# XÂY DỰNG HỆ THỐNG LUYỆN THI IELTS TRỰC TUYẾN TÍCH HỢP CHẤM ĐIỂM TỰ ĐỘNG CHO KỸ NĂNG VIẾT VÀ NÓI BẰNG MÔ HÌNH HỌC SÂU

**Lê Đỗ Nghiêm[a], Nguyễn Xuân Trường[b], Trà Nguyễn Văn Nhựt[c], Hoàng Hải Long[d], Nguyễn Quốc Vương[e]**

**Tóm tắt:**

Trong nghiên cứu này, chúng tôi trình bày quá trình xây dựng và đánh giá một hệ thống luyện thi IELTS trực tuyến tích hợp các mô đun chấm điểm tự động cho hai kỹ năng Writing và Speaking. Hệ thống được thiết kế cho mục đích luyện tập, cho phép người học gửi bài, nhận band điểm dự báo theo từng tiêu chí và theo dõi biến động kết quả qua nhiều lượt làm bài, không nhằm thay thế giám khảo trong bối cảnh khảo thí chính thức. Đối với Writing, đề bài và bài làm được mã hóa bằng all-MiniLM-L6-v2, sau đó kết hợp với 52 chỉ báo ngôn ngữ tại tầng hợp nhất để dự đoán bốn tiêu chí Task Response/Task Achievement (TR/TA), Coherence and Cohesion (CC), Lexical Resource (LR) và Grammatical Range and Accuracy (GRA). Đối với Speaking, mỗi mẫu được biểu diễn bằng 420 đặc trưng đa nguồn rút ra từ transcript, biểu diễn ngữ nghĩa, tín hiệu lưu loát và một số đặc trưng âm học; bốn bộ hồi quy XGBoost được sử dụng để dự đoán Fluency and Coherence (FC), Lexical Resource (LR), Grammatical Range and Accuracy (GRA) và Pronunciation (PR). Thực nghiệm được tiến hành trên 2.162 bài viết và 281 mẫu nói. Kết quả cho thấy mô đun Writing đạt Quadratic Weighted Kappa (QWK) tổng thể 0,6482 trên tập kiểm tra sau khi lấy trung bình đầu ra của ba mô hình, trong khi mô đun Speaking đạt QWK trung bình 0,6240 theo đánh giá 5-fold cross-validation. Các kết quả bước đầu cho thấy việc tích hợp mô hình chấm điểm tự động vào nền tảng luyện tập trực tuyến là khả thi. Tuy nhiên, quy mô dữ liệu Speaking còn nhỏ, quy trình gán nhãn chưa được chuẩn hóa đầy đủ và hệ thống chưa sinh phản hồi định tính gắn trực tiếp với rubric IELTS. Do đó, các hướng phát triển tiếp theo gồm mở rộng dữ liệu, kiểm soát rò rỉ dữ liệu, chuẩn hóa nhãn và bổ sung cơ chế phản hồi định tính trước khi triển khai ở quy mô lớn.

**Từ khóa:** *IELTS, dự báo band điểm, chấm điểm tự động, học sâu, học máy, xử lý ngôn ngữ tự nhiên, luyện tập trực tuyến.*

[a] Khoa Công nghệ Thông tin, Trường Đại Học Đông Á. e-mail: nghiem101397@donga.edu.vn
[b] Khoa Công nghệ Thông tin, Trường Đại Học Đông Á. e-mail: truong99073@donga.edu.vn
[c] Khoa Công nghệ Thông tin, Trường Đại Học Đông Á. e-mail: nhut98840@donga.edu.vn
[d] Khoa Công nghệ Thông tin, Trường Đại Học Đông Á. e-mail: long98386@donga.edu.vn
[e] Khoa Công nghệ Thông tin, Trường Đại Học Đông Á. e-mail: vuongnq@donga.edu.vn

# BUILDING AN ONLINE IELTS PRACTICE SYSTEM INTEGRATED WITH AUTOMATIC SCORING FOR WRITING AND SPEAKING USING DEEP LEARNING MODELS

## Abstract:

In this study, we present the development and evaluation of an online IELTS practice system integrated with automatic scoring modules for the Writing and Speaking skills. The system is designed for practice purposes, allowing learners to submit responses, receive criterion-based predicted band scores, and monitor score changes across multiple practice attempts, rather than replacing human examiners in official assessment contexts. For the Writing module, the prompts and responses are encoded using all-MiniLM-L6-v2 and then combined with 52 linguistic indicators at a fusion layer to predict four criteria: Task Response/Task Achievement (TR/TA), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). For the Speaking module, each sample is represented by 420 multi-source features extracted from transcripts, semantic representations, fluency-related signals, and selected acoustic features; four XGBoost regressors are used to predict Fluency and Coherence (FC), Lexical Resource (LR), Grammatical Range and Accuracy (GRA), and Pronunciation (PR). Experiments were conducted on 2,162 writing responses and 281 speaking samples. The results show that the Writing module achieved an overall Quadratic Weighted Kappa (QWK) of 0.6482 on the test set after averaging the outputs of three models, while the Speaking module obtained an average QWK of 0.6240 under 5-fold cross-validation. These preliminary findings indicate that integrating automatic scoring models into an online practice platform is feasible. However, the Speaking dataset remains small, the labelling process has not yet been fully standardized, and the system has not generated qualitative feedback directly linked to the IELTS rubric. Therefore, future work should focus on expanding the dataset, controlling data leakage, standardizing labels, and incorporating qualitative feedback mechanisms before large-scale deployment.

**Keywords:** *IELTS, band score prediction, automatic scoring, deep learning, machine learning, natural language processing, online practice.*

## Đặt vấn đề

Trong luyện IELTS, người học thường gặp khó nhất ở hai kỹ năng cần sản sinh ngôn ngữ. Một bài viết hoặc bài nói không được đánh giá theo đáp án đúng - sai đơn giản, mà theo mức hoàn thành yêu cầu, tổ chức thông tin, vốn từ, độ chính xác ngữ pháp, khả năng duy trì diễn ngôn và phát âm. Theo tài liệu mô tả của IELTS, Writing sử dụng các tiêu chí TA/TR, CC, LR và GRA; Speaking sử dụng FC, LR, GRA và PR (IELTS, 2023, 2026a, 2026b). Vì vậy, một công cụ luyện tập hữu ích cần cho người học biết điểm yếu nằm ở tiêu chí nào, không chỉ trả về một con số tổng quát.

Trong lĩnh vực đánh giá ngôn ngữ bằng máy, các nghiên cứu ban đầu thường lượng hóa bài làm thông qua đặc trưng thủ công. Những đặc trưng này bao gồm độ dài, lỗi ngữ pháp, độ phong phú từ vựng, dấu hiệu lưu loát hoặc tín hiệu phát âm (Attali & Burstein, 2004; Burstein & Chodorow, 1999; Fant, 2009; Zechner et al., 2015). Mặc dù cách tiếp cận này không còn mới, nó vẫn phù hợp với bối cảnh dữ liệu nhỏ vì nhà nghiên cứu có thể đối chiếu đặc trưng với rubric và kiểm soát hành vi mô hình dễ hơn (Hussein et al., 2019).

Sự xuất hiện của Transformer và các mô hình ngôn ngữ tiền huấn luyện giúp bài toán chấm điểm có thêm khả năng biểu diễn nội dung ở mức ngữ cảnh. Các công trình gần đây khai thác học sâu, học đa nhiệm hoặc biểu diễn mạch lạc để dự đoán điểm cho văn bản dài (Alikaniotis et al., 2016; Cummins et al., 2016; Devlin et al., 2019; Farag et al., 2018; Jin et al., 2018; Ke & Ng, 2019; Reimers & Gurevych, 2019; Taghipour & Ng, 2016). Tuy nhiên, một hệ thống dùng trong học tập cần được đánh giá rộng hơn chỉ số mô hình. Độ tin cậy của nhãn, nguy cơ học lệch nguồn dữ liệu, tính ổn định khi gặp người học mới và giá trị sư phạm của phản hồi đều là các yếu tố phải được xem xét (Yannakoudakis & Cummins, 2015).

Trên cơ sở đó, nghiên cứu này xây dựng một nguyên mẫu có hai nhánh: nhánh Writing kết hợp embedding câu với đặc trưng ngôn ngữ, và nhánh Speaking khai thác bộ đặc trưng đa nguồn trong điều kiện số mẫu còn hạn chế. Đóng góp của nghiên cứu không nằm ở việc khẳng định hệ thống có thể thay thế giám khảo, mà ở việc thử nghiệm một kiến trúc có thể đưa kết quả mô hình vào luồng luyện tập trực tuyến: nộp bài, suy luận điểm, lưu lịch sử và hỗ trợ người học nhận biết xu hướng tiến bộ.

## Đối tượng và phương pháp nghiên cứu

### *Đối tượng, phạm vi và định hướng triển khai*

Nghiên cứu tập trung vào lõi dự đoán điểm nằm trong một nền tảng luyện IELTS trên web. Về chức năng, hệ thống nhận bài làm của người học, chuẩn hóa dữ liệu, kích hoạt mô hình phù hợp với kỹ năng được chọn và hiển thị điểm theo từng tiêu chí. Hai kỹ năng được chọn là Writing và Speaking vì chúng cần chấm thủ công nhiều lần trong quá trình luyện, đồng thời yêu cầu phản hồi chi tiết hơn so với các câu hỏi trắc nghiệm.

Phạm vi của nghiên cứu được giới hạn ở mức nguyên mẫu thực nghiệm. Các chỉ số báo cáo trong bài chỉ dùng để xem xét tính khả thi của giải pháp trong môi trường luyện tập, không được diễn giải như bằng chứng đủ cho bối cảnh khảo thí chính thức. Cách giới hạn này đặc biệt quan trọng vì dữ liệu IELTS có nhãn tiêu chí đáng tin cậy rất khó tiếp cận ở quy mô lớn.

### *Dữ liệu, nguồn nhãn và cách chia tập*

Bộ dữ liệu Writing gồm 2.162 mẫu; mỗi mẫu có đề bài, câu trả lời, điểm thành phần và Overall. Dữ liệu được chia thành 1.513 mẫu train, 324 mẫu validation và 325 mẫu test. Bộ dữ liệu Speaking nhỏ hơn, gồm 281 mẫu đã có véc-tơ 420 chiều cùng điểm FC, LR, GRA, PR và Overall. Sự chênh lệch này phản ánh thực tế thu thập: văn bản có thể chuẩn hóa dễ hơn, còn dữ liệu nói cần đồng thời có âm thanh, transcript và nhãn theo tiêu chí.

Dữ liệu Writing đến từ ba nhóm nguồn: bài luyện trên hệ thống, tài liệu luyện thi công khai có band tham khảo và phần nhãn do nhóm nghiên cứu chuẩn hóa theo band descriptors. Tập này bao gồm cả Academic Task 1 và Task 2. Để thống nhất đầu ra, TA của Task 1 và TR của Task 2 được gộp thành TR/TA. Cách làm này giúp tăng số mẫu cho huấn luyện đa nhiệm, nhưng cũng là giới hạn vì hai dạng bài có logic đánh giá khác nhau: một bên thiên về mô tả dữ liệu, một bên thiên về lập luận.

Dữ liệu Speaking được tạo từ các bài thi thử, bài luyện nói và mẫu công khai có band score hoặc có thể đối chiếu với mô tả tiêu chí. Vì chưa có kho audio IELTS chuẩn kèm nhãn tiêu chí đủ lớn, nghiên cứu sử dụng đặc trưng đã trích xuất thay vì học trực tiếp từ tín hiệu âm thanh thô. Nhãn Speaking do đó cần được xem là kém chắc chắn hơn nhãn Writing và kết quả ở phần này chủ yếu chứng minh tính khả thi ban đầu.

Nghiên cứu chưa có điều kiện báo cáo đầy đủ độ tin cậy liên chủ thể. Trong các bước tiếp theo, dữ liệu cần được xử lý theo quy trình chặt hơn: ẩn danh hóa người học, kiểm tra quyền sử dụng, loại bỏ bài trùng giữa các tập chia, chấm độc lập bởi ít nhất hai người và ghi rõ cách xử lý khi hai điểm chênh lệch quá ngưỡng.

Ở bước làm sạch, các bản ghi quá ngắn, thiếu đề bài hoặc thiếu điểm thành phần cần được loại trước khi huấn luyện. Với Writing, các cặp đề - bài nên được kiểm tra trùng lặp theo nội dung gần đúng để tránh cùng một bài xuất hiện ở cả train và test. Với Speaking, các mẫu từ cùng một người nói hoặc cùng một video cần được gom nhóm khi chia fold trong các nghiên cứu tiếp theo để giảm nguy cơ mô hình nhận diện nguồn thay vì học năng lực ngôn ngữ.

**Bảng 1. Tóm tắt: dữ liệu và thiết lập thực nghiệm**

| Nội dung | Writing | Speaking |
|---|---|---|
| Số mẫu | 2.162 bài viết | 281 mẫu nói |
| Đầu vào mô hình | Đề bài, bài làm và 52 đặc trưng ngôn ngữ | 420 đặc trưng đa nguồn đã trích xuất |
| Tiêu chí dự đoán | TR/TA, CC, LR, GRA | FC, LR, GRA, PR |
| Mô hình chính | MiniLM + 52 feature + ordinal heads | XGBoost đa đầu ra |
| Chỉ số đánh giá | QWK, RMSE, MAE, Pearson, Spearman | QWK 5-fold |
| Chiến lược thực nghiệm | 3 seed + ensemble | StratifiedKFold 5 fold |

### *Chỉ số đánh giá và nguyên tắc đọc kết quả*

QWK được chọn vì điểm IELTS có thứ bậc: sai lệch 1,5 band nghiêm trọng hơn sai lệch 0,5 band. Với Writing, nghiên cứu còn trình bày RMSE, MAE, Pearson và Spearman để nhìn đồng thời độ lệch tuyệt đối, độ lệch bình phương và mức tương quan giữa dự đoán với nhãn. Trước khi hiển thị, điểm dự báo được đưa về nấc 0,5 band để phù hợp hơn với cách người học quen đọc kết quả IELTS (IELTS, 2026b; Yannakoudakis & Cummins, 2015).

Cần lưu ý rằng số QWK của hai kỹ năng không dùng để xếp hạng lẫn nhau. Writing dùng một tập test cố định sau train/validation, còn Speaking dùng 5-fold vì số mẫu nhỏ. Do thiết kế đánh giá khác nhau, mỗi kết quả chỉ phản ánh hiệu năng trong phạm vi dữ liệu của chính mô đun đó.

## Mô hình và hệ thống đề xuất

### *Kiến trúc chức năng của hệ thống*

Nguyên mẫu được triển khai theo ba tầng chức năng. Tầng giao diện nhận thao tác của người học như đăng ký, chọn dạng bài, nhập bài viết hoặc tải audio. Tầng ứng dụng thực hiện tiền xử lý, gọi mô hình, tính điểm tổng hợp và trả kết quả. Tầng dữ liệu lưu bài làm, đặc trưng, điểm dự báo và lịch sử luyện tập để phục vụ phân tích tiến bộ cũng như tái huấn luyện sau này.

Về vận hành, hệ thống lưu riêng ba loại thông tin: dữ liệu gốc do người học gửi, đặc trưng trung gian dùng cho mô hình và kết quả dự báo sau khi làm tròn band. Cách tách này giúp quá trình kiểm thử dễ truy vết hơn: khi điểm bất thường xuất hiện, nhóm phát triển có thể kiểm tra lỗi đến từ bài nộp, bước tiền xử lý hay mô hình dự đoán.

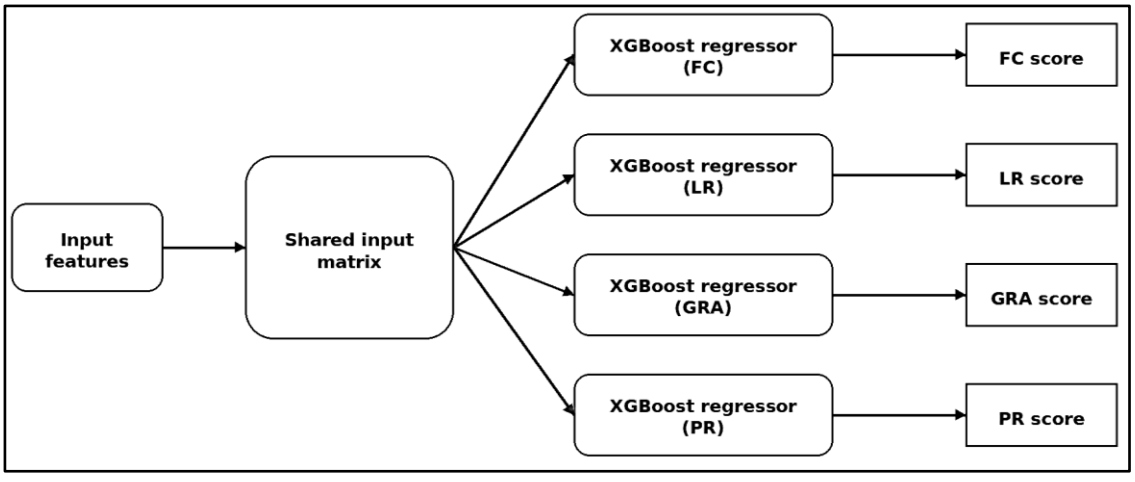
Ở nhánh Writing, đầu vào gồm đề bài và câu trả lời; hệ thống chuẩn hóa văn bản, rút trích đặc trưng, tạo embedding rồi dự đoán bốn tiêu chí. Ở nhánh Speaking, định hướng triển khai đầy đủ là audio → transcript → đặc trưng → điểm. Riêng trong thực nghiệm này, nghiên cứu chỉ đánh giá bước từ đặc trưng sang điểm nhằm tách chất lượng mô hình dự đoán khỏi sai số của khâu nhận dạng tiếng nói.

### *Mô đun Speaking*

Mô đun Speaking được thiết kế theo phương án nhẹ để phù hợp với 281 mẫu hiện có. Mỗi phản hồi nói được chuyển thành một véc-tơ 420 chiều gồm thông tin từ vựng, cú pháp, lưu loát, liên quan chủ đề, mạch lạc, âm học, tương tác chéo và embedding ngữ nghĩa. Bốn bộ hồi quy XGBoost xử lý cùng ma trận đầu vào và lần lượt sinh FC, LR, GRA và PR. Cách tách đầu ra này giữ được cấu trúc tiêu chí của IELTS nhưng vẫn đủ gọn cho hạ tầng thử nghiệm (Fouladvand et al., 2023).

**Bảng 2. Nhóm đặc trưng đầu vào của mô đun Speaking**

| Nhóm | Số lượng | Nguồn | Công cụ | Ý nghĩa |
|---|---|---|---|---|
| Lexical Richness | 10 | Transcript | NLTK | Mô tả độ phong phú từ vựng qua TTR, Hapax Legomena, entropy và đa dạng n-gram |
| Grammar Complexity | 10 | Transcript | spaCy | Phản ánh độ dài câu, cấu trúc phụ thuộc, dependency length và clause ratio |
| Fluency | 6 | Transcript | NLTK, Regex | Ghi nhận số từ, số câu và biến thiên độ dài phát ngôn |
| Topic Relevance | 2 | Question, Answer | Sentence-BERT | Đo mức gần nghĩa giữa câu hỏi và câu trả lời |
| Discourse Coherence | 2 | Transcript | Sentence-BERT | Ước lượng sự liên kết giữa các câu liên tiếp |
| Acoustic Features | 3 | Audio | Librosa | Gồm speech rate, pause ratio và energy |
| Cross Features | 3 | Transcript, Audio | Custom | Kết hợp thông tin lexical, grammar, coherence và fluency |
| Semantic Embedding | 384 | Transcript | Sentence-BERT | Biểu diễn ngữ nghĩa của câu trả lời |
| Tổng cộng | 420 | | | |




**Hình 1. Kiến trúc mô đun ước lượng điểm Speaking**

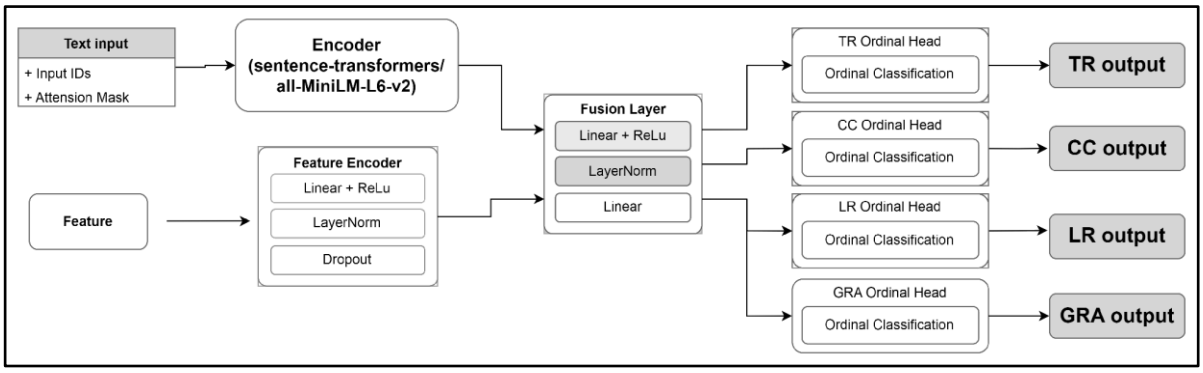
Overall của Speaking được suy ra từ bốn điểm thành phần để người học nhìn thấy lý do tương đối phía sau điểm tổng. Điểm liên tục sau suy luận được làm tròn về nấc 0,5 band. Điểm yếu của thiết kế hiện tại nằm ở phát âm: ba đặc trưng âm học còn quá ít để mô tả ngữ điệu, trọng âm, trường độ, âm vị và các lỗi phát âm tinh vi.

### *Mô đun Writing*

Với định hướng triển khai trong lớp học, đầu ra theo tiêu chí quan trọng hơn một điểm Overall đơn lẻ. Người học có thể đạt cùng band tổng nhưng khác nguyên nhân: một bài yếu ở TR/TA cần luyện phân tích đề và phát triển luận điểm, trong khi bài yếu ở GRA cần tập trung kiểm soát cấu trúc câu và lỗi ngữ pháp. Vì vậy, mô hình được giữ ở dạng đa đầu ra thay vì chỉ tối ưu điểm tổng.

Nhánh Writing có hai đường xử lý song song. Đường thứ nhất dùng all-MiniLM-L6-v2 để mã hóa quan hệ giữa đề bài và bài làm. Đường thứ hai đưa vào 52 chỉ báo thủ công, ví dụ độ dài câu, độ đa dạng từ vựng, dấu hiệu liên kết, lỗi ngữ pháp và chỉ số dễ đọc. Hai nguồn tín hiệu được ghép tại tầng fusion rồi chuyển tới bốn đầu ra ordinal cho TR/TA, CC, LR và GRA.

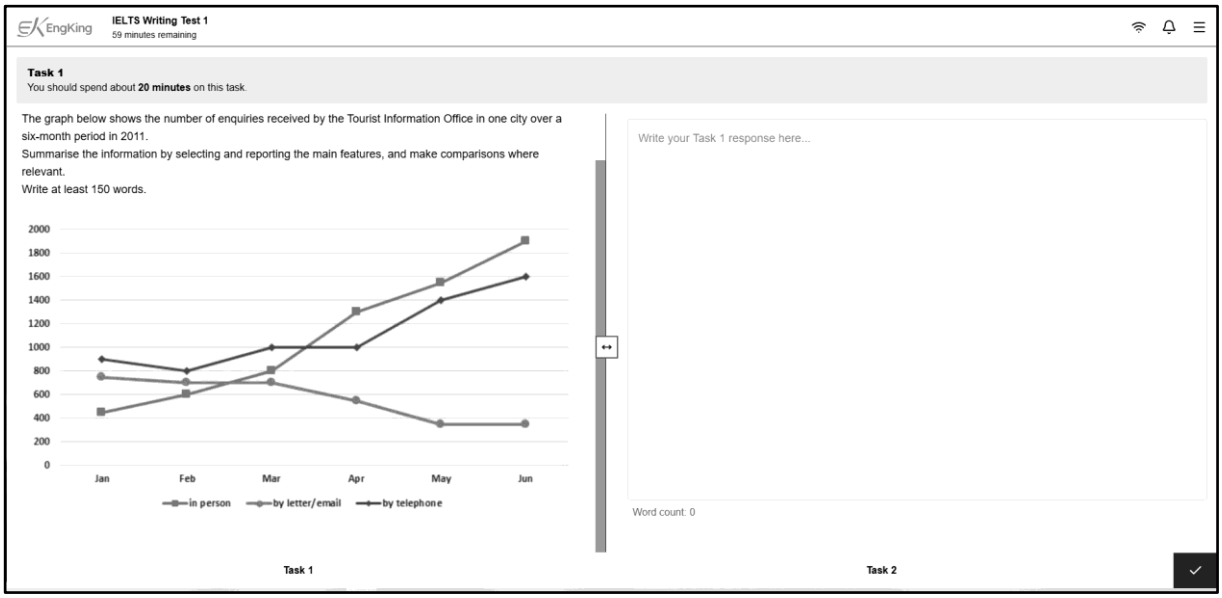
Ý tưởng của mô hình là không phụ thuộc hoàn toàn vào một kiểu tín hiệu. Embedding giúp nắm bắt nội dung và quan hệ ngữ nghĩa; đặc trưng thủ công giữ lại các chỉ báo dễ giải thích, nhất là khi dữ liệu huấn luyện chưa lớn. Để giảm ảnh hưởng của khởi tạo ngẫu nhiên, nhóm huấn luyện ba phiên bản với seed 42, 52 và 62 rồi lấy trung bình đầu ra khi suy luận.



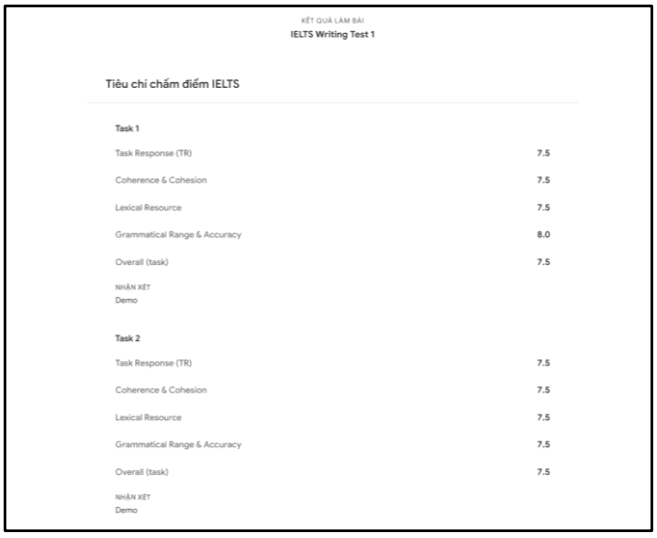

**Hình 2. Kiến trúc mô đun ước lượng điểm Writing**

### *Giao diện nguyên mẫu*

Các mô đun đã được đưa vào một ứng dụng web thử nghiệm. Người học có thể nhập bài Writing, gửi bài Speaking, xem điểm theo tiêu chí và truy lại lịch sử luyện tập. Ở phiên bản hiện tại, đầu ra chủ yếu là định lượng. Phần còn thiếu là cơ chế chuyển điểm số thành hướng dẫn học tập cụ thể, chẳng hạn nêu lỗi thường gặp, chỉ ra tiêu chí yếu và đề xuất cách luyện tiếp theo.




**Hình 3. Giao diện làm bài Writing**




**Hình 4. Giao diện hiển thị kết quả chấm điểm**

## Kết quả và thảo luận

### *Kết quả đối với mô đun Writing*

Ở nhánh Writing, ba lần huấn luyện cho QWK validation lần lượt là 0,5906, 0,6014 và 0,5939 tương ứng với seed 42, 52 và 62. Sau khi lấy trung bình dự đoán của ba mô hình, QWK trên tập test đạt 0,6482. Mức này cho thấy cấu trúc hybrid có thể khai thác được cả thông tin ngữ nghĩa và đặc trưng thủ công trên bộ dữ liệu hiện có.

**Bảng 3. Kết quả huấn luyện và ensemble của mô đun Writing**

| Mô hình/seed | Số epoch | Val QWK tốt nhất | Train QWK cuối | Test QWK |
|---|---|---|---|---|
| Seed 42 | 15 | 0,5906 | 0,8733 | - |
| Seed 52 | 15 | 0,6014 | 0,8754 | - |
| Seed 62 | 13 | 0,5939 | 0,8548 | - |
| Ensemble | - | - | - | 0,6482 |

Theo từng tiêu chí, CC đạt 0,6592; LR đạt 0,6512; GRA đạt 0,6466; TR/TA đạt 0,6360. Việc TR/TA thấp nhất không bất ngờ, vì tiêu chí này đòi hỏi mô hình đánh giá mức độ xử lý yêu cầu đề và chất lượng triển khai ý. Những yếu tố đó khó lượng hóa hơn so với các tín hiệu hình thức như từ vựng, cấu trúc câu hoặc lỗi bề mặt (Farag et al., 2018; Ke & Ng, 2019; Taghipour & Ng, 2016).

**Bảng 4. Hiệu năng của mô đun Writing trên tập kiểm tra**

| Tiêu chí | QWK | RMSE | MAE | Pear. | Spear. |
|---|---|---|---|---|---|
| TR/TA | 0,6360 | 0,1405 | 0,1110 | 0,6546 | 0,6301 |
| CC | 0,6592 | 0,1391 | 0,1094 | 0,6687 | 0,6525 |
| LR | 0,6512 | 0,1423 | 0,1126 | 0,6662 | 0,6490 |
| GRA | 0,6466 | 0,1413 | 0,1130 | 0,6587 | 0,6349 |
| Overall | 0,6482 | 0,1408 | 0,1115 | 0,6624 | 0,6417 |

QWK trên train cao hơn nhiều so với validation, nhất là ở seed 42 và 52. Khoảng cách này cho thấy mô hình có dấu hiệu học quá sát dữ liệu huấn luyện. Vì vậy, trong bối cảnh sử dụng thực tế, điểm dự báo nên được dùng như tín hiệu tham khảo để theo dõi tiến bộ, không dùng như kết luận đánh giá chính thức (Yannakoudakis & Cummins, 2015).

**Bảng 5. So sánh mô hình Writing với baseline**

| Mô hình | TR/TA | CC | LR | GRA | Overall |
|---|---|---|---|---|---|
| Text-only Regression | 0,3415 | 0,3986 | 0,3963 | 0,3601 | 0,3741 |
| Feature-only Regression | 0,6069 | 0,6335 | 0,6119 | 0,6119 | 0,6160 |
| Hybrid Ordinal Ensemble | 0,6360 | 0,6592 | 0,6512 | 0,6466 | 0,6482 |

Mô hình hybrid vượt hai baseline trong bảng so sánh. Kết quả này gợi ý rằng 52 đặc trưng thủ công vẫn mang thông tin hữu ích, còn embedding từ Transformer bổ sung lớp thông tin ngữ nghĩa mà mô hình chỉ dựa vào feature thủ công hoặc chỉ dựa vào text chưa khai thác đầy đủ.

### *Kết quả đối với mô đun Speaking*

Với Speaking, QWK trung bình qua 5 fold là 0,6240. FC đạt 0,6551, LR đạt 0,6404, GRA đạt 0,6238 và PR đạt 0,5766. Kết quả thấp ở PR phù hợp với thiết kế hiện

tại, vì phát âm chỉ được biểu diễn bằng một số đặc trưng âm học đơn giản, trong khi các tiêu chí như FC hoặc LR liên quan nhiều hơn đến transcript và độ dài phát ngôn.

**Bảng 6. Kết quả 5-fold cross-validation của mô đun Speaking**

| Tiêu chí | QWK trung bình | QWK thấp nhất | QWK cao nhất |
|---|---|---|---|
| Fluency and Coherence | 0,6551 | 0,5868 | 0,7484 |
| Lexical Resource | 0,6404 | 0,5220 | 0,7218 |
| Grammatical Range and Accuracy | 0,6238 | 0,5359 | 0,6981 |
| Pronunciation | 0,5766 | 0,4554 | 0,6283 |
| Overall | 0,6240 | 0,5250 | 0,6835 |

Phân tích feature importance cho thấy các biến về số từ, biến thiên độ dài phát ngôn, chỉ báo lưu loát, entropy từ vựng và tốc độ nói đóng góp nổi bật. Như vậy, mô đun Speaking hiện thiên về đánh giá tổ chức lời nói và độ phong phú ngôn ngữ hơn là đánh giá phát âm ở mức âm vị. Muốn cải thiện PR, cần bổ sung đặc trưng âm thanh sâu hơn hoặc mô hình học trực tiếp từ biểu diễn phổ âm (Fant, 2009; Loukina et al., 2017; Yu et al., 2015; Zechner et al., 2011).

**Bảng 7. So sánh mô hình Speaking với baseline**

| Mô hình | FC | LR | GRA | PR | Overall |
|---|---|---|---|---|---|
| Random Forest | 0,6127 | 0,6290 | 0,5460 | 0,5349 | 0,5806 |
| LightGBM | 0,6653 | 0,6565 | 0,6219 | 0,5776 | 0,6303 |
| Ridge Regression | 0,6391 | 0,6706 | 0,6209 | 0,6055 | 0,6340 |
| XGBoost tích hợp thử nghiệm | 0,6551 | 0,6404 | 0,6238 | 0,5766 | 0,6240 |

Bảng 7 cho thấy XGBoost đạt mức cạnh tranh nhưng không đứng đầu theo Overall QWK. Ridge Regression đạt 0,6340 và LightGBM đạt 0,6303, đều nhỉnh hơn XGBoost. Vì vậy, XGBoost trong bài này chỉ nên được xem là lựa chọn tích hợp thử nghiệm. Phiên bản tiếp theo cần bổ sung độ lệch chuẩn, kiểm định khác biệt và ablation theo nhóm đặc trưng trước khi quyết định mô hình triển khai cuối cùng.

Một hướng triển khai khả thi là chia phản hồi thành hai tầng. Tầng thứ nhất trả điểm nhanh ngay sau khi nộp bài, giúp người học có tín hiệu tức thời. Tầng thứ hai chạy chậm hơn để tạo nhận xét theo rubric, đối chiếu bài làm với lỗi điển hình và đề xuất bài luyện bổ sung. Cách chia tầng này phù hợp hơn với hệ thống trực tuyến vì không buộc mọi thành phần phản hồi phải hoàn tất trong cùng một thời điểm.

### *Thảo luận về giá trị hệ thống và giới hạn*

Hai mô đun thể hiện hai chiến lược khác nhau trong cùng một hệ thống. Writing tận dụng mô hình ngôn ngữ để xử lý văn bản dài và quan hệ đề - bài làm; Speaking dùng bộ đặc trưng đã thiết kế sẵn để phù hợp với dữ liệu nhỏ. Cách kết hợp

này thực tế đối với nghiên cứu sinh viên, nơi dữ liệu, thời gian gán nhãn và tài nguyên tính toán đều hạn chế (Loukina & Cahill, 2016).

Về ứng dụng, hệ thống nên được xem như trợ giảng số. Điểm theo tiêu chí có thể giúp người học biết phần cần cải thiện sau mỗi lượt luyện, nhưng bản thân điểm số chưa đủ tạo tác động sư phạm. Giá trị sử dụng sẽ tăng khi hệ thống tạo được phản hồi có hành động, ví dụ chỉ rõ câu thiếu liên kết, lập luận chưa trả lời đề, từ vựng dùng sai ngữ cảnh hoặc phát âm thiếu rõ theo rubric (Loukina et al., 2018; Nebhi & Szaszák, 2023).

Nghiên cứu còn ba giới hạn. Một là dữ liệu IELTS có nhãn chi tiết không phổ biến công khai, nhất là Speaking. Hai là mô hình Writing có dấu hiệu overfitting nên cần regularization, chống trùng lặp và kiểm tra data leakage tốt hơn. Ba là tầng phản hồi mới dừng ở điểm số, chưa sinh nhận xét định tính. Đây là các việc cần giải quyết trước khi mở rộng hệ thống.

## Kết luận

Nghiên cứu đã giới thiệu một nguyên mẫu luyện IELTS trực tuyến có hai mô đun dự báo điểm cho Writing và Speaking. Với Writing, all-MiniLM-L6-v2 được ghép với 52 đặc trưng ngôn ngữ và ensemble ba seed, đạt QWK 0,6482 trên tập test. Với Speaking, XGBoost đa đầu ra trên 420 đặc trưng đa nguồn đạt QWK trung bình 0,6240 trong đánh giá 5-fold.

Các kết quả cho thấy hệ thống có thể cung cấp tín hiệu định lượng ban đầu cho người học luyện tập thường xuyên. Tuy nhiên, do dữ liệu nhãn còn hạn chế, có dấu hiệu overfitting và chưa có phản hồi định tính, hệ thống chưa phù hợp để thay thế người chấm. Hướng tiếp theo là mở rộng dữ liệu, chuẩn hóa quy trình gán nhãn, báo cáo độ tin cậy giữa người chấm, kiểm soát rò rỉ dữ liệu, bổ sung đặc trưng âm thanh cho Speaking và xây dựng phản hồi bám sát từng tiêu chí IELTS.

## Tài liệu tham khảo

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 3-29. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 715-725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1068

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of the Workshop on Computer-Mediated Language Assessment and Evaluation in Natural Language Processing* (pp. 68-75). Association for Computational Linguistics. https://aclanthology.org/W99-0509/

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Cummins, R., Zhang, M., & Briscoe, T. (2016). Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 789-799). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1075

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Evanini, K., Xie, S., & Zechner, K. (2013). Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 157-162). Association for Computational Linguistics. https://aclanthology.org/W13-1721/

Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 263-271). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1024

Hussein, M. A., Hassan, H. A., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

IELTS. (2023). *IELTS writing band descriptors*. https://ielts.org/cdn/Guides/ielts-writing-band-descriptors.pdf

IELTS. (n.d.). *IELTS Academic format: Writing*. Retrieved April 17, 2026, from https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-writing

IELTS. (n.d.). *IELTS scoring in detail*. Retrieved April 17, 2026, from https://ielts.org/take-a-test/your-results/ielts-scoring-in-detail

Jin, C., He, B., Hui, K., & Sun, L. (2018). TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1088-1097). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1100

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300-6308). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/879

Loukina, A., & Cahill, A. (2016). Automated scoring across different modalities. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 130-135). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0514

Loukina, A., Madnani, N., & Cahill, A. (2017). Speech- and text-driven features for automated scoring of English speaking tasks. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing* (pp. 67-77). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4609

Loukina, A., Zechner, K., Bruno, J., & Beigman Klebanov, B. (2018). Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1-12). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0501

Nebhi, K., & Szaszak, G. (2023). Automatic assessment of spoken English proficiency based on multimodal & multitask transformers. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 769-776). INCOMA Ltd. https://doi.org/10.26615/978-954-452-092-2_083

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882-1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193

Yannakoudakis, H., & Cummins, R. (2015). Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 213-223). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0625

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., Ivanou, A., & Qian, Y. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level Abstract:ions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 338-345). IEEE. https://doi.org/10.1109/ASRU.2015.7404814

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S. Y. (2015). Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT). *ETS Research Report Series, 2015*(2), 1-17. https://doi.org/10.1002/ets2.12080

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883-895. https://doi.org/10.1016/j.specom.2009.04.001

Zechner, K., Wang, X. (2013). Automated content scoring of spoken responses in an assessment for teachers of English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 73-81). Association for Computational Linguistics. https://aclanthology.org/W13-1709/

Zechner, K., Xi, X., & Chen, L. (2011). Evaluating prosodic features for automated scoring of non-native read speech. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 461-466). IEEE. https://doi.org/10.1109/ASRU.2011.6163975