# Can GPT-4 do L2 analytic assessment?

**Stefano Bannò, Hari K. Vydana, Kate M. Knill, Mark J. F. Gales**
ALTA Institute, Department of Engineering, University of Cambridge (UK)
{sb2549,hkv21,kmk1001,mjfg100}@cam.ac.uk



## Abstract

Automated essay scoring (AES) to evaluate second language (L2) proficiency has been a firmly established technology used in educational contexts for decades. Although holistic scoring has seen advancements in AES that match or even exceed human performance, analytic scoring still encounters issues as it inherits flaws and shortcomings from the human scoring process. The recent introduction of large language models presents new opportunities for automating the evaluation of specific aspects of L2 writing proficiency. In this paper, we perform a series of experiments using GPT-4 in a zero-shot fashion on a publicly available dataset annotated with holistic scores based on the Common European Framework of Reference and aim to extract detailed information about their underlying analytic components. We observe significant correlations between the automatically predicted analytic scores and multiple features associated with the individual proficiency components.

## 1 Introduction

Automated essay scoring (AES) of second language (L2) proficiency is a well-established technology in educational settings, involving the automatic scoring and evaluation of learners' written productions through computer programs (Shermis and Burstein, 2003).

Originating in the 1960s, the roots of AES can be traced back to the development of Project Essay Grade (PEG) (Page, 1966, 1968), an automatic system which evaluated writing skills based only on proxy traits: hand-written texts had to be manually entered into a computer, and a scoring algorithm then quantified superficial linguistic features, such as essay length, average word length, count of punctuation, count of pronouns and prepositions, etc. Across the following decades, as natural language processing (NLP) technologies have advanced and increased their power (Landauer, 2003), the field of AES has expanded and improved, and more significant studies have been conducted from the 1990s and early 2000s. The most widely known automated scoring systems for essays include the e-rater®, developed by Educational Testing Service (Burstein, 2002; Attali and Burstein, 2006), IntelliMetric™ by Vantage Learning (Rudner et al., 2006), and the Intelligent Essay Assessor™, built at Pearson Knowledge Technologies (Landauer et al., 2002).

In recent years, deep neural network (DNN) approaches have brought significant improvements (Alikaniotis et al., 2016), and especially the advent of transformer-based architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) which took the world of NLP and, consequently, AES by storm, outperforming classic feature-based systems (Rodriguez et al., 2019). Yet, the most recent breakthrough has been brought by large language models (LLMs), such as the GPT models (Brown et al., 2020; OpenAI, 2023), which might revolutionise the world of AES, not only from the NLP experts' and language testers' perspective, but also considering the users' point of view due to GPT's extremely accessible and intuitive interface. In the context of L2 writing assessment, previous studies have employed GPT-3.5 (Mizumoto and Eguchi, 2023) and GPT-4 (Yancey et al., 2023), obtaining promising results.

Although LLMs have been employed for holistic scoring (i.e., assessing the overall quality of a composition as a whole, considering various aspects such as vocabulary, grammar, coherence, etc. altogether), to the best of our knowledge, so far they have not been investigated for the task of analytic scoring (i.e., breaking down a composition into specific components or criteria and assigning separate scores or ratings to each component).[1] Offering L2

[1] Naismith et al. (2023) investigated the use of GPT-4 on a proprietary dataset annotated with specific scores targeting coherence only.

learners specific analytic proficiency scores is crucial for delivering insightful and effective feedback, emphasising both their strengths and weaknesses to facilitate improvement.

For holistic scoring, previous works have shown that state-of-the-art automatic techniques can reach near-human results (Alikaniotis et al., 2016; Taghipour and Ng, 2016) or even outperform them (Rodriguez et al., 2019). This is, at least in part, ascribable to the fact that holistic scores are generally easier to obtain for human evaluators (see Section 2). Conversely, assessing analytic aspects of language proficiency is generally considered to be more difficult, time-consuming, and cognitively demanding for human evaluators, and, as a result, "noisy" ground truth scores are harder to learn and predict for automatic systems (see Section 2).

Starting from these premises, in this paper, we conduct a series of exploratory experiments on a publicly available dataset annotated with holistic scores according to the Common European Framework of Reference (CEFR) (Council of Europe, 2001, 2020) using GPT-4 in a zero-shot fashion, and aim to extract specific information about their underlying analytic components. Although ground truth analytic scores are not available, we find significant correlations between the analytic scores predicted by the model and several features related to the analytic scores.

## 2 Holistic versus analytic scoring

### 2.1 Human assessment

Holistic and analytic approaches to assessing L2 proficiency are commonly utilised, differing in scoring methods, underlying assumptions, and practical application. While holistic assessment consists of assigning a single overall numerical score to a specific performance based on a singular set of rating criteria, analytic assessment involves providing various sub-scores to the performance based on multiple sets of criteria. As a result, there are conceptual differences between the two approaches (Barkaoui, 2011). Holistic assessment typically assumes that the construct being evaluated is a unitary entity and can be represented on a single scale. While this approach acknowledges that the construct may consist of various elements, it implies that development across various aspects of proficiency is uniform. Conversely, analytic assessment views the construct as multi-dimensional and advocates for a multi-faceted assessment, recognising that development across various aspects may be irregular. For instance, the levels of the CEFR are structured according to 'can-do' descriptors of language proficiency outcomes and expect evaluators to grade proficiency by means of holistic assessments. Nonetheless, the CEFR levels do have a modularisable structure with multiple underlying components (e.g., vocabulary range, vocabulary control, grammatical accuracy, etc.), acknowledging that a learner may be more proficient in certain aspects than others (Council of Europe, 2001, 2020).

When we consider assessment strictly from a human perspective, holistic assessment is considered highly practical as it is more time-efficient per se and in relation to rater training (White, 1984), less cognitively demanding (Xi, 2007), and generally has a higher inter-annotator agreement (Weigle, 2002) than analytic assessment. On the other hand, holistic scoring may suffer from lack of clarity regarding how different aspects are prioritised, which may vary among evaluators (Weigle, 2002; Xi, 2007), the risk that evaluators might primarily concentrate on candidates' strengths rather than their weaknesses (Bacha, 2001), and the potentially erroneous assumption that various aspects of proficiency develop uniformly over time (Kroll, 1990).

Analytic assessment allows for a more detailed and systematic evaluation and is supposed to provide much more detailed feedback to L2 learners, by highlighting their fortes and their weaknesses (Hamp-Lyons, 1995) in addition to enhancing scoring validity. However, it is not a panacea. Analytic scores may be psychometrically redundant (Lee et al., 2009) due to a halo effect (Engelhard, 1994), whereby raters fail to distinguish between different aspects of learners' performances but assess all or some of them with similar scores. For example, when assessing grammatical accuracy, raters might be influenced by the score previously assigned to vocabulary range. On top of this, raters might confuse analytic criteria in the phase of assessment due to high cognitive load (Underhill, 1987; Cai, 2015) or, more simply, to indefiniteness of the analytic criteria (Douglas and Smith, 1997). The difficulty in providing analytic scores — especially for a large number of written productions — is evident in the total absence of publicly available L2 English learner datasets annotated in this way[2] and the fact that the primary emphasis in AES

[2]To the best of our knowledge, the only formerly publicly

research has been on holistic scoring.

### 2.2 Automatic assessment

The introduction of automatic assessment techniques — and especially their recent advancements — have started to change the game. For holistic scoring, DNN-based systems reached near-human performances (Alikaniotis et al., 2016; Taghipour and Ng, 2016), and the application of transformers-based architectures even beat human inter-annotator agreement (Rodriguez et al., 2019). However, a notorious problem lies in the impossibility to enter the black box of neural scoring models, and this poses a challenge for explainability and interpretability of the machine-generated holistic scores. Even more so, it is important to explore the ability of automatic models to evaluate specific aspects of language proficiency through analytic scoring: if it is not possible to decompose the holistic assessment process by peeking inside the black box, it may be possible to reconstruct holistic scores starting from their analytic components (with the caveat that we should keep in mind the potential unreliability of human analytic scores, as discussed above). In this regard, automatic systems have been found to be generally better at evaluating specific linguistic phenomena, whilst humans tend to focus on more general aspects of proficiency. For example, Enright and Quinlan (2010) suggested that human raters might achieve higher results when assessing ideas, content, and organisation, whereas automatic systems might have better performances when evaluating microfeatures at the grammatical, syntactic, lexical, and discourse levels. It should be noted, however, that these limitations attributed to automatic systems may no longer necessarily be true in light of the recent advancements involving neural systems, which can be used quite effectively also to assess higher-level aspects of proficiency. For example, previous studies have focused on specific traits of written productions, such as organisation, content, word choice, sentence fluency, narrativity, etc. (Hussein et al., 2020; Mathias and Bhattacharyya, 2020; Ridley et al., 2021), but they have used the ASAP dataset, which is problematic for reproducibility and only features essays written by L1 English speakers (see note 2). For L2 speaking assessment, the initial study by Bannò et al. (2022) investigated the use of multiple different graders, each of which focused on a different set of features related to a specific proficiency aspect.

The introduction of LLMs could be a further game-changer, considering their outstanding results in a broad range of tasks.

To sum up, given that:

- holistic scores are generally easier to obtain both from human and automatic graders and generally have a higher inter-annotator agreement, hence higher reliability;
- analytic scores are difficult to obtain and might not always be sufficiently reliable;
- more often than not, L2 learner datasets are annotated with holistic scores only;
- LLMs have been proven to be extremely powerful tools in many NLP tasks;

we pose the following research question:

> is it possible to extract information about analytic aspects from L2 learner essays and their assigned holistic scores using GPT-4?

Figure 1 shows the pipeline adopted in this study, which will be illustrated in detail in Section 4.

## 3 Data

### 3.1 Write & Improve

Write & Improve (W&I) is an online platform where L2 learners of English can practise their writing skills (Yannakoudakis et al., 2018). Users can submit their compositions in response to different prompts, and the W&I automatic system provides assessment and feedback. Some of these compositions have been manually annotated with CEFR levels and grammatical error corrections since 2014, resulting in a corpus of 3,300 texts, partitioned into a training set of 3,000 and a validation set of 300 essays.[3] The proficiency scale ranges from A1 to C2 but also has intermediate levels, resulting in 12 levels, that we arranged on a scale from 1 to 6.5, where 1 is A1, 1.5 is A1+, 2 is A2, 2.5 is A2+, etc., as shown in Table 5 (see Appendix D).

available dataset annotated with analytic scores is the ASAP dataset (kaggle.com/c/asap-aes/data), but the test data are no longer available for evaluation and comparison with previous work. Furthermore and most importantly, it contains essays written by L1 English speakers.

[3]The dataset can be downloaded from this link: huggingface.co/datasets/wi_locness.

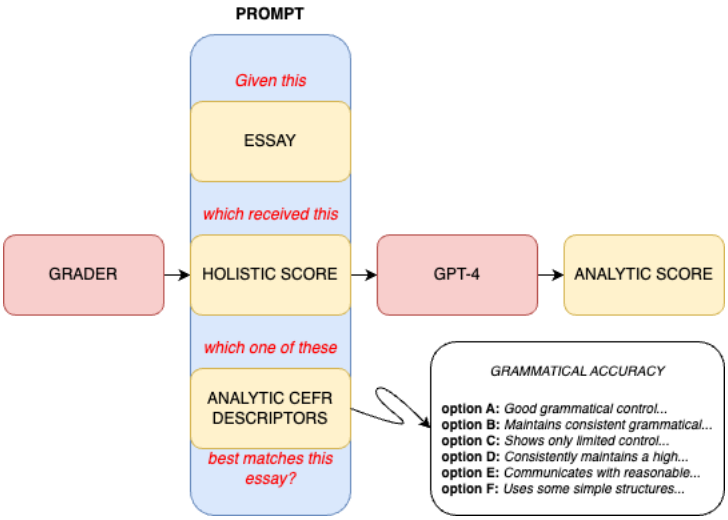


Figure 1: The pipeline presented in this study. Grammatical accuracy is only one of the aspects considered.

### 3.2 EFCAMDAT

Arguably the largest publicly available[4] L2 learner corpus, the second release of EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013; Huang et al., 2017, 2018) comprises 1,180,310 scripts written by 174,743 L2 learners as assignments to Englishtown, an online English language school. The compositions are annotated with a score on a scale from 0 to 100 and a proficiency level from 1 to 16 (mapped to CEFR levels from A1 to C2).[5] In order to align them to the proficiency levels in the W&I dataset, we normalised the scores as described in Table 5 (see Appendix D). For our experiments, we selected a subset of data consisting of 753,508 essays for the training set and 7612 for the validation set, following a similar process to Bannò et al. (2023).

## 4 Experimental setup

### 4.1 Longformer-based holistic grader

Following the pipeline illustrated in Figure 1, we start our experiments from training a holistic grader, which consists of a Longformer model (Beltagy et al., 2020) in the version provided by the HuggingFace Transformer Library,[6] a dropout layer, a dense layer of 768 nodes, a dropout layer, another dense layer of 128 nodes, and finally, the output layer. The baseline model (W&I) is trained on the W&I training data and optimised on the W&I validation data using an Adam optimiser (Kingma and Ba, 2014) for 3 epochs with batch size 16, learning rate 1e-6 and mean squared error as loss, but our best-performing model — which is the one we will use in the following steps of our pipeline — is trained on the EFCAMDAT training set and optimised on the validation data from the same dataset for 0.5 epochs with batch size 16 and learning rate 1e-5, and subsequently fine-tuned on the W&I training data and optimised on the W&I validation data for 4 epochs.

To evaluate the holistic grader performance, we use Pearson's correlation coefficient (PCC), Spearman's rank coefficient (SRC), and root-mean-square error (RMSE).

### 4.2 GPT-4-based analytic graders

Once we obtain the holistic scores from the Longformer-based model, we move on to feeding them into GPT-4 (*"gpt-4-1106-preview"*) to extract analytic scores. Specifically, the analytic scores are related to 9 proficiency aspects as described in Council of Europe (2020), reported in Appendix A. Five of them compose the linguistic competence: *general linguistic range, vocabulary range, grammatical accuracy, vocabulary control,* and *orthographic control*; while the remaining four form the pragmatic competence: *flexibility, thematic development, coherence and cohesion,* and *propositional precision*.

We excluded sociolinguistic appropriateness because it is not consistently elicited in the W&I essays, as well as the aspects strictly related to speaking proficiency (i.e., phonological control, turntaking, and fluency) for obvious reasons.

[4] ef-lab.mmll.cam.ac.uk/EFCAMDAT.html

[5] englishlive.ef.com/en/how-it-works/levels-and-certificates/

[6] huggingface.co/allenai/longformer-base-4096

The prompt given to GPT-4 can be found in Appendix C. To exclude potential biases, the holistic scores are fed in their numerical form (i.e., from 1 to 6.5) instead of the original CEFR notation (i.e., from A1 to C2+), and the analytic CEFR descriptors are provided in random order and, obviously, without any reference to the CEFR levels. For completeness, we also try this experiment without giving GPT-4 the holistic score.

At the end of the process, the option selected by GPT-4 is mapped back to its respective CEFR level.

### 4.3 Explanation of the features

As mentioned in Section 1, the W&I dataset does not include analytic scores, but we find significant correlations with relevant features extracted from the essays (see Tables 3 and 4).

**%gram.** refers to the grammatical error rate, which is the number of grammatical error edits divided by the number of words in the essay. These edits are extracted by feeding the original and corrected versions of the W&I essays into the ERRor ANnotation Toolkit (ERRANT) (Bryant et al., 2017).

**#dif.wds.** is the number of unique difficult words extracted with `textstat`.[7]

**#unq.wds.** refers to the number of unique words.

**%l.d.t.** is the percentage of text types that are content words obtained using TAACO (Tool for the Automatic Analysis of Text Cohesion) 2.0 (Crossley et al., 2019).

**#unq.n.chunks** refers to the number of unique noun chunks identified and extracted using spaCy. [8]

**#unq.q.m.a.** refers to the number of unique qualifiers, modality markers, and ambiguity indicators identified and extracted using spaCy.

**fl.-kinc.** is the Flesch Kincaid readability score (Kincaid et al., 1975), obtained using `textstat`.

**w2v** is the average word2vec (Mikolov et al., 2013) similarity score between all adjacent paragraphs, extracted with TAACO 2.0.[9]

**av.s.ln.** is the average sentence length.

The correlations between these features and the analytic scores are evaluated using SRC since we do not necessarily expect a linear correlation between the two. For example, it is well-known that certain grammatical errors are absent or rare in the A1 level, increase after B1, and then decline again by C2 (Hawkins and Buttery, 2010).

## 5 Experimental results

### 5.1 Holistic scoring

Table 1 shows the results of the Longformer-based holistic graders on the W&I validation set in terms of PCC, SRC, and RMSE. The model pre-trained on EFCAMDAT and fine-tuned on the W&I training set outperforms the baseline across all metrics as expected. These results should confirm that holistic grading is a relatively easy task and, since the training data are fully publicly available, potentially within everyone's reach.

| Model | PCC | SRC | RMSE |
|---|---|---|---|
| W&I | 0.707 | 0.772 | 1.267 |
| EFC+W&I | 0.866 | 0.874 | 0.786 |

Table 1: Holistic scoring results on W&I validation set.

### 5.2 Holistic score reconstruction

Once we obtain the holistic scores from the Longformer-based grader, we are ready to feed them into GPT-4. However, before moving on to the analysis of the individual analytic scores, we first calculate the correlation between the average of the predicted analytic scores — when providing GPT-4 with the holistic scores from the ground truth (GT) or the Longformer-based grader (EFC+W&I), or with no holistic score (-) — and the holistic scores, both the ground truth (GT) and the scores automatically predicted by the Longformer-based grader (EFC+W&I), as shown in Table 2.

| GPT-4 Prompt Holistic Score | Reference | |
|---|---|---|
| | GT | EFC+W&I |
| GT | 0.904 | 0.874 |
| EFC+W&I | 0.828 | 0.898 |
| - | 0.797 | 0.827 |

Table 2: SRC correlation between the average of the predicted analytic scores and the holistic scores.

The first result that catches the eye is that GPT-4 reaches a significant correlation of 0.797 when it is not provided with additional information about holistic scores (-), although this does not necessarily mean that all the underlying analytic scores

[7] github.com/textstat/textstat

[8] spacy.io/

[9] Initially, we also extracted the similarity score using Latent Semantic Analysis (Landauer et al., 1998) and Latent Dirichlet Allocation (Blei et al., 2003), which showed similar figures, but we did not include them due to reasons of space.

are effectively targeting their respective proficiency aspects, as we will discuss in the next section. Secondly, it is interesting to observe that the two sources of holistic score in the prompts (i.e., GT and EFC+W&I) result in the information derived from these scores being used in a non-deterministic fashion, introducing a certain degree of variability.

### 5.3 Analytic scoring

We can now move on to discussing the results of analytic scoring. Table 3 shows the correlation results in terms of SRC between the predicted analytic scores and several relevant features for each proficiency aspect. Table 4 does the same but giving GPT-4 the ground truth holistic scores instead of the scores predicted by the holistic grader. Particularly in the latter, when focusing on the results highlighted in bold, we can observe a broad trend towards an approximate diagonal which passes through most of the proficiency aspects of the linguistic (Lng.) and pragmatic (Prg.) competences on the y-axis and the relevant features on the x-axis. For completeness, in Table 6 (see Appendix D), we also report the results obtained without giving GPT-4 the holistic score, but the correlations are not as significant as the ones shown in Tables 3 and 4 as the holistic score seems to work as a guide for analytic scoring. Furthermore, as expected, the correlations between each individual predicted analytic score and the holistic scores are significantly lower than the ones reported in Tables 3 and 4. Therefore, our analysis in the following lines will not dwell on these results.

As expected, grammatical error rate (%gram.) shows the highest correlations with the aspects of grammatical accuracy and orthographic control both on Tables 3 and 4.

The number of unique difficult words (#dif.wds.) seems to be a suitable feature to measure vocabulary control, e.g., if we compare the A2 level (i.e., "Can control a narrow repertoire dealing with concrete, everyday needs.") and the C1 level (i.e., "Uses less common vocabulary idiomatically and appropriately."), as described in Council of Europe (2020, pp. 132-133) (see Appendix A). Indeed, this feature shows the highest correlation with the score related to vocabulary control.

If we look at the results obtained giving the ground truth holistic scores to GPT-4 shown in Table 4, we can see that the number of unique words (#unq.wds.), the percentage of lexical density types (%l.d.t.), and the number of unique noun chunks (#unq.n.cks.), which are all related to lexicality, have their highest correlation with the two scores related to vocabulary. As expected, the same features have slightly weaker — but still relevant — correlations when we use the automatically predicted holistic scores, as shown in Table 3.

The number of unique qualifiers, modality markers, and ambiguity indicators (#unq.q.m.a.) is supposed to be a measure for propositional precision since, for example, as shown in Appendix A, a C1-level learner "[c]an qualify opinions and statements precisely in relation to degrees of, for example, certainty/uncertainty, belief/doubt, likelihood, etc" and "[c]an make effective use of linguistic modality to signal the strength of a claim, an argument or a position", and a C2-level learner "[c]an convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of qualifying devices [...]" and "[c]an give emphasis, differentiate and eliminate ambiguity" (Council of Europe, 2020, p. 141). As can be observed in Table 4, this feature has the second-highest correlation with the propositional precision score and the highest correlation with the score related to vocabulary control, with which it is in fact connected. Similarly to what we observed about the lexical features, the results of the fully-automated pipeline for this feature are less evident, but we can still see a rather high correlation with propositional precision.

Given its emphasis on precision and clarity, we thought that also the Flesch-Kincaid readability score (fl.kinc.) would be a suitable feature to measure these. We found that the highest correlation was exactly with propositional precision followed by vocabulary control on both Tables 3 and 4.

Furthermore, we considered two features for the pragmatic competence, especially in relation to cohesion and coherence. The first one is the average word2vec similarity score between all adjacent paragraphs (w2v), which shows the highest correlations on propositional precision and cohesion and coherence in Table 4. The second is average sentence length (av.s.ln.), which should be an indicator of higher use of subordination and cohesive devices (i.e., longer sentences should generally be more complex). This feature shows similar results, as shown in Table 4. When using the scores provided by the automatic holistic grader, the results on both features are also slightly weaker (see Table 3), as observed already for other features above.

It is rather difficult to provide a precise and exhaustive explanation of the results for the general

| | score | %gram. | #dif.wds. | #unq.wds. | %l.d.t. | #unq.n.cks. | #unq.q.m.a. | fl.-kinc. | w2v | av.s.ln. | holistic |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lng.** | gen. lin. | 0.695 | 0.584 | 0.514 | 0.400 | 0.493 | 0.527 | 0.259 | 0.258 | 0.143 | 0.765 |
| | gramm. | **0.698** | 0.505 | 0.469 | 0.370 | 0.423 | 0.468 | 0.189 | 0.265 | 0.134 | 0.737 |
| | orth. | **0.718** | 0.395 | 0.317 | 0.244 | 0.291 | 0.350 | 0.155 | 0.206 | 0.073 | 0.652 |
| | voc. ctrl. | 0.652 | **0.638** | **0.580** | **0.445** | 0.537 | **0.600** | **0.263** | 0.291 | **0.189** | 0.779 |
| | voc. rg. | 0.651 | **0.621** | 0.568 | 0.424 | **0.548** | 0.576 | 0.254 | **0.339** | 0.177 | 0.749 |
| **Prg.** | propos. | 0.601 | 0.607 | 0.545 | 0.389 | 0.528 | 0.568 | **0.294** | **0.351** | **0.202** | 0.702 |
| | coh. | 0.662 | **0.621** | **0.574** | 0.410 | **0.551** | **0.588** | 0.248 | 0.336 | 0.180 | 0.774 |
| | flexib. | 0.424 | 0.414 | 0.390 | 0.291 | 0.367 | 0.412 | 0.178 | 0.195 | 0.125 | 0.443 |
| | themat. | 0.584 | 0.544 | 0.527 | **0.428** | 0.516 | 0.534 | 0.203 | 0.287 | 0.145 | 0.650 |
| | holistic | 0.732 | 0.640 | 0.665 | 0.451 | 0.623 | 0.637 | 0.178 | 0.364 | 0.141 | 1.000 |

Table 3: SRC correlation of the GPT-4 predicted scores and relevant linguistic features (**using holistic scores predicted by the Longformer-based grader**). The *holistic* entry refers to the ground-truth holistic scores. In bold the two highest correlations columnwise.

| | score | %gram. | #dif.wds. | #unq.wds. | %l.d.t. | #unq.n.cks. | #unq.q.m.a. | fl.-kinc. | w2v | av.s.ln. | holistic |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lng.** | gen. lin. | 0.726 | 0.574 | 0.541 | 0.414 | 0.522 | 0.519 | 0.197 | 0.267 | 0.129 | 0.814 |
| | gramm. | **0.731** | 0.472 | 0.464 | 0.363 | 0.433 | 0.450 | 0.100 | 0.286 | 0.030 | 0.791 |
| | orth. | **0.726** | 0.436 | 0.398 | 0.310 | 0.354 | 0.427 | 0.146 | 0.203 | 0.060 | 0.729 |
| | voc. ctrl. | 0.674 | **0.640** | **0.621** | **0.453** | **0.591** | **0.624** | **0.243** | 0.319 | 0.179 | 0.854 |
| | voc. rg. | 0.672 | **0.624** | **0.582** | **0.452** | **0.563** | 0.573 | 0.218 | 0.280 | 0.134 | 0.816 |
| **Prg.** | propos. | 0.600 | **0.624** | 0.581 | 0.417 | 0.560 | **0.593** | **0.261** | **0.353** | **0.190** | 0.771 |
| | coh. | 0.702 | 0.555 | 0.534 | 0.372 | 0.511 | 0.535 | 0.238 | **0.339** | **0.201** | 0.827 |
| | flexib. | 0.425 | 0.370 | 0.368 | 0.249 | 0.357 | 0.368 | 0.140 | 0.163 | 0.104 | 0.488 |
| | themat. | 0.639 | 0.514 | 0.504 | 0.413 | 0.492 | 0.483 | 0.224 | 0.264 | 0.179 | 0.745 |
| | holistic | 0.732 | 0.640 | 0.665 | 0.451 | 0.623 | 0.637 | 0.178 | 0.364 | 0.141 | 1.000 |

Table 4: SRC correlation of the GPT-4 predicted scores and relevant linguistic features (**using ground truth holistic scores**). The *holistic* entry refers to the ground-truth holistic scores. In bold the two highest correlations columnwise.

linguistic range score, which is a broad indicator by definition since it includes elements of grammatical accuracy, syntactic complexity, and vocabulary, and, as a result, shows strong correlations with multiple features. On the other hand, the aspect of flexibility seems to be a little problematic with respect to both the features and the holistic score, probably also due to its "longitudinality", since it seems to be evaluated in relation to previous performances, according to its descriptors (see Appendix A).

Finally, we selected some essays in which there was a large discrepancy between two or more analytic scores, and we evaluated them impressionistically. One example can be found in Appendix B. If we focus on the highest and lowest scores, we notice vocabulary range and orthographic control on one hand, and coherence and cohesion on the other hand. Although quite extreme, this discrepancy makes sense, considering that the learner uses almost no connectors at all and mostly uses coordinating clauses (or even parataxis), but has quite a rich vocabulary and makes no orthographic errors (except for punctuation).

### 5.4 Statistical tests

Additionally, we explore the relationships among analytic scores using a repeated measures design in order to assess whether there are significant differences among them. While the repeated measures analysis of variance (rANOVA) is a widely known approach for such designs, our data fail to meet the assumptions of sphericity and normality required for its application. Hence, we employ the Friedman test (Friedman, 1937), known as the non-parametric equivalent of rANOVA. This test assesses whether there are significant differences in ranks among multiple paired groups. With a significant $p$-value obtained, we confirm significant differences among the analytic scores. To determine which scores show significant differences, we conduct post-hoc multiple comparisons using the Ne-

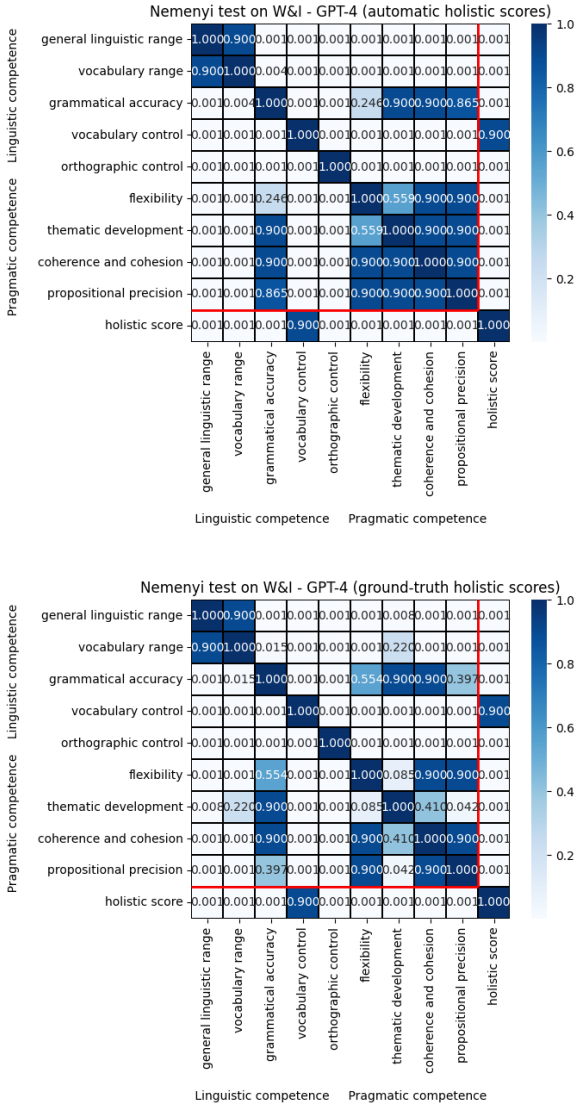


Figure 2: Results of the post-hoc Nemenyi test.

menyi test (Nemenyi, 1963), whose results are reported in Figure 2. The majority of the paired comparisons, even those with the holistic score (except when paired with vocabulary control), show significant differences (i.e., $p$-value<0.05) both when we provide the ground truth and the automatic holistic scores to GPT-4. In addition to the pairs "general linguistic range - vocabulary range" and "thematic development - vocabulary range", which have some clear overlaps in their descriptors, there seem be non-significant differences over the group of aspects related to the pragmatic competence (i.e., flexibility, thematic development, coherence and cohesion, and propositional precision) and the aspect of grammatical accuracy. While we could expect to see non-significant differences among the aspects related to the pragmatic competence due to their frequent overlaps, the non-significant differences of these with grammatical accuracy might be explained with the fact that not only do its descriptors stress the importance of correctness but, as shown in Appendix A, they also emphasise complexity (e.g., for A1: "Shows only limited control of a few simple grammatical structures [...]"; for B2: "Has a good command of simple language structures and some complex grammatical forms [...]"), which is inherently connected to aspects such as thematic development and coherence and cohesion (Purpura, 2004). In this regard, it is also worth noting that the coherence and cohesion score is the third most correlated with grammatical error rate.

To sum up, under ideal conditions, GPT-4 appears to produce analytic scores that are very reasonably related to the proficiency aspects they are expected to evaluate. The fully-automated pipeline is not always consistent with the ideal system but generates results that are mostly in line with it. This is especially evident for the scores pertaining to grammar and vocabulary.

## 6 Conclusions

In this paper, we have conducted an initial study on the use of GPT-4 for assessing 9 individual aspects of L2 writing underlying the CEFR proficiency levels in a zero-shot fashion. To do this, we used a holistic grading system on the essays of the W&I validation set and, subsequently, fed them with their respective holistic scores into GPT-4, asking to assess one individual aspect at a time. Although the ground truth analytic scores are not available, we have obtained significant correlations between the predicted analytic scores and various features linked to the componential aspects of the CEFR levels. Beyond its immediate implications for computer-assisted language learning applications, we believe that our exploratory experiments may hold promise as valuable contributions to theoretical studies on construct validity in the broader field of language testing and assessment, given the inclusion of CEFR descriptors in our study.

In order to collect further evidence to support our findings, we plan to deploy this system, use it in educational settings, and evaluate its effectiveness by monitoring learners' progress in relation to each specific aspect of proficiency. Future work will also explore the use of multi-modal systems, such as the one presented in Tang et al. (2023), for assessing L2 speech in a similar fashion.

## Limitations

The main limitation of this study is clearly the lack of ground truth analytic scores. The reader should keep in mind, however, that, as mentioned in Section 2, human analytic scoring is often an extremely difficult process, which might not produce completely reliable information. As evidence of this, the absence of publicly available L2 English learner datasets annotated with analytic scores speaks loud and clear and is not only an issue for the objectives of this paper, but for the whole scientific community.

## Acknowledgements

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. The authors would like to thank the ALTA Spoken Language Processing Technology Project Team for general discussions and contributions to the evaluation infrastructure.

## References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 715–725. Association for Computational Linguistics (ACL).

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Nahla Bacha. 2001. Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, 29(3):371–383.

Stefano Bannò, Bhanu Balusu, Mark J. F. Gales, Kate M. Knill, and Konstantinos Kyriakopoulos. 2022. View-specific assessment of L2 spoken English. In *Proceedings of Interspeech 2022*, pages 4471–4475.

Stefano Bannò, Michela Rais, and Marco Matassoni. 2023. Grammatical Error Correction for L2 Speech Using Publicly Available Data. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 136–140.

Khaled Barkaoui. 2011. Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18(3):279–293.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Jill Burstein. 2002. The e-rater scoring engine: automated essay scoring with natural language processing. In M. D. Shermis and J. Burstein, editors, *Automated essay scoring: a cross-disciplinary perspective*, pages 113–122. Routledge, New York.

Hongwen Cai. 2015. Weight-Based Classification of Raters and Rater Cognition in an EFL Speaking Test. *Language Assessment Quarterly*, 12(3):262–282.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment - Companion volume*. Council of Europe, Strasbourg.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51(1):14–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Douglas and Jan Smith. 1997. *Theoretical underpinnings of the Test of Spoken English revision project*. Educational Testing Service Princeton, NJ.

George Engelhard. 1994. Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2):93–112.

Mary K. Enright and Thomas Quinlan. 2010. Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 3(27):317–334.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*, pages 240–254, Somerville. Cascadilla Proceedings Project.

Liz Hamp-Lyons. 1995. Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, 29(4):759–762.

John A. Hawkins and Paula Buttery. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1:1–23.

Yan Huang, Jeroen Geertzen, Rachel Baker, Anna Korhonen, and Theodora Alexopoulou. 2017. The EF Cambridge Open Language Database (EFCAMDAT): Information for users.

Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.

Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5).

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report, Tech. Rep. Naval Technical Training Command - Millington TN Research Branch.

Diederick P. Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

Barbara Kroll. 1990. *Second Language Writing (Cambridge Applied Linguistics): Research Insights for the Classroom*. Cambridge Applied Linguistics. Cambridge University Press.

Thomas K. Landauer. 2003. Automatic essay assessment. *Assessment in education: principles, policy and practice*, 10(3):295–308.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse processes*, 25(2-3):259–284.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2002. Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis and J. Burstein, editors, *Automated essay scoring: a cross-disciplinary perspective*, pages 87–112. Routledge, New York.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2009. Toward Automated Multi-trait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics*, 31(3):391–417.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA → Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Peter B. Nemenyi. 1963. *Distribution-free Multiple Comparisons*. Ph.D. thesis, Princeton University.

OpenAI. 2023. GPT-4 Technical Report.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Ellis B. Page. 1968. The use of the computer in analyzing student essays. *International Review of Education*, 14(2):210–225.

James E. Purpura. 2004. *Assessing Grammar*. Cambridge Language Assessment. Cambridge University Press.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring.

Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Mark D. Shermis and Jill Burstein. 2003. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Routledge, New York.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. SALMONN: Towards Generic Hearing Abilities for Large Language Models.

Nic Underhill. 1987. *Testing spoken language: A handbook of oral testing techniques*. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge Language Assessment. Cambridge University Press.

Edward M. White. 1984. Holisticism. *College Composition and Communication*, 35(4):400–409.

Xiaoming Xi. 2007. Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing*, 24(2):251–286.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

# A Appendix A

## LINGUISTIC COMPETENCE

### General linguistic range

**A1:** Has a very basic range of simple expressions about personal details and needs of a concrete type. Can use some basic structures in one-clause sentences with some omission or reduction of elements.

**A2:** Has a repertoire of basic language which enables them to deal with everyday situations with predictable content, though they will generally have to compromise the message and search for words/signs. Can produce brief, everyday expressions in order to satisfy simple needs of a concrete type (e.g. personal details, daily routines, wants and needs, requests for information). Can use basic sentence patterns and communicate with memorised phrases, groups of a few words/signs and formulae about themselves and other people, what they do, places, possessions, etc. Has a limited repertoire of short, memorised phrases covering predictable survival situations; frequent breakdowns and misunderstandings occur in non-routine situations.

**B1:** Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and film. Has enough language to get by, with sufficient vocabulary to express themselves with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel and current events, but lexical limitations cause repetition and even difficulty with formulation at times.

**B2:** Can express themselves clearly without much sign of having to restrict what they want to say. Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words/signs, using some complex sentence forms to do so.

**C1:** Can use a broad range of complex grammatical structures appropriately and with considerable flexibility. Can select an appropriate formulation from a broad range of language to express themselves clearly, without having to restrict what they want to say.

**C2:** Can exploit a comprehensive and reliable mastery of a very wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No signs of having to restrict what they want to say.

### Vocabulary range

**A1:** Has a basic vocabulary repertoire of words/signs and phrases related to particular concrete situations.

**A2:** Has sufficient vocabulary to conduct routine everyday transactions involving familiar situations and topics. Has sufficient vocabulary for the expression of basic communicative needs. Has sufficient vocabulary for coping with simple survival needs.

**B1:** Has a good range of vocabulary related to familiar topics and everyday situations. Has sufficient vocabulary to express themselves with some circumlocutions on most topics pertinent to their everyday life such as family, hobbies and interests,

work, travel and current events.
**B2:** Can understand and use the main technical terminology of their field, when discussing their area of specialisation with other specialists. Has a good range of vocabulary for matters connected to their field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. Can produce appropriate collocations of many words/signs in most contexts fairly systematically. Can understand and use much of the specialist vocabulary of their field but has problems with specialist terminology outside it.
**C1:** Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Can select from several vocabulary options in almost all situations by exploiting synonyms of even words/ signs less commonly encountered. Has a good command of common idiomatic expressions and colloquialisms; can play with words/signs fairly well. Can understand and use appropriately the range of technical vocabulary and idiomatic expressions common to their area of specialisation.
**C2:** Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.

### Grammatical accuracy

**A1:** Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.
**A2:** Uses some simple structures correctly, but still systematically makes basic mistakes; nevertheless, it is usually clear what they are trying to say.
**B1:** Communicates with reasonable accuracy in familiar contexts; generally good control, though with noticeable mother-tongue influence. Errors occur, but it is clear what they are trying to express. Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.
**B2:** Good grammatical control; occasional "slips" or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect. Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding. Has a good command of simple language structures and some complex grammatical forms, although they tend to use complex structures rigidly with some inaccuracy.
**C1:** Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.
**C2:** Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).

### Vocabulary control

**A1:** No descriptors available.
**A2:** Can control a narrow repertoire dealing with concrete, everyday needs.
**B1:** Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations. Uses a wide range of simple vocabulary appropriately when discussing familiar topics.
**B2:** Lexical accuracy is generally high, though some confusion and incorrect word/sign choice does occur without hindering communication.
**C1:** Uses less common vocabulary idiomatically and appropriately. Occasional minor slips, but no significant vocabulary errors.
**C2:** Consistently correct and appropriate use of vocabulary.

### Orthographic control

**A1:** Can copy familiar words and short phrases, e.g. simple signs or instructions, names of everyday objects, names of shops, and set phrases used regularly. Can spell their address, nationality and other personal details. Can use basic punctuation (e.g. full stops, question marks).
**A2:** Can copy short sentences on everyday subjects, e.g. directions on how to get somewhere. Can write with reasonable phonetic accuracy (but not necessarily fully standard spelling) short words that are in their oral vocabulary.
**B1:** Can produce continuous writing which is generally intelligible throughout. Spelling, punctuation and layout are accurate enough to be followed most of the time.
**B2:** Can produce clearly intelligible, continuous writing which follows standard layout and paragraphing conventions. Spelling and punctuation are reasonably accurate but may show signs of mother-tongue influence.
**C1:** Layout, paragraphing and punctuation are consistent and helpful. Spelling is accurate, apart from occasional slips of the pen.

**C2:** Writing is orthographically free of error.

## PRAGMATIC COMPETENCE

### Flexibility

**A1:** No descriptors available.
**A2:** Can adapt well-rehearsed, memorised, simple phrases to particular circumstances through limited lexical substitution. Can expand learnt phrases through simple recombinations of their elements.
**B1:** Can adapt their expression to deal with less routine, even difficult, situations. Can exploit a wide range of simple language flexibly to express much of what they want.
**B2:** Can adjust what they say and the means of expressing it to the situation and the recipient and adopt a level of formality appropriate to the circumstances. Can adjust to the changes of direction, style and emphasis normally found in conversation. Can vary formulation of what they want to say. Can reformulate an idea to emphasise or explain a point.
**C1:** Can make a positive impact on an intended audience by effectively varying style of expression and sentence length, use of advanced vocabulary and word order. Can modify their expression to express degrees of commitment or hesitation, confidence or uncertainty.
**C2:** Shows great flexibility in reformulating ideas in differing linguistic forms to give emphasis, differentiate according to the situation, interlocutor, etc. and to eliminate ambiguity.

### Thematic development

**A1:** No descriptors available.
**A2:** Can tell a story or describe something in a simple list of points. Can give an example of something in a very simple text using "like" or "for example".
**B1:** Can clearly signal chronological sequence in narrative text. Can develop an argument well enough to be followed without difficulty most of the time. Shows awareness of the conventional structure of the text type concerned when communicating their ideas. Can reasonably fluently relate a straightforward narrative or description as a sequence of points.
**B2:** Can develop an argument systematically with appropriate highlighting of significant points, and relevant supporting detail. Can present and respond to complex lines of argument convincingly. Can follow the conventional structure of the communicative task concerned when communicating their ideas. Can develop a clear description or narrative, expanding and supporting their main points with relevant supporting detail and examples. Can develop a clear argument, expanding and supporting their points of view at some length with subsidiary points and relevant examples. Can evaluate the advantages and disadvantages of various options. Can clearly signal the difference between fact and opinion.
**C1:** Can use the conventions of the type of text concerned to hold the target reader's attention and communicate complex ideas. Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion. Can write a suitable introduction and conclusion to a long, complex text. Can expand and support the main points at some length with subsidiary points, reasons and relevant examples.
**C2:** Can use the conventions of the type of text concerned with sufficient flexibility to communicate complex ideas in an effective way, holding the target reader's attention with ease and fulfilling all communicative purposes.

### Propositional precision

**A1:** Can communicate basic information about personal details and needs of a concrete type in a simple way.
**A2:** Can communicate what they want to say in a simple and direct exchange of limited information on familiar and routine matters, but in other situations they generally have to compromise the message.
**B1:** Can explain the main points in an idea or problem with reasonable precision. Can convey simple, straightforward information of immediate relevance, getting across the point they feel is most important. Can express the main point they want to make comprehensibly.
**B2:** Can pass on detailed information reliably. Can communicate the essential points even in more demanding situations, though their language lacks expressive power and idiomaticity.
**C1:** Can qualify opinions and statements precisely in relation to degrees of, for example, certainty/uncertainty, belief/doubt, likelihood, etc. Can make effective use of linguistic modality to signal the strength of a claim, an argument or a position.
**C2:** Can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of qualifying devices (e.g. adverbs expressing degree,

clauses expressing limitations). Can give emphasis, differentiate and eliminate ambiguity.

**Coherence and cohesion**

**A1:** Can link words/signs or groups of words/signs with very basic linear connectors (e.g. "and" or "then").
**A2:** Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points. Can link groups of words/signs with simple connectors (e.g. "and", "but" and "because").
**B1:** Can introduce a counter-argument in a simple discursive text (e.g. with "however"). Can link a series of shorter, discrete simple elements into a connected, linear sequence of points. Can form longer sentences and link them together using a limited number of cohesive devices, e.g. in a story. Can make simple, logical paragraph breaks in a longer text.
**B2:** Can use a variety of linking expressions efficiently to mark clearly the relationships between ideas. Can use a limited number of cohesive devices to link their utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution. Can produce text that is generally well-organised and coherent, using a range of linking expressions and cohesive devices. Can structure longer texts in clear, logical paragraphs.
**C1:** Can produce clear, smoothly flowing, well-structured language, showing controlled use of organisational patterns, connectors and cohesive devices. Can produce well-organised, coherent text, using a variety of cohesive devices and organisational patterns.
**C2:** Can create coherent and cohesive text making full and appropriate use of a variety of organisational patterns and a wide range of cohesive devices.

## B Appendix B

*I deal with consulting and sales of financial products and services to an international bank, in the mass-market and small-business. I follow the relationship with customers from acquisition to the advise until the realization of contracts, building and maintaining relationships after-sales in the aim of customer satisfaction*

*I also worked with large and small teams in back-offices, managed many administrative activities related to mortages, personal loans, contability and investments too.*

*I worked for several years to the acquisition of new customers, to provide them with a complete service, from the account to insurance products, investment products, personal loans, revolving credit, and cross-selling products. In many years of work I have honed my skills in managing non-standard situations, analyzing the problem, finding and implementing practical and easy solutions. non-standard situations, analyzing the problem, finding and implementing practical and easy solutions.*

*I have faced several situations always work with serenity and enthusiasm, I like to work in a multicultural and dynamic.*

*I'm careful to meet the goals of the team in which I work, cooperating with colleagues to achieve the goals by providing my skills, always willing to learn, respecting other points of view together finding ways to deal. I work for the same large company for 25 years, now is the time to change and find new job opportunities. Needs to work my husband has been living in Zaandam, I want to find a new job in Holland to rejoin our family.*

*I like sports such as skiing, riding and swimming. I've also got the rescue licence, I worked as a lifeguard in the summer studying for the patent padi dive master*

The holistic score is 3.5 (B1+), and GPT-4 provided these analytic scores:

- general linguistic range: 3
- vocabulary range: 4
- grammatical accuracy: 2
- vocabulary control: 3
- orthographic control: 4
- flexibility: 2
- thematic development: 2
- coherence and cohesion: 1
- propositional precision: 3

## C Appendix C

When we include the holistic score, the prompt given to GPT-4 is the following:

```
Consider the following essay:
[ESSAY]
```

It has been given this score on a scale from 1 to 6.5: [HOLISTIC SCORE].

I want you to assess it only considering the aspect of [ASPECT], for which you have 6 different feedback options, that you will have to accept or reject: [ANALYTIC CEFR DESCRIPTORS]

ONLY ONE option can be accepted and is the option you will have to output by only selecting the option letter in the following format: 'option A/B/C/D/E/F'[10] WITHOUT ANY ADDITIONAL OBSERVATION, COMMENT, NOTE, EXPLANATION, CLARIFICATION, OR JUSTIFICATION OF ANY SORT.

Your answer:

When we do not provide GPT-4 with the holistic score, the prompt is the following:

Consider the following essay: [ESSAY]

I want you to assess it only considering the aspect of [ASPECT], for which you have 6 different feedback options, that you will have to accept or reject: [ANALYTIC CEFR DESCRIPTORS]

ONLY ONE option can be accepted and is the option you will have to output by only selecting the option letter in the following format: 'option A/B/C/D/E/F'[11] WITHOUT ANY ADDITIONAL OBSERVATION, COMMENT, NOTE, EXPLANATION, CLARIFICATION, OR JUSTIFICATION OF ANY SORT.

Your answer:

# D Appendix D

### Score alignment

Table 5 shows the holistic score normalisation process for EFCAMDAT.

[10]The aspects of vocabulary control, flexibility, and thematic development only have options A-E since no descriptors are available for the A1 level.

[11]See note 10.

| CEFR | W&I | EFCAMDAT |
|---|---|---|
| A1 | A1 (1) | 1,2 |
| | A1+ (1.5) | 3 |
| A2 | A2 (2) | 4,5 |
| | A2+ (2.5) | 6 |
| B1 | B1 (3) | 7,8 |
| | B1+ (3.5) | 9 |
| B2 | B2 (4) | 10,11 |
| | B2+ (4.5) | 12 |
| C1 | C1 (5) | 13,14 |
| | C1+ (5.5) | 15 |
| C2 | C2 (6) | 16 (score$<$85) |
| | C2+ (6.5) | 16 (score$\geq$85) |

Table 5: Score alignment.

### Additional experimental results

Table 6 reports the results of the experiment conducted when no holistic scores are given to GPT-4.

| | score | %gram. | #dif.wds. | #unq.wds. | %l.d.t. | #unq.n.cks. | #unq.q.m.a. | fl.-kinc. | w2v | av.s.ln. | holistic |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lng.** | gen. lin. | 0.643 | 0.622 | 0.547 | 0.471 | 0.526 | 0.565 | 0.268 | 0.275 | 0.148 | 0.739 |
| | gramm. | **0.707** | 0.408 | 0.365 | 0.284 | 0.324 | 0.364 | 0.151 | 0.170 | 0.099 | 0.692 |
| | orth. | **0.730** | 0.362 | 0.290 | 0.234 | 0.259 | 0.309 | 0.133 | 0.166 | 0.068 | 0.653 |
| | voc. ctrl. | 0.697 | 0.391 | 0.363 | 0.305 | 0.331 | 0.369 | 0.153 | 0.102 | 0.107 | 0.654 |
| | voc. rg. | 0.529 | 0.539 | 0.456 | **0.410** | 0.450 | 0.452 | **0.247** | 0.241 | 0.131 | 0.616 |
| **Prg.** | propos. | 0.432 | 0.510 | 0.442 | 0.341 | 0.430 | 0.492 | **0.246** | **0.304** | 0.145 | 0.539 |
| | coh. | 0.602 | **0.601** | **0.542** | 0.379 | **0.533** | **0.571** | 0.244 | 0.299 | **0.162** | 0.729 |
| | flexib. | 0.307 | 0.361 | 0.363 | 0.282 | 0.348 | 0.346 | 0.202 | 0.149 | **0.160** | 0.330 |
| | themat. | 0.425 | **0.612** | **0.587** | **0.496** | **0.576** | **0.583** | 0.242 | **0.333** | 0.150 | 0.543 |
| | holistic | 0.732 | 0.640 | 0.665 | 0.451 | 0.623 | 0.637 | 0.178 | 0.364 | 0.141 | 1.000 |

Table 6: SRC correlation of the GPT-4 predicted scores and relevant linguistic features (**without giving GPT-4 the holistic score**). The *holistic* entry refers to the ground-truth holistic scores. In bold the two highest correlations columnwise.