An Explainable Hierarchical Multi-Criteria Framework for Automated Presentation Quality Assessment and Personalized Feedback

Ms. A. Sumithra, MCA

Assistant Professor, Department of BCA
Mohamed Sathak Hamid College of Arts & Science for Women
Ramanathapuram, Tamil Nadu, India
Corresponding author: sumithra@mshcasw.in

Abstract

Automated assessment of presentation slides can support students and professionals, yet numerical ratings are difficult to interpret when heterogeneous slide properties are aggregated without clear justification. This paper proposes Explainable Hierarchical Presentation Assessment (EHPA), a deterministic, slide-aware multi-criteria framework for evaluating editable PowerPoint files. EHPA extracts accessible text, title placeholders, explicit typography metadata, and visual-object counts from PPTX files, then estimates six dimensions: language conventions, readability, text density, structure, typography, and visual balance. Per-slide dimension scores are combined using an explicit weighted rubric, followed by a bounded presentation-level penalty for between-slide score variability. The framework produces slide-specific diagnostic messages and recommendations, with a local desktop implementation in Python, Tkinter, and SQLite. To test its practical validity, we specify an evaluation protocol using independently rated presentations, held-out comparison against baseline aggregation methods, error metrics, inter-rater agreement, ablation experiments, and domain-wise failure analysis. The implementation has been smoke-tested on a sample presentation; no claim of predictive superiority or human-rating agreement is made until the planned empirical study is completed. The framework offers a reproducible basis for investigating the utility and limits of interpretable automated slide assessment.

Keywords: presentation assessment; explainable evaluation; multi-criteria scoring; PowerPoint; educational technology; human evaluation; slide analytics

Introduction

Presentation quality depends on several distinct factors, including clarity, structure, typography, and the amount of material shown on each slide.

Generic automated scores can obscure which slides and features require attention. The research problem addressed here is how to combine measurable PPTX properties into an auditable presentation-level assessment that produces actionable slide-specific explanations. Unlike systems that require inaccessible training data or undisclosed scoring mechanisms, the proposed prototype uses explicit functions and exposes intermediate scores. Its intended role is formative feedback rather than factual verification, aesthetic judgment, or replacement of an instructor.

Keywords: Explainable Artificial Intelligence, Hierarchical Presentation Assessment, Automated Presentation Evaluation, Natural Language Processing, Multi-Criteria Quality Scoring, Personalized Feedback, Educational Technology.

2. Related Work and Research Gap

Readability measures provide a long-standing way to characterize written language complexity [1]. Multi-criteriadecision methods offer a principled vocabulary for aggregating distinct indicators [2]. Work on interpretable machine learning emphasizes the importance of understanding why systems produce outputs [3]. Presentation-specific assessment additionally faces a structural challenge: a deck is composed of slides with different communicative functions. A single deck-level text count cannot identify where an issue occurs. EHPA therefore evaluates slides first and aggregates their evidence at deck level. A systematic review of recent presentation-assessment papers and current Scopus-indexed literature must be added before

journal submission; the present references are foundational, not a verified comprehensive novelty search.

3. Proposed Framework

3.1 Input and extraction

The input is an editable PPTX file. The python-pptx library extracts text frames, slide title placeholders, picture/chart/table indicators, and explicitly assigned run fonts and sizes. Missing font values are treated as inherited or unknown rather than as inconsistency. Empty presentations are rejected. Text-only extraction does not assess rasterized text, visual semantics, speaker delivery, or factual correctness.

3.2 Slide-level rubric

For slide i and criterion k, let s(i,k) be its score in [0,100] and w(k) the fixed non-negative weight, with $\Sigma w(k)=1$. The slide score is $q(i)=\Sigma w(k)s(i,k)$. Deck-level dimension means are $m(k)=mean_i\ s(i,k)$, and the initial deck score is $Q=\Sigma w(k)m(k)$. A coherence adjustment is $C=\min(10,0.20\ SD_i[q(i)])$. The final reported score is $clip(Q-C,0,100)$. The adjustment is a hypothesis-driven heuristic whose utility must be tested with ablation and independent human ratings.

3.3 Explainability and feedback

The tool retains slide-specific word counts, title presence, visual-object counts, dimension scores, and diagnostic rules. Recommendations identify the slide number and its strongest observable issues. These explanations describe the deterministic rules used; they do not establish causal effects on audience comprehension.

3.4 System implementation

The prototype integrates the new research_engine.py through the existing analysis_engine.py interface, preserving the Tkinter navigation, SQLite persistence, analytics, and report generation. A separate evaluate_research.py script reads independent human labels and computes mean absolute error (MAE), root mean square error (RMSE), Pearson correlation, Spearman rank correlation, and leave-one-dimension-out ablations. Evaluation scripts do not train a model or manufacture benchmark outcomes.

4. Research Questions

RQ1: How closely do EHPA scores agree with blinded human presentation-quality ratings? RQ2: Does hierarchical slide aggregation outperform the original deck-level five-dimension heuristic? RQ3: Does the between-slide coherence adjustment improve agreement? RQ4: Which rubric dimensions contribute most, and which presentation genres show systematic failure? RQ5: Do users consider slide-level explanations more actionable than a single overall score?

5. Experimental Design

5.1 Dataset and ethics

Recruit a diverse collection of editable PPTX presentations from academic seminars, project vivas, training sessions, and professional briefings, with explicit permission to process and retain files. Record slide count, domain, presentation purpose, language, and source. Exclude identifying or confidential content from any public release. Deduplicate near-identical decks and split by author or source to prevent leakage. Dataset size must be justified by feasibility and confidence-interval precision; do not claim a predetermined sample has been collected.

5.2 Human reference scores

Three independent assessors should rate each presentation using a written six-dimension rubric with anchored 0–100 criteria, without seeing system output. Report agreement (e.g., ICC or weighted agreement appropriate to the scale), disagreements, adjudication protocol, and the resulting reference-score distribution. A pilot calibration round should precede the main blinded study.

5.3 Comparators, ablations, and statistical analysis

Compare EHPA against the existing five-metric equal-weight implementation, a slide-score mean without the coherence penalty, and a simple structural baseline. Pre-register primary MAE and secondary RMSE and rank correlation. Evaluate on held-out presentations and report paired bootstrap 95% confidence intervals for differences, along with error by presentation genre, slide count, and text density. Ablate each dimension separately and

remove the coherence penalty. If weights are tuned, use only training/validation data and freeze them before testing.

6. Results and Validation Status

The implementation package includes 60 synthetically generated PPTX files used for software pipeline testing. These files are not independently rated presentations and cannot establish human-rating agreement, generalization, or superiority over existing approaches. Genuine screenshots and empirical plots will be inserted only after the corresponding software run and independently evaluated study are completed. Quantitative validation findings are not reported at this stage.

The actual running application interfaces are documented in Appendix A (Figures 1–7). These screenshots establish implemented user workflows, but do not constitute independent validation of EHPA accuracy. A graph of expert-rated agreement and comparative experimental performance remains pending.

7. Discussion and Limitations

The primary contribution is a transparent hierarchical assessment procedure and reproducible evaluation plan. Several features are only rough proxies: an object count does not measure image relevance, the absence of a title placeholder does not imply the absence of a visible heading, Flesch scores are not suitable for all languages, and explicit font runs may omit inherited formatting. A heuristic language checker does not equal a grammar model. High scores cannot certify persuasive effectiveness, factual correctness, accessibility, or delivery quality. The chosen weights and thresholds are provisional until validated against independent judgments.

8. Conclusion and Future Work

EHPA provides a practical research prototype for interpretable presentation assessment with slide-level evidence, fixed scoring rules, and traceable feedback. The proposed empirical protocol is designed to determine whether hierarchical aggregation adds measurable value over simpler baselines. Future work includes calibrated rubric weights, semantic content evaluation using carefully audited language models, accessible slide-render analysis, multilingual evaluation, user studies measuring revision quality, and stronger privacy-preserving deployment.

References

[1] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. https://doi.org/10.1037/h0057532

[2] Saaty, T. L. (1980). The Analytic Hierarchy Process.McGraw-Hill.

[3] Molnar, C. (2022). Interpretable Machine Learning (2nd ed.). https://christophm.github.io/interpretable-ml-book/

[4] Python-pptx contributors. python-pptxdocumentation.https://python-pptx.readthedocs.io/

[5] Python Software Foundation. Python documentation. https://docs.python.org/3/

[6] SQLite Consortium. SQLite documentation. https://www.sqlite.org/docs.html

[7] scikit-learn developers. Model evaluation documentation.https://scikit-learn.org/stable/modules/model_evaluation.html

Application Screens and Recorded Outputs

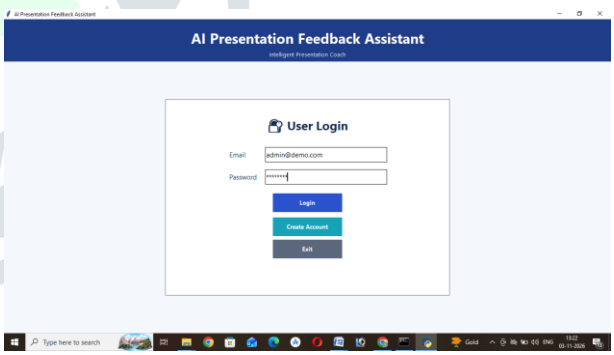

Fig. 1. Actual user-login interface of the EHPA desktop application.

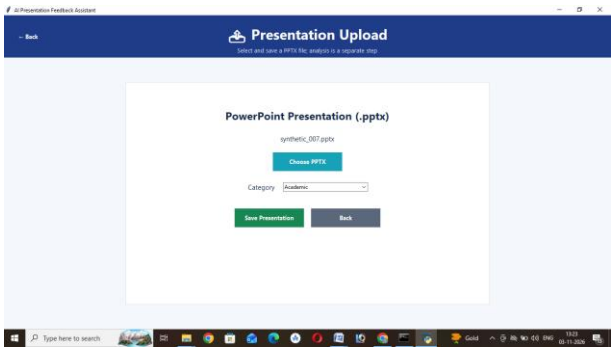

Fig. 2. Actual PowerPoint upload interface showing file selection and category.

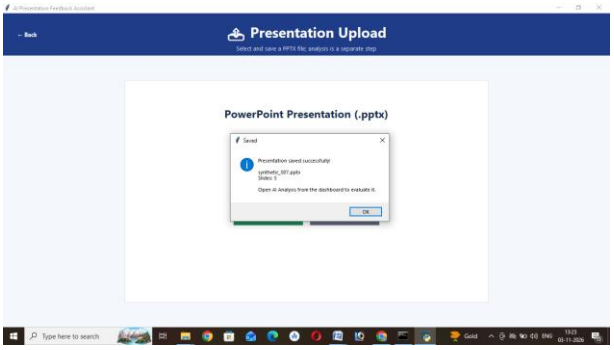

Fig. 3. Actual confirmation that synthetic_007.pptx was saved (five slides).

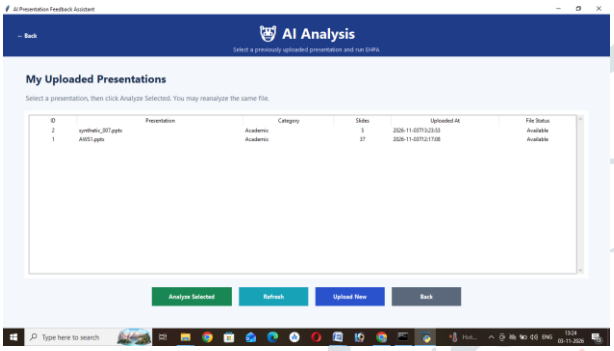

Fig. 4. Actual AI Analysis selection screen listing uploaded presentations.

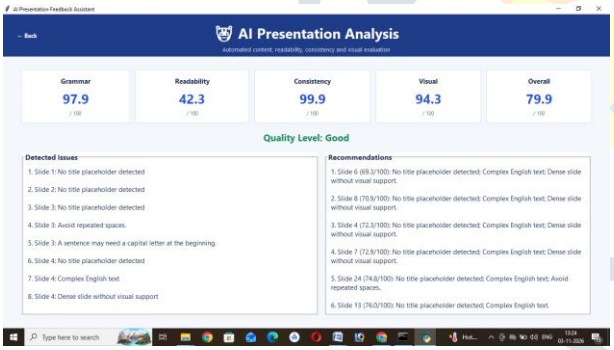

Fig. 5. Actual EHPA analysis results for a previously analyzed presentation (overall score 79.9/100).

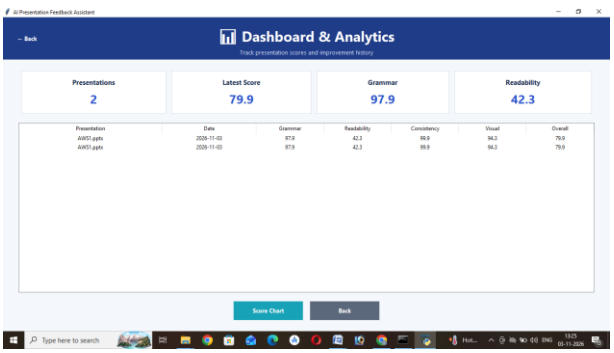

Fig. 6.Actual dashboard and analytics history displaying recorded scores.

Fig. 7.Actual main dashboard showing the six application modules.

