KEYSCORE — Keystroke-enhanced Automated Essay Scoring

Nils-Jonathan Schaller[1], Daniel Mora[1],
Thorben Jansen[1], Olaf Köller[1], Andrea Horbach[1,2]
[1]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2]Kiel University, Kiel, Germany
schaller@leibniz-ipn.de

Abstract

We investigate the predictive power of keystroke logging data for automated essay scoring using the newly collected PISA FLA writing process dataset. Based on 3,882 writing sessions, we extract a comprehensive set of keystroke-based process features, including temporal measures, pause and burst patterns, deletion behavior, production efficiency, and navigation activity and evaluate their ability to predict holistic essay scores on a 0–5 scale. We specifically compare process-feature-based models with content-based scoring approaches trained on data written with and without the help of an AI chatbot, and investigate how predictive power evolves over the course of a writing session by training models at multiple time thresholds. Our analysis reveals that keystroke features provide genuine early predictive signal, capturing aspects of writing fluency and revision behavior that distinguish writers before their texts are long enough to score conventionally. Additionally, our results suggest that process-based scoring is a viable complement to product-based approaches, with promise for formative, real-time feedback during writing.

1 Introduction

Feedback is most effective when it is timely and targets the writing process rather than only the finished product (Hattie and Timperley, 2007; Graham et al., 2015; Shute, 2008). However, much of the traditional automated assessment literature focuses on evaluating completed texts and their summative outcomes rather than providing in-process feedback (Cotos, 2015).

One way to achieve such feedback is by constant monitoring of the writing process, which is often hindered by a lack of suitable writing process data. To address this research gap, we present a new dataset PISA FLA, consisting of almost 4000 English-as-a-foreign-language essay writing sessions from German participants in the 2025 PISA data collection. In this dataset, students responded to two individual argumentative writing prompts and produced texts under two conditions: with and without the assistance of a generative AI chatbot (hereafter referred to as *Chat* and *NoChat*, respectively). This dataset thus offers ample opportunities to investigate both how students write and how essays can be automatically scored already during the writing process.

We investigate the latter question in three different setups: By means of traditional transformer-based essay scoring models, through keylog-based process features and by combining the two approaches. We further address the question of how early on (within the 30 minutes of a single writing session) reliable summative feedback can be given, i.e. feedback with an accuracy similar to that of a model trained on full texts. We achieve this by comparing performance at certain points in time of the writing session. Finally, we investigate the influence of chatbot support on automated scoring.

In doing so, our paper makes the following contributions:

- We present a new dataset of EFL essays together with detailed keylog information making the writing process transparent.
- We show that reliable automated feedback does not require a complete text, but that there is a tradeoff between essay scoring performance and writing time.
- We show that integrating process features into a transformer-based model improves prediction at early writing stages.

The data is available by request until September 2027 after which it will be openly published: https://github.com/darius-ipn/keystroke_pisa2025_fla.

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 347–357

After 5 minutes

Disagree:
– more important that the students learn sth about the subject
– the teacher doesnt have to be in a good relationship with the students, the teacher only has to be polite
– the students dont learn anything if the teacher is polite

After 10 minutes

Disagree:
– more important that the students learn sth about the subject
– the teacher doesnt have to be in a good relationship with the students, the teacher only has to be polite
– the students dont learn anything if the teacher is polite but doesnt know sth about the topic

Argumentation:
Today we are talking about the question if a teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught.

Figure 1: Writing snapshots from a single participant. After 5 minutes, the student has only produced planning notes in bullet-point form. By 10 minutes, an introductory sentence has been added, illustrating the transition from planning to text production.

2 Related Work

Our work builds on two lines of research: keystroke logging as a window into writing processes, and automated essay scoring. We review relevant work in each area before describing how our approach connects them.

2.1 Writing Process Research and Keystroke Logging

Writing research has long established that composing involves planning, producing text, and revising, and that writers switch between these processes rather than completing them in a fixed order (Flower and Hayes, 1981). Later models distinguish between the cognitive effort of formulating ideas and typing them out (Hayes, 2012). Keystroke logging offers a way to observe these processes (Leijten and Van Waes, 2013).

Recent work has demonstrated that providing students with keystroke-based feedback on their writing process can improve text quality (Vandermeulen et al., 2020). Allen et al. (2016) showed that keystroke indices, particularly total number of keystrokes, explain a large proportion ($R^2 = .76$) of variance in essay scores for L1 English undergraduates writing in an intelligent tutoring system. However, Conijn et al. (2022) investigated which keystroke features are useful for predicting writing quality during the writing process but found the relationship to be limited. They did show that feature importance shifts over the course of writing, an observation our work further addresses at larger scale and with complementary text-based models.

Several keystroke corpora with quality ratings have recently become available, including KLiCKe by Tian et al. (2025) with ∼5,000 argumentative essays from L1/L2 adults and KUPA-KEYS by Velentzas et al. (2024) with 1,006 EFL essays graded on the CEFR scale. Our dataset complements these resources by focusing on adolescent EFL learners in a standardised assessment context and by including an AI-assisted writing condition.

2.2 Automatic Essay Scoring

Recent work on automated essay scoring has moved toward transformer-based architectures (Li and Ng, 2024). DeBERTa has shown strong performance in this context, outperforming BERT and RoBERTa on the ASAP dataset (Susanto et al., 2024) and serving as the model of choice in top entries of the Kaggle Automated Essay Scoring 2.0 competition, where winning solutions used ensembles of DeBERTa-v3-large with two-staged training (Learning Agency Lab, 2024). Based on these results, we adopt DeBERTa-v3-base as the text encoder in our experiments.

However, many transformer-based AES systems rely solely on the learned text representation for scoring. Uto et al. (2020) demonstrated that concatenating essay-level features (e.g., readability indices, syntactic counts) with a transformer's `[CLS]` representation before the scoring layer yields consistent gains, with minimal additional parameters. Building on this, Lohmann et al. (2025) systematically compared feature-based, embedding-based, and hybrid models for analytic trait scoring on argumentative essays from both L1 and L2 learners, including essays from the MEWS corpus (Keller et al., 2020) that shares our writing prompts. Their hybrid architecture, combining 220 linguistic features with DistilBERT embeddings, consistently outperformed the feature-only and embedding-only models, with the largest gains for content and organization assessment.

Additionally, they showed that linguistic features and contextual embeddings capture partially differ-

ent aspects of text quality, supporting architectures that combine both sources. More broadly, Li and Ng (2024) argue that AES research should move beyond optimizing scores on standard benchmarks toward new task settings and formative applications — a direction our work pursues by evaluating scoring across writing time thresholds.

Our work extends this line of research in two directions. To our knowledge, no prior work has integrated keystroke process features directly into a transformer's prediction architecture.

First, we replace the linguistic feature vector with keystroke process features, testing whether signals from the writing process can complement a transformer's text representation in the same way that linguistic features do. Second, we evaluate this combined model not only on completed essays but across multiple time thresholds during writing, addressing the question of how early in the writing process reliable scoring becomes feasible.

3 Data

In the following we present the PISA 2025 FLA dataset used in our experiments. The dataset was collected in the 2025 PISA assessment as an additional task for foreign language assessment in German High schools on students at the age of 15. It comprises initially 4,091 writing sessions, in which students responded to argumentative writing prompts in English as a foreign language. Students were assigned to one out of two groups: Those that had access to an AI chatbot for writing support and those without. The chatbot could be used unconditional for all kinds of prompts. The underlying model was gpt-4o-mini. After data cleaning (see Section 3.1) we are left with 3,882 essays, which are distributed roughly evenly across the two conditions, as displayed in Table 1. The two conditions differ substantially in writing outcomes. Students with chat access produced considerably longer texts (1,813 vs. 1,138 characters on average; 274 vs. 200 words) and received higher holistic scores (mean 3.2 vs. 2.1 on a 0–5 scale), with 20.2% of chat-assisted essays reaching the maximum score of 5 compared to just 2.4% without chat. Notably, chat-assisted students achieved these longer, higher-scored texts with fewer keystrokes (1,666 vs. 1,951), suggesting substantial use of copy-pasting from the chatbot. The writing was limited to 30 minutes for all conditions although the majority stopped at approximately 25 minutes.

	NoChat	*Chat*
N (scored)	1,911	1,971
Avg. text length (chars)	1,138	1,813
Avg. text length (words)	200	274
Avg. sentences	8.9	14.0
Score (mean)	2.1	3.2
Score (median)	2.2	3.2
% scoring 5	2.4	20.2
Avg. writing time (min)	24.7	25.2
Avg. active writing ratio	0.5	0.4
Avg. keystrokes	1,951	1,666

Table 1: Descriptive statistics by condition. Scores are on a 0–5 holistic scale. Active writing ratio is 1 minus the proportion of break time (pauses ≥ 2 s) relative to total session time.

The active writing ratio was slightly higher for *NoChat* (0.5 vs. 0.4), consistent with *Chat* students spending time interacting with the chatbot interface rather than typing continuously.

3.1 Data Cleanup

To ensure data quality, we applied two filtering criteria: (1) submissions with fewer than 10 characters were excluded as empty or trivial responses, and (2) submissions exceeding 10,000 characters (above the 99th percentile) were excluded as they reflected excessive copy-pasting behaviour rather than authentic writing processes.

Of the initial N = 4,091 submissions, 130 were excluded for insufficient length and 35 for excessive length, resulting in a sample of N = 3,926 (96.0% retention). Of these, 3,882 had valid holistic scores and are the final analytical sample.

3.2 Keystroke Measures

Keystrokes and texts were collected within a self-developed web interface. From each writing session we extract 25 process features covering six behavioural dimensions. **Temporal measures** capture total writing time (first to last keystroke) and initial pause (time from task start to first keystroke). **Pause and break patterns** characterize interruptions in typing: we count all pauses $\geq 2{,}000$ ms as breaks and derive their total duration, mean duration, and the ratio of break time to total writing time. **Burst characteristics** describe the production sequences between breaks or revisions, including burst count, mean burst length (in characters), and mean burst duration. **Deletion behaviour** is captured through the total number of delete/backspace keystrokes, the deletion ratio (deletions / total keystrokes) and total characters deleted. **Produc-**

tion efficiency includes final text length (characters and words), characters per minute (relative to active writing time), total keystrokes, and the process-to-product ratio (total characters typed / final text length), which indicates how much text was produced and subsequently removed. Finally, **navigation activity** is measured through copy-paste count, a linearity index reflecting the proportion of forward cursor transitions (1 = fully sequential writing), and area switches between interface regions. For sessions in the *Chat* condition, we additionally capture time allocation across four interface areas (text editor, task text, chat window, chat prompt) and the number of area switches; these features are set to zero for NoChat. All features are z-score normalised within each time threshold (see Section 3.4), so that distributions reflect the writing stage rather than the completed essay. See also Appendix A, Table 7.

The features were selected to cover established behavioral dimensions in keystroke logging research such as temporal, pause, burst, deletion, production, and navigation measures. Note that the full dataset also tracks mouse movements and other non-keyboard events, which we partially integrate such as cursor reposition events, paste events, interface area switches, and time in chat/task/editor/no area.

3.3 Dataset Annotation

In the following we describe the automatic and manual annotation efforts to obtain the holistic scores used in our scoring experiments.

3.3.1 Creating a silver standard

For our current study, we were not able to obtain human annotations for the full dataset. As a proxy for a human gold standard, we employ an automated scoring model which has been trained on essays from the MEWS dataset, i.e., essays answering the same writing task but from a slightly different learner population.

Specifically, a linear regression model was trained on the MEWS dataset (Keller et al., 2020, 2024) to predict both holistic scores and individual trait scores. The model uses 220 linguistic features extracted in the same way as in Lohmann et al. (2025). These features capture various aspects of the text, including lexical variety, syntactic structure, cohesion, as well as error-, frequency- and length-based indicators. Before training, all features are scaled using a min-max scaler, and multicollinearity is reduced by removing highly correlated features (correlation > 0.9).

This silver standard is only used as training material. All models in this study are evaluated against the smaller human-annotated test dataset described in the following, ensuring that reported results reflect agreement with human judgments.

3.3.2 Gold Standard Scores

To obtain a gold standard test set, we drew a stratified random sample from the corpus. Strata were defined by domain (Teacher vs. Advertisement), experimental condition (*Chat*/*NoChat*), and a quality profile constructed from tertile bins (Low / Medium / High) of silver-standard analytic scores for Content, Organization, and Language Quality, yielding up to 27 performance clusters per domain–condition cell. Allocation was proportional to population frequency with a minimum-count floor to retain rare profiles.

The final sample comprises 429 essays (212 *NoChat*, 217 *Chat*), each double-annotated by trained raters (Krippendorff's $\alpha = .79$).

Four student annotators with a background in education were trained on the original MEWS scoring rubrics. Each essay was randomly assigned and annotated twice. Inter-annotator agreement reached a Krippendorff's α of 0.79 (interval metric), indicating substantial agreement. Final gold scores were obtained by averaging.

These annotations were also used to validate the silver standard: We evaluated the automatic scoring model (trained on MEWS data) against the gold annotations and obtained a QWK of 0.76 between the automatically annotated silver standard and the human gold standard. This is comparable to the inter-annotator agreement, supporting the use of the silver labels as training signal.

The gold-annotated subset serves as the held-out test set in all experiments.

3.4 Time Threshold Splits on Test Set

One core research question we aim to answer is: How early in the writing process can we give meaningful feedback to learners? To this end, we reconstruct from the keylog data for each essay its state after 5, 10, 15, 20 and 25 minutes of writing time, reflecting planning notes, partial drafts, or incomplete revisions rather than a finished product (see Figure 1 for an illustration). A small number of writing sessions (1–2 per threshold) with no keystroke activity up to that time, i.e. still empty

texts, were excluded. Note that our goal is to judge the assumed quality of the finished essay early in the writing process. Therefore we do not collect new human judgements for these incomplete texts, but propagate the label from the full essay to all earlier stages of that essay.

4 Method

In the following, we describe first the different model infrastructures for scoring, then detail the shared experimental setup covering data splits, time thresholds, and evaluation.

4.1 Scoring Models

We approach in-process essay scoring through three complementary model families: (1) process-feature models that predict scores exclusively from keystroke behaviour, without access to the essay text; (2) transformer-based models that rely on the essay text; and (3) an architecture that combines both information sources. All models are trained and evaluated separately for *Chat* and *NoChat* and at six time thresholds throughout the 30-minute writing session, allowing us to trace how prediction quality evolves over the course of composition.

4.1.1 Process-Feature Baselines

To evaluate the predictive power of keystroke features independently of essay content, we train gradient-boosted regression trees (XGBoost) that operate exclusively on the process features described in Section 3.2. Each model is trained in one of three feature configurations: (1) *all* keystroke features, (2) *length only*, using only the text length in characters at the respective time threshold as a single predictor, and (3) *no length*, using all keystroke features except text length. This decomposition disentangles the contribution of text length, a known dominant predictor in automated essay scoring (Allen et al., 2016), from the behavioural signal captured by the remaining process features such as pause patterns, burst characteristics, and deletion behaviour. Hyperparameters are selected via grid search over the number of estimators (100, 200, 500), maximum tree depth (3, 5, 7), learning rate (0.01, 0.05, 0.1), and L1 regularisation strength (0, 0.1, 1.0), with subsampling rate and column sampling fixed at 0.8. The combination maximising QWK on the dev set is selected. We also trained ordinary least squares regression models in the same configurations; results are reported in Appendix A Table 6.

4.1.2 Text-Based Scoring with DeBERTa

We fine-tune `DeBERTa-v3-base` (He et al., 2021) with a single-output regression head predicting the holistic score on a continuous 0–5 scale. Input texts are tokenised with the model's default tokeniser. Due to hardware restrictions it was necessary to truncate essays to a maximum of 512 tokens. In the training set, this led to truncation in 8.0% of texts, with truncated samples losing on average 24% of their tokens. At intermediate timepoints, truncation was less frequent, as texts were naturally shorter.

Training uses AdamW with a learning rate of 2×10^{-5} and a batch size of 8 for up to 10 epochs with early stopping.

We evaluate two training variants:

Full-text model. The model is trained exclusively on complete essays. At inference, it receives either the full essay or an intermediate text snapshot at a given time threshold. This reveals how well a model trained on finished writing generalises to incomplete drafts.

Hybrid model. The model is trained on the hybrid training set described in Section 4.2.1, which combines complete essays with intermediate snapshots, exposing the model to both complete and in-progress texts.

4.1.3 Combined Text and Process Features

Following Uto et al. (2020), who concatenated handcrafted linguistic features with a transformer's text representation, we inject keystroke process features into the DeBERTa prediction head. We concatenate DeBERTa's pooled text representation with a keystroke feature vector and train a joint regression head on both information sources simultaneously. Concretely, we obtain the pooled CLS representation from DeBERTa-v3-base, a 768-dimensional vector produced by the context pooler, which applies a learned linear projection and activation on top of the `[CLS]` token. This vector is concatenated with a 32-dimensional side vector composed of two parts: (1) the 25 z-scored keystroke process features described in Section 3.2, and (2) a 6-dimensional one-hot encoding of the source timepoint (*5 min*, *10 min*, *15 min*, *20 min*, *25 min*, *full*). The timepoint indicator is necessary because the keystroke features are z-score normalised *within* each time threshold. Without this signal, the model cannot distinguish between normalisation groups. Because z-score normalisation on small per-timepoint groups can produce

	NoChat			*Chat*		
Score	Train	Dev	Test	Train	Dev	Test
0.0	134	15	16	60	7	5
0.5	82	9	18	34	4	5
1.0	122	13	18	47	5	10
1.5	205	23	25	94	10	12
2.0	241	27	32	142	16	20
2.5	289	32	36	212	24	18
3.0	217	24	29	218	24	39
3.5	119	13	16	179	20	23
4.0	69	8	10	153	17	23
4.5	17	2	4	119	13	19
5.0	34	4	8	320	36	43
N	1,529	170	212	1,578	176	217

Table 2: Distribution of holistic scores (binned to nearest 0.5) across train/dev/test splits for both conditions.

extreme values, we clip all keystroke z-scores to the range $[-5, +5]$ before training. This prevents rare outliers from destabilising gradient updates, an issue we observed during initial experiments. The combined 800-dimensional vector ($768 + 32$) is passed through a two-layer feed-forward network ($800 \rightarrow 128 \rightarrow 1$) with GeLU activation and dropout (p=0.1), producing a continuous regression score. For ordinal metrics (QWK, macro F_1), predictions are clipped to $[0, 5]$ and rounded to the nearest 0.5, yielding 11 categories. The entire model is trained end-to-end with the same hyperparameters as the text-only DeBERTa model.

4.2 Experimental Setup

4.2.1 Data Splits

We use the 429 gold-annotated essays described in Section 3.3.2 as our fixed test set. The remaining data is split at a 90/10 ratio using stratified sampling on holistic scores binned to the nearest 0.5. Rare score bins with fewer than 2 instances are merged into their nearest neighbour before stratification. Table 2 shows the resulting distributions. Scores in the *NoChat* condition concentrate around 2.0–3.0, whereas the *Chat* condition is right-skewed, with the 5.0 bin alone accounting for roughly 20 % of sessions.

Hybrid training set. Following Schaller et al. (2025), we construct a hybrid training set by pairing each full essay with one randomly sampled intermediate snapshot (drawn uniformly from the five time thresholds), approximately doubling the training set from 3,107 to 6,199 instances (see Table 3). This exposes the model to both complete and in-progress texts, reducing its reliance on structural cues present only in finished essays.

4.2.2 Summary

All models are trained and evaluated at six time thresholds (5, 10, 15, 20, 25 minutes, and full essay). For the process-feature models, we train one XGBoost model per combination of condition (*Chat*, *NoChat*), feature configuration (*all*, *length only*, *no length*), and time threshold, yielding 36 combinations. For the text-based models, we evaluate three DeBERTa-v3-base variants: *full-text*, *hybrid* (Section 4.2.1), and *combined*. All predictions are clipped to [0,5] and rounded to the nearest 0.5 for QWK computation.

	NoChat	*Chat*	**Total**
Full texts	1,529	1,578	3,107
5 min	306 (20.0%)	329 (21.0%)	635
10 min	266 (17.4%)	315 (20.2%)	581
15 min	330 (21.6%)	310 (19.8%)	640
20 min	303 (19.8%)	294 (18.8%)	597
25 min	324 (21.2%)	315 (20.2%)	639
Hybrid total	**3,058**	**3,141**	**6,199**

Table 3: Distribution of training samples in the Hybrid training set. Each essay appears once as its full text and once as a randomly sampled intermediate snapshot.

5 Results

In the following, we first present results on the process-feature-based models, then the text-based transformers and finally the combined architecture.

5.1 Process-Feature-Based Scoring

Table 4 presents results for the three keylog-based feature configurations: all keystroke features (ALL), text length only (LEN), and all features except length (NOLEN), separately for the two experimental conditions (*NoChat*, *Chat*)

Unsurprisingly, prediction quality improves steadily over the course of the writing session, rising from QWK .32 at 5 minutes to .76 at completion for *NoChat* (all features), and from .59 to .91 for *Chat*. The steepest gains occur between 5 and 15 minutes, after which performance improvements level off, suggesting that the first half of the writing session already captures most of the predictive signal. The relative contribution of text length and behavioural process features shifts markedly over time. At 5 minutes, LEN alone is a weak predictor (QWK of .14 and .31), while NOLEN already achieve substantially higher performance (.38 and

.55). This pattern reverses later: by completion, length alone reaches .77 (*NoChat*) and .92 (*Chat*), matching the full feature set. This confirms the well-established role of text length as a dominant predictor in automated essay scoring (Allen et al., 2016) while demonstrating that its dominance only holds once sufficient text has been produced.

Notably, the NOLEN configuration performs competitively with ALL across both conditions, matching it at most time thresholds, with small fluctuations in either direction that likely reflect estimation noise given the test set size. This suggests that explicit text length is largely redundant given the remaining process features.

Model performance was consistently higher in the *Chat* condition, with QWK reaching .91 at completion compared to .76 *NoChat*. This likely reflects two compounding factors: *Chat*-assisted students produced longer, more content-rich texts earlier in the session, providing a stronger signal for scoring models, and the right-skewed score distribution — with 20.2% of essays at the maximum — may additionally increase class separability across all models. A more fine-grained analysis of individual AI usage strategies, including copy-paste behaviour and prompt intent, is left for future work. These results suggest that keystroke process features provide genuine early predictive signal, capturing aspects of typing fluency, planning behaviour, and revision activity that distinguish writers before their texts are long enough to score conventionally. Linear regression models trained in the same configurations showed comparable patterns at lower overall performance (see Appendix 6), suggesting that these findings are robust to model choice.

To further investigate which features drive prediction, Appendix A Table 8 reports Pearson correlations of each feature with text length and holistic score, separately by condition. Text length is the strongest predictor of score ($r = .73$ in both conditions). That the NOLEN XGBoost configuration matches or exceeds ALL at most time thresholds (Table 4) indicates, that the combination of features can match or exceed the length signal as a predictor of score. The correlation differs between conditions: `total_keystrokes` and `break_count` are positive predictors in *NoChat* but show near-zero or negative correlations in *Chat* ($r = -.01$ and $-.14$), likely because higher keystroke counts in Chat reflect chatbot interaction rather than productive writing. `initial_pause` is the most condition-stable predictor ($r_{\text{score}} = -.19$ NoChat, $-.16$ Chat), indicating that students who begin writing sooner tend to score higher regardless of final text length.

	NoChat			*Chat*		
	All	Len.	NoLen.	All	Len.	NoLen.
5 min	.32	.14	**.38**	**.59**	.31	.55
10 min	.52	.35	**.58**	**.71**	.55	**.71**
15 min	.57	.59	**.62**	**.79**	.74	.78
20 min	.65	.64	**.68**	**.88**	.86	.86
25 min	**.76**	.74	**.76**	**.91**	**.91**	.90
FullTx	.76	**.77**	.75	.91	**.92**	**.91**

Table 4: QWK for XGBoost models at each time threshold. ALL = all keystroke features; LEN = text length only; NOLEN = all features except length. Best per condition and threshold in bold.

5.2 Text-Based Scoring

Table 5 reports QWK for the two DeBERTa models (FULL and HYBRID) evaluated at each time threshold and on complete essays, separately for *NoChat* and *Chat*.

	n	Full	Hyb.	Δ_H	Combi.	Δ_C
NoChat						
5 min	212	.15	.33	+.19	**.43**	+.28
10 min	212	.42	.53	+.11	**.63**	+.21
15 min	212	.65	.63	−.03	**.69**	+.04
20 min	212	**.77**	.67	−.10	.72	−.05
25 min	212	**.86**	.74	−.13	.77	−.09
FullTx	212	**.86**	.75	−.12	.77	−.10
All	1272	.58	.62	+.04	**.67**	+.09
Chat						
5 min	217	.21	.40	+.19	**.58**	+.37
10 min	217	.41	.60	+.19	**.73**	+.32
15 min	217	.62	.73	+.11	**.83**	+.21
20 min	217	.84	.86	+.02	**.90**	+.05
25 min	217	**.95**	.90	−.05	.94	−.01
FullTx	217	**.95**	.90	−.05	.94	−.01
All	1302	.57	.73	+.16	**.82**	+.26
Chat+NoChat						
5 min	429	.22	.43	+.21	**.58**	+.36
10 min	429	.45	.62	+.17	**.73**	+.28
15 min	429	.67	.73	+.06	**.81**	+.13
20 min	429	.84	.81	−.03	**.85**	+.01
25 min	429	**.92**	.85	−.07	.88	−.04
FullTx	429	**.93**	.86	−.07	.88	−.04
All	2574	.62	.73	+.11	**.79**	+.18

Table 5: QWK by time threshold for FULL, HYBRID, and COMBI. DeBERTa models. Δ_H and Δ_C denote differences to FULL.

Full-text model. When trained exclusively on complete essays, DeBERTa achieves strong performance at full length (QWK of .86 and .95 for

the *NoChat* and *Chat* conditions, respectively), outperforming the best XGBoost keystroke-feature models (.77 and .92; Table 4). However, performance degrades sharply on incomplete drafts: at 5 minutes, QWK drops to .15 (*NoChat*) and .21 (*Chat*). This is expected: in the training data, a short, structurally incomplete text is often a weak finished essay, whereas in the test data it may be an early snapshot of an ultimately high-scoring one.

Hybrid model. Hybrid training yields clear improvements at early time thresholds: QWK rises by +.19 at 5 minutes in both conditions, and gains persist through 10 minutes (+.11 / +.19) and, in the *Chat* condition, through 15 minutes (+.11). This is consistent with Schaller et al. (2025): training on a mix of complete and incomplete texts helps the model generalise to intermediate writing stages.

However, the hybrid model incurs a cost at later stages: from 15–20 minutes onward, QWK falls below the full-text model, with the largest gap at 25 minutes in the *NoChat* condition (−.13). The full-essay ceiling is also lower (.75 vs. .86 *NoChat*; .90 vs. .95 *Chat*), indicating that the mixed training signal dilutes the model's ability to fully exploit the richer information available in complete texts.

Condition differences. The crossover point, where the hybrid model transitions from outperforming to underperforming the full-text model, differs between conditions. For *NoChat*, the hybrid model falls behind already at 15 minutes; for *Chat*, it remains competitive through 20 minutes and only drops at 25 minutes. This asymmetry likely reflects the role of AI chat scaffolding: chat-assisted students produce more structured, content-rich text earlier in the session, providing the hybrid model with more "scorable" material at intermediate time points. This is reinforced by consistent larger hybrid gains at early thresholds in the *Chat* condition.

Comparison with keystroke-feature models. Compared to the XGBoost keystroke-feature models, the DeBERTa models achieve higher QWK once sufficient text is available: at full length, DeBERTa (full) reaches .86 / .95 vs. .76 / .91 for the best XGBoost configuration. At early stages the picture reverses: the XGBoost all-features model achieves QWK of .59 at 5 minutes in the *Chat* condition, outperforming both DeBERTa (full) at .21 and DeBERTa (hybrid) at .40. In the *NoChat* condition, it outperforms the full-text model (.32 vs .15) but is on par with the hybrid model (.32 vs .33). This confirms that keystroke and text features capture complementary signals: process features provide more robust early prediction when text is minimal, while text-based models dominate once enough content has been produced.

5.3 Combined Text and Process Feature Scoring

Table 5 extends the comparison with COMBINED, which concatenates DeBERTa text embeddings, keystroke features, and a timepoint one-hot encoding (see also Figure 2).

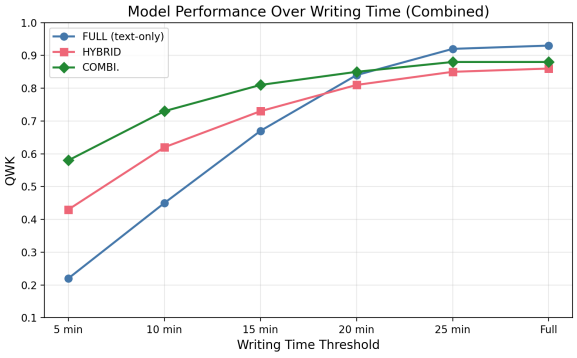

Figure 2: Performance of all 3 DeBERTa models.

Performance across writing stages. At early time thresholds, where essay text is still sparse, COMBINED consistently outperforms both FULL and HYBRID. At 5 minutes, the gains over FULL are substantial (+.28 *NoChat*, +.37 *Chat*), confirming that keystroke process features compensate for what the transformer cannot yet extract from minimal text. Even compared to HYBRID, COMBINED adds +.10 to +.18 QWK at the 5-minute mark. As essays grow longer, the advantage narrows: by 20 minutes in the *NoChat* condition, FULL overtakes COMBINED, and at essay completion, FULL achieves the highest QWK in *NoChat* (.86 vs .77), while in *Chat* the two models perform comparably (.95 vs .94). This mirrors the HYBRID trade-off: process-enriched training slightly reduces discriminative power on finished texts.

Aggregate performance. Across all time thresholds, COMBINED achieves the highest QWK in every condition (.67 *NoChat*, .82 *Chat*, .79 combined), outperforming HYBRID (.62/.73/.73) and FULL (.58/.57/.62). This is most relevant for practical applications where a single model serves writers at any stage of composition.

Condition asymmetry. The advantage of COMBINED over FULL is consistently larger in the *Chat* condition (+.26 vs. +.09 in aggregate). This likely reflects both the richer keystroke features available in the *Chat* condition (e.g., interface switches, time allocation across areas) and the right-skewed score distribution, which may amplify the benefit of supplementary process information for distinguishing among high-scoring essays.

6 Conclusion

We investigated whether keystroke logging data can contribute to automated essay scoring beyond its established role in describing writing processes. Additionally, we present the first scoring experiments on the PISA 2025 FLA dataset, using nearly 4,000 EFL writing sessions to compare process-based, text-based, and combined approaches across writing time thresholds, demonstrating its utility for writing process research.

Our results show that process features carry genuine predictive signal for essay quality. At early writing stages, when texts are too short for conventional scoring, process features distinguish writers at performance levels that text-based models cannot yet reach. At 5 minutes, process-feature models achieve QWK scores of .32 (*NoChat*) and .59 (Chat), substantially outperforming DeBERTa trained on full texts (.15 and .21). This advantage diminishes as essays grow longer: at completion, text-based scoring clearly dominates (.86/.95 vs. .76/.91), showing that process features complement rather than replace content-based assessment.

The combined text-and-process model offers the best trade-off across the writing timeline. In the aggregate evaluation across all time thresholds, it outperforms both the text-only and hybrid models, achieving the highest QWK in every condition. This extends reliable prediction: in *Chat*, the combined model reaches QWK above .70 at 10 minutes, versus 20 minutes for the text-only model.

Across all models and time thresholds, prediction quality was consistently higher in the *Chat* condition, likely because chat-assisted students produced more structured, content-rich text earlier in the session, improving scoring at intermediate stages. In addition, the right-skewed score distribution in this condition may also contribute to the observed differences.

While our results demonstrate that process features improve predictive accuracy at intermediate writing stages, whether these predictions translate into effective formative feedback remains an open question for future intervention studies. Further work should investigate which process features drive prediction at different writing stages and evaluate the approach on human-annotated training data.

Limitations

Our models are trained on silver-standard labels from a linear regression model rather than human annotations, although all evaluations are against human-annotated gold scores. The consistent performance advantage in the *Chat* condition may partly reflect its right-skewed score distribution rather than richer behavioural signal alone. Our dataset comprises German high school students writing argumentative EFL essays in a standardised assessment context; generalisability to other populations, tasks, and settings remains to be established. Finally, we demonstrate predictive accuracy at intermediate time points, not the pedagogical effectiveness of delivering such predictions as formative feedback in practice.

Acknowledgments

We used large language model assistants for code development support and language editing. All experimental design, analysis, and scientific interpretation are our own.

References

Laura K. Allen, Matthew E. Jacovina, Mihai Dascalu, Rod D. Roscoe, Kevin M. Kent, Aaron D. Likens, and Danielle S. McNamara. 2016. {ENTER}ing the time series {space}: Uncovering the writing process through keystroke analyses. In *Proceedings of the 9th International Conference on Educational Data Mining 2016*, pages 22–29.

Rianne Conijn, Christine Cook, Menno van Zaanen, and Luuk Van Waes. 2022. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866.

Elena Cotos. 2015. Automated writing analysis for writing pedagogy: From healthy tension to tangible prospects. *Writing and Pedagogy*, 7(2-3):263–281.

Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.

Steve Graham, Michael Hebert, and Karen R. Harris. 2015. Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4):523–547.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

John R. Hayes. 2012. Modeling and remodeling writing. *Written Communication*, 29(3):369–388.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Stefan D. Keller, Johanna Fleckenstein, Maleika Krüger, Olaf Köller, and André A. Rupp. 2020. English writing skills of students in upper secondary education: Results from an empirical study in switzerland and germany. *Journal of Second Language Writing*, 48:100700.

Stefan D. Keller, Julian Lohmann, Ruth Trüb, Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, and Jens Möller. 2024. Language quality, content, structure: What analytic ratings tell us about efl writing skills at upper secondary school level in germany and switzerland. *Journal of Second Language Writing*, 65:101129.

Learning Agency Lab. 2024. Automated essay scoring 2.0. Kaggle Competition. 1st-place writeup: https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2/discussion/516571.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research. *Written Communication*, 30:358–392.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Julian F. Lohmann, Fynn Junge, Jens Möller, Johanna Fleckenstein, Ruth Trüb, Stefan Keller, Thorben Jansen, and Andrea Horbach. 2025. Neural networks or linguistic features? Comparing different machine-learning approaches for automated assessment of text quality traits among L1- and L2-Learners' argumentative essays. *International Journal of Artificial Intelligence in Education*, 35:1178–1217.

Nils-Jonathan Schaller, Yuning Ding, Thorben Jansen, and Andrea Horbach. 2025. Don't score too early! evaluating argument mining models on incomplete essays. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 345–355, Vienna, Austria. Association for Computational Linguistics.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Hansel Susanto, Alexander Agung Santoso Gunawan, and Muhammad Fikri Hasani. 2024. Development of automated essay scoring system using DeBERTa as a transformer-based language model. In *Data Analytics in System Engineering. CoMeSySo 2023*, volume 935 of *Lecture Notes in Networks and Systems*, pages 202–215, Cham. Springer.

Yu Tian, Scott Crossley, and Luuk Van Waes. 2025. The klicke corpus: Keystroke logging in compositions for knowledge evaluation. *Journal of Writing Research*, 17:23–60.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nina Vandermeulen, Mariëlle Leijten, and Luuk Van Waes. 2020. Reporting writing process feedback in the classroom. using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, 12:109–140.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

A Appendix

	w/o Chat			*w/ Chat*		
	All	Len.	NoLen.	All	Len.	NoLen.
5 min	.25	.05	**.25**	.52	.25	**.52**
10 min	**.42**	.26	.38	**.52**	.42	.48
15 min	**.55**	.42	.48	**.64**	.57	.55
20 min	**.57**	.56	.50	**.71**	.62	.60
25 min	**.70**	.66	.65	**.77**	.74	.58
Full	**.72**	.68	.67	**.80**	.77	.63

Table 6: QWK for linear regression baselines at each time threshold. ALL = all keystroke features; LEN = text length only; NOLEN = all features except length.

Category	Feature	Definition
Temporal	`write_time`	First to last keystroke (s)
	`init_pause`	Task start to first key (ms)
Pauses	`brk_count`	Pauses $\geq$2 000 ms
	`brk_total`	Sum of pause durations (s)
	`brk_mean_dur`	Mean pause duration (ms)
	`brk_ratio`	Break / total time (0–1)
Bursts	`bst_count`	Production sequences
	`bst_mean_len`	Mean chars per burst
	`bst_mean_dur`	Mean burst duration (ms)
Deletions	`del_count`	Delete/backspace events
	`del_ratio`	Deletions / total keys (0–1)
	`del_chars`	Total chars removed
Production	`key_count`	Total key-down events
	`txt_len_chr`	Char count at threshold
	`chars_per_min`	Chars / active time (cpm)
	`pp_ratio`	Process-to-product ratio
Navigation	`nav_count`	Cursor reposition events
	`cp_count`	Paste events
	`lin_index`	Forward cursor transitions (0–1)
	`tgt_switches`	Interface area switches
Chat†	`t_editor`	Time in text editor (s)
	`t_task`	Time on task description (s)
	`t_chat`	Time in chat window (s)
	`t_prompt`	Time in chat prompt (s)
	`t_none`	Time in no interface area (s)

Table 7: Keystroke features. †*Chat* only; set to zero for *NoChat*.

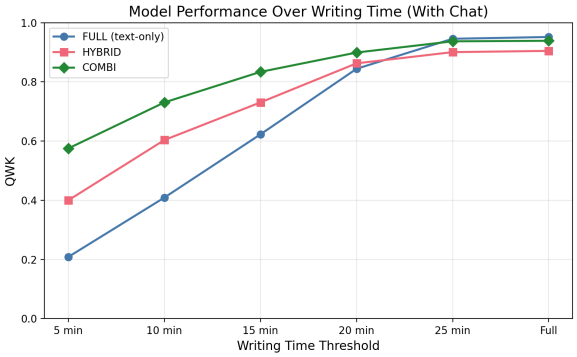

Figure 3: Performance of all 3 DeBERTa models only on *Chat* condition.

Table 8: Pearson correlations of keystroke features with text length (r_{len}) and holistic score (r_{score}) for the NoChat and Chat conditions (all tasks combined, $N = 3{,}882$). Features are sorted by $|r_{\text{score}}|$ in the NoChat condition. † denotes the LEN baseline feature.

	NoChat		Chat	
Feature	r_{len}	r_{score}	r_{len}	r_{score}
total_keystrokes	.44	.37	−.02	−.01
break_count	.29	.34	−.17	−.14
burst_count	.25	.28	−.21	−.20
total_writing_time	.26	.28	−.02	.03
process_product_ratio	−.24	−.27	−.24	−.32
linearity_index	.22	.23	−.07	−.08
burst_mean_length_char	.22	.20	.11	.11
initial_pause	.05	−.19	−.03	−.16
break_ratio	−.24	−.18	.10	.06
break_mean_duration	.01	−.15	.09	.14
deletion_ratio	.02	−.08	.02	−.02
burst_mean_duration	.03	−.04	.02	.04
txt_len_chr†	1.00	.73	1.00	.73
chars_per_minute	.56	.41	.63	.46

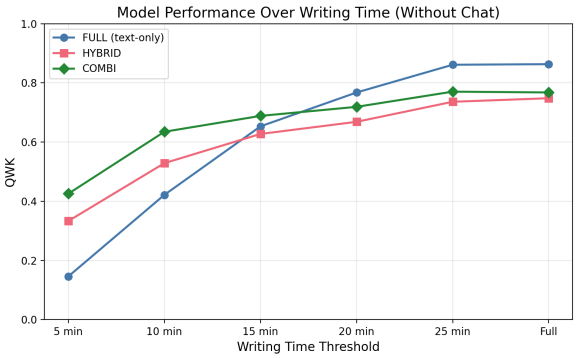

Figure 4: Performance of all 3 DeBERTa models only on *NoChat* condition.