### Research Summary: AI-Powered Essay Evaluation Systems

#### 1. Commercial and Institutional Tools

**Turnitin**  
Website: [https://www.turnitin.com](https://www.turnitin.com/)  
Offers rubric-based grading suggestions, grammar and citation checking. Widely adopted in academic environments with full LMS integration.

**Gradescope**  
Website: [https://www.gradescope.com](https://www.gradescope.com/)  
Supports AI-assisted grading, rubric-based scoring, assignment review, and PDF scanning. Useful for managing large-scale assessments with instructor-student feedback.

**APEUni**  
Website: <https://www.apeuni.com/en/pte/user_center/home?type=home>  
Offers essay and listening evaluation for test prep (PTE-style), including scoring of Read Aloud, Summarize Spoken Text, Re-tell Lecture, and Answer Short Question.

**Eklavvya**  
Website: [https://www.eklavvya.com](https://www.eklavvya.com/)  
Provides onscreen evaluation, AI-based response assessments, automatic grading, and paper uploads.

**Crowdmark**  
Website: [https://crowdmark.com](https://crowdmark.com/)  
Supports digital evaluation with LMS (Canvas, Moodle) integration. Allows collaborative grading, question-specific scoring, graphical reports, and cloud-based student portfolios.

**ETS (Educational Testing Service)**  
Website: [https://www.ets.org](https://www.ets.org/)  
Uses NLP via the E-rater system to analyze essays for grammar, coherence, content relevance, and structure. Applies statistical algorithms to ensure scoring reliability and simulate human grading consistency.

#### 2. Open-Source Projects and Code Repositories

**Automatic Essay Scoring**  
Repository: <https://github.com/sankalpjain99/Automatic-Essay-Scoring>  
Uses LSTM and Flask with HP Essays dataset for scoring essays. Effective for developing baseline AES models.

**EssayScore\_FYP**  
Repository: <https://github.com/tingwei3931/EssayScore_FYP>  
A final year project on neural network-based essay scoring with limitations in usability.

**DeepEssay**  
Repository: <https://github.com/Logisx/DeepEssay>  
Focused on grading IELTS-style writing tasks.

**Automated Essay Grading System**  
Repository: <https://github.com/hazem-alabiad/automated-essay-grading-system>  
Built using ML, DL, and NLP with Scikit-learn, Gensim, Keras, NLTK, etc.

**Essay Grading IELTS**  
Repository: <https://github.com/ADHIKSHA/Essay-Grading-IELTS>  
Not usable due to dependency and compatibility issues.

**Automatic Essay Grading and Feedback**  
Repository: <https://github.com/aniruddhadave/Automatic-Essay-Grading-and-Feedback>  
Emphasizes generating AI feedback with real-time scoring.

**English Language Learning Prediction with AI**  
Repository: <https://github.com/suhasmaddali/English-Language-Learning-Prediction-with-AI-and-Machine-Learning>  
Uses topic modeling, sentiment analysis, and NLP to evaluate writing quality.

#### 3. Datasets Consulted and Used

**HP Essays Dataset**  
Source: Kaggle  
Used in multiple projects including LSTM-based AES training.

**Custom English Writing Dataset**  
Link: <https://drive.google.com/file/d/1E2PVYltjXNKxYwr7IrVgjh4FL4CaIWZy/view>  
Used for retraining LSTM, GPT-2, BERT, and ML-based classifiers.

#### 4. Models and Architectures Explored

**Traditional Models:**  
Logistic Regression, Random Forest, AdaBoost, Support Vector Classifier, K-Nearest Neighbors. Accuracy ranged between 0.42 to 0.59 depending on features and data.

**Deep Learning and Neural Networks:**  
LSTM and CNN models for sequential text evaluation and feature extraction.

**Transformer-Based Models:**  
BERT, DeBERTa, RoBERTa, DistilBERT, and Deepseek. BERT models with feature augmentation showed moderate performance (~0.54–0.56). Deepseek reached as high as 0.85 accuracy on internal datasets. LLaMA models (specifically 70B variant) reached up to 0.75 and demonstrated strong generalization and feedback generation capabilities.

#### 5. Research Papers Studied

**Paper 1: Comparing AI with Human Evaluation (ChatGPT)**  
Link: <https://www.sciencedirect.com/science/article/pii/S2666920X24000274>  
ChatGPT scored reflective essays with 34.9% (single) and 72.8% (average) accuracy. Demonstrated potential of generative AI in improving scoring consistency and speed. Emphasized the need to address academic integrity and educator awareness.

**Paper 2: Hybrid AI-Based Essay Scoring in Physics Education**  
Link: <https://www.ijiet.org/vol14/IJIET-V14N6-2113.pdf>  
Combined Cosine Similarity, SVR, and KNN to evaluate students’ conceptual understanding. Accuracy reached 40.4%, aligning closely with human grading. Showed utility in STEM education assessment.

**Paper 3: Human-AI Collaboration with LLMs (GPT-3.5, GPT-4, LLaMA3)**  
Link: <https://arxiv.org/abs/2401.06431>  
Introduced dual-process human-AI grading model. Highlighted LLaMA3’s 70% accuracy and consistent, explainable feedback. Found that AI support improved human grader performance and confidence.

#### 6. Final Model Selection: LLaMA3-70B

LLaMA3-70B was chosen as the final model after comparing over 15+ ML and Transformer-based models.

It satisfies multiple core requirements:

* High accuracy (~70%) validated in academic and practical contexts.
* Excellent generalization across test types and writing styles.
* Ability to provide rubric-specific scoring and detailed improvement suggestions.
* Strong performance in low-confidence cases and ambiguous answers.
* Fully open-source with scalable deployment support.

Supporting Research: <https://arxiv.org/abs/2401.06431>