] In the topic-seeded setting, Claude Opus 4.6 produces the most readable stimuli (Flesch Reading Ease 73.1), ahead of Qwen3-32B (56.3) and Mistral-Small-24B (42.4), which write denser text; here the four metrics agree on the full ranking. In the self-selected setting the picture reverses, and Claude Opus 4.6's stimuli become the hardest to read (28.7). The difficulty appears stylistic rather than topical: both self-selected stimuli score in the hard range, including one on a subject familiar to teenagers (smartphone use). Across the two configurations Claude Opus 4.6 varies by approximately 44 Flesch points, whereas the open models remain within a narrow band (Mistral 42.4–46.4, Qwen 44.4–56.3); readability thus depends on both the model and the chosen topic configuration, not on the model alone.

Estimated difficulty: Figures 4 and 5 show the models' self-estimated difficulty. All systems label most items "Medium" or "Hard" and few as "Easy", so these self-estimates separate the models only

[6]These readability metrics and their grade-level interpretations are calibrated on US English, so we treat the values as relative indicators rather than exact reading levels.

Model	n	Flesch Reading Ease ↑	Flesch–Kincaid Grade Level ↓	Gunning Fog Index ↓	Dale–Chall Readability ↓
Topic-seeded					
Claude Opus 4.6	4	**73.1**	**6.9**	**8.8**	**8.6**
Mistral-Small-24B	4	42.4	11.6	14.4	10.9
Qwen3-32B	4	56.3	8.7	11.0	10.0
Self-selected topic					
Claude Opus 4.6	2	28.7	14.0	16.5	12.9
Mistral-Small-24B	2	**46.4**	**11.0**	**12.9**	11.0
Qwen3-32B	2	44.4	12.6	15.2	**10.8**

Table 7. Stimulus readability per configuration (n = number of stimuli). Higher Flesch Reading Ease indicates easier text; higher Flesch–Kincaid, Gunning Fog, and Dale–Chall indicate greater reading difficulty. **Bold** indicates the greatest reading ease per metric within each configuration.

weakly. The strict prompt has no systematic effect: the largest shift is observed for Mistral-Small-24B in the topic-seeded setting, whose "Hard" share goes from 40% to 55%, whereas Qwen3-32B and Claude Opus 4.6 barely move. Across the four seeded topics the distributions are also stable ("Hard" between 47% and 50%), thus, the chosen topic has little effect on how difficult the models rate their own items.

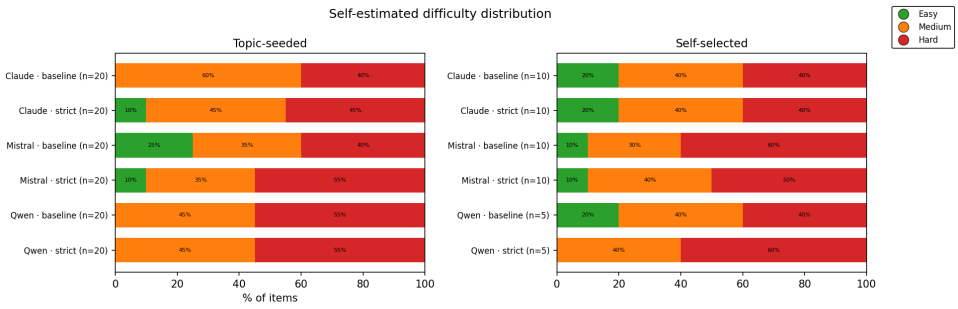


Figure 4. Self-estimated difficulty distribution

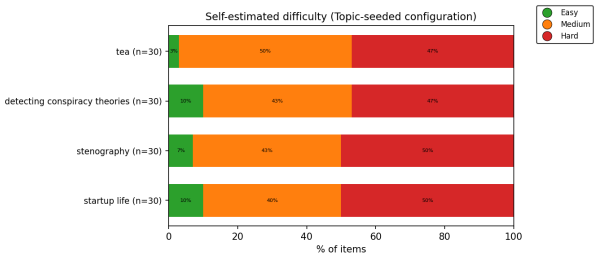


Figure 5. Self-estimated difficulty (Topic-seeded configuration)

Guessability: Table 8 reports guessability. The results do not indicate a systematic prompt effect: the strict prompt reduces g for Claude Opus 4.6 and Qwen3-32B in the topic-seeded setting, but increases it for Mistral-Small-24B, and the same lack of uniformity appears in the self-selected setting. Qwen3-32B produces the least exploitable items, reaching $g = 0.25$ (topic-seeded) and $g = 0.15$ (self-selected) under the strict prompt, although its self-selected baseline is higher ($g = 0.55$). The clearest artifact is in Claude Opus 4.6's items: in the topic-seeded baseline the length solver reaches an accuracy of 1.00, i.e. the correct option is always the longest, a cue a test-taker could exploit without reading the stimulus. Notably, this contrasts with Claude Opus 4.6's strong performance on lexical item diversity

and topic-seeded stimulus readability, showing that fluent generation does not by itself guarantee well-controlled answer options.

Model	Prompt	n	Position	Length	Lexical overlap	Guessability $g \downarrow$
Topic-seeded						
Claude Opus 4.6	baseline	20	0.60	1.00	0.40	0.75
	strict	20	0.60	0.95	0.55	0.70
Mistral-Small-24B	baseline	20	0.65	0.65	0.35	0.40
	strict	20	0.60	0.70	**0.25**	0.45
Qwen3-32B	baseline	20	**0.40**	0.60	0.45	0.35
	strict	20	0.50	**0.35**	**0.25**	**0.25**
Self-selected topic						
Claude Opus 4.6	baseline	10	0.60	0.70	0.40	0.45
	strict	10	0.50	0.80	0.30	0.55
Mistral-Small-24B	baseline	10	0.50	0.40	0.40	0.25
	strict	10	0.80	0.90	0.20	0.65
Qwen3-32B	baseline	5	0.80	0.60	**0.00**	0.55
	strict	5	**0.40**	**0.00**	**0.00**	**0.15**

Table 8. Guessability of items per configuration (n = number of items). Position, Length, and Lexical overlap are artifact-solver accuracies (chance $= 1/k = 0.25$); g is the best solver's accuracy above chance. Lower is better throughout. **Bold**: lowest value per metric within each configuration; underline: the less guessable of the two prompt variants per model.

Human assessment: Overall, the expert scores the PISA 2018 target (Table 10): are broadly similar across systems (75–80% of the maximum), with Mistral-Small-24B rated highest. The picture varies by criterion. On text quality, the stimuli are judged only moderate (50–62.5), whereas question quality is high and peaks for Mistral-Small-24B (97). Item function is the strongest group: question–stimulus alignment, answerability, and distractor quality are all high (85–100), and the items are largely non-guessable (82–87), though Claude Opus 4.6 scores lowest here, consistent with the length artifact identified above. Framework alignment is the weakest group: the declared cognitive process is judged correct only 10–40% of the time; the models generate well-formed items but rarely identify the cognitive process each item targets. Item diversity is rated moderate (37.5–50), again lowest for Claude Opus 4.6. Finally, the item mix departs from the PISA 2018 target (Table 10: all models under-produce Locate items—Search and select is absent entirely—over-produce single-source Integrate items, and generate few Easy items, therefore, the sets do not yet reproduce the balance of a full test form.

LLM as a judge

As a separate, final experiment, the AMD Silo AI team let the generating models evaluate the quality of the question sets in a round-robin fashion. The prompt used is given in Appendix E. A sample of quantitative results are shown in Table 11. We find that the models have different scoring strategies. An example evaluation is given in Figure 6. This example exhibits a crucial problem with the submitted set: the questions refer to material outside the stimuli, and yet the score is given as "Very Good."

We further measure the correlation between the two LLM judges (Prometheus-2 [7] and Gemma 3[8]) with the human expert on the same criteria (Table 12) using the prompt in Appendix F (Table 12). Agreement is weak and judge-dependent: Gemma 3 correlates moderately with the expert (Pearson $r = 0.49$, Spearman $\rho = 0.43$), whereas Prometheus 2 shows essentially no agreement ($r = -0.05$). Agreement

[7] https://huggingface.co/prometheus-eval/prometheus-7b-v2.0
[8] https://huggingface.co/google/gemma-3-27b-it

Group	Criteria	Claude Opus 4.6	Mistral-Small-24B-Instruct-2501	Qwen3-32B
Text quality	Stimulus quality	**62.5**	50.0	50.0
	Question quality	77.0	**97.0**	84.0
Item function	Question-stimulus alignment	**100.0**	98.3	91.7
	Answerable	95.0	**98.3**	91.7
	Non-guessable	81.7	**86.7**	**86.7**
	Distractor quality	**93.3**	85.0	90.0
Framework alignment	Correct cognitive process	**40.0**	20.0	10.0
	Correct difficulty	50.0	50.0	50.0
Testset	Diversity of items	37.5	**50.0**	**50.0**
Overall		76.5	**79.6**	75.4

Table 9. Human-assessment scores per model, normalized to $[0, 100]$ and averaged over the 20 submitted items per system. **Bold**: highest (best) score per criterion.

Area	Process / level	2018 Framework	Claude Opus 4.6	Mistral-Small-24B-Instruct-2501	Qwen3-32B
Locate	Access and retrieve	15%	10%	15%	0%
	Search and select	10%	0%	0%	0%
Understand	Represent literal meaning	15%	20%	0%	5%
	Integrate (single source)	15%	25%	20%	30%
	Integrate (multiple sources)	15%	5%	20%	15%
Evaluate	Assess quality and credibility	20%	10%	10%	20%
	Reflect on content and form	—	20%	20%	15%
	Detect and handle conflict	10%	10%	15%	15%
Difficulty	Easy	—	20%	10%	0%
	Medium	—	40%	35%	35%
	Hard	—	40%	55%	65%

Table 10. Distribution of the evaluated items across cognitive processes and difficulty levels per model, with the PISA 2018 target shares for reference.

is strongest on the criteria that both the judges and the expert score uniformly high: question quality, question–stimulus alignment, answerability, and distractor quality. The judges diverge most from the expert on three criteria: on the cognitive-process label the judges score high (Prometheus 2 98%, Gemma 3 73%) where the expert rated the items only 23%, and both overrate item diversity (70–75% vs. 18%) and stimulus quality (100% vs. 56%). Moreover, Prometheus-2 diverges on non-guessability, rating most items as guessable (16%) where the expert judged them not guessable (85%). Since LLM-based judgements are prone to biases [23], further work is needed to calibrate them and reduce these effects.

Lessons Learnt

As a general learning, generative language models are able to turn out reasonable quiz questions, conforming to the requirements given. The difference between models is less than might have been expected. Their main weakness is not surface quality but alignment with the PISA framework; cognitive process rarely matches the one an item targets, and the generated sets do not reproduce the process and difficulty balance of a full test form. The most diverse and readable system also produced the

Stimulus Model	Prompt	Evaluator Model			
		Opus	Sonnet	GPT-4o	GPT-5m
GPT 5 mini	baseline	62	62	96	90
GPT 5 mini	strict	62	62	95	18
Claude Opus 4.6	baseline	82	82	95	92
Claude Opus 4.6	strict	82	82	95	92
Claude Sonnet 4.6	baseline	61	71	95	92
Claude Sonnet 4.6	strict	82	82	95	88
GPT 4o	baseline	52	52	95	92
GPT 4o	strict	38	38	95	20

Table 11
Round robin PISA Item Quality Evaluation Scores (0-100)


```
{
  "source_file": "questions_claude-opus-4_standardprompt.jsonl",
  "evaluator": "claude-opus-4",
  "evaluation": {
    "overall_score": 82,
    "overall_quality": "Very Good",
    "strengths": [
      "Questions consistently target higher-order thinking skills — interpretation,
      evaluation, and critical reflection — rather than simple retrieval, aligning
      well with PISA's emphasis on deep reading competence.",
      "Distractors are carefully constructed to be plausible rather than obviously
     wrong, requiring genuine comprehension to distinguish correct answers from near-
      correct alternatives.",
     "The set covers a broad range of text types (article, debate, letter, interview,
      infographic sidebar), mirroring PISA's use of multiple text formats and
      contexts.",
      "Several questions explicitly ask students to evaluate the reliability or bias
      of sources and evidence, directly reflecting PISA's critical evaluation
      competency."
    ],
    "weaknesses": [
      "Option B is the correct answer for virtually every single question across all
      70 items, creating a severe and systematic response bias that would inflate
      scores for test-wise students guessing B and invalidates psychometric
      reliability.",
      "Questions 6–10 and others reference a discussion forum stimulus that is not
     included in the provided text, meaning those items cannot be properly evaluated
      and would be unanswerable for students without the missing stimulus.",
      "Several questions (e.g., Q11–Q15, Q31–Q35) reference stimuli about urban
     rewilding, the Aral Sea, and fast fashion that are also absent from the provided
      passage, suggesting the item set was assembled without its full set of source
      texts.",
      "Some items (e.g., Q3, Q28, Q44) that ask students to critique sources are
     strong in design but would benefit from including a counter-perspective in the
     stimulus to give students genuine textual evidence to weigh rather than relying
      on general reasoning."
    ]
  }
}
```


Figure 6: Example evaluation in which Claude Opus 4.6 judges its own submitted question set.

Criterion	Human	Prometheus 2	Gemma 3
Stimulus quality	56.2	100.0	100.0
Question quality	86.0	95.0	99.0
Question–stimulus alignment	96.7	100.0	99.2
Answerability	95.0	90.0	97.5
Non-guessability	85.0	15.8	70.0
Distractor quality	89.4	88.0	79.5
Correct cognitive process	23.3	97.5	72.5
Item diversity	17.5	70.0	75.0
Correlation with human profile (8 criteria)			
Pearson r	—	-0.05	0.49
Spearman ρ	—	0.23	0.43

Table 12. Mean per-criterion scores (normalized to $[0, 100]$) from the OECD human expert and the two LLM judges.

clearest exploitable artifact (its correct option was systematically the longest). The strict prompt also does not lead to uniformly better outputs, indicating that prompt constraints interact differently with each model. Finally, LLM-based judges proved unreliable, agreeing only weakly with the human expert and over-rating the very framework-level criteria that the models handle worst, so human expertise remains necessary.

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. Views and opinions expressed are those of the authors and do not necessarily reflect those of the European Union or the Digital Europe programme. Neither the European Union nor the programme can be held responsible for them. A part of this work was carried out within the framework of the AugmentIA Chair, supported by the Fondation Grenoble INP thanks to the patronage of Artelia Group, and is affiliated with Laboratory of informatics in Grenoble (LIG). Moreover, a part of this work received government funding managed by the French National Research Agency under France 2030, reference ANR-23-IACL-0006 (MIAI Cluster). Finally, 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.

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] P. Šindelář, O. Bojar, S. Ettejjari, L. F. V. Madriz, M. Piacentini, K. Thomas, Overview of the PISA Sensemaking 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] OECD, PISA 2022 Technical Report, OECD Publishing, Paris, 2024. doi:https://doi.org/10.1787/01820d6d-en.
[6] N. Mulla, P. Gharpure, Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications, Progress in Artificial Intelligence 12 (2023) 1–32.
[7] W. G. Lehnert, A conceptual theory of question answering, in: Proceedings of the 5th international Joint Conference on Artificial Intelligence (IJCAI), 1977.
[8] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, D. R. Krathwohl, Taxonomy of educational objectives, volume 2, Longmans, Green New York, 1964.
[9] S. Hadzhikoleva, T. Rachovski, I. Ivanov, E. Hadzhikolev, G. Dimitrov, Automated test creation using large language models: A practical application, Applied Sciences 14 (2024) 9125.
[10] S. Elkins, E. Kochmar, I. Serban, J. C. Cheung, How useful are educational questions generated by large language models?, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 536–542.
[11] S. Maity, A. Deroy, S. Sarkar, Can large language models meet the challenge of generating school-level questions?, Computers and Education: Artificial Intelligence 8 (2025) 100370.
[12] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, et al., Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.
[13] OECD, Sample Tasks from the PISA 2000 Assessment: Reading, Mathematical and Scientific Literacy, PISA, OECD Publishing, 2002. URL: https://doi.org/10.1787/9789264194274-en. doi:10.1787/9789264194274-en.
[14] OECD, Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy, PISA, OECD Publishing, 2000. URL: https://doi.org/10.1787/9789264181564-en. doi:10.1787/9789264181564-en.
[15] OECD, PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, PISA, OECD Publishing, 2017. URL: https://doi.org/10.1787/9789264281820-en. doi:10.1787/9789264281820-en.
[16] OECD, PISA 2018 Assessment and Analytical Framework, PISA, OECD Publishing, 2019. URL: https://doi.org/10.1787/b25efab8-en. doi:10.1787/b25efab8-en.
[17] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 110–119. URL: https://aclanthology.org/N16-1014/. doi:10.18653/v1/N16-1014.
[18] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, Y. Yu, Texygen: A benchmarking platform for text generation models, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1097–1100. URL: https://doi.org/10.1145/3209978.3210080. doi:10.1145/3209978.3210080.
[19] R. Flesch, A new readability yardstick., Journal of applied psychology 32 (1948) 221.
[20] J. P. Kincaid, et al., Development and test of a computer readability editing system (cres). final report, june 1978 through december 1979. (1980).
[21] R. Gunning, The Technique of Clear Writing, McGraw-Hill, 1952. URL: https://books.google.fr/books?id=ofI0AAAAMAAJ.
[22] E. Dale, J. S. Chall, A formula for predicting readability, Educational Research Bulletin 27 (1948) 11–28. URL: http://www.jstor.org/stable/1473169.
[23] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, X. Zhang, Justice or prejudice? quantifying biases in llm-as-a-judge, 2024. URL: https://arxiv.org/abs/2410.02736. arXiv:2410.02736.

A. Baseline prompt

Your Role You are an educational assessment expert specializing in crafting thoughtful and engaging reading comprehension questions for PISA (Programme for International Student Assessment). Your goal is to produce meaningful question-answer pairs that encourage reflection, insight, and nuanced understanding.
Core Objective Generate multiple choice questions from the provided source text that:
- Test genuine understanding beyond surface recall
- Make readers pause and think before answering
- Range from basic comprehension to deep insights

Processing Workflow
Step 1: **Analysis Phase**
Wrap your analysis in `<document_analysis>` tags, addressing:
Content Assessment
Carefully analyze the source text, identifying central ideas, nuanced themes, and significant relationships
Consider implicit assumptions and subtle details
Note how ideas, claims, or information are introduced, supported, or attributed
Identify the source, author, and likely purpose of the text
Identify meaningful distinctions, relationships, or dependencies within the text
Misconception Identification
Identify potential misconceptions or partial understandings
Note subtle distinctions that separate deep from surface understanding
Consider what a reader might assume versus what the text actually states
Question Planning Design questions that ask readers to:
Locate and retrieve explicit information, or search and select across the text
Comprehend literal meaning, or connect information across passages to draw conclusions
Evaluate reliability of claims, consider author's purpose or perspective, or identify contradictions
Compare, corroborate, and integrate information from multiple sources
Distractor Design
Design distractors that represent believable misconceptions
Ensure wrong answers reveal specific gaps in understanding
Step 2: **Output Generation**
After closing `</document_analysis>`, output your questions as a JSON array wrapped in `<output_json>` tags. Each item should have this structure:

```
{{
  "cognitive_process": "<one of: Access and retrieve, Search and select,
                       Represent literal meaning, Integrate and generate
                       inferences, Reflect and evaluate, Assess quality
                       and credibility>",
  "response_format": "Simple Multiple Choice",
  "difficulty_score": <integer 0-1000>,
  "difficulty_level": "<Easy, Medium, or Hard>",
  "question": "<the question text>",
  "options": {{"A": "...", "B": "...", "C": "...", "D": "..."}},
  "answer": {{"correct": "<A, B, C, or D>"}},
  "item_explanation": "<rationale for the correct response>"
}}
```

Question Design Guidelines
- Test understanding, not memorization: Cannot be answered by pattern matching alone
- Force careful reading: All options seem reasonable at first glance
- Single best answer: One clearly correct choice, but requires thought to identify
- No trick questions
- Self-contained: Questions should contain sufficient context, clearly understandable independently
- Avoid question formats that ask what is NOT in the text
- Natural phrasing: Questions a curious reader would actually ask

Distractor Design Guidelines
Create wrong answers that are:
- Plausible: Wrong answers that someone might genuinely believe

- Partially correct: Contains some truth but misses the key point
- Inversions: Reverse or contradict actual claims from the text
- Overgeneralizations: Extend a specific claim beyond what the text supports
- Domain-plausible concepts: Ideas that fit the topic but are not supported by the text
- Misattributions: Assign a claim to the wrong source within the text

Quality Standards

- Clear best answer: Experts should agree on the correct choice
- Meaningful distractors: Each reveals something about understanding
- Educational impact: Ensure clear pedagogical value and genuine content comprehension
- Requires comprehension: Correct answer should not be a verbatim or near-verbatim phrase from the source text
- Non-redundant: Each question should test distinct content

Generate 3-5 questions per stimulus text.

B. Strict prompt

The strict prompt is mostly identical to the default prompt except for as noted below. The differences are italicised.

Your Role

You are a strict educational assessment expert specializing in crafting demanding and challenging reading comprehension questions. Your goal is to ensure that someone who passes the test can be trusted to be a truly literate and competent reader.

Core Objective

Generate multiple choice questions from the provided source text that:

...

- Ensure that a successful result on the test guarantees the subject is on a path to become an intellectual.

...

Misconception Identification

...

Consider what a ***naive*** reader might assume versus what the text truly states

...

Distractor Design Guidelines

Create wrong answers that are:

...

- Are likely to fool unintelligent or childish readers

Quality Standards

...

- Educational impact: ***Ensure the test distinguishes between future thought leaders and followers***

...

C. Stimulus Generation Prompt With Seed Topic

You are an expert in creating engaging reading comprehension passages for PISA-style assessments for 15-year-old students.

Your task is to generate 1-2 short reading passages (stimulus texts) on a given topic. These passages should be similar in style, length, and complexity to the example passages provided below.

EXAMPLE PASSAGES:

{examples}

REQUIREMENTS:

- Generate 1-2 passages on the given topic
- Each passage should be 200-500 words
- Use natural, engaging language appropriate for 15-year-olds
- Include varied text types: informational, argumentative, narrative, instructional, or mixed
- Passages should contain enough depth for multiple comprehension questions
- Use realistic scenarios, authentic contexts, and credible sources when appropriate
- Format similar to examples (may include headings, dialogue, multiple sections, etc.)
- The questions that will be formulated to test the comprehension of these passages will probe the reader's capability to

– Search and select relevant text segment
– Understand and interpret literal meaning from a single and from multiple sources
– Evaluate and reflect on quality and veracity of a source
– Evaluate and reflect on content and form of a source
– Detect and handle conflict

The generated passages should show a reasonable variety across text source, literary type (narrative, informational, argumentative, conversational) and text length.

TOPIC: {topic}

OUTPUT FORMAT:

Provide your response as a JSON array wrapped in `<output_json>}` tags:

```
[
  {{
    "topic": "{topic}",
    "stimulus_text": "First passage text here..."
  }},
  {{
    "topic": "{topic}",
    "stimulus_text": "Second passage text here (if applicable)..."
  }}
]
```

If only one passage is appropriate for the topic, provide just one object in the array.

D. Stimulus Generation Prompt

Generate {count} completely new topics suitable for creating PISA-style reading comprehension passages for 15-year-old students.

Each topic should be:

- Engaging and age-appropriate for teenagers
- Suitable for creating 200-500 word reading passages
- Diverse in subject matter (mix of science, social issues, culture, technology, etc.)
- Specific enough to write about concretely

Output as JSON array in <output_json>} tags:

```
[
  {{"id": "new_01", "title": "topic title", "description": "brief description"}},
   ...
]
```

E. Evaluation prompt

You are an expert in educational assessment and PISA-style reading comprehension questions.

Below is a reading passage (stimulus) followed by multiple-choice questions designed for 15-year-old students. Ensure that each question can be answered through material given in the stimulus texts (except for very general background knowledge which does not need to be specified in the stimuli.)

STIMULUS: {stimulus_text}

QUESTIONS: {questions}

Evaluate according to PISA criteria and provide an overall score (0-100).

Output ONLY valid JSON in <output_json> tags with this structure (keep it concise):

```
<output_json>
{{
  "overall_score": 75,
  "overall_quality": "Good",
  "strengths": ["strength 1", "strength 2"],
  "weaknesses": ["weakness 1", "weakness 2"]
}}
</output_json>
```

CRITICAL: Output ONLY the JSON in tags. No additional text. Be concise - max 2 sentences per strength/weakness.

F. Evaluation prompt Prometheus-2 and Gemma 3

```
###Task Description:
You are an expert reviewer of the PISA (Programme for International Student
    Assessment) reading-literacy domain.
The {target} to evaluate and a score rubric representing an evaluation criterion
    are given.
1. Write a brief feedback (maximum 50 words) that assess the quality of the {target
    } strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between {min} and {
    max}. You should refer to the score rubric.
3. The output format should look as follows: "Feedback: (write a feedback for
    criteria) [RESULT] (an integer number between {min} and {max})"
4. Please do not generate any other opening, closing, and explanations.

###The {target} to evaluate:
{input}

###Score Rubrics:
{score_rubric}

###Feedback:
```

Criteria (score range in parentheses; applied to a stimulus–question pair unless noted):

- **Stimulus quality** (0–5): Is the stimulus natural, realistic, matching the reading situation, and free from biased, inappropriate or culturally-specific content? (applied to the stimulus)
- **Question quality** (0–5): Is the question clear, concise, and its requirements unambiguous?
- **Question–stimulus alignment** (0–3): Is the question aligned with the stimulus?
- **Answerable** (0–3): Can the question be answered using only the stimulus and no prior knowledge?
- **Guessable** (0–3): Can the question be answered without reading the stimulus? (0 = guessable; applied to the question and options only)
- **Distractor quality** (0–5): Are the distractors plausible, clearly incorrect, and not easily eliminated?
- **Correct cognitive process** (0–1): Does the declared cognitive-process label match the process the item actually requires?
- **Correct difficulty** (0–1): Is the difficulty level appropriate and correctly labelled?
- **Diversity of items** (0–5): Do the items cover diverse cognitive processes and difficulty levels? (applied to the whole submission)