DOI: 10.15514/ISPRAS-2026-38(4)-30

# A Hybrid Approach to Zero-Shot Cross-Corpus Essay Scoring via Rubric-Prompted LLM Agents and Deterministic Linguistic Features

*Chikake T. M., ORCID: 0009-0009-9183-0958 <tendaichikake@phystech.edu>*
*Mihailov A.A., ORCID: 0009-0000-6013-7744 <andrey20034ad@gmail.com>*

*Center for Applied Linguistic Research and Testing,*
*Moscow Institute of Physics and Technology,*
*9, Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia.*

**Abstract.** Automated Essay Scoring systems depend on scored training data for every new corpus. We investigate whether this dependency can be reduced by combining two complementary scoring paradigms: rubric-prompted large language model agents that evaluate essays without any training examples, and deterministic linguistic features that capture vocabulary, morphology, and syntax independently of any rubric. On the ASAP 2.0 benchmark (17,307 essays scored 1–6), a hybrid Random Forest regressor fusing both paradigms achieves Quadratic Weighted Kappa of 0.765 on the full corpus, approaching the supervised state of the art (0.841). Features alone reach 0.719 without neural networks. A greedy ensemble of 15 zero-shot agents scores 0.650 with no training at all. Population-matched rubrics significantly outperform cross-population rubrics in zero-shot mode, confirming that rubric design, rather than model size, determines scoring quality. We formalise these findings as an assessment portability spectrum: deterministic features transfer most reliably, followed by population-matched and then cross-population rubrics.

**Keywords:** automated essay scoring; zero-shot evaluation; cross-corpus transfer; linguistic features; LLM agents; ASAP; assessment portability.

**For citation:** Chikake T.M., Mihailov A.A. A Hybrid Approach to Zero-Shot Cross-Corpus Essay Scoring via Rubric-Prompted LLM Agents and Deterministic Linguistic Features. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 4, pp. 257–266. DOI: 10.15514/ISPRAS-2026-38(4)-30.

**Acknowledgments.** The authors acknowledge the support of this research by the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-03-2026-305 (January 16, 2026), associated with project “Applied Research on the Implementation of Artificial Intelligence Technologies in Higher Education” (project code: FSMG-2025-0086). The ASAP 2.0 dataset was provided by the Learning Agency Lab via the Kaggle platform. The EFCAMDAT corpus was made available by the EF-Cambridge research partnership.





# Гибридный подход к оценке эссе без обучения на целевом корпусе с помощью LLM-агентов на основе рубрик и детерминированных лингвистических признаков

*Чикаке Т. М., ORCID: 0009-0009-9183-0958 <tendaichikake@phystech.edu>*
*Михайлов А.А., ORCID: 0009-0000-6013-7744 <andrey20034ad@gmail.com>*

*Центр прикладных лингвистических исследований и тестирования,*
*Московский физико-технический институт,*
*141701, Россия, Московская обл., г. Долгопрудный, Институтский пер., 9*

**Аннотация.** Системы автоматической оценки эссе требуют размеченных данных для каждого нового корпуса. В данной работе исследуется, можно ли снизить эту зависимость за счёт объединения двух парадигм: агентов на основе больших языковых моделей (БЯМ), оценивающих эссе без обучения на целевых данных, и детерминированных лингвистических признаков, описывающих лексику, морфологию и синтаксис независимо от рубрики. На эталонном наборе ASAP 2.0 (17 307 эссе, шкала 1-6) гибридный регрессор на основе случайного леса достигает квадратично взвешенной каппы (QWK) = 0,765 на полном корпусе, приближаясь к лучшим результатам методов обучения с учителем (0,841). Признаки в отдельности дают 0,719 без нейросетей. Жадный ансамбль из 15 агентов набирает QWK = 0,650 без обучения. Рубрики, согласованные с целевой популяцией, значимо превосходят межпопуляционные, подтверждая ведущую роль построения рубрик. Результаты формализованы как «спектр переносимости оценки»: детерминированные признаки переносимы наиболее надёжно, за ними следуют рубрики для целевой популяции и межпопуляционные рубрики.

**Ключевые слова:** автоматическая оценка эссе; оценка без обучения; межкорпусный перенос; лингвистические признаки; агенты на основе БЯМ; набор данных ASAP; переносимость оценки.

**Для цитирования:** Чикаке Т. М., Михайлов А.А. Гибридный подход к оценке эссе без обучения на целевом корпусе с помощью LLM-агентов на основе рубрик и детерминированных лингвистических признаков. Труды ИСП РАН, 2026, том 38, вып. 4, часть 2, стр. 257-266 (на английском языке). DOI: 10.15514/ISPRAS-2026-38(4)-30.

**Благодарности.** Авторы выражают благодарность за поддержку данного исследования Министерству науки и высшего образования Российской Федерации (соглашение № 075-03-2026-305 от 16.01.2026), в рамках проекта «Прикладные исследования по внедрению технологий искусственного интеллекта в высшем образовании» (код проекта: FSMG-2025-0086). Набор данных ASAP 2.0 предоставлен организацией Learning Agency Lab через платформу Kaggle. Корпус EFCAMDAT предоставлен в рамках партнёрства EF–Cambridge.

## *1. Introduction*

Page’s Project Essay Grade [1] launched the automated essay scoring (AES) field over six decades ago. What followed was a progression from handcrafted feature systems like e-rater [2], through statistical approaches surveyed by Shermis and Burstein [3], to the BERT-era [4] transformer models which now reach near-human agreement on standard benchmarks [5]. These supervised systems share a costly dependency: they need substantial scored training data for every new corpus, rubric, and prompt. Where labelled data is scarce, they cannot be deployed.

Two paradigms sidestep this constraint. Rubric-prompted LLM evaluation encodes scoring rubrics directly into the prompt, turning large language models into zero-shot evaluators without scored examples [6-7]. Deterministic linguistic feature extraction computes interpretable text metrics (vocabulary sophistication [8, 9], syntactic complexity [10-11], and morphological diversity) which capture writing quality independent of any rubric or scoring scale [12].

Neither paradigm is guaranteed to transfer. A rubric built for second-language (L2) learners may fail to discriminate quality among native (L1) speakers. Features calibrated on one corpus may lose their

predictive power on another. The question is how portable these assessment methods really are, and whether combining them yields something stronger than either alone.

We map this assessment portability spectrum through a cross-corpus evaluation on ASAP 2.0 [13-14], a standard AES benchmark of 17,307 essays written by North American students in grades 7–10, scored holistically 1–6 by dual human raters. The research questions are:

1. How does rubric alignment affect zero-shot LLM scoring?
2. Can deterministic linguistic features achieve competitive AES performance without neural networks?
3. Do features and LLM agents capture complementary quality dimensions?
4. Can linguistic features recover meaningful quality signal without any labelled data?

## *2. Related Work*

### 2.1 Supervised AES

Page [1] first showed that proxy features like essay length correlate with human holistic scores. The e-rater system [2] turned this into a feature-engineering paradigm which dominated AES for two decades [3]. The current state of the art spans a wide performance range: BERT-based few-shot learning reaches QWK = 0.974 on the original ASAP [5], a DeBERTa ensemble achieves QWK = 0.841 on ASAP 2.0 [15], cross-prompt trait scoring via ProTACT achieves QWK = 0.592 on the original ASAP [16], and zero-shot open-source LLM scoring remains below QWK = 0.55 [6]. The gap between supervised methods (QWK $>$ 0.8) and zero-shot approaches (QWK $<$ 0.55) remains wide.

### 2.2 LLM-Based Writing Assessment

Lee et al. [7] showed that chain-of-thought prompting with GPT-3.5/GPT-4 improves science assessment accuracy. Mizumoto and Eguchi [6] tested LLM-based essay scoring in a language learning context, with mixed results. Our approach differs: we use smaller, locally-deployed models (Qwen3:14b [17], 14B parameters) with methodology-specific rubric prompts. Companion work [18] has validated this approach across 17 international testing systems on EFCAMDAT [19].

### 2.3 Feature Engineering for Writing Assessment

Linguistic features have been central to writing assessment for decades [12]. Syntactic complexity measures [10, 11], type–token ratio variants [8], and Academic Word List coverage [9] all predict writing quality. The feature architecture we evaluate has two tiers: 95 deterministic features (Tier 1) and 33 ML-based features (Tier 2). On 3,198 L2 texts from EFCAMDAT [19], prior work [18] reports 66.72% exact accuracy for six-class CEFR [20] classification. Whether these features generalize to L1 essay quality has not been tested. Tier 2 features are excluded from the present study to isolate the contribution of fully deterministic measures.

## *3. Methodology*

### 3.1 ASAP 2.0 Dataset

The ASAP 2.0 dataset [13, 14] comprises 17,307 essays written by North American students in grades 7–10, scored holistically from 1 to 6 by dual human raters. The score distribution is heavily skewed: 1 (7.2%), 2 (27.3%), 3 (36.3%), 4 (22.7%), 5 (5.6%), 6 (0.9%).





### 3.2 Approach 1: LLM Agent Scoring (Zero-Shot)

Seventy rubric-prompted criterion agents are evaluated from 17 frameworks, including IELTS [21] (4 agents), CEFR [20] (4), TOEFL [22] (7), Cambridge B2/C1/C2 [23] (12), PTE (4), APTIS (5), EGE (5), and two ASAP-native holistic agents.

All agents use Qwen3:14b [17] (Q4_K_M quantization, 9.3 GB) at temperature 0 via the istok-agents framework, deployed on NVIDIA V100-32GB and RTX 3090 GPUs via Ollama 0.6. Agent scores are normalized to 0–1 and mapped to the ASAP 1–6 scale. No ASAP-scored essays are used for calibration; this is strictly zero-shot evaluation. The rubric prompts are available in the code repository [24].

### 3.3 Approach 2: Tier 1 Feature Classification

Ninety-five Tier 1 deterministic linguistic features are extracted: lexical sophistication (21 features: TTR variants [8], CEFR vocabulary distribution, AWL coverage [9]), morphological complexity (64 features: POS distribution, derivational complexity, enhanced morphosyntactic patterns), and syntactic complexity (10 features: parse tree depth [10], clause density, T-unit length [11]).

A Random Forest classifier [25] (200 trees, max depth 20, balanced class weights) is trained using 5-fold stratified CV. Feature extraction is CPU-based and processes 17,307 essays in 116 mins.

### 3.4 Approach 3: Hybrid (Features + LLM Scores)

The hybrid approach augments the 95 Tier 1 features with normalized agent scores and confidence values. In the full configuration, 70 agents contribute 140 Tier 3 features (score + confidence each), yielding 235 total features. All 70 agents have been evaluated on all 17,307 essays (1,211,490 evaluations; ≈ 9,400 GPU-hours on NVIDIA V100 and RTX 3090 via Ollama 0.6).

A greedy forward-selection procedure identifies the best agent subset: starting from the single best-performing agent, each step adds the agent whose inclusion maximally improves ensemble QWK via simple score averaging. The procedure converges at 15 agents (QWK = 0.650), after which additional agents reduce performance.

No data leakage occurs: Tier 3 agent scores are computed in zero-shot mode without access to ASAP labels. The 5-fold stratified CV applies only to the RF training stage.

### 3.5 Evaluation Metrics

Quadratic Weighted Kappa (QWK) [26] is the primary metric, supplemented by Spearman $\rho$ and exact/adjacent accuracy. All metrics are computed via 5-fold stratified CV.

## *4. Results*

### 4.1 LLM Agent Scoring

Table 1 reports zero-shot performance on 198 class-balanced essays with bootstrap 95% CIs.

**Rubric alignment matters.** Native ASAP agents (QWK = 0.474, CI [0.41, 0.53]) nearly double L2 agents (0.21–0.25) with non-overlapping CIs. The two native agents have overlapping CIs, suggesting that population alignment matters more than rubric variant.

The **L2 ensemble** (QWK = 0.100) performs worse than individual L2 agents because averaging near-constant predictions (range 3–5) further collapses variance. All agents exhibit **leniency bias**: none predicts score 1, systematically over-scoring low-quality essays.

*Table 1. Zero-shot LLM agent performance on ASAP 2.0 (198 class-balanced essays, 33 per score). Bootstrap 95% CIs from 10,000 resamples.*

| Agent | Rubric | QWK | 95% CI | $\rho$ | Range |
|---|---|---|---|---|---|
| ASAP (source) | Native | 0.474 | [0.41, 0.53] | 0.70 | 2–5 |
| ASAP (indep.) | Native | 0.449 | [0.38, 0.51] | 0.67 | 2–6 |
| IELTS coher. | L2 0–9 | 0.253 | [0.21, 0.30] | 0.64 | 4–5 |
| CEFR prod. | L2 0–5 | 0.243 | [0.18, 0.30] | 0.50 | 2–5 |
| IELTS lexical | L2 0–9 | 0.210 | [0.16, 0.26] | 0.50 | 4–5 |
| *Ens: native (2)* | | 0.432 | [0.36, 0.50] | 0.66 | 2–6 |
| *Ens: L2 (3)* | | 0.100 | [0.06, 0.15] | 0.32 | 3–5 |
| *Ens: all (5)* | | 0.299 | [0.23, 0.36] | 0.55 | 3–5 |

## 4.2 Feature-Only Classification

On the full dataset, 95 deterministic features achieve **QWK = 0.719** (RF Regressor), exceeding published non-BERT baselines. 97.4% of predictions land within one score point (Table 2). Scores 2–4 (86% of data) reach 64–66% exact accuracy; extremes are harder due to class imbalance.

*Table 2. Tier 1 feature classification on ASAP 2.0 (Random Forest, 5-fold stratified CV).*

| Method | QWK | $\rho$ | Exact | Adj. | $N$ |
|---|---|---|---|---|---|
| T1 features (full) | **0.719** | **0.739** | 58.6% | 97.4% | 17,307 |
| T1 features (200) | 0.590 | 0.646 | 56.9% | 93.1% | 200 |
| T1 features (48) | 0.398 | 0.434 | 37.5% | 87.5% | 48 |

## 4.3 Hybrid Approach

**Scaling from 13 to 70 agents improves full-corpus QWK by 0.012** (0.753 → 0.765), with diminishing returns beyond 20 agents (Table 3). **Class balancing transforms performance**: T1 features alone reach QWK = 0.828 (vs. 0.719 on skewed data). The 70-agent T1+T3 regressor reaches **QWK = 0.870** on the balanced subset and **0.765 on the full corpus**, approaching the DeBERTa supervised winner [15] (0.841 on the same full corpus). Features and agents capture complementary quality dimensions: the hybrid improves over features alone by 0.046 QWK on the full corpus.

## 4.4 Score Boundary Plateau Analysis

`content_word_count` has the largest effect size at four of five boundaries ($d = 0.91$–$1.22$). The $1 \leftrightarrow 2$ exception is dominated by syntactic features (modal verb ratio, verb phrase density), reflecting a shift from minimal to functional syntax (Table 4).





*Table 3. Hybrid results on full corpus and balanced subset (5-fold stratified CV, RF Regressor). The 70-agent hybrid reaches QWK = 0.765 on the full corpus and 0.870 under class balancing.*

| Method | Feat. | QWK | $\rho$ | Exact | Adj. |
|---|---|---|---|---|---|
| *Full corpus (17,307 essays)* | | | | | |
| T1 features (regressor) | 95 | 0.719 | 0.739 | 58.6% | 97.4% |
| T3 all 70 agents | 140 | 0.658 | 0.672 | 52.9% | 96.3% |
| **T1+T3 hybrid (70 agents)** | **235** | **0.765** | **0.776** | **61.9%** | **98.3%** |
| T1+T3 hybrid (13 agents) | 121 | 0.753 | 0.764 | 60.6% | 98.1% |
| *Balanced subset (5,006 essays, up to 970/class)* | | | | | |
| T1 features (regressor) | 95 | 0.828 | 0.839 | 52.4% | 95.0% |
| T3 all 70 agents | 140 | 0.791 | 0.805 | 46.3% | 93.3% |
| **T1+T3 hybrid (70 agents)** | **235** | **0.870** | **0.876** | **57.9%** | **97.0%** |
| T1+T3 hybrid (13 agents) | 121 | 0.860 | 0.867 | 56.2% | 96.7% |

*Table 4. Per-boundary classification. content_word_count dominates every boundary except $1 \leftrightarrow 2$, where syntactic features lead. Cohen's $d$ in parentheses.*

| Boundary | Acc. | QWK | $N$ | Top Feature ($d$) |
|---|---|---|---|---|
| $1 \leftrightarrow 2$ | 79.8% | 0.110 | 5,975 | `pos_modal_ratio` (0.53) |
| $2 \leftrightarrow 3$ | 77.0% | 0.521 | 11,003 | `content_word_count` (0.98) |
| $3 \leftrightarrow 4$ | 76.2% | 0.489 | 10,206 | `content_word_count` (1.18) |
| $4 \leftrightarrow 5$ | 83.5% | 0.356 | 4,896 | `content_word_count` (1.22) |
| $5 \leftrightarrow 6$ | 86.9% | 0.139 | 1,126 | `content_word_count` (0.91) |

## 4.5 Length Ablation

Removing all length features reduces QWK (Table 5) by only 0.011 (from 0.719 to 0.708). Although `content_word_count` accounts for 51.6% of RF split importance, this information is largely redundant with correlated features (TTR, vocabulary breadth, syntactic density). Within-quartile evaluation confirms quality discrimination at fixed length bands (QWK = 0.20–0.47).

*Table 5. Length ablation (RF Regressor, 5-fold CV, 17,307 essays). Removing length features reduces QWK by only 0.011. Bootstrap CIs from 10,000 resamples.*

| Feature Set | $N_f$ | QWK | 95% CI |
|---|---|---|---|
| All features (baseline) | 95 | **0.719** | [0.712, 0.726] |
| Without length (strict) | 91 | 0.708 | [0.700, 0.715] |
| Without length (extended) | 89 | 0.706 | [0.699, 0.713] |

### 4.6 Unsupervised and Regression Approaches

The RF Regressor (QWK = 0.719) edges out the classifier (0.709) by treating score as continuous (Table 6). Purely unsupervised K-Means reaches QWK = 0.623 with no labels at all, confirming that Tier 1 features capture substantial quality signal even without supervision.

*Table 6. Unsupervised and regression results on 17,307 ASAP essays. *PCA uses score distribution as prior only.*

| Method | Superv. | QWK | MAE | Adj. |
|---|---|---|---|---|
| RF Regressor | CV | **0.719** | 0.498 | 97.4% |
| Ridge Regression | CV | 0.706 | 0.517 | 97.1% |
| K-Means $k = 6$ | None | 0.623 | – | 94.7% |
| PCA projection | None* | 0.280 | – | 80.1% |

### 4.7 Comparison with Baselines

The T1+T3 hybrid regressor achieves QWK = 0.765 on the full 17,307-essay corpus, approaching the supervised DeBERTa winner [15] (0.841) on the same corpus (Table 7). On a balanced subset, it reaches 0.870. The zero-shot greedy ensemble of 15 agents (QWK = 0.650) substantially exceeds prior zero-shot baselines without any training.

## *5. Discussion*

### 5.1 The Assessment Portability Spectrum

The results point to a clear portability hierarchy: (1) **deterministic features** (most portable, QWK = 0.719 on full corpus) measure text properties directly without reference to any rubric; (2) **population-matched rubrics** (QWK = 0.474) enable meaningful zero-shot scoring when descriptors align with the target population; (3) **cross-population rubrics** (QWK = 0.21–0.25) correctly rank essays but compress scores. For cross-corpus deployment, the hybrid approach (QWK = 0.765, approaching the DeBERTa supervised winner at 0.841) confirms that features and agents should be combined.

### 5.2 Complementary Strengths across Corpora

On ASAP (L1 data), the 70-agent ensemble alone reaches QWK = 0.658, while features alone reach 0.719; combined, they yield 0.765. On EFCAMDAT (L2 data), features dominate agents [18]. The





hybrid gain (+0.046 over features alone on the full corpus) confirms complementarity: agents capture argumentation and rhetorical sophistication, while features detect measurable surface-level deficiencies at the scale extremes that agents refuse to recognize.

*Table 7. Comparison with baselines. All results on the full ASAP 2.0 corpus (17,307 essays) except †balanced subsets. [a]Results on the original ASAP (2012), not directly comparable with ASAP 2.0.*

| Method | Training | QWK | Type |
|---|---|---|---|
| BERT few-shot [5] | 120/prompt | 0.974 | Supervised[a] |
| T1+T3 Regr. (ours)† | CV + z-shot | 0.870 | Hybrid |
| DeBERTa ens. [15] | Full ASAP 2.0 | 0.841 | Supervised |
| BERT Base [5] | Full corpus | 0.837 | Supervised[a] |
| **T1+T3 Regr. (ours)** | **CV + z-shot** | **0.765** | **Hybrid** |
| RF Regr. (ours) | CV only | 0.719 | Interpretable |
| Ensemble (greedy 15) | None | 0.650 | Zero-shot |
| K-Means (ours) | None | 0.623 | Unsupervised |
| ProTACT [16] | Cross-prompt | 0.592 | Supervised[a] |
| ASAP native (ours)† | None | 0.474 | Zero-shot |

### 5.3 The Length–Quality Confound

Although `content_word_count` has the largest feature importance (51.6%), the length ablation shows that removing all four length-related features reduces QWK by only 0.011. The remaining 91 features preserve nearly all predictive power. Length is a legitimate quality signal on ASAP (longer timed essays demonstrate more evidence and argumentation), but the model captures substantially more than volume alone. This contrasts with EFCAMDAT, where morphological diversity [18] is the primary signal.

### 5.4 Classification versus Regression

The RF Regressor (QWK = 0.719) outperforms the RF Classifier (0.709) because regression preserves ordinal structure. Ridge Regression (QWK = 0.706) confirms the advantage is not model-specific.

### 5.5 The Leniency Problem

LLM agents are systematically lenient at the scale bottom [7, 6]. Features correct for this: score 1 essays have measurable deficiencies [10, 20] that features detect reliably, overriding the LLM's generosity.

### 5.6 Limitations

All LLM evaluations use Qwen3:14b [17]; larger models might achieve higher zero-shot performance. Scaling from 13 to 70 agents yields diminishing returns (+0.012 QWK on the full corpus), suggesting that agent diversity matters more than quantity. The feature set was designed for CEFR classification [20] and its transfer to ASAP (QWK = 0.719) needs validation on additional datasets.

## *6. Conclusion*

We evaluated assessment portability on ASAP 2.0 with 70 rubric-prompted agents from 17 frameworks across all 17,307 essays. Five contributions emerge: (1) the hybrid T1+T3 regressor achieves QWK = 0.765 on the full corpus (0.870 on a balanced subset), approaching the supervised DeBERTa winner (0.841); (2) features alone reach 0.719 on the full corpus, exceeding published non-BERT baselines; (3) a greedy ensemble of 15 zero-shot agents reaches QWK = 0.650 without any training; (4) an assessment portability spectrum is identified; (5) the dominant quality signal is corpus-dependent (length on ASAP vs. morphological diversity on EFCAMDAT). Feature engineering, when fused with rubric-prompted LLM agents, approaches supervised deep learning while retaining interpretability and cross-corpus robustness.

## *References*

[1]. Page E.B. The Imminence of Grading Essays by Computer. Phi Delta Kappan, 1966, vol. 47, no. 5, pp. 238-243.
[2]. Attali Y., Burstein J. Automated Essay Scoring With e-rater V.2. Journal of Technology, Learning, and Assessment, 2006, vol. 4, no. 3.
[3]. Shermis M.D., Burstein J. Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, NJ, 2003.
[4]. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171-4186.
[5]. Amin T., Tanoli Z., Aadil F., Awan K.M., Lim S. Enhancing Essay Scoring: An Analytical and Holistic Approach with Few-Shot Transformer-Based Models. IEEE Access, 2025, vol. 13. DOI: 10.1109/ACCESS.2025.3530272.
[6]. Mizumoto A., Eguchi M. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. Research Methods in Applied Linguistics, 2023, vol. 2, no. 2, pp. 100050. DOI: 10.1016/j.rmal.2023.100050.
[7]. Lee G., Latif E., Wu X., Liu N., Zhai X. Applying Large Language Models and Chain-of-Thought for Automatic Scoring. Computers and Education: Artificial Intelligence, 2024, vol. 6, pp. 100213. DOI: 10.1016/j.caeai.2024.100213.
[8]. Jarvis S. Short Texts, Best-Fitting Curves and New Measures of Lexical Diversity. Language Testing, 2002, vol. 19, no. 1, pp. 57-84. DOI: 10.1191/0265532202lt220oa.
[9]. Coxhead A. A New Academic Word List. TESOL Quarterly, 2000, vol. 34, no. 2, pp. 213-238. DOI: 10.2307/3587951.
[10]. Lu X. Automatic Analysis of Syntactic Complexity in Second Language Writing. International Journal of Corpus Linguistics, 2010, vol. 15, no. 4, pp. 474-496. DOI: 10.1075/ijcl.15.4.02lu.
[11]. Kyle K., Crossley S.A. Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. Modern Language Journal, 2018, vol. 102, no. 2, pp. 333-349. DOI: 10.1111/modl.12468.
[12]. Crossley S.A. Linguistic Features in Writing Quality and Development: An Overview. Journal of Writing Research, 2020, vol. 11, no. 3, pp. 415-443. DOI: 10.17239/jowr-2020.11.03.01.
[13]. Crossley S.A., Baffour P., Tian Y., Holden L., Betts A. A Large-Scale Corpus for Assessing Source-Based Writing Quality: ASAP 2.0. Assessing Writing, 2025, vol. 63, pp. 100912. DOI: 10.1016/j.asw.2025.100912.
[14]. Crossley S., Baffour P., King J., Burleigh L., Reade W., Demkin M. Learning Agency Lab – Automated Essay Scoring 2.0, 2024. Available at: https://kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2, accessed 01.06.2026.





[15]. Learning Agency Lab. Learning Agency Lab – Automated Essay Scoring 2.0: Competition Results, 2024. Available at: https://kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2/leaderboard, accessed 01.06.2026.
[16]. Do H., Kim Y., Lee G.G. Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring. Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 1538-1551.
[17]. Qwen Team. Qwen3, 2025. Available at: https://qwenlm.github.io/blog/qwen3/, accessed 01.06.2026.
[18]. Chikake T.M., Bazanova E.M., Gorizontova A.V. Multi-tier linguistic feature engineering for CEFR classification: a comprehensive analysis of deterministic and machine learning-based features. Modeling and Analysis of Information Systems, 2026, vol. 33, no. 1, pp. 6-29. DOI: 10.18255/1818-1015-2026-1-6-29.
[19]. Geertzen J., Alexopoulou T., Korhonen A. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). Proceedings of the 31st Second Language Research Forum, 2013, pp. 240-254.
[20]. Council of Europe. CEFR Companion Volume with New Descriptors. Council of Europe Publishing, Strasbourg, 2020.
[21]. British Council, IDP, Cambridge Assessment English. IELTS Writing Band Descriptors (Task 2). 2023. Available at: https://www.ielts.org, accessed 21.04.2026.
[22]. Educational Testing Service. TOEFL iBT Writing Rubrics. 2024. https://www.ets.org/toefl, accessed 21.04.2026.
[23]. Cambridge Assessment English. Cambridge English Writing Assessment: Teacher Guides, 2024.
[24]. Available at: https://github.com/deeppavlov/istok-platform, accessed 27.06.2026.
[25]. Breiman L. Random Forests. Machine Learning, 2001, vol. 45, no. 1, pp. 5-32. DOI: 10.1023/A:1010933404324.
[26]. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, vol. 70, no. 4, pp. 213-220. DOI: 10.1037/h0026256.

## *Информация об авторах / Information about authors*

Тендаи Мапунгвана ЧИКАКЕ – младший научный сотрудник Центра прикладных лингвистических исследований и тестирования «ИСТОК» МФТИ. Область научных интересов: автоматическая оценка текстов, обработка естественного языка, агентные системы на основе больших языковых моделей, псевдобулевы полиномы.

Tendai Mapungwana CHIKAKE is a junior researcher at the Center for Applied Linguistic Research and Testing “ISTOK” of Moscow Institute of Physics and Technology (MIPT). Research interests: automated text assessment, natural language processing, LLM-based agent systems, pseudo-boolean polynomials.

Андрей Алексеевич МИХАЙЛОВ – сотрудник Центра прикладных лингвистических исследований и тестирования «ИСТОК» МФТИ. Область научных интересов: машинное обучение, обработка естественного языка, автоматическая оценка текстов.

Andrey Alekseevich MIHAILOV works at the Center for Applied Linguistic Research and Testing “ISTOK” of Moscow Institute of Physics and Technology (MIPT). Research interests: machine learning, natural language processing, automated text assessment.