Develop a Generic Essay Scorer for Practice Writing Tests of Statewide Assessments

Yi Gui
University of Iowa
yi-gui@uiowa.edu

Abstract

This study examines whether NLP transfer learning techniques, specifically BERT, can be used to develop prompt-generic AES models for practice writing tests. Findings reveal that fine-tuned DistilBERT, without further pre-training, achieves high agreement (QWK ≈ 0.89), enabling scalable, robust AES models in statewide K-12 assessments without costly supplementary pre-training.

1 Introduction

Currently, Automated Essay Scoring (AES) is widely utilized in large-scale standardized tests with writing assessments in the US. However, there are some notable limitations in the current major AES engines that are used for many high-stakes writing assessments, such as the annual statewide assessments in K-12 education. These limitations prevent the provision of instantaneous online essay scoring services in writing practice tests of those statewide assessments for students' daily exercise.

One major limitation of AES algorithms trained with traditional machine learning (ML) approaches is the substantial sample size required for training sets with essays scored by human raters. The random assignment of prompts in practice tests results in some prompts having too few essay samples to effectively train a scoring model using traditional ML methods. For instance, Intelligent Essay Assessor (IEA), a major AES engine developed by Pearson which is used in many operational tests, including several statewide assessments, requires a sample of approximately 500 student responses evaluated by human raters to score essays on a specific prompt in high-stakes assessments (Foltz et al., 2013). While it also scores essays in MyLab Writing online services instantly with immediate overall evaluations, it still needs hundreds of submissions scored by human raters to build scoring models for each prompt (Pearson Inc., 2010).

A precursor area with this frequent lack of "labelled" data quandary in ML is the image classification problem through computer vision. The traditional ML model needs to be trained for a specific task of image classification with the target data from scratch, making no use of the knowledge previously learned from similar tasks. To deal with this predicament, transfer learning is applied because it is able to build accurate models even without enough labeled data from the target domain (Rawat & Wang, 2017). With transfer learning, the model-building process starts from the "knowledge" that has been learned previously instead of zero, when solving relevant problems in the past.

Thus, the purpose of the study is to develop a generic essay scorer generalizable to essays on any prompts in the target domain with Google's BERT (Bidirectional Encoder Representations from Transformers), one of state-of-the-art NLP transfer learning techniques, for low-stakes online writing practice tests of those statewide student assessments, even if there is not enough essay sample to train scoring algorithms with traditional ML approaches. With such a generic essay scorer, students' routine practice essays can be scored similarly to those assessment essays even outside the annual test windows, providing students with timely and meaningful feedback during their preparation.

Transfer learning using Google's BERT revolutionizes traditional ML approaches by leveraging pre-trained models on extensive datasets to improve performance on specific downstream tasks. BERT is pre-trained on a large

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 58–81

corpus of human language text materials, including the entirety of Wikipedia (comprising roughly 2.5 billion words) and the BookCorpus dataset (comprising approximately 800 million words). This pre-training method is particularly advantageous as it allows BERT to generate deep contextualized word embeddings that capture nuanced relationships within the text and be fine-tuned with minimal labeled target data to develop high-performing models in target domains. Thus, this study seeks to investigate how BERT can be utilized to help develop generic AES models and examine how different treatments of BERT's pre-training affect the models' scoring performances in an AES research experiment designed to answer these research questions. Moreover, an analytic essay scoring method focusing on specific writing traits has been selected in this research. The four traits to be scored are development, organization, language use, and prompt task, based on the ELA Common Score Standards of writing, and the scoring rubrics of the SWAS essays used as the target data in the study.

In this research, the following research questions are expected to be answered:

1) How many hyperparameter settings of the original BERT model, when fine-tuned on target data, achieve a Quadratic Weighted Kappa (QWK) value greater than 0.7 for each writing trait (development, organization, prompt task, language use) without additional pre-training?

2) How many hyperparameter settings result in QWK values greater than 0.7 when a pre-trained BERT model undergoes further pre-training on either "within-task" or "in-domain" materials, followed by fine-tuning? Additionally, do these settings outperform the original BERT model in terms of performance?

3) What is the performance rank orders of fine-tuned scoring models for various writing traits when using the same hyperparameter settings, and what are the implications?

The target domain consists of essays written by high school students, while the scoring results produced by the AES engine, IEA, for the available SWAS essays in the study serve as the reference against which the study's scoring results are compared. The flowchart in Figure 1 illustrates the research design and the experimental procedures of the study.

2 Related Work

Automated Essay Scoring (AES) systems have historically depended on handcrafted linguistic features coupled with traditional machine-learning methods. Early influential systems like Project Essay Grade (PEG) used simple textual proxies—such as sentence length or vocabulary—to approximate human grades (Page, 1966). Later, more sophisticated AES engines, notably IntelliMetric and E-rater, employed extensive feature engineering, including grammar accuracy, lexical diversity, and structural coherence (Attali & Burstein, 2006; Shermis & Burstein, 2013). These approaches established AES as a viable alternative for essay scoring, yet their accuracy and adaptability heavily depended on the quality and quantity of manually crafted features and extensive prompt-specific training data.

The release of the Automated Student Assessment Prize **(**ASAP**)** dataset (Shermis & Burstein, 2013) significantly advanced AES research by offering a standardized evaluation benchmark. With this dataset, neural network methods emerged, notably recurrent neural networks **(**RNNs**)** and convolutional neural networks **(**CNNs**),** which automatically learned textual representations rather than relying solely on manual features. Taghipour and Ng (2016) demonstrated that simple CNN-RNN hybrids could surpass traditional AES baselines by directly learning meaningful text patterns from essays. Still, these early neural models struggled to effectively represent complex, long-range discourse structures characteristic of persuasive and argumentative essays.

The advent of pretrained transformer-based language models, particularly BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), dramatically shifted the AES paradigm. These models, pretrained on massive textual corpora, offered deep contextualized embeddings capable of capturing semantic and syntactic nuances beyond the reach of simpler neural architectures (Devlin et al., 2019). Mayfield and Black (2020) provided an influential early evaluation of fine-tuning BERT for AES, showing that transformer models could achieve accuracy comparable to highly-engineered feature-based systems, although computational demands were notably higher. Their work demonstrated transformers' potential for AES, while also highlighting practical trade-offs in model deployment.

To better exploit transformers' strengths, researchers developed specialized fine-tuning methods. Yang et al. (2019) proposed combining a traditional regression loss with a ranking loss, guiding transformer models toward learning not only accurate score predictions but also correct relative ordering of essay quality. This dual-objective approach improved Quadratic Weighted Kappa (QWK)—a standard AES performance metric—by approximately 2–3 percentage points over standard fine-tuning, demonstrating that carefully crafted training objectives can significantly enhance transformer-based AES.

AES research has also addressed the perennial challenge of data scarcity through domain adaptation and multi-task learning. Typically, each essay prompt has limited training data, posing significant risks of overfitting. Cao et al. (2020) presented a domain-adaptive framework combining adversarial training and auxiliary self-supervised tasks (e.g., sentence-order prediction) to learn prompt-invariant essay representations. Their approach not only improved performance on previously unseen prompts but also established a practical methodology for mitigating prompt-specific data shortages through domain transfer. Similarly, Muangkammuen and Fukumoto (2020) employed multi-task learning by integrating an auxiliary sentence-level sentiment analysis task alongside AES. This hierarchical joint training improved QWK scores, illustrating that complementary learning tasks could enrich the representation learned by AES models, enhancing their generalizability.

Holistic essay scoring, while common, limits the detailed feedback educators desire. Thus, recent AES research emphasizes analytic scoring, separately evaluating distinct writing traits (e.g., organization, content, grammar). Historically, separate models were developed independently for each trait, ignoring the natural correlations among writing dimensions. For example, early analytic scoring models, like those by Persing and Ng (2015, 2016), modeled traits like argument strength or organization independently with trait-specific features and classifiers. More recently, Do et al. (2024) proposed Autoregressive Score Generation for Multi-trait Scoring (ArTS), using a transformer-based T5 model to sequentially generate scores for multiple traits. This innovative framework explicitly modeled trait dependencies, significantly improving trait-level AES performance and marking a notable advancement in providing nuanced formative feedback to students.

Evaluation methods have also become standardized with AES advancements. Quadratic Weighted Kappa (QWK) remains a widely adopted metric, penalizing larger scoring errors more heavily and thus closely aligning automated evaluations with human judgments. Current transformer-based AES models routinely achieve QWK scores around 0.75 to 0.80 on standard benchmarks like ASAP, nearing human inter-rater agreement levels (~0.80–0.85; Mayfield & Black, 2020; Yang et al., 2019). This demonstrates substantial progress in AES technology toward human-level reliability.

Overall, AES research has evolved significantly—from feature-engineered regressors to sophisticated transformer-based methods—driven by transformer architectures, specialized training strategies, multi-task learning, and domain adaptation. These advances collectively address critical challenges such as data scarcity and trait-specific feedback, facilitating robust, reliable, and informative automated scoring systems. This literature provides a robust foundation for the current study's exploration of developing prompt-generic AES models for statewide educational assessments, emphasizing transformer-based methods' potential to improve scoring quality, reduce data requirements, and enhance educational feedback.

3 Method

A distilled version of BERT (DistilBERT) was employed to develop prompt-generic essay scoring models. Three variants were compared:

Group 1 (Baseline): DistilBERT fine-tuned directly on SWAS essays.

Group 2 (ASAP-pretrained): DistilBERT further pre-trained on the ASAP corpus, then fine-tuned on SWAS.

Group 3 (SWAS-pretrained): DistilBERT further pre-trained on a 500-essay "within-task" SWAS subset, then fine-tuned on SWAS.

3.1 Data Preparation

Two corpora were used. The SWAS corpus originally contained 4,500 essays (1,500 per grade for grades 9–11). Handwritten submissions (n = 1,203) were excluded, leaving 3,297 typed essays (Figure 2). A random sample of 500 typed essays was reserved for within-task pre-training. The ASAP corpus, comprising 12,970 essays across eight prompts and two genres (Table 1), was used for in-domain pre-training.

To mitigate score-level imbalance from handwritten-essay removal, RandomOverSampler was applied separately to each analytic trait. The balance improvements were confirmed via stacked-bar plots and annotated tables (Figures 3 and 4), though downstream benefits were minimal. Oversampled sets were used only for diagnostics.

3.2 Model Pre-training and Fine-tuning

DistilBERT weights (66 M parameters) were loaded from the Hugging Face "distilbert-base-uncased" checkpoint. In Groups 2 and 3, intermediate pre-training was performed using a learning rate of 5×10^{-4} and batch sizes of 16 and 32. All pre-training ran for a uniform number of epochs, ensuring each variant saw equal exposure to its respective corpora.

Subsequently, each variant was fine-tuned on SWAS essays using an Ordinal Logistic Regression (OLR) classifier built on DistilBERT embeddings. Hyper-parameters for fine-tuning were selected via grid search over three regularization strengths ($\alpha \in \{0.01, 0.10, 1.00\}$) and three maximum-iteration ceilings ({100, 500, 1000}), yielding nine distinct configurations.

3.3 Evaluation Protocol

Model evaluation employed a leave-one-grade-out design: in three rounds, essays from two grades were used for training and the remaining grade served as the test set (Tables 2–4). Within each round, five-fold cross-validation was executed, and the entire process was repeated with three random seeds to assess stability. Aggregate statistics across folds and seeds were computed for: **Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Exact Accuracy, Adjacent Accuracy (predictions within ±1 score point), Precision, recall, and F1** were calculated per score point (1–5) (see Figures 8–10 for accuracy, Figures 11–12 for precision, recall, and F1).

By systematically comparing baseline and pre-trained variants under consistent optimization settings and a robust leave-one-grade-out protocol, this method section demonstrates how prompt-generic essay scoring can be realized with minimal reliance on prompt-specific labeled data. The design ensures fairness across groups, repeatability via multiple seeds, and comprehensive trait-level analysis through detailed metric computation and visualization.

4 Results

4.1 Agreement and Accuracy Across Splits

Table 2–4 report mean Quadratic Weighted Kappa (QWK) results for each leave-one-grade-out split. When trained on grades 9 & 10 and tested on grade 11 (Table 2), mean QWK ranged from 0.889 to 0.893 across the best hyper-parameter settings. Similar stability was observed for the other splits: training on grades 9 & 11 (Table 3) yielded QWK near 0.892, and training on grades 10 & 11 (Table 4) yielded QWK near 0.893. Exact accuracy, summarized in Tables 5–7, consistently hovered around 0.68–0.69 for all splits. Aggregating across splits (Table 8) confirms mean QWK ≈ 0.89 and mean accuracy ≈ 0.68, demonstrating that two-grade training provides robust linguistic coverage for scoring the held-out grade.

4.2 Impact of Pre-training

Supplementary pre-training did not yield a uniform advantage; effects depended on split, α, and trait. At $\alpha = 1.0$, with the strongest regularization, the no-pretraining baseline (Group I) achieved the highest QWK across all traits in the train 9&11 → test 10 design (Table 3). In other splits, leadership shifted: for train 9&10 → test 11 (Table 2), Group II (ASAP-pretrained) led Organization, Prompt Task, and Development, while Group III (SWAS-pretrained) led Language Use; for train 10&11 → test 9 (Table 4), Group II dominated most traits, with all groups performing similarly on Prompt Task. At lower α, leadership occasionally changed by trait but without clear consistency. Overall, even at $\alpha = 1.0$, where performance was most stable, relative rankings fluctuated across splits, showing that train–test design substantially shaped outcomes and prevented conclusive judgments of model performance.

4.3 Trait-Level Performance

Figures 5–7 plot QWK trajectories across max_iter for each trait in the three splits. Organization consistently scored highest (peak QWK ≈ 0.93), followed by Language Use and Development (≈ 0.90), with Prompt Task trailing (≈ 0.86). Even the most challenging trait, Prompt Task, exceeded the operational QWK threshold of 0.70 in every configuration (Figures 5–7). These rankings held irrespective of pre-training group, confirming a stable hierarchy of trait difficulty. Macro-average F1 scores per trait across splits are summarized in Table 9.

4.4 Precision, Recall, and F1 by Score Point

Figures 11–12 show per-score precision, recall, and F1 for the 9+10→11 split (and supplementary figures for the other splits). All groups peak at the extreme scores (1 & 5) and dip in the mid-range (2–4) for precision and recall, reflecting both data imbalance and inherent scoring difficulty. No group gains a systematic edge from extra pre-training.

4.5 Hyper-parameter Fine-tuning

Hyper-parameter sweeps confirm that regularization strength α = 1.0 combined with at least 500 training iterations produces the most stable and highest-performing models. Early stopping at 100 iterations dropped QWK by roughly 0.005–0.006 (see Tables 2–4), and increasing beyond 1,000 iterations yielded diminishing returns. Lower α values (0.01, 0.10) led to mild over-fitting, indicated by higher training QWK but lower test QWK and increased variance across seeds.

4.6 Oversampling Correction

Although RandomOverSampler successfully equalized class frequencies (supplemental bar plots), oversampling did not materially improve modeling outcomes. Precision and recall at rare score points improved slightly in some configurations, but aggregate QWK and accuracy remained unchanged or marginally worse when oversampled sets were used for training.

4.7 Macro-Average F_1 Summary

To condense all per-score results, Table 9 reports the macro-averaged F_1 (mean over score points 1–5) for each trait, group, and leave-one-grade-out split. Together, Tables 2–9 and Figures 5–12 show that a baseline DistilBERT (G1) fine-tuned on two-grade SWAS essays yields high agreement (QWK ≈ 0.89), accuracy (≈ 0.68), and F_1 across traits, without the need for extra pre-training or oversampling.

Overall, these results demonstrate that a baseline DistilBERT model—fine-tuned exclusively on two-grade SWAS data—achieves high agreement (QWK ≈ 0.89) and accuracy (≈ 0.68) across grade splits and analytic traits without requiring additional pre-training or extensive oversampling (Tables 2–9, Figures 5–12).

5 Discussion

The stability of model performance across all three leave-one-grade-out splits suggests that DistilBERT's pre-trained language representations are highly adaptable to essay scoring—even without extensive prompt-specific data. Training on any two adjacent grades yielded nearly identical agreement (QWK ≈ 0.89), exact accuracy (~ 0.68), and Adjacent Accuracy (> 98 %), confirming that essays from two grades supply sufficient linguistic and rhetorical variety to generalize to a held-out grade.

Perhaps most surprisingly, neither large-scale in-domain pre-training on ASAP nor "within-task" pre-training on a SWAS subset produced consistent gains. As Table 9's macro-average F_1 summary shows, the baseline model (Group 1) ties or outperforms both ASAP-pretrained (Group 2) and SWAS-pretrained (Group 3) variants in every trait and split. For instance, Prompt Task F_1 on 9+10→11 is 0.714 for Group 1 versus 0.706 (Group 2) and 0.698 (Group 3). This counter-intuitive result implies that when the BERT's original pretraining corpus is already massive and representative enough, further pre-training can introduce stylistic noise or domain drift instead of strengthening task alignment.

Hyper-parameter analysis reinforces the need for careful regularization and adequate training steps. Models with α = 1.0 and at least 500 (ideally 1,000) iterations consistently achieve the highest and most reproducible QWK. Lower α values permit mild over-fitting—evident in higher training QWK but lower test QWK—while very short runs (100 iterations) leave a nontrivial 0.005–0.006 QWK gap compared to longer runs.

Trait-level performance reveals a stable hierarchy of difficulty. Organization is most easily predicted (peak QWK ≈ 0.93), followed by Development and Language Use (≈ 0.90), with Prompt Task trailing (≈ 0.86). Crucially, even the most challenging trait exceeds the operational QWK threshold of 0.70, indicating that all four analytic dimensions can be scored with confidence.

Finally, oversampling to correct class imbalance offered minimal benefit. Although frequency distributions were equalized, aggregate QWK, accuracy, and micro-F_1 remained flat or dipped slightly, suggesting that model capacity and the breadth of cross-grade coverage outweigh precise score-level balance when fine-tuning transformer embeddings.

Taken together, these findings validate a lightweight, prompt-agnostic AES pipeline: fine-tune a standard DistilBERT checkpoint on a representative two-grade corpus with $\alpha = 1.0$ and 500–1,000 iterations, and skip costly intermediate pre-training or complex oversampling. This approach simplifies system development, reduces computational overhead, and still delivers robust, reproducible scoring across multiple writing traits and grade levels.

6 Conclusion

This study has demonstrated that prompt-generic automated essay scoring (AES) can be achieved efficiently by fine-tuning DistilBERT on representative two-grade essay sets, without the need for extensive prompt-specific pre-training or elaborate data balancing. Across three leave-one-grade-out splits and nine hyper-parameter configurations, baseline DistilBERT models consistently achieved strong agreement (QWK ≈ 0.89), exact accuracy (~0.68), and adjacent accuracy (> 98%). These results challenge conventional assumptions, showing that DistilBERT's general-domain representations suffice for robust scoring when paired with straightforward fine-tuning.

A particularly striking finding was the observation of "knowledge collapse": applying supplementary pre-training settings to overwrite existing parameters paradoxically diminished downstream scoring performance. This counter-intuitive effect—where newly acquired "knowledge" impaired rather than enhanced task ability—underscores the critical need to avoid equating machine learning processes with human learning, and suggests that care must be taken to preserve previously learned representations during transfer learning.

From a practical standpoint, clear hyper-parameter guidelines have emerged: a regularization strength of $\alpha = 1.0$ and a training horizon of 500–1 000 iterations reliably maximize performance and model stability. This simple recipe offers a low-overhead path to deploying AES in educational contexts, minimizing both computational cost and engineering complexity.

Nonetheless, certain limitations temper the generalizability of these conclusions. The within-task pre-training set was limited to 500 essays covering a single prompt per grade, which may have constrained the potential benefits of task-specific pre-training. Exclusion of 1203 handwritten essays—due to transcription challenges—introduced moderate score-level imbalance and restricted the training corpus's representativeness. Finally, employing a single scoring rubric across all prompts may have simplified the generalization challenge.

To address these gaps, future work should explore larger, more diverse essay collections spanning multiple prompts, genres, and rubrics to assess how prompt variety and score distribution affect adaptability. Alternative machine-learning frameworks beyond ordinal logistic regression—such as ensemble methods or neural classifiers—should be evaluated for further performance gains. It will also be important to develop transfer-learning strategies that explicitly guard against "knowledge collapse," preserving core representations while incorporating new domain information. Integrating advanced handwriting recognition technologies remains essential for inclusive AES that covers all response formats.

In closing, this research provides compelling evidence that a lightly fine-tuned DistilBERT model can serve as a scalable, reliable AES engine for formative writing practice, dramatically reducing the data and computational burdens. By recommending concrete hyper-parameter settings and highlighting the nuanced effects of further pre-training, this work lays a pragmatic foundation for the next generation of accessible, robust AES tools in K-12 education.

A Appendices

ASAP Dataset	Topics
Prompt 1	The effects computers have on people
Prompt 2	Censorship in the libraries
Prompt 3	Respond to an extract about how the features of a setting affected a cyclist
Prompt 4	Explain why an extract from *Winter Hibiscus* by Minfong Ho was concluded in the way the author did
Prompt 5	Describe the mood created by the author in an extract from *Narciso Rodriguez* by Narciso Rodriguez
Prompt 6	The difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there
Prompt 7	Write a story about patience
Prompt 8	The benefits of laughter

Table 1: Topics of Eight Prompts in ASAP Dataset

	Fine-tuning Parameter: Alpha is set to be the same across three groups No. of essays for training=2251 No. of essays for test=1046						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	QWK in training	QWK in testing	QWK in training	QWK in testing	QWK in training	QWK in testing
1.0	Language Use	0.940	0.887	0.940	0.876	0.943	0.922
	Organization	0.938	0.851	0.944	0.920	0.946	0.908
	Prompt Task	0.939	0.914	0.940	0.922	0.942	0.917
	Development	0.935	0.897	0.938	0.925	0.940	0.903
0.1	Language Use	0.944	0.873	0.941	0.897	0.938	0.902
	Organization	0.945	0.930	0.942	0.928	0.944	0.893
	Prompt Task	0.937	0.895	0.930	0.896	0.940	0.847
	Development	0.938	0.884	0.939	0.923	0.938	0.876
0.01	Language Use	0.938	0.851	0.934	0.895	0.931	0.886
	Organization	0.940	0.915	0.935	0.912	0.936	0.884
	Prompt Task	0.931	0.863	0.930	0.896	0.925	0.881
	Development	0.933	0.867	0.933	0.904	0.931	0.839

Table 2: Mean QWK Results vs. Alpha for Train on Grade 9&10 and Test on Grade 11

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2170 No. of essays for test=1127						
	maxiter=500	Group I (No Further Pre-training)		Group II (Further Pre-training with ASAP Essays)		Group III (Further Pre-training with SWAS Essays)	
Alpha	Trait	QWK in training	QWK in testing	QWK in training	QWK in testing	QWK in training	QWK in testing
1	Language Use	0.935	0.922	0.938	0.918	0.941	0.871
	Organization	0.939	0.935	0.944	0.879	0.941	0.853
	Prompt Task	0.941	0.921	0.946	0.911	0.948	0.871
	Development	0.934	0.926	0.940	0.903	0.937	0.871
0.1	Language Use	0.939	0.914	0.936	0.909	0.937	0.882
	Organization	0.941	0.932	0.944	0.827	0.941	0.865
	Prompt Task	0.944	0.919	0.945	0.899	0.945	0.879
	Development	0.936	0.923	0.941	0.895	0.936	0.893
0.01	Language Use	0.933	0.890	0.931	0.886	0.928	0.867
	Organization	0.938	0.924	0.938	0.789	0.935	0.855
	Prompt Task	0.940	0.907	0.939	0.887	0.938	0.858
	Development	0.935	0.902	0.936	0.869	0.929	0.877

Table 3: Mean QWK Results vs. Alpha for Train on Grade 9&11 and Test on Grade 10

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2173 No. of essays for test=1124						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	QWK in training	QWK in testing	QWK in training	QWK in testing	QWK in training	QWK in testing
1.0	Language Use	0.946	0.911	0.948	0.913	0.949	0.908
	Organization	0.949	0.922	0.948	0.927	0.949	0.925
	Prompt Task	0.936	0.890	0.940	0.891	0.939	0.892
	Development	0.938	0.903	0.941	0.917	0.940	0.888
0.1	Language Use	0.949	0.909	0.947	0.901	0.946	0.912
	Organization	0.949	0.920	0.945	0.914	0.946	0.900
	Prompt Task	0.936	0.824	0.938	0.858	0.933	0.854
	Development	0.941	0.900	0.940	0.902	0.938	0.870
0.01	Language Use	0.945	0.894	0.941	0.888	0.939	0.874
	Organization	0.944	0.910	0.938	0.881	0.936	0.872
	Prompt Task	0.926	0.815	0.928	0.852	0.924	0.840
	Development	0.936	0.872	0.937	0.881	0.927	0.834

Table 4: Mean QWK Results vs. Alpha for Train on Grade 10 & 11 and Test on Grade 9

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2251 No. of essays for test=1046						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing
1	Language Use	0.793	0.722	0.792	0.575	0.799	0.760
	Organization	0.791	0.760	0.795	0.697	0.795	0.722
	Prompt Task	0.766	0.718	0.769	0.699	0.776	0.718
	Development	0.762	0.722	0.773	0.738	0.775	0.722
0.1	Language Use	0.805	0.680	0.798	0.667	0.783	0.702
	Organization	0.793	0.702	0.789	0.741	0.789	0.680
	Prompt Task	0.761	0.682	0.761	0.701	0.764	0.682
	Development	0.774	0.675	0.776	0.746	0.771	0.675
0.01	Language Use	0.786	0.658	0.773	0.681	0.755	0.663
	Organization	0.779	0.663	0.763	0.707	0.765	0.658
	Prompt Task	0.744	0.627	0.731	0.659	0.725	0.627
	Development	0.762	0.633	0.765	0.705	0.758	0.633

Table 5: Mean Accuracy Results vs. Alpha Configurations for Train on Grade 9&10 and Test on Grade 11

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2170 No. of essays for test=1127						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing
1.0	Language Use	0.781	0.762	0.789	0.752	0.799	0.633
	Organization	0.793	0.760	0.805	0.627	0.795	0.595
	Prompt Task	0.781	0.704	0.796	0.684	0.803	0.603
	Development	0.769	0.736	0.788	0.697	0.778	0.643
0.1	Language Use	0.792	0.738	0.782	0.725	0.784	0.672
	Organization	0.794	0.755	0.802	0.547	0.791	0.621
	Prompt Task	0.790	0.702	0.794	0.659	0.791	0.630
	Development	0.774	0.732	0.794	0.683	0.781	0.684
0.01	Language Use	0.773	0.662	0.767	0.670	0.754	0.650
	Organization	0.785	0.732	0.786	0.517	0.775	0.606
	Prompt Task	0.776	0.674	0.771	0.637	0.791	0.630
	Development	0.783	0.675	0.788	0.640	0.778	0.659

Table 6: Mean Accuracy Results vs. Alpha Configurations for Train on Grade 9&11 and Test on Grade 10

	Fine-tuning Hyperparameter: Alpha is set to be the same across three groups No. of essays for training=2173 No. of essays for test=1124						
	maxiter=1000	Group I (No Further Pre-training)		Group II (Further Pre-training With ASAP Essays)		Group III (Further Pre-training With SWAS Essays)	
Alpha	Trait	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing	Accuracy in training	Accuracy in testing
1.0	Language Use	0.806	0.734	0.808	0.740	0.813	0.728
	Organization	0.800	0.739	0.794	0.759	0.801	0.750
	Prompt Task	0.760	0.698	0.768	0.673	0.768	0.691
	Development	0.767	0.678	0.776	0.740	0.773	0.631
0.1	Language Use	0.814	0.726	0.804	0.722	0.801	0.681
	Organization	0.797	0.736	0.783	0.735	0.791	0.699
	Prompt Task	0.758	0.576	0.760	0.612	0.751	0.627
	Development	0.776	0.687	0.770	0.725	0.765	0.609
0.01	Language Use	0.797	0.677	0.785	0.686	0.775	0.611
	Organization	0.797	0.736	0.756	0.669	0.760	0.588
	Prompt Task	0.729	0.570	0.730	0.618	0.718	0.594
	Development	0.767	0.625	0.761	0.701	0.739	0.563

Table 7: Mean Accuracy Results vs. Alpha Configurations for Train on Grade 10&11 and Test on Grade 9

Train → Test	Mean QWK	Mean Accuracy
G9 & G10 → G11	0.893	0.687
G10 & G11 → G9	0.889	0.679
G9 & G11 → G10	0.892	0.673

Table 8: Average Performance by Leave-One-Grade-Out Split

Trait	Model Group	9+10 → 11	9+11 → 10	10+11 → 9
Prompt Task	Baseline (G1)	0.714	0.714	0.645
	ASAP-pretrained (G2)	0.706	0.702	0.630
	SWAS-pretrained (G3)	0.698	0.710	0.626
Organization	Baseline (G1)	0.753	0.741	0.648
	ASAP-pretrained (G2)	0.742	0.725	0.622
	SWAS-pretrained (G3)	0.741	0.730	0.642
Development	Baseline (G1)	0.714	0.722	0.705
	ASAP-pretrained (G2)	0.704	0.710	0.695
	SWAS-pretrained (G3)	0.698	0.716	0.686
Language Use	Baseline (G1)	0.767	0.773	0.762
	ASAP-pretrained (G2)	0.758	0.764	0.752
	SWAS-pretrained (G3)	0.753	0.760	0.740

Table 9: Macro-Average F_1 by Trait, Model Group, and Leave-One-Grade-Out Split

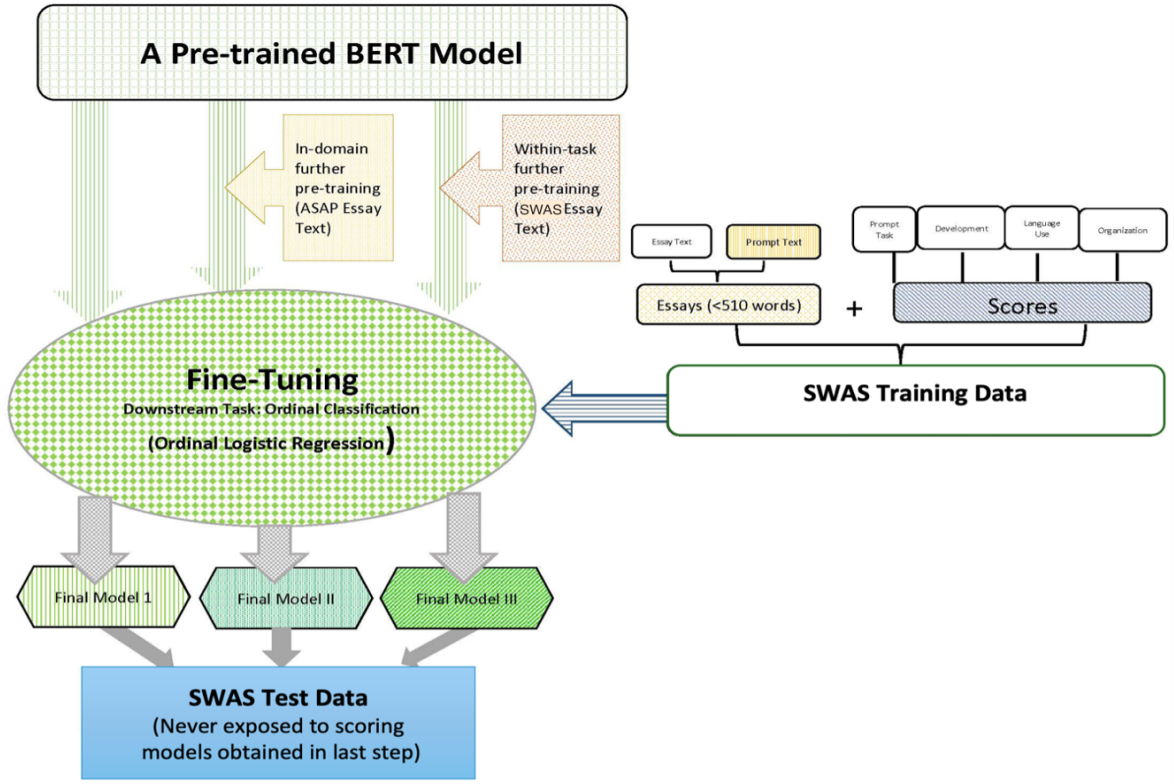

Figure 1: Research Design of the Study

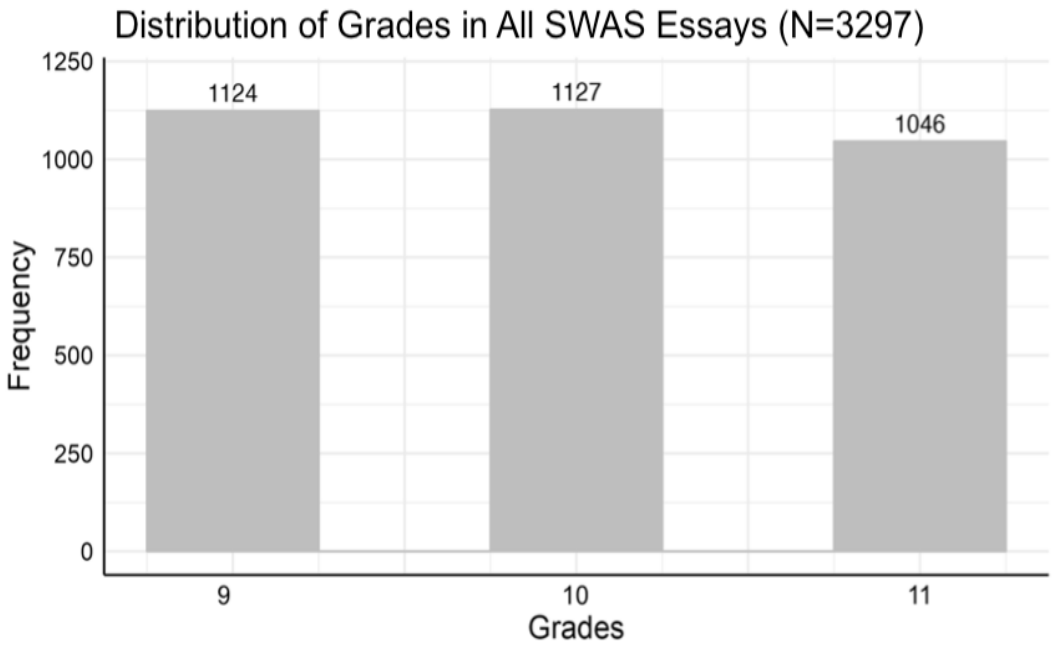

Figure 2: Grade Level Distribution of Available SWAS Essays

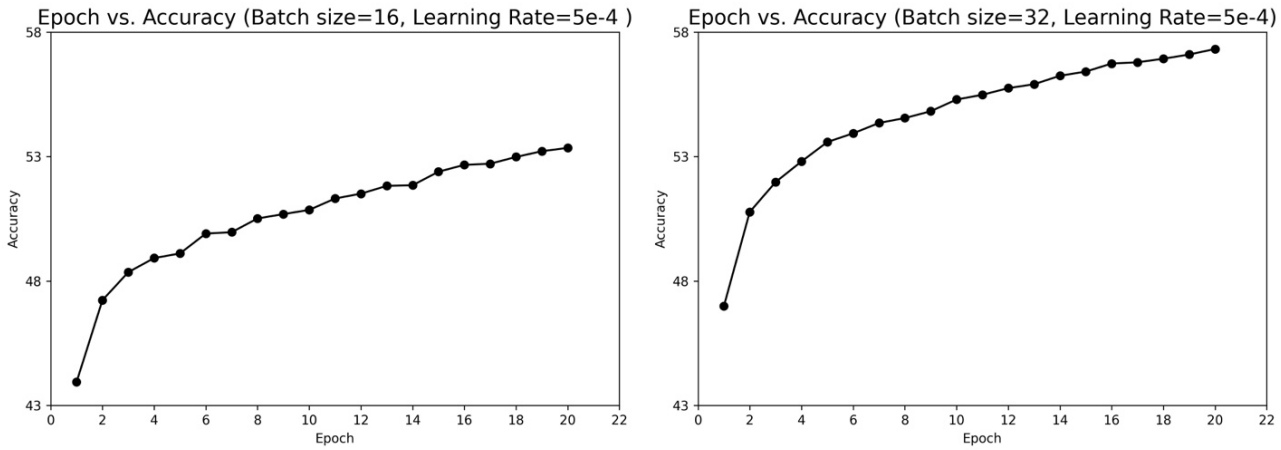

Figure 3: Accuracy in Further Pre-training with ASAP Essays

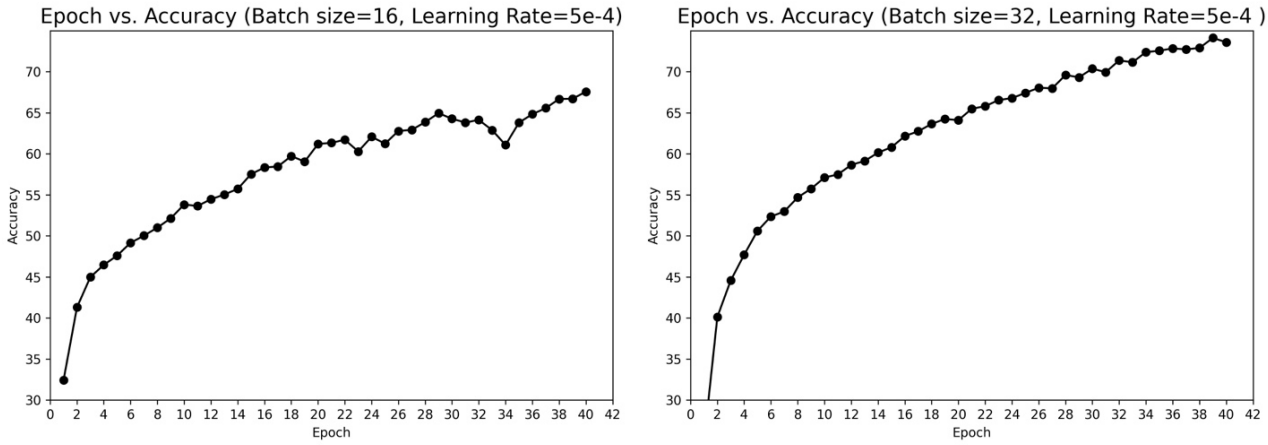

Figure 4: Accuracy vs. Epoch in Further Pre-training with 500 SWAS Essays

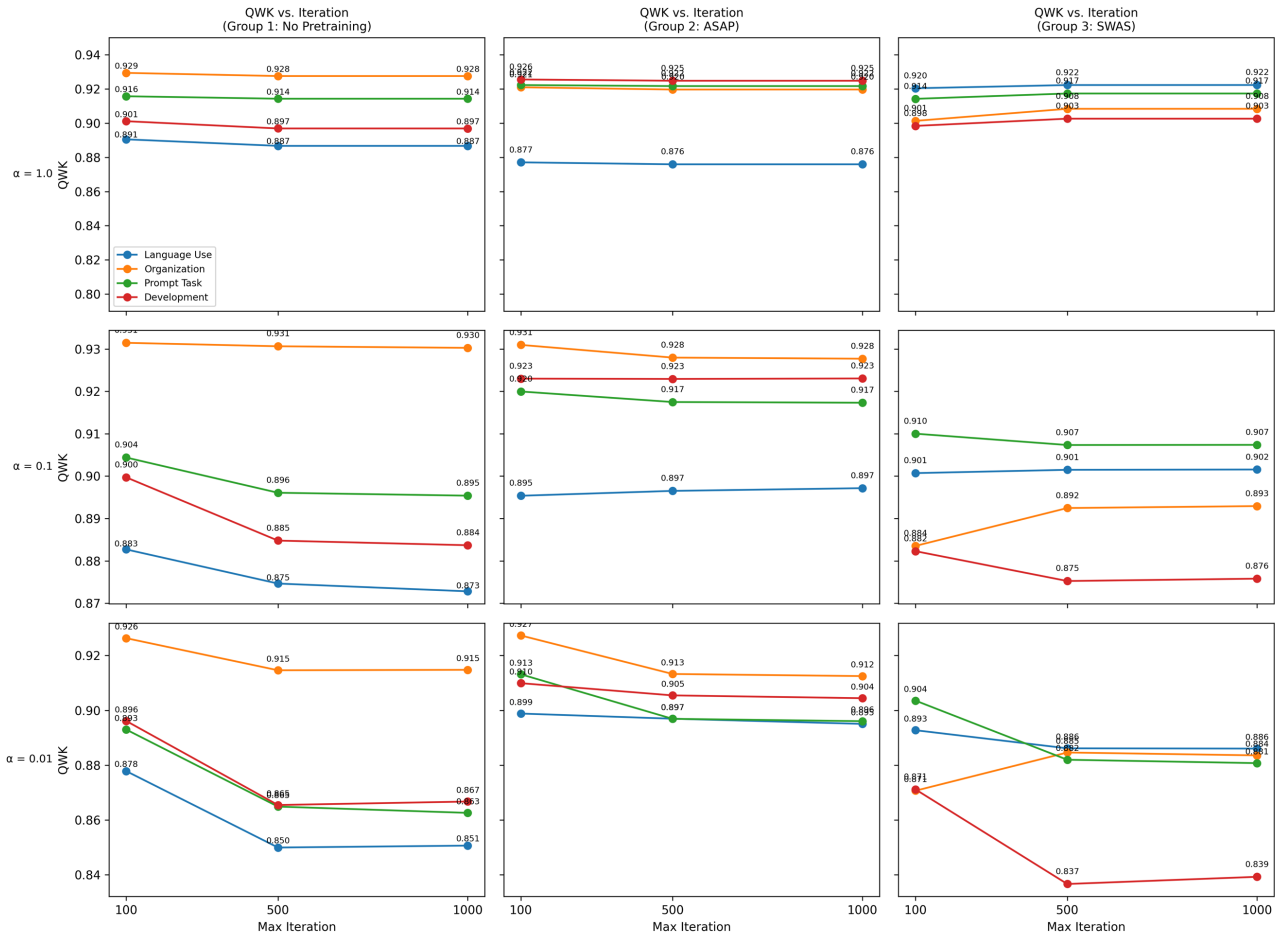

Figure 5: Comparisons of Mean QWK vs. Maxiter in 4 Trait Scores for Train on Grade 9&10 and Test on Grade 11

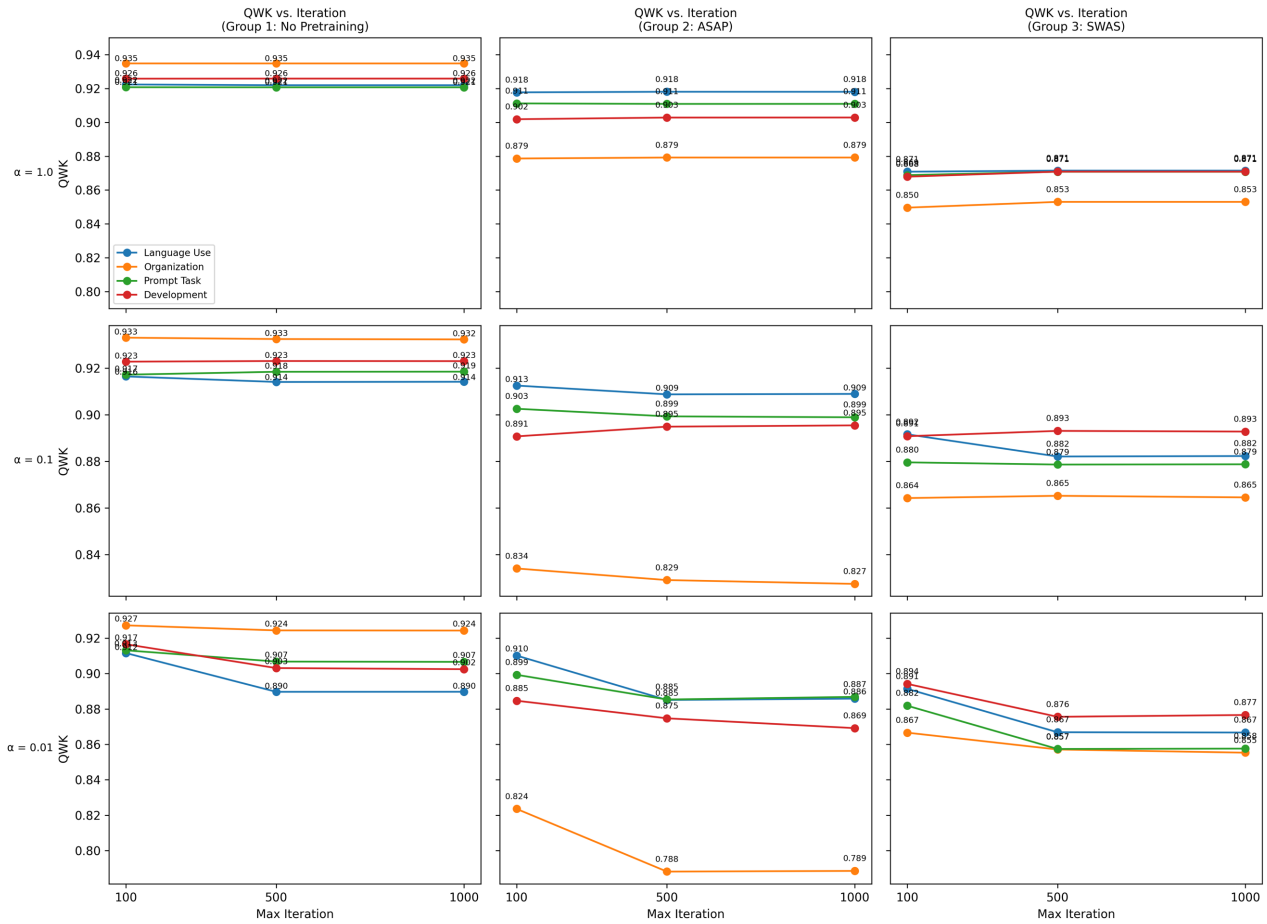

Figure 6: Comparisons of Mean QWK vs. Maxiter in 4 Trait Scores for Train on Grade 9&11 and Test on Grade 10

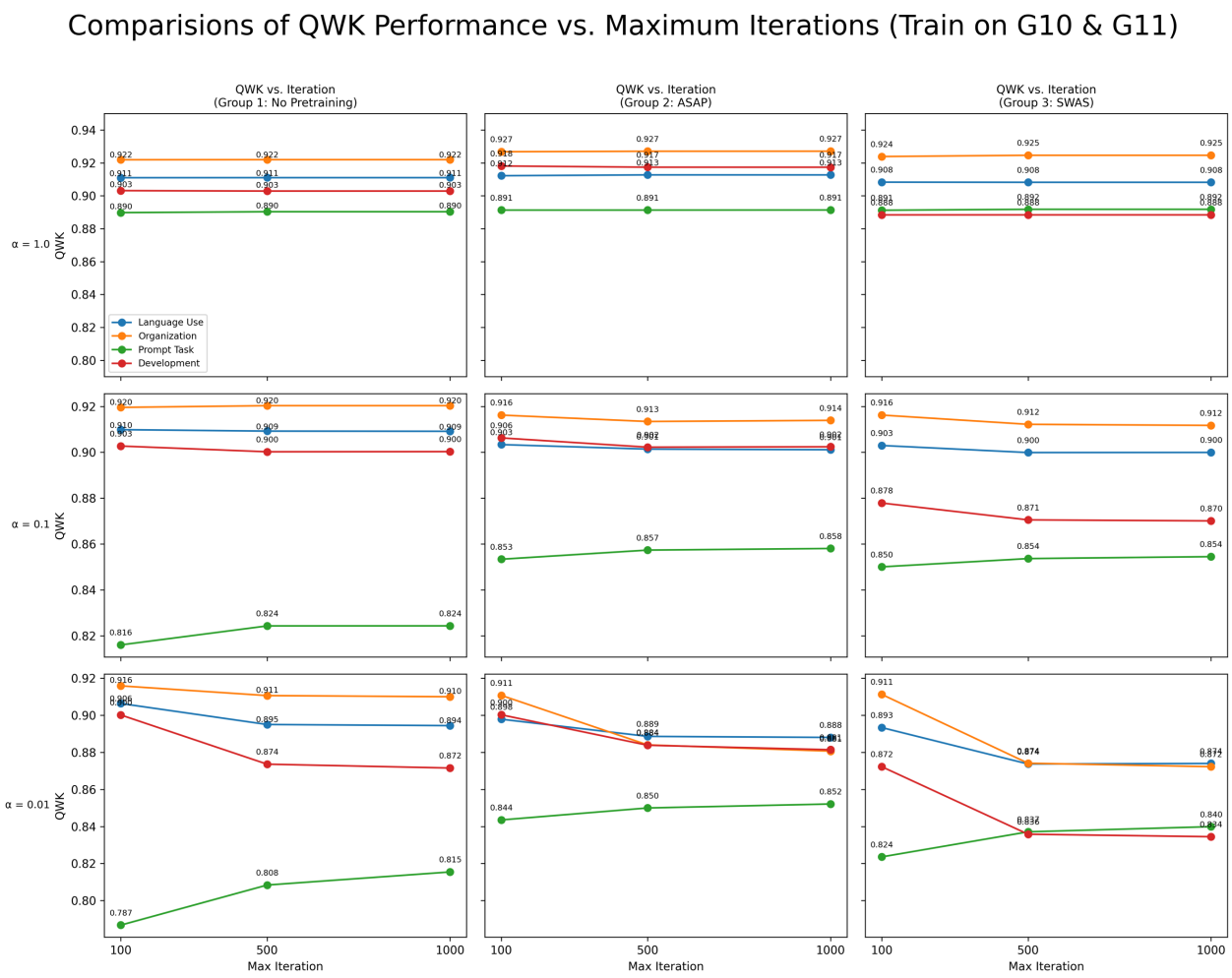

Figure 7: Comparisons of Mean QWK vs. Maxiter in 4 Trait Scores for Train on Grade 10&11 and Test on Grade 9

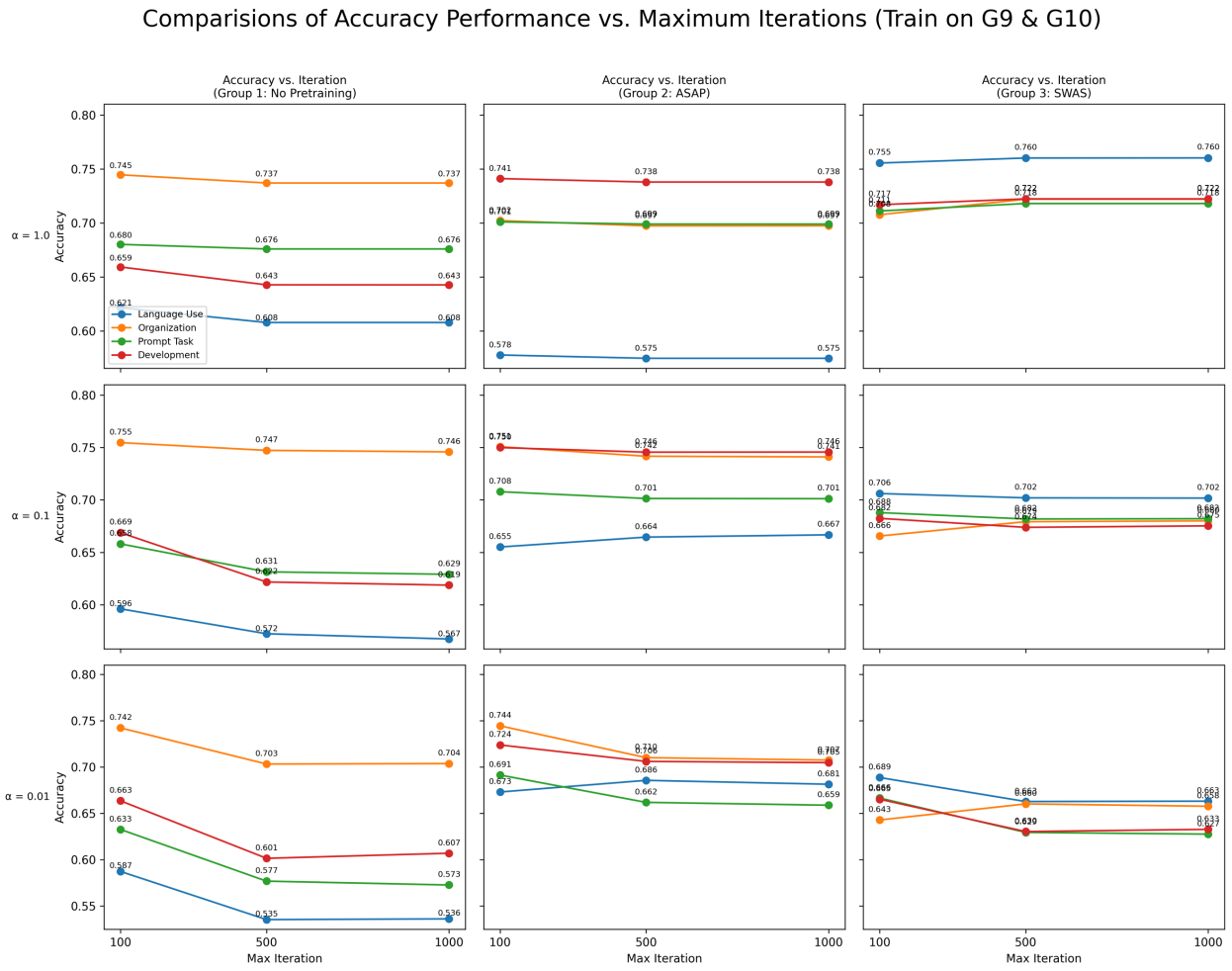

Figure 8: Comparisons of Mean Accuracy Performance vs. Maxiter in 3 Groups for Train on Grade 9&10 and Test on Grade 11

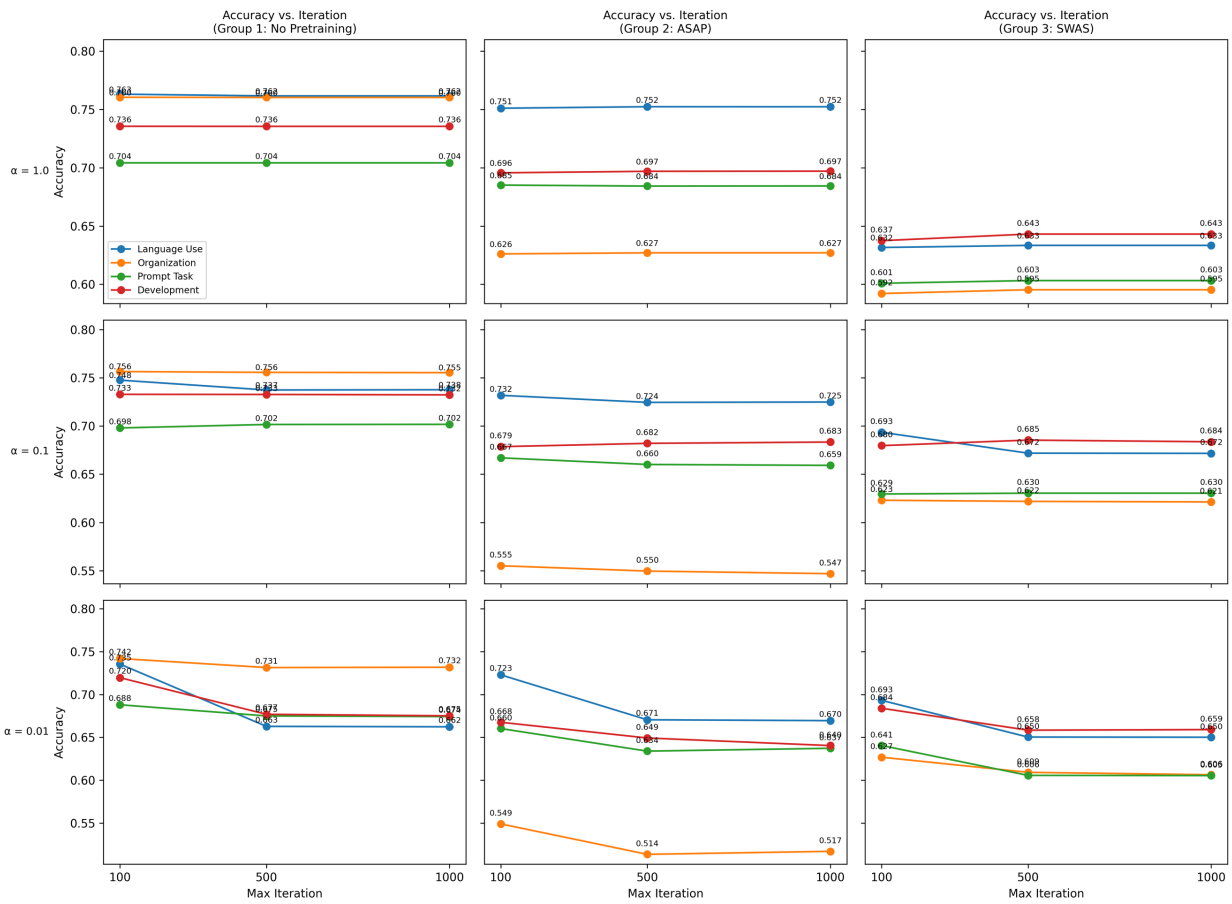

Figure 9: Comparisons of Mean Accuracy Performance vs. Maxiter in 3 Groups for Train on Grade 9&11 and Test on Grade 10

Comparisions of Accuracy Performance vs. Maximum Iterations (Train on G10 & G11)

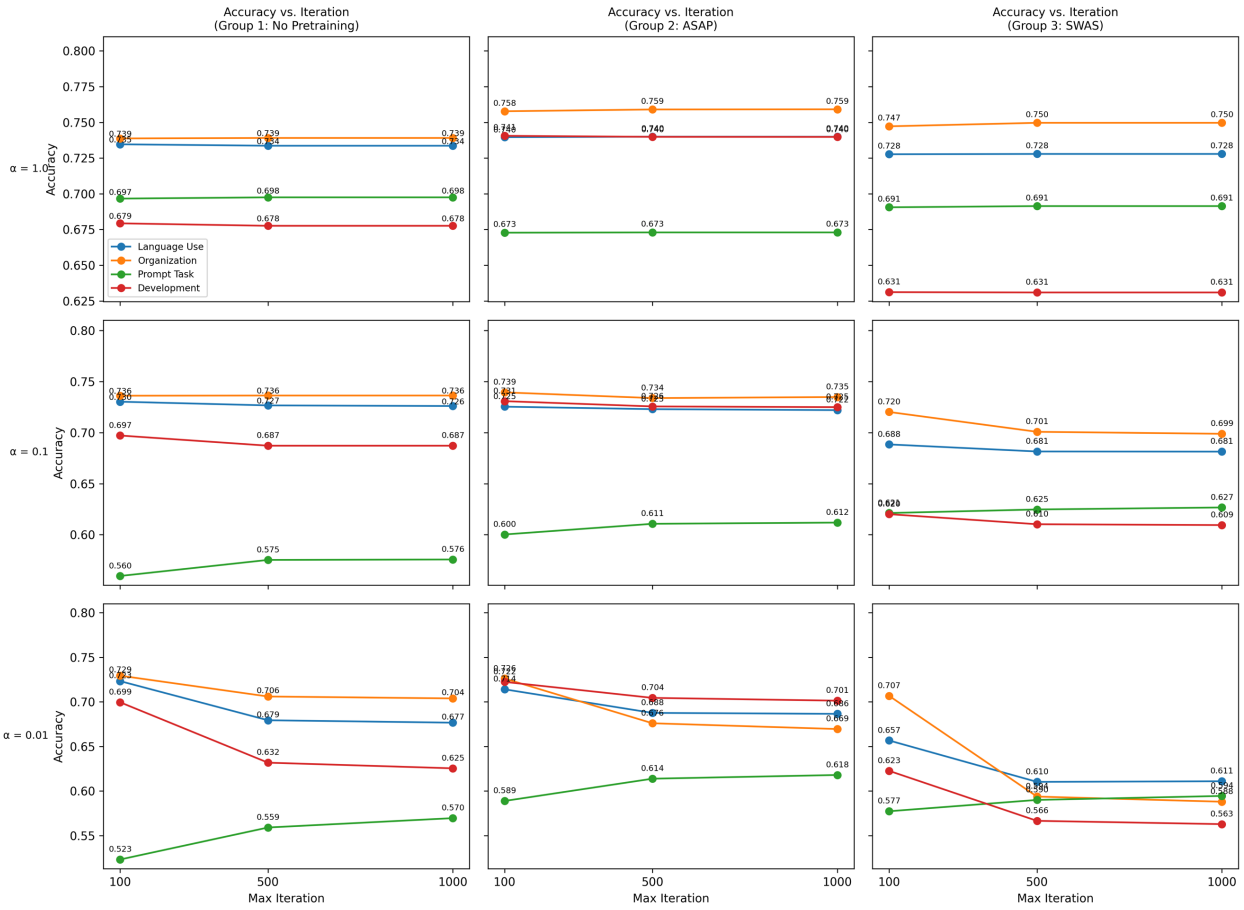

Figure 10: Comparisons of Mean Accuracy Performance vs. Maxiter in 3 Groups for Train on Grade 10&11 and Test on Grade 9

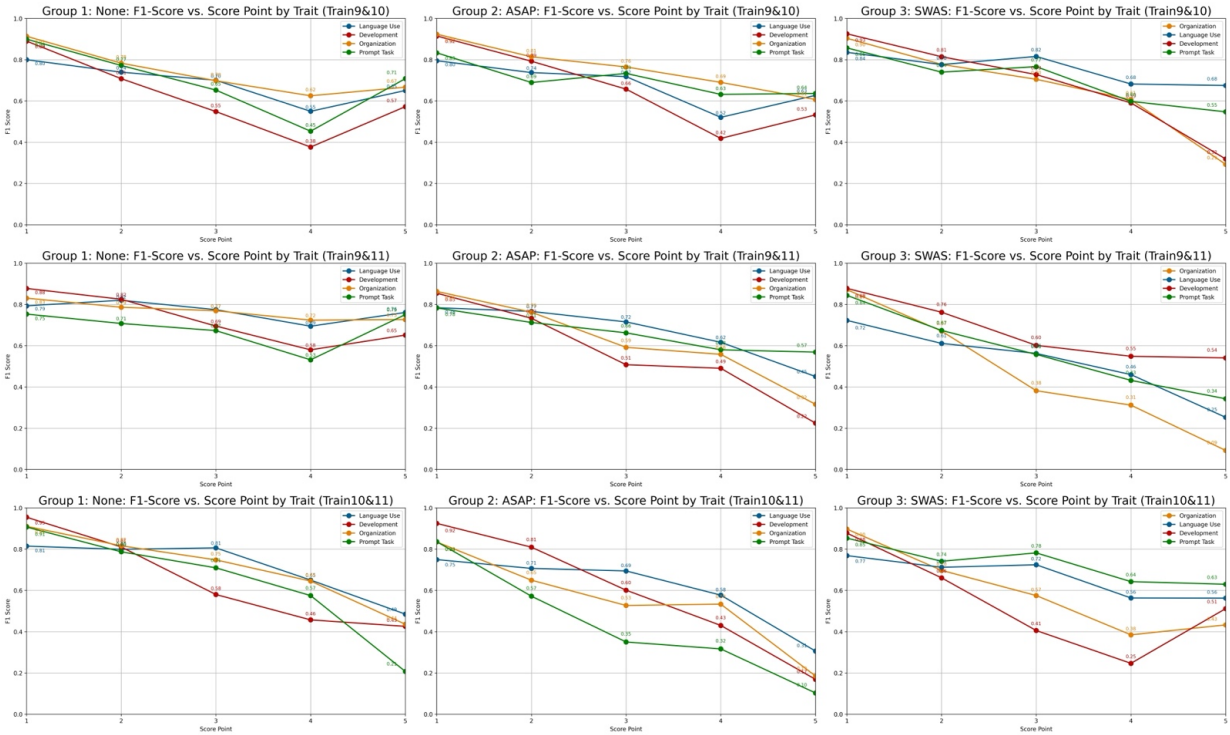

Figure 11: Mean F1-scores in All Score Levels of the 3 Groups across Trait Scores

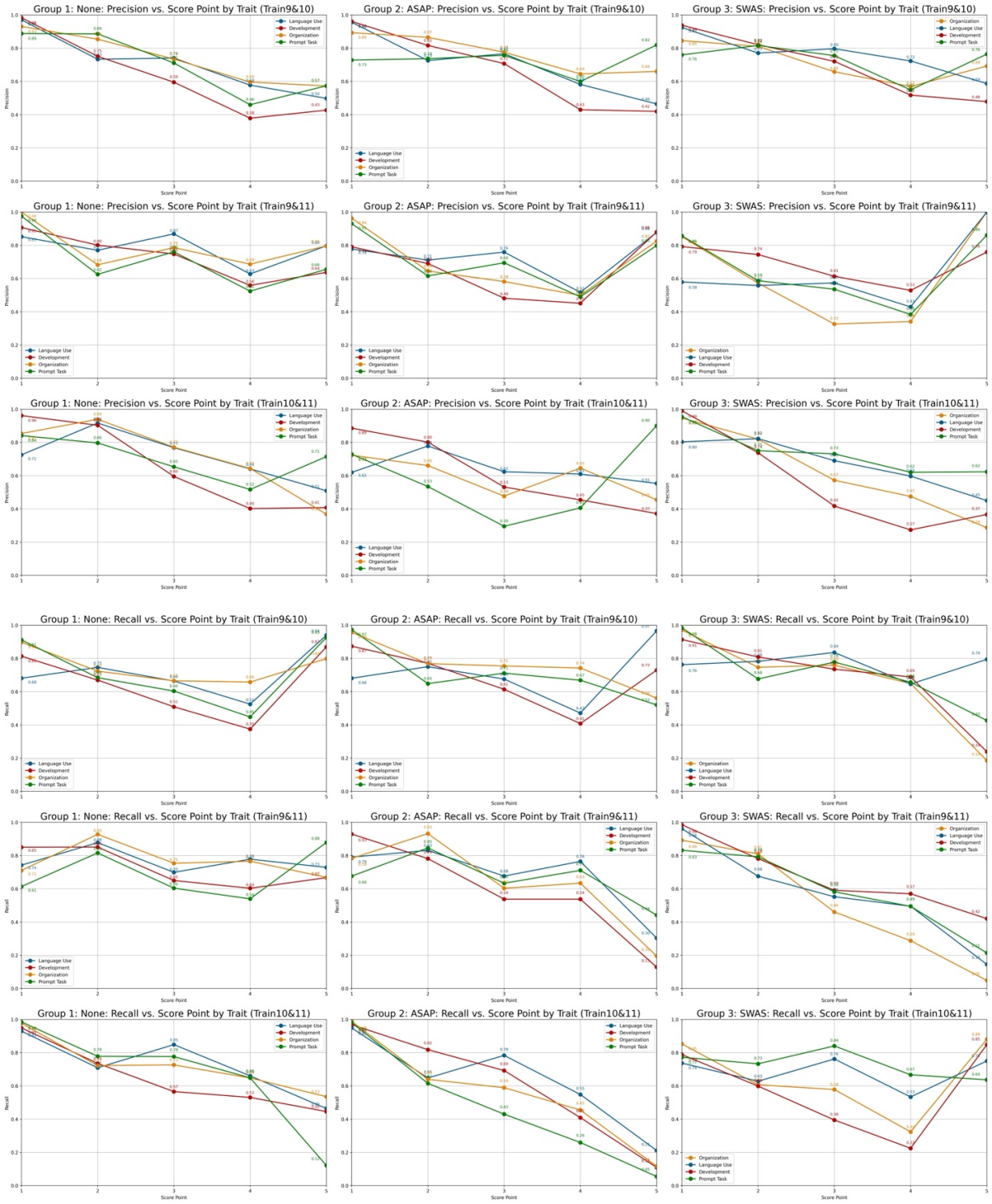

Figure 12: Mean Precision and Recall in All Score Levels of the 3 Groups across Trait Scores

References

Adepoju, A., and K. Adeleke. 2010. Ordinal logistic regression model: An application to pregnancy outcomes. *Journal of Mathematics and Statistics* 6:279–285.

Ajafabadi, M. M. N., F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2:1–21.

Allwright, S. 2022. What is a good F1 score and how do I interpret it? *stephenallwright.com*. Available at https://stephenallwright.com/good-f1-score/

American Educational Research Association, American Psychological Association, and NCME. 2014. Standards for educational and psychological testing. *American Educational Research Association*.

Anyoha, R. 2017. The history of artificial intelligence. Available at https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Attali, Y., and J. Burstein. 2006. Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment* 4(3).

Ayari, R. 2020. NLP: Word embedding techniques demystified—Bag-of-Words vs TF-IDF vs Word2Vec vs Doc2Vec vs Doc2VecC. *towardsdatascience.com*. Available at https://towardsdatascience.com/nlp-embedding-techniques-51b7e6ec9f92

Baccianella, S., A. Esuli, and F. Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications*, 283–287. IEEE.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* 3(Feb):1137–1155.

Buhl, N. 2023. F1 score in machine learning. *encord.com*. Available at https://encord.com/blog/f1-score-in-machine-learning/

Calvo, M. R. 2018. Dissecting BERT Part 1: The encoder. *medium.com*. Available at https://medium.com/dissecting-bert/dissecting-bert-part-1-d3c3d495cdb3

Cao, Y., H. Jin, X. Wan, and Z. Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1011–1020.

Center for Applied Linguistics. 2018. Annual technical report for ACCESS for ELLs 2.0 English language proficiency test, series 401 online, 2016–2017 administration (WIDA consortium annual technical report no. 13B). Washington, DC.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4):213–220.

Collobert, R., and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, vol. 1, 160–167.

Conneau, A., and D. Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1699–1704. European Language Resources Association.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Delua, J. 2021. Supervised versus unsupervised learning: What's the difference? Available at https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Devlin, J., and M.-W. Chang. 2018. Open sourcing BERT: State-of-the-art pre-training for natural language processing. *Google AI Language*. Available at https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019, Volume 1: Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dhami, D. 2020. Understanding BERT — word embeddings. *medium.com*. Available at https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca

Diederich, P. B. 1974. Measuring growth in English. *National Council of Teachers of English*.

Dodge, J., G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. 2020. Fine-tuning pre-trained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Dong, L., N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems* 32.

Donges, N. 2023. Introduction to natural language processing (NLP): The ultimate goal of natural language processing is to help computers understand language as well as we do. *builtin.com*. Available at https://builtin.com/data-science/introduction-nlp

Faigley, L. 1985. Assessing writers' knowledge and processes of composing. *Ablex Publishing Corporation*.

Feathers, T. 2019. Flawed algorithms are grading millions of students' essays. *Vice*. Available at https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays

Fiacco, J., E. Cotos, and C. Rosé. 2019. Towards enabling feedback on rhetorical structure with neural sequence models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 310–319. Association for Computing Machinery.

Fleiss, J., B. Levin, and M. Paik. 2004. *Statistical Methods for Rates and Proportions*, 3rd edn. Wiley-Interscience.

Fleiss, J. L., and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3):613–619.

Foltz, P. W., L. A. Streeter, K. E. Lochbaum, and T. K. Landauer. 2013. Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 68–88. Routledge.

Gere, A. R. 1980. Written composition: Toward a theory of evaluation. *College English* 42(1):44–58.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Google Inc. 2020. BERT. *Hugging Face*. Available at https://huggingface.co/docs/transformers/model_doc/bert

Graham, P., and R. Jackson. 1993. The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology* 46(9):1055–1062.

Hamner, B., and M. D. Shermis. 2013. Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 314–346. Routledge.

Hearst, M. A. 2000. The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications* 15(5):22–37.

High, P. 2017. Carnegie Mellon dean of computer science on the future of AI. *Forbes*. Available at https://www.forbes.com/sites/peterhigh/2017/10/30/carnegie-mellon-dean-of-computer-science-on-the-future-of-ai/?sh=1e8c39b72197

Hinton, G., O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Houlsby, N., A. Giurgiu, S. Jastrzębski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2790–2799. PMLR.

Howard, J., and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Long Papers*, 328–339. Association for Computational Linguistics.

huggingface.co. 2020. Summary of the Tokenizers—transformers 4.3.0 documentation. *huggingface.co*. Available at https://huggingface.co/docs/transformers/tokenizer_summary

Huh, M., P. Agrawal, and A. A. Efros. 2016. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Katyal, N. K. 2003. The promise and precondition of educational autonomy. *Hastings Constitutional Law Quarterly* 31:557–613.

Keskar, N. S., D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint abs/1609.04836*.

Khanna, C. 2021. WordPiece: Subword-based tokenization algorithm: Understand subword-based tokenization algorithm used by state-of-the-art NLP models — WordPiece. *towardsdatascience.com*. Available at https://towardsdatascience.com/wordpiece-subword-based-tokenization-algorithm-1fbd14394ed7

Komatsuzaki, A. 2019. One epoch is all you need. *arXiv preprint arXiv:1906.06669*.

Kortschak, H. 2020. Attention and transformer models: A complex algorithm, simply explained. *towardsdatascience.com*. Available at https://towardsdatascience.com/attention-and-transformer-models-fe667f958378

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Landauer, T. K., and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

Landauer, T. K., D. Laham, and P. Foltz. 2000. The Intelligent Essay Assessor. *Intelligent Systems, IEEE* 15:27–31.

LeCun, Y., L. Bottou, G. B. Orr, and K.-R. Müller. 2002. Efficient backprop. In *Neural Networks: Tricks of the Trade*, 9–50. Springer.

Lee, C., K. Cho, and W. Kang. 2019. Mixout: Effective regularization to finetune large-scale pre-trained language models. In *International Conference on Learning Representations*. Available at https://arxiv.org/abs/1909.11299

Liu, J., Y. Xu, and L. Zhao. 2019. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

Mayfield, E., and A. Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 151–162. Association for Computational Linguistics.

Megumi, K.-M. 2003. E-rater software. *Association for Language Teaching*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* 26.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Muangkammuen, P., and F. Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. *AACL*.

Murphy, R. F. 2019. Artificial intelligence applications to support K-12 teachers and teaching. *Rand Corporation*.

Nadeem, F., H. Nguyen, Y. Liu, and M. Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications at ACL 2019*, 484–493.

National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. Common Core State Standards for English Language Arts. Washington, DC.

Oquab, M., L. Bottou, I. Laptev, and J. Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1717–1724. https://doi.org/10.1109/CVPR.2014.222

Page, E. B. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan* 47(5):238–245.

Pearson Inc. 2010. Reliable automated writing assessment. Available at https://mlm.pearson.com/northamerica/mywritinglab/educators/features/writing-practice/index.html

Pedregosa, F., F. Bach, and A. Gramfort. 2017. On the consistency of ordinal regression methods. *Journal of Machine Learning Research* 18(55):1–35.

Pennington, J., R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Persing, I., and V. Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 543–552. Association for Computational Linguistics.

Persing, I., and V. Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1384–1394.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In M. Walker, H. Ji, and A. Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, 2227–2237. Association for Computational Linguistics.

Peters, M. E., S. Ruder, and N. A. Smith. 2019. To tune or not to tune? Adapting pre-trained representations

to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 7–14. Association for Computational Linguistics.

Phang, J., T. Févry, and S. R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Rawat, W., and Z. Wang. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation* 29(9):2352–2449.

Rennie, J., and N. Srebro. 2005. Loss functions for preference levels: Regression with discrete ordered labels. *IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, Edinburgh, Scotland.

Riley, J. C. 2019. The Massachusetts Board of Elementary and Secondary Education update on automated test scoring. Massachusetts Department of Elementary and Secondary Education. Available at https://www.doe.mass.edu/bese/docs/fy2019/2019-01/spec-item2.html

Rodriguez, P. U., A. Jafari, and C. M. Ormerod. 2019. Language models and automated essay scoring. *International Journal of Assessment Tools in Education* 10(3):149–163.

Ruder, S., M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. 2015. ImageNet large-scale visual recognition challenge. *International Journal of Computer Vision* 115:211–252.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Saravia, E. 2018. Deep learning for NLP: An overview of recent trends. *medium.com*. Available at https://medium.com/dair-ai/deep-learning-for-nlp-an-overview-of-recent-trends-d0d8f40a776d

Scherer, D. L. 1985. Measuring the measurements: A study of evaluation of writing: An annotated bibliography.

Schuster, C. 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement* 64(2):243–253.

Shermis, M., J. Burstein, D. Higgins, and K. Zechner. 2010. Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, 20–26.

Shermis, M. D., and J. Burstein (eds.). 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 1st edn. Routledge.

Shermis, M., H. Mzumara, and J. Olson. 2001. On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education* 26.

Sheshadri, A. K., A. R. Vijjini, and S. Kharbanda. 2021. Wer-BERT: Automatic WER estimation with BERT in a balanced ordinal classification paradigm. *arXiv preprint arXiv:2101.05478*.

Sun, C., X. Qiu, Y. Xu, and X. Huang. 2019. How to fine-tune BERT for text classification? In *Proceedings of Chinese Computational Linguistics: 18th China National Conference*, 194–206. Springer.

Taghipour, K., and H. T. Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891.

Vanbelle, S. 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika* 81(2):399–410.

Vantage Learning. 2001. A preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments. *Vantage Learning*.

Vantage Learning. 2002. A study of IntelliMetric™ scoring for responses. *Vantage Learning*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems* 30.

Veal, L. R., and S. A. Hudson. 1983. Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English* 17(3):290–296.

White, E. M. 1985. Teaching and assessing writing: Recent advances in understanding and improving student performance. *ERIC*.

Williamson, D., X. Xi, and F. J. Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice* 31(1):2–13.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized autoregressive pre-training for language understanding. *Advances in Neural Information Processing Systems* 32.

Yannakoudakis, H., and R. Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 213–223.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 27.

Young, T., D. Hazarika, S. Poria, and E. Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13(3):55–75.

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems* 32.