

PHÁT HIỆN VĂN BẢN TẠO SINH BỞI TRÍ TUỆ NHÂN TẠO TRONG ĐỒ ÁN TIẾNG VIỆT: THỰC NGHIỆM TRÊN PHOBERT VÀ XLM-ROBERTA VỚI CHIẾN LƯỢC PHÂN ĐOẠN TỪ

Nguyễn Thành Huy^{1*}, Trần Duy Hùng¹,

Nguyễn Thành Tùng¹, Trần Tiến Dũng¹

*Tác giả liên hệ, Email: huynt@hou.edu.vn. ORCID: 0009-0000-6315-3061

Ngày tòa soạn nhận được bài báo: 15/06/2026

Ngày phản biện đánh giá: 22/06/2026

Ngày bài báo được duyệt đăng: 14/09/2026

DOI: 10.59266/houjs.2026.1423

Tóm tắt: Các mô hình ngôn ngữ lớn (Large Language Models - LLMs) như ChatGPT, Gemini, DeepSeek, Claude AI đã thay đổi cách thức học tập và nghiên cứu của sinh viên. Việc lạm dụng văn bản tạo sinh thiếu kiểm chứng vào trong đồ án, luận văn, các nơi mà tính học thuật cần được đảm bảo và minh bạch rõ ràng đã và đang xuất hiện ngày một nhiều. Trong nghiên cứu này, chúng tôi thực nghiệm khung nghiên cứu, kế thừa từ khung công nghệ tiên tiến dùng để phân tích và phân loại văn bản (Fine-Grained AI-Generated Text Detection - FAID), một hệ thống phát hiện văn bản tạo sinh Tiếng Việt trong miền thông tin đồ án và khóa luận tốt nghiệp. Bài báo thực nghiệm với mô hình PhoBERT, thay thế cho mô hình đa ngôn ngữ XLM-RoBERTa. Bên cạnh đó, chúng tôi đề xuất 14 quy tắc học thuật thường xuất hiện trong các đồ án, luận văn để tiền xử lý văn bản Tiếng Việt với mục tiêu cải thiện hiệu quả của PhoBERT. Chúng tôi trích rút từ bộ dữ liệu của FAID để thu được 46,790 mẫu Tiếng Việt làm dữ liệu cho thực nghiệm. Kết quả cho thấy PhoBERT không vượt XLMRoBERTa về Accuracy/Macro F1; cấu hình tốt nhất trên hai chỉ số này vẫn là XLMRoBERTa không phân đoạn từ (Experiment I-EXPI: 95,76%/95,99%). Đóng góp của 14 quy tắc thể hiện rõ nhất ở việc hạ tỉ lệ cáo buộc sai (False Accusation Rate-FAR) xuống mức thấp nhất (8,73%) trong khi giữ Accuracy/Macro F1 tương đương, phù hợp với bối cảnh ưu tiên giảm nguy cơ buộc tội oan.

Từ khóa: học đối lập đa tầng, ô hình ngôn ngữ lớn, PhoBERT, phân đoạn từ, văn bản tạo sinh

I. Đặt vấn đề

Nghiên cứu về kiến trúc “tự chú ý” của (Vaswani và cộng sự, 2017) the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on

two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task,

¹ Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.”,”container-title”:”Advances in Neural Information Processing Systems 30 (NeurIPS 2017 về một mạng nơ-ron nhân tạo với kiến trúc “tự chú ý” (Attention) *đã thay đổi cách vận hành của các hệ thống trí tuệ nhân tạo (AI), tạo tiền đề cho các mô hình ngôn ngữ lớn (LLMs) như ChatGPT-4o* (OpenAI và cộng sự, 2024). Các mô hình ngôn ngữ lớn hiện nay có thể sinh ra các văn bản trôi chảy, mạch lạc gần giống với người viết. Các LLMs hiện có thể sinh văn bản trôi chảy, mạch lạc gần giống người viết, khiến việc xác định nguồn gốc văn bản (người viết, AI tạo sinh, hay kết hợp cả hai) trở thành bài toán thách thức các nhà nghiên cứu trong những năm gần đây (Soni và cộng sự, 2025)

Tại Việt Nam, tình trạng lạm dụng AI trong viết luận văn ngày càng phổ biến, trong khi mức độ hoàn thiện của LLMs khiến văn bản do AI tạo sinh hoặc chỉnh sửa ngày càng khó phân biệt với văn bản do con người viết. Phức tạp hơn, việc AI hỗ trợ người viết mở rộng ý tưởng tạo ra dạng văn bản kết hợp mà cách tiếp cận nhị phân “người viết/AI tạo sinh” truyền thống khó phân biệt triệt để. Trước thách thức này, một số hướng tiếp cận đã được đề xuất: phát hiện dựa trên thống kê (zero-shot detection) (Soni và cộng sự, 2025), học có giám sát dựa trên RoBERTa (Abassy và cộng sự, 2024)

và học đối lập đa tầng (Multi-Level Contrastive Learning) (Guo và cộng sự, 2024). Tuy nhiên, các nghiên cứu này tập trung chủ yếu vào Tiếng Anh với bài toán phân loại nhị phân hoặc bổ sung nhãn “kết hợp người-AI”, và chưa được áp dụng cho văn bản Tiếng Việt.

Năm 2026, nhóm nghiên cứu (Ta và cộng sự, 2026) công bố khung học đối lập đa tầng kết hợp phân loại đa nhiệm - FAID trên bộ dữ liệu đa ngôn ngữ FAIDSet (bao gồm Tiếng Việt), đạt độ chính xác 94.42% trên tập kiểm tra Tiếng Việt. Tuy nhiên, FAID sử dụng mô hình đa ngôn ngữ XLM-RoBERTa (Conneau và cộng sự, 2020) không chuyên biệt cho Tiếng Việt nên chưa tối ưu cho một số đặc thù ngôn ngữ và văn bản Tiếng Việt.

Từ khoảng trống nghiên cứu trên, bài báo đề xuất khung FAID-VN, thích nghi từ FAID cho văn bản Tiếng Việt, với các đóng góp:

- Thay thế XLM-RoBERTa bằng PhoBERT-base-v2 (huấn luyện trước trên 140GB văn bản Tiếng Việt), đánh giá trên tập văn bản Tiếng Việt thuần túy trích từ FAIDSet;

- Đề xuất chiến lược phân đoạn từ với 14 quy tắc xử lý văn bản Tiếng Việt nhằm tăng hiệu quả nhận diện cho PhoBERT (mô hình không có pha tiền phân đoạn từ);

- Thực nghiệm đánh giá tác động của lựa chọn mô hình LLMs và chiến lược phân đoạn từ trên tập dữ liệu trích rút;

II. Cơ sở lý luận

Nhiều nghiên cứu trên thế giới đã đề xuất các hướng tiếp cận khác nhau cho bài toán phát hiện văn bản sinh bởi AI, có thể phân thành ba nhóm chính:

• Nhóm thứ nhất là các phương pháp thống kê phi giám sát (zero-shot detection), dựa trên phân phối xác suất token để phân biệt văn bản AI và người viết mà không cần nhãn huấn luyện, tiêu biểu như DetectLLM (Su và cộng sự, 2023) và (Hans và cộng sự, 2024) as both LLMs and humans can exhibit a wide range of complex behaviors. However, we find that a score based on contrasting two closely related language models is highly accurate at separating human-generated and machine-generated text. Based on this mechanism, we propose a novel LLM detector that only requires simple calculations using a pair of pre-trained LLMs. The method, called Binoculars, achieves state-of-the-art accuracy without any training data. It is capable of spotting machine text from a range of modern LLMs without any model-specific modifications. We comprehensively evaluate Binoculars on a number of text sources and in varied situations. Over a wide range of document types, Binoculars detects over 90% of generated samples from ChatGPT (and other LLMs.

• Nhóm thứ hai tinh chỉnh mô hình ngôn ngữ lớn cho bài toán phân loại: LLM-DetectAIve (Abassy và cộng sự, 2024) tinh chỉnh RoBERTa cho phân loại bốn lớp, bổ sung cả trường hợp văn bản được AI chỉnh sửa; T5-Sentinel (Chen và cộng sự, 2023) dùng kiến trúc mã hóa - giải mã T5 để dự đoán token theo ngữ cảnh; (Sentence-Level AI-Generated Text Detection - SeqXGPT) (Wang và cộng sự, 2023) raising concerns about the abuse of

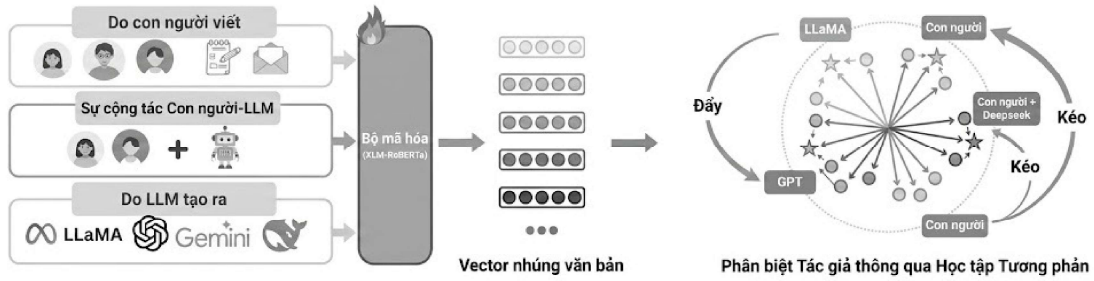
LLMs. Therefore, it is important to build strong AI-generated text (AIGT kết hợp xác suất token từ nhiều LLMs với cơ chế tự chú ý để phát hiện ở mức câu.

Yêu cầu chặt chẽ hơn trong đánh giá học thuật đặt ra nhu cầu bổ sung nhãn “kết hợp người - AI”, dẫn đến nhóm thứ ba: các khung phân loại đa lớp. Bộ dữ liệu MixSet (Mixcase Dataset - MixSet) (Zhang và cộng sự, 2024) gán nhãn phân biệt văn bản AI-sinh-rồi-người-sửa và người-viết-rồi-AI-sửa. Cả hai đều giới hạn ở Tiếng Anh, chưa có phiên bản song ngữ hoặc đa ngôn ngữ.

Một hướng tiếp cận có ảnh hưởng lớn là học đối lập đa tầng, coi LLM như một “người viết” có phong cách riêng và tổ chức lại không gian ẩn sao cho embedding cùng nguồn gần nhau hơn, tiêu biểu là DeTeCtive (Detecting AI-generated Text via Multi-Level Contrastive Learning) (Guo và cộng sự, 2024). Khung FAID (Ta và cộng sự, 2026) mở rộng ý tưởng của DeTeCtive sang bài toán ba nhãn (người, AI, kết hợp người - AI) trên bộ dữ liệu đa ngôn ngữ FAIDSet (Anh - Việt), đạt độ chính xác 94,42% trên tập kiểm tra Tiếng Việt với mô hình XLM-RoBERTa (Conneau và cộng sự, 2020).

Đối với Tiếng Việt (Tran và cộng sự, 2024) công bố bộ dữ liệu ViDetect gồm 6.800 mẫu luận văn (3.400 mẫu người viết, còn lại do LLMs sinh), so sánh PhoBERT, ViT5, BARTpho và mDeBERTa cho phân loại nhị phân, nhưng chưa xem xét dùng PhoBERT (huấn luyện trên văn bản đã phân đoạn từ) thay cho XLM-RoBERTa. Ngoài

ra, công cụ phân đoạn từ underthesea² chưa cập nhật một số thuật ngữ, ký hiệu học thuật gần đây (như “GPT-4o”, “fine-tune”), dễ làm sai lệch dữ liệu đầu vào của PhoBERT. Đến nay, chưa có nghiên cứu nào đề xuất chiến lược phân đoạn từ chuyên biệt cho văn bản học thuật Tiếng Việt trong miền luận văn, đồ án kỹ thuật, cũng như chưa có đánh giá định lượng về tác động của chiến lược phân đoạn từ đến hiệu năng phát hiện.



Hình 1. Kiến trúc huấn luyện FAID

Học đối lập đa tầng (Multi-level Contrastive Learning): Mỗi mẫu T_i được gán nhãn ba cấp độ $T_i(x_i, y_i, z_i)$ trong đó:

- x_i : là nhãn cho biết có phải LLM hay không
- y_i : là nhãn cho biết có phải người viết hay không
- z_i : chỉ số họ LLM

Bộ mã hóa (Encoder) mã hóa không gian biểu diễn phân cấp theo quy tắc các văn bản cùng họ LLM sẽ về gần nhau hơn và khác xa văn bản do người viết và văn bản kết hợp người chỉnh sửa bởi AI. Hàm mất mát đối lập cho mẫu q với K^+ và K^- là các tập mẫu tích cực, tiêu cực được định nghĩa :

$$\mathcal{L}_q = -\log \frac{\exp(\sum_{k \in K^+} \frac{S(q, k)}{\tau} / N_{K^+})}{\exp(\sum_{k \in K^+} \frac{S(q, k)}{\tau} / N_{K^+}) + \sum_{k \in K^-} \exp(\frac{S(q, k)}{\tau})} \quad (1)$$

Hàm mất mát của tổng hợp đa tầng

$$\mathcal{L}_{mcl} = \sum_{i=1}^N [x_i(\alpha \mathcal{L}_{qi,1} + \beta \mathcal{L}_{qi,2}) + (1 - x_i)(\gamma \mathcal{L}_{qi,3} + \delta \mathcal{L}_{qi,4} + \varsigma \mathcal{L}_{qi,5})] \quad (2)$$

² <https://github.com/undertheseanlp/underthesea>

III. Phương pháp tiếp cận

3.1. Khung nhận dạng FAID

FAID được xây dựng dựa trên kiến trúc DeTeCtive và giải quyết bài toán phân loại ba lớp:

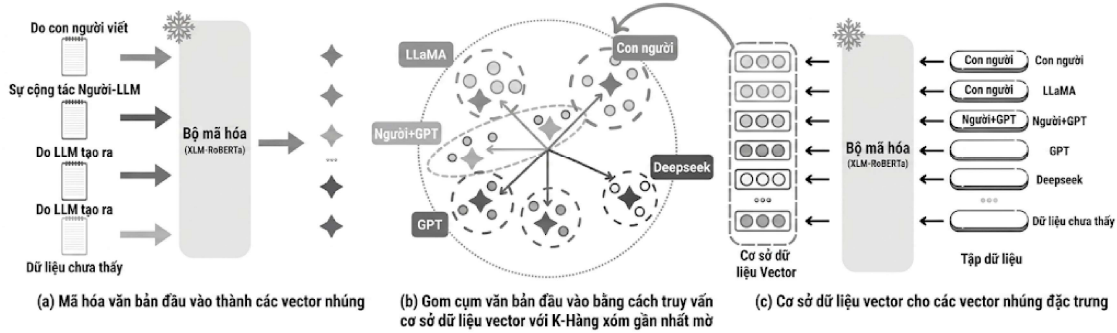
- Văn bản do người viết (nhãn 0).
- Văn bản do AI tạo sinh (nhãn 1).
- Văn bản kết hợp người-AI (nhãn 2).

Khung FAID được thực hiện thông qua hai thành phần chính là học đối lập đa tầng và phân loại đa nhiệm phụ trợ (multi-task auxiliary classification).

Phân loại đa nhiệm là một mạng MLP được gắn vào đầu ra của bộ mã hóa, huấn luyện bằng hàm mất mát cross-entropy $\mathcal{L}_{CE}$. Hàm mất mát tổng hợp chung được tính bằng:

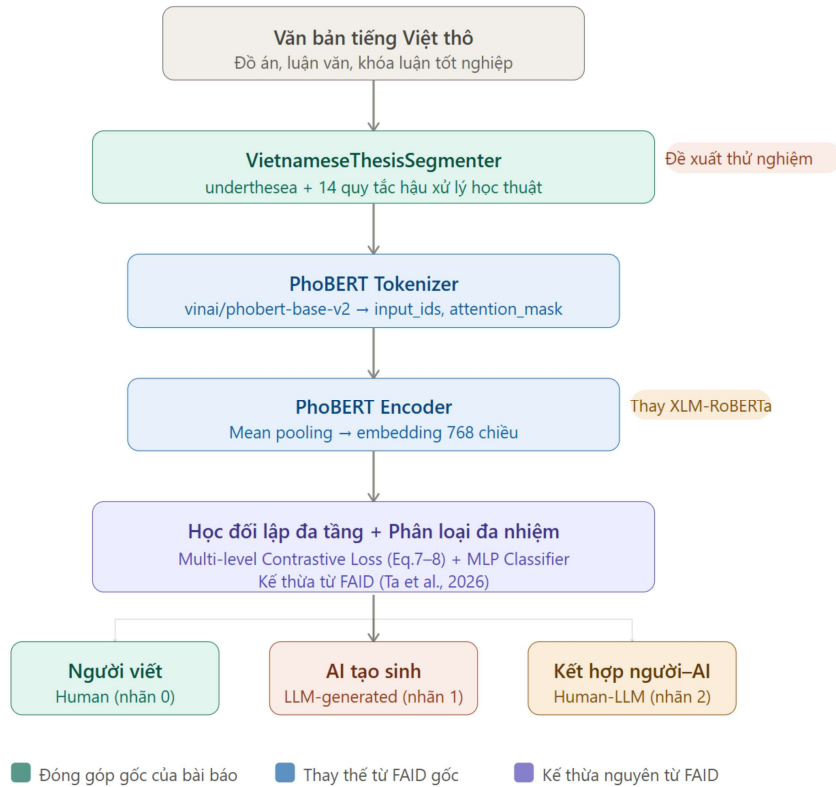
$$\mathcal{L} = \mathcal{L}_{mcl} + \mathcal{L}_{CE} \quad (3)$$

Các hàm Loss và tham số $\alpha, \beta, \gamma, \delta, \varsigma$, được tái tạo lại và thừa kế từ khung FAID gốc. Trong pha suy luận FAID đã xây dựng cơ sở dữ liệu vector lưu toàn bộ embedding của tập huấn luyện và kiểm định. Nhân dự đoán cho mẫu mới được xác định bằng gom nhóm đa số k láng giềng gần nhất với thuật toán Fuzzy kNN (k=50).



Hình 2. Kiến trúc suy luận của FAID

3.2. Mô hình đề xuất



Hình 3. Sơ đồ hoạt động của FAID-VN

Bài báo đề xuất khung hệ thống FAID-VN được thừa kế từ khung FAID (Ta và cộng sự, 2026) trong đó giữ nguyên toàn bộ kiến trúc của FAID trong cả hai pha huấn luyện và suy diễn. Để thực nghiệm chúng tôi thực hiện:

- Trích xuất tập dữ liệu song ngữ FAIDSet sang bộ dữ liệu thuần Tiếng Việt.
- Thay thế mô hình đa ngôn ngữ XLM-RoBERTa bằng mô hình đơn ngữ PhoBERT (PhoBERT-base-v2)
- Đề xuất 14 quy tắc cho bộ phân đoạn từ hậu xử lý chuyên biệt cho văn bản học thuật.

Sơ đồ hoạt động của khung hoạt động FAID-VN được biểu diễn ở hình 3.

PhoBERT-base-v2³ (Nguyen và Nguyen, 2020) được tiền huấn luyện trên 140 GB văn bản Tiếng Việt với lượng từ vựng 64.000 token mã hóa cặp byte. PhoBERT đã chứng minh có kết quả vượt

trội trong các tác vụ xử lý ngôn ngữ tự nhiên trên Tiếng Việt (Nguyen và Nguyen, 2020).

PhoBERT được tiền huấn luyện trên văn bản đã phân đoạn từ nên đầu vào cũng nên được phân đoạn để nhất quán. Tuy nhiên, giả thuyết này cần được kiểm chứng bằng thực nghiệm, bởi bộ phân đoạn tổng quát (underthesesa) có thể phân đoạn sai chính những thuật ngữ kỹ thuật như: *GPT4o*, *finetune*, *XLMRoBERTa*, vốn là tín hiệu phân biệt quan trọng cho bài toán phát hiện văn bản AI. Do đó chúng tôi đặt hai câu hỏi:

- (i) Phân đoạn từ tổng quát có thực sự cải thiện hiệu năng không;
- (ii) Nếu bổ sung quy tắc chuyên biệt có khắc phục được các lỗi phân đoạn đó không.

Từ những nhận định trên, bài báo đề xuất 14 quy tắc hậu xử lý các thuật ngữ thường hay xuất hiện trong luận văn, đồ án tiếng Việt được đề cập trong bảng sau:

Bảng 1. Danh sách các quy tắc xử lý hậu phân đoạn từ

Nhóm	Quy tắc	Mô tả	Ví dụ
A. Thuật ngữ	R1	Giữ nguyên cụm tiếng Anh có dấu “-”	fine-tune: giữ nguyên
Ký hiệu phần trăm	R2	Bảo vệ và chuẩn hóa số kèm ký hiệu %	94.42 % → 94.42%
Tên mô hình AI	R3	Bảo vệ tên kiến trúc và mô hình ngôn ngữ	PhoBERT, GPT-4o
Tên người Việt	R4	Ghép âm tiết họ tên Việt bằng gạch dưới	Nguyễn Văn An → Nguyễn_Văn_An
URL	R5	Bảo vệ địa chỉ web http/; https://www	http://hou.edu.vn
Chuỗi phiên bản	R6	Bảo vệ định danh phiên bản phần mềm	v2, 2.7.0+cu128
Độ đo đánh giá	R7	Bảo vệ ký hiệu chỉ số hiệu năng	F1-macro, MCC
Tên thuật toán	R8	Bảo vệ tên thuật toán ghép gạch nối	Fuzzy-kNN, AdamW
Ký hiệu toán học	R9	Bảo vệ biến số và công thức (dạng tex)	L_{mcl} , Eq.(8), $\sigma(i,j)$
Từ viết tắt tổ chức	R10	Bảo vệ chuỗi chữ in hoa 2+ ký tự	IEEE, ACM
Trích dẫn tài liệu	R11	Bảo vệ cú pháp trích dẫn APA/ACL	Wang và cộng sự (2024d)
DOI / arXiv	R12	Bảo vệ mã định danh tài liệu	arXiv:2505.14271v3
Tên dataset	R13	Bảo vệ tên tập dữ liệu chuẩn	FAIDSet
Đường dẫn file	R14	Bảo vệ tên file và đường dẫn module	config_exp5.json

³ <https://github.com/VinAIRsearch/PhoBERT>

IV. Kết quả thực nghiệm

4.1. Dữ liệu thực nghiệm

Chúng tôi sử dụng tập dữ liệu đa ngôn ngữ FAIDSet (Ta và cộng sự, 2026) gồm 83.350 mẫu văn bản học thuật bằng tiếng Anh và tiếng Việt trên hai miền đồ án sinh viên và tóm tắt bài báo khoa học. Tập con mẫu văn bản tiếng Việt được lọc ra từ FAIDSet bằng các nhận dạng dấu thanh điệu và các từ đặc trưng tiếng Việt. Nhãn của

mẫu dữ liệu được gán tương ứng từ thư mục nguồn theo quy tắc:

- Human - text: nhãn 0
- gpt-text, gemini-text, deepseek-text: nhãn 1
- human -- gpt - text: nhãn 2.

Tập con tiếng Việt thu được gồm 46,790 sau khi đã bóc tách từ FAIDSet. Ta nhận thấy dữ liệu bị mất cân bằng ở các nhãn AI chiếm ~20%, Người chiếm ~ 28% và ~52% tỉ lệ cao là nhãn kết hợp người - AI.

Bảng 2. Thống kê tập dữ liệu FAIDSET-VN

Tập dữ liệu	Người (0)	AI (1)	Kết hợp (2)	Tổng
Huấn luyện	9,152	6,589	17,202	32,943
Kiểm định	2,003	1,320	3,616	6,939
Kiểm tra	1,971	1,329	3,608	6,908
Tổng	13,126	9,238	24,426	46,790

4.2. Thiết kế thực nghiệm

Để tìm hiểu tác động của từng thành phần như so sánh tác động về việc thay đổi XLM-RoBERTa bằng PhoBERT, ảnh hưởng của phân đoạn từ so với không sử

dụng phân đoạn từ và việc đề xuất 14 quy tắc phân đoạn từ trên tập dữ liệu tiếng Việt, chúng tôi thiết kế các kịch bản đánh giá được ký hiệu EXP-0 đến EXP-5 với EXP-0 là kịch bản tái tạo của FAID gốc.

Bảng 3. Các kịch bản thử nghiệm

Thực nghiệm	Encoder	Phân đoạn từ	Dữ liệu huấn luyện	Mục đích
EXP-0	XLM-RoBERTa	Không	EN + VN	Baseline FAID gốc
EXP-1	XLM-RoBERTa	Không	VN	Baseline tiếng Việt
EXP-2	XLM-RoBERTa	underthesea	VN	Tác động phân đoạn từ (XLM-R)
EXP-3	PhoBERT	underthesea	VN	Tác động phân đoạn từ (PhoBERT)
EXP-4	PhoBERT	Không	VN	So sánh encoder thuần túy
EXP-5	PhoBERT	+14 quy tắc	VN	Đóng góp đề xuất

Các kịch bản được tái tạo trên GPU RTX 5060Ti 16GB nên một số tham số điều chỉnh lại cho phù hợp với cấu hình

V-RAM. Các tham số được cấu hình của thực nghiệm liệt kê ở bảng sau.

Bảng 4. Bảng tham số huấn luyện thực nghiệm

Tham số	FAID-gốc	FAID-VN
top-K (Fuzzy kNN)	50	50
Temperature	0.07	0.07
Batch size	64	8
lr	2×10^{-5}	2×10^{-5}
Weight decay	10^{-4}	10^{-4}
Epochs	50	50
Warmup steps	2.000	2.000
GPU	A100 40GB	RTX 5060 Ti 16GB

4.3. Độ đo và đánh giá

Trong bài toán gốc ngoài các độ đo được sử dụng như Accuracy, Macro F1, WeightF1 và hệ số tương quan Matthews (Matthews Correlation Coefficient-MCC), để phù hợp với đặc trưng của bài toán phát hiện văn bản sinh bởi AI bài báo sẽ thực hiện thống kê trên 2 thang đo bổ sung gồm:

Tỉ lệ kết luận sai-FAR (False Accusation Rate): là tỉ lệ mẫu văn bản người viết nhưng bị phân loại thành AI tạo sinh hoặc kết hợp người - AI

$$FAR = \frac{FP_{0 \rightarrow 1} + FP_{0 \rightarrow 2}}{N_0} \quad (4)$$

Với và là số lượt mẫu người viết bị nhầm sang nhãn 1 và nhãn 2, N_0 là tổng số mẫu người viết trong tập kiểm tra.

Tỉ lệ bỏ sót AI-MAR(Missed AI Rate - MAR): là tỉ lệ mẫu AI tạo sinh

nhưng bị hệ thống phân loại nhầm thành văn bản người viết

$$MAR = \frac{FP_{1 \rightarrow 0} + FP_{2 \rightarrow 0}}{N_1 + N_2} \quad (5)$$

• $FP_{1 \rightarrow 0}$ là tổng số mẫu AI tạo sinh bị phân loại nhầm thành nhãn 0, N_1 là tổng số mẫu AI trong tập kiểm tra.

• $FP_{2 \rightarrow 0}$ là tổng số mẫu Người+AI tạo sinh bị phân loại nhầm thành nhãn 0, N_2 là tổng số mẫu Người+AI trong tập kiểm tra.

V. Kết quả và thảo luận

5.1. Kết quả tổng hợp

Sau khi thực hiện 06 kịch bản theo hướng đề xuất trên, kịch bản EXP-0 được sử dụng làm baseline để so sánh, kết quả thu được trong các kịch bản với XLM-RoBERTa và PhoBERT như sau:

Bảng 5. Tổng hợp kết quả thực nghiệm

EXP	Encoder	Phân đoạn từ	Accuracy	Macro F1	Weighted F1	MCC	FAR	MAR
EXP-0	XLM-R	Không	96.40%	96.27%	96.38%	93.96%	8.76%	1.08%
EXP-1	XLM-R	Không	95.76%	95.99%	95.73%	93.03%	9.44%	1.60%
EXP-2	XLM-R	underthesea	95.37%	95.55%	95.33%	92.40%	10.55%	1.54%
EXP-3	PhoBERT	underthesea	95.32%	95.53%	95.29%	92.33%	9.89%	1.82%
EXP-4	PhoBERT	Không	95.51%	95.77%	95.49%	92.63%	8.73%	2.19%
EXP-5	PhoBERT	14 quy tắc	95.64%	95.87%	95.62%	92.84%	8.73%	2.01%

Ghi chú: EXP-Experiment: thực nghiệm; Encoder: Bộ mã hóa; Accuracy: Độ chính xác; Macro F1: Độ đo F1 trung bình đa lớp; Weighted F1: Độ đo F1 theo tỉ trọng mẫu

ACC ổn định giữa các kịch bản Tiếng Việt (FAID+XLM-RoBERTa). FAR thấp nhất (8,73%) ở EXP-4, EXP-5 (PhoBERT), nhưng đa seed không xác nhận ý nghĩa thống kê Bảng 5 chưa đủ cơ sở xếp hạng.

Underthesea giảm hiệu năng hai bộ mã hóa, trái kỳ vọng; 4 quy tắc đề xuất cũng không khác biệt so với không phân đoạn. Giá trị chiến lược là duy trì

ổn định, tránh suy giảm do phân đoạn tổng quát, không phải cải thiện độ chính xác.

5.2. So sánh các chiến lược kịch bản

Để đánh giá hiệu quả hoạt động giữa PhoBERT và XLM-RoBERTa ở cả hai kịch bản có sử dụng phân đoạn từ và không sử dụng phân đoạn từ với underthesea ở bảng sau:

Bảng 6. So sánh giữa PhoBERT và XLM-RoBERTa

So sánh	Δ Accuracy	Δ Macro F1	Δ FAR	Δ MAR
EXP-4 - EXP-1 (không phân đoạn)	-0.0025	-0.0022	-0.0071	+0.0059
EXP-3 - EXP-2 (có phân đoạn)	-0.0005	-0.0002	-0.0066	+0.0028

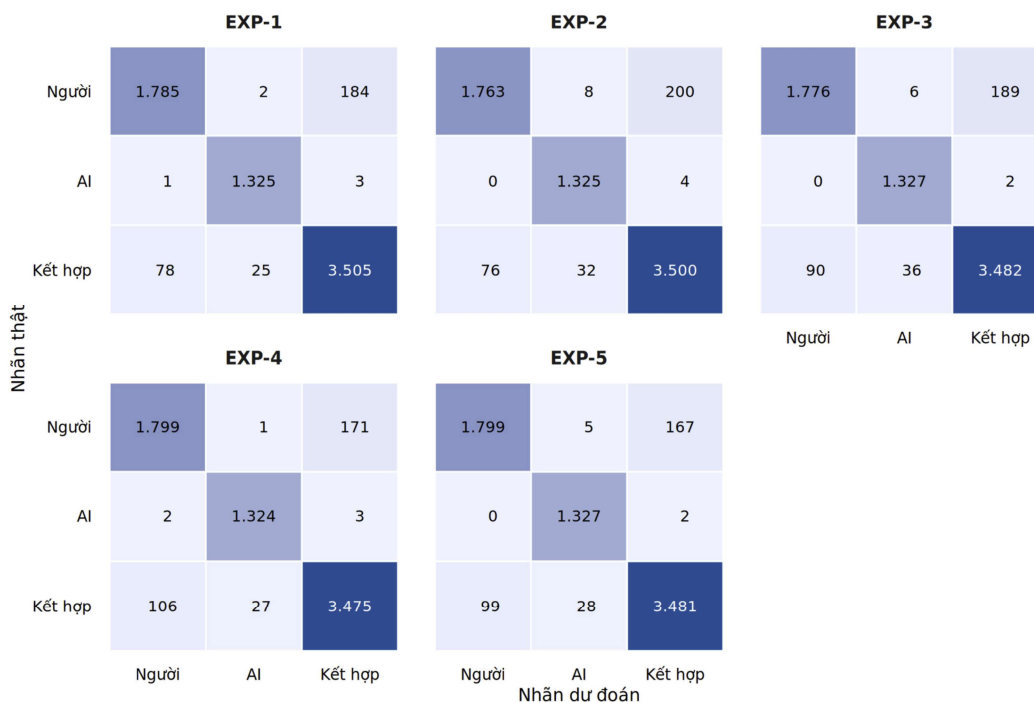
Theo bảng kết quả trên PhoBERT luôn cho kết quả với chỉ số FAR nhỏ hơn XLM-RoBERTa với chi phí đánh đổi làm tăng MAR qua đó cho thấy nếu chúng ta

quan tâm nhiều hơn đến kết luận oan sai thì PhoBERT chiếm ưu thế trong trường hợp nếu quan tâm đến việc bỏ sót phân loại thì XLM-RoBERT là lựa chọn phù hợp.

Bảng 7. Thống kê kết quả theo lớp nhãn

EXP	Lớp	Precision	Recall	F1
EXP-1	Người / AI / Kết hợp	0.9576 / 0.9800 / 0.9493	0.9056 / 0.9970 / 0.9715	0.9309 / 0.9884 / 0.9603
EXP-2	Người / AI / Kết hợp	0.9587 / 0.9707 / 0.9449	0.8945 / 0.9970 / 0.9701	0.9255 / 0.9837 / 0.9573
EXP-3	Người / AI / Kết hợp	0.9518 / 0.9693 / 0.9480	0.9011 / 0.9985 / 0.9651	0.9257 / 0.9837 / 0.9565
EXP-4	Người / AI / Kết hợp	0.9434 / 0.9793 / 0.9523	0.9127 / 0.9962 / 0.9631	0.9278 / 0.9877 / 0.9577
EXP-5	Người / AI / Kết hợp	0.9478 / 0.9757 / 0.9537	0.9127 / 0.9985 / 0.9648	0.9300 / 0.9870 / 0.9592

Để có kết quả phân tích rõ hơn, thực nghiệm cũng đã lập ra bảng ma trận nhầm lẫn qua từng kịch bản. Các ma trận tương ứng với từng kịch bản được biểu diễn dưới đây.



Hình 4. Ma trận nhầm lẫn trong qua 5 kịch bản

Phân tích F1 theo từng nhãn cho thấy lớp Người có Precision/Recall thấp nhất trong mọi kịch bản (89,5-91,3%), trong khi lớp AI đạt Recall trên 99,6%, cho thấy văn bản AI thuần túy được phân biệt mạnh; nhãn kết hợp Người-AI vẫn là bài toán khó, là thách thức cho nghiên cứu sau.

So sánh giữa các kịch bản, EXP-4 và EXP-5 có tỉ lệ nhầm sang nhãn kết hợp thấp nhất (8,7% và 8,5%), nhất quán FAR thấp nhất; EXP-2 (underthesesa, XLM-R) cao nhất (10,1%), phản ánh tác động tiêu cực của phân đoạn từ tổng quát lên nhận diện văn bản người. Mẫu lỗi gợi ý mô hình có xu hướng diễn giải đoạn văn phong cách không đồng nhất (đặc điểm thường gặp trong luận văn sinh viên tổng hợp từ nhiều nguồn) như dấu hiệu

kết hợp người-AI, dù toàn bộ văn bản do người viết.

5.3. Kiểm định thống kê đa seed giữa EXP-4 và EXP-5

Kết quả kiểm định ở bảng 8 bổ sung với 5 seed {7, 42, 1337, 2024, 20260625}, thiết kế bắt cặp giữa EXP-4 và EXP-5 để loại nhiễu khởi tạo, cho kết quả: không độ đo nào đạt ý nghĩa thống kê ($p < 0,05$), thấp nhất ở FAR ($p = 0,518$); kích thước hiệu ứng d_z dao động 0,04-0,32 (không đáng kể đến nhỏ theo Cohen). Trên bốn độ đo tổng thể, EXP-4 vượt EXP-5 ở 3/5 seed, EXP-5 nhỉnh hơn 2/5 seed còn lại, cho thấy thứ hạng Bảng 5 không ổn định. McNemar theo từng seed hội tụ ($p = 0,219-0,903$), không seed nào khác biệt có ý nghĩa, khẳng định chênh lệch ban đầu do nhiễu khởi tạo, không phải hiệu quả thực của 14 quy tắc phân đoạn..

Bảng 8. So sánh EXP-4 và EXP-5 trên 5 seed

Độ đo	EXP-4	EXP-5	Δ	KTC 95% của Δ	p_value	d_z
Accuracy	95.61 $\pm$ 0.14	95.65 $\pm$ 0.33	0.03	[-0.23; 0.29]	0.729	0.17
Macro F1	95.89 $\pm$ 0.14	95.90 $\pm$ 0.34	0.01	[-0.26; 0.28]	0.935	0.04
Weighted F1	95.59 $\pm$ 0.15	95.63 $\pm$ 0.34	0.03	[-0.23; 0.29]	0.731	0.17
MCC	92.79 $\pm$ 0.23	92.85 $\pm$ 0.54	0.06	[-0.37; 0.48]	0.736	0.16
FAR	9.30 $\pm$ 1.37	9.03 $\pm$ 1.09	-0.27	[-1.35; 0.80]	0.518	-0.32
MAR	1.99 $\pm$ 0.47	1.93 $\pm$ 0.20	-0.06	[-0.50; 0.38]	0.739	-0.16

Từ các kết quả trên, chúng tôi rút ra hai hệ quả cho đóng góp của nghiên cứu như sau:

- Thứ nhất, không có cơ sở để khẳng định EXP-5 vượt trội EXP-4; hai cấu hình cho hiệu năng tương đương trên toàn bộ sáu độ đo được xem xét.

- Thứ hai, giá trị thực của 14 quy tắc phân đoạn đề xuất không cải thiện hiệu năng so với việc không sử dụng phân đoạn từ mà nằm ở việc duy trì hiệu năng bằng ngưỡng trong khi tránh được

mức suy giảm quan sát. Ngưỡng FAT thấp trong Bảng 5 là đặc trưng gắn với PhoBERT không phân đoạn chứ không do kết quả của 14 quy tắc phân đoạn từ tạo ra.

VI. Kết luận và kiến nghị

Bài báo trình bày FAID-VN, hệ thống phát hiện văn bản AI trong đề án, khóa luận Tiếng Việt thích nghi từ FAID. Qua sáu thực nghiệm, ba phát hiện chính: (1) underthesesa làm giảm hiệu năng trên cả hai bộ mã hóa; (2)

chiến lược đề xuất không suy giảm hiệu năng EXP-5 đạt Accuracy/Macro F1 cao nhất (95,64%/95,87%) và FAR thấp nhất (8,73%), song 14 quy tắc chưa thể hiện rõ vai trò; (3) PhoBERT không vượt XLM-RoBERTa như kỳ vọng, dù FAR thấp hơn. EXP-5 được khuyến nghị làm cấu hình mặc định, phù hợp khi chi phí buộc tội oan nghiêm trọng hơn bỏ sót.

Hạn chế: lỗi nhầm người thành kết hợp người-AI ở mọi kịch bản cho thấy giới hạn bài toán ba lớp; phân đoạn đề xuất mới xử lý mẫu hình từ điển, chưa bao quát biến thể ngôn ngữ học thuật phi cấu trúc.

Tài liệu tham khảo

- Abassy, M., Elozeiri, K., Aziz, A., Ta, M. N., Tomar, R. V., Adhikari, B., Ahmed, S. E. D., Wang, Y., Mohammed Afzal, O., Xie, Z., Mansurov, J., Artemova, E., Mikhailov, V., Xing, R., Geng, J., Iqbal, H., Mujahid, Z. M., Mahmoud, T., Tsvigun, A., ... Nakov, P. (2024). LLM-DetectAIve: A Tool for Fine-Grained Machine-Generated Text Detection. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 336-343. <https://doi.org/10.18653/v1/2024.emnlp-demo.35>
- Bao, G., Rong, L., Zhao, Y., Zhou, Q., & Zhang, Y. (2025). *Decoupling Content and Expression: Two-Dimensional Detection of AI-Generated Text* (arXiv:2503.00258). arXiv. <https://doi.org/10.48550/arXiv.2503.00258>
- Chen, Y., Kang, H., Zhai, V., Li, L., Singh, R., & Raj, B. (2023). Token Prediction as Implicit Classification to Identify LLM-Generated Text. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13112-13120. <https://doi.org/10.18653/v1/2023.emnlp-main.810>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., ... Zhang, Z. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645(8081), 633-638. <https://doi.org/10.1038/s41586-025-09422-z>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). *The Llama 3 Herd of Models* (arXiv:2407.21783). arXiv. <https://doi.org/10.48550/arXiv.2407.21783>
- Guo, X., Zhang, S., He, Y., Zhang, T., Feng, W., Huang, H., & Ma, C. (2024). DeTeCtive: Detecting AI-generated Text via Multi-Level Contrastive Learning. *Advances in Neural Information Processing Systems* 37, 88320-88347. <https://doi.org/10.52202/079017-2802>
- Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024). *Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text* (arXiv:2401.12070). arXiv. <https://doi.org/10.48550/arXiv.2401.12070>

- Nguyen, D. Q., & Nguyen, A. T. (2020). *PhoBERT: Pre-trained language models for Vietnamese* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2003.00744>
- OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Ćwady, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., ... Malkov, Y. (2024). *GPT-4o System Card* (arXiv:2410.21276). arXiv. <https://doi.org/10.48550/arXiv.2410.21276>
- Soni, R., Misra, R., & Mukopadhyay, S. (2025). DetectLLM: A Multimodal Fusion Approach for Detecting LLM-Generated Text. *2025 5th International Conference on AI-ML-Systems (AIMLSystems)*, 50-57. <https://doi.org/10.1109/AIMLSystems67835.2025.11387030>
- Su, J., Zhuo, T., Wang, D., & Nakov, P. (2023). DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12395-12412. <https://doi.org/10.18653/v1/2023.findings-emnlp.827>
- Ta, M. N., Van, D. C., Hoang, D.-A., Le-Anh, M., Nguyen, T., Nguyen, M. A. T., Wang, Y., Nakov, P., & Dinh, S. (2026). *FAID: Fine-Grained AI-Generated Text Detection Using Multi-Task Auxiliary and Multi-Level Contrastive Learning* (arXiv:2505.14271). arXiv. <https://doi.org/10.48550/arXiv.2505.14271>
- Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., ... Vinyals, O. (2025). *Gemini: A Family of Highly Capable Multimodal Models* (arXiv:2312.11805). arXiv. <https://doi.org/10.48550/arXiv.2312.11805>
- Tran, Q.-D., Nguyen, V.-Q., Pham, Q.-H., Nguyen, K. B. T., & Do, T.-H. (2024). *Vietnamese AI Generated Text Detection* (arXiv:2405.03206). arXiv. <https://doi.org/10.48550/arXiv.2405.03206>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 6000-6010. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, P., Li, L., Ren, K., Jiang, B., Zhang, D., & Qiu, X. (2023). *SeqXGPT: Sentence-Level AI-Generated Text Detection* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2310.08903>
- Zhang, Q., Gao, C., Chen, D., Huang, Y., Huang, Y., Sun, Z., Zhang, S., Li, W., Fu, Z., Wan, Y., & Sun, L. (2024). LLM-as-a-Coach: Can Mixed Human-Written and Machine-Generated Text Be Detected? *Findings of the Association for Computational Linguistics: NAACL 2024*, 409-436. <https://doi.org/10.18653/v1/2024.findings-naacl.29>

AI - GENERATED TEXT DETECTION IN A VIETNAMESE THESIS: EXPERIMENTS ON PHOBERT AND XLM- ROBERTA WITH WORD SEGMENTATION STRATEGY

Thanh Huy Nguyen¹, Tran Duy Hung¹,
Nguyen Thanh Tung¹, Tran Tien Dung¹

Abstract: *Large Language Models (LLMs) such as ChatGPT, Gemini, DeepSeek, and Claude AI have transformed the way students approach learning and research. There is an increasing trend of using unverified generative text in academic projects and theses-contexts where academic integrity and transparency are paramount. In this study, we empirically evaluate the FAID-VN framework-derived from the original FAID system-which is designed to detect Vietnamese generative text within the domain of student projects and graduation theses. We conducted experiments using the PhoBERT model as a replacement for the multilingual XLM-RoBERTa model. Additionally, we proposed 14 academic rules-based on conventions common in student projects and theses-to preprocess Vietnamese text with the aim of enhancing PhoBERT's performance. We extracted 46,790 Vietnamese samples from the FAIDSet dataset for our experiments. The results indicate that PhoBERT did not outperform XLM-RoBERTa in terms of Accuracy or Macro F1; the best configuration for these metrics remained XLM-RoBERTa without word segmentation (EXP 1: 95.76% / 95.99%). The impact of the 14 rules was most evident in reducing the False Acceptance Rate (FAR) to its lowest level (8.73%) while maintaining comparable Accuracy and Macro F1 scores-a result well-suited to a context that prioritizes minimizing the risk of false accusations.*

Keywords: *large language models, multi-level contrastive learning, PhoBERT, text generation, word segmentation*

¹ Hanoi Open University, Hanoi, Vietnam