WHITEPAPER

Veribias

Phương pháp luận có cấu trúc
cho kiểm chứng nội dung tự động và hỗ trợ biên tập

Verify the truth. Detect the bias.

Phiên bản	1.0
Ngày	03/07/2026
Tác giả	Zen (Nguyễn Hiếu Thượng)
Tổ chức	Veribias

Mục lục

Tóm tắt

Các hệ thống kiểm chứng nội dung tự động (Automated Fact Checking, AFC) trong tài liệu nghiên cứu thường được mô hình hóa thành chuỗi ba bước: phát hiện claim đáng kiểm chứng, truy xuất bằng chứng, và dự đoán tính xác thực (Guo và cộng sự, 2022). Cách tiếp cận này gặp ba hạn chế khi áp dụng vào thực tiễn biên tập: nó đánh giá từng claim rời rạc thay vì cấu trúc lập luận của văn bản, nó có xu hướng luôn đưa ra kết luận kể cả khi bằng chứng mỏng, và verdict của nó khó kiểm toán khi bị chất vấn.

Veribias đề xuất một phương pháp luận nhằm xử lý cả ba hạn chế trên, xây dựng quanh ba cơ chế: (1) **taxonomy hai chiều** phân loại mỗi claim theo vai trò trong cấu trúc lập luận (trung tâm, hỗ trợ, phụ) và theo bản chất kiểm chứng (kiểm chứng được, diễn giải); (2) **chuỗi phụ thuộc logic** (dependency chain) mô hình hóa quan hệ giữa các claim, cho phép verdict phản ánh sự sụp đổ dây chuyền khi tiền đề cốt lõi sai; (3) **quy tắc abstention** buộc hệ thống từ chối kết luận khi tỷ lệ đơn vị phân tích thiếu bằng chứng vượt ngưỡng định trước.

Toàn bộ đầu ra tuân theo một schema dữ liệu chuẩn hóa (Veribias Report Schema) với mã định danh báo cáo, mức tin cậy cho từng claim, và ngày truy cập cho từng nguồn, phục vụ ba mục tiêu thiết kế: giải thích được (explainability), truy vết được (auditability), và có trách nhiệm (accountability). Veribias được thiết kế cho vai trò draft verdict trong quy trình biên tập có con người giám sát, không nhằm thay thế thẩm định con người ở các quyết định rủi ro cao, và được tối ưu cho tiếng Việt cùng hệ thống văn bản pháp luật Việt Nam.

1. Bối cảnh và vấn đề

Chi phí sản xuất nội dung đã giảm về gần bằng không nhờ các mô hình ngôn ngữ lớn, trong khi chi phí kiểm chứng nội dung hầu như không đổi: vẫn cần người có chuyên môn, thời gian, và khả năng truy nguồn. Sự bất đối xứng này đặt các tòa soạn, cơ quan chính sách và bộ phận truyền thông doanh nghiệp vào thế bị động, nơi khối lượng nội dung cần thẩm định tăng nhanh hơn năng lực thẩm định.

Nghiên cứu về AFC đã phát triển hơn một thập kỷ, từ bài toán phát hiện claim đáng kiểm chứng (Hassan và cộng sự, 2017) đến các bộ dữ liệu kiểm chứng quy mô lớn (Thorne và cộng sự, 2018) và các hệ thống kiểm chứng với bằng chứng thu thập từ web mở (Schlichtkrull và cộng sự, 2023). Các khảo sát tổng quan (Thorne và Vlachos, 2018; Guo và cộng sự, 2022; Nakov và cộng sự, 2021) chỉ ra một số khoảng trống lặp lại giữa nghiên cứu và nhu cầu của người kiểm chứng chuyên nghiệp, trong đó ba khoảng trống trực tiếp thúc đẩy thiết kế của Veribias:

Thứ nhất, thiếu tầng phân tích cấu trúc lập luận. Pipeline AFC phổ biến đánh giá từng claim như một đơn vị độc lập. Nhưng một bài viết không phải tập hợp các câu rời: nó là một cấu trúc lập luận, nơi một tiền

đề sai có thể kéo sập nhiều kết luận xây trên nó, và ngược lại, một chi tiết phụ sai không đáng làm mất giá trị toàn bài. Biên tập viên chuyên nghiệp đánh giá cấu trúc; phần lớn hệ thống tự động thì chưa.

Thứ hai, xu hướng ép kết luận. Các mô hình phân loại và các mô hình ngôn ngữ lớn đều có thiên hướng đưa ra một nhãn cho mọi đầu vào. Trong kiểm chứng, đây là khuyết tật nghiêm trọng: một verdict tự tin trên nền bằng chứng mỏng gây hại nhiều hơn một lời từ chối kết luận có giải thích. Bài toán từ chối dự đoán khi thiếu chắc chắn đã được nghiên cứu dưới tên selective prediction (Kamath và cộng sự, 2020), nhưng hiếm khi được đưa thành quy tắc bắt buộc trong các hệ thống AFC ứng dụng.

Thứ ba, verdict khó kiểm toán. Khi một verdict bị chất vấn (bởi tác giả bài viết, bởi pháp chế, hoặc bởi chính độc giả), tổ chức sử dụng công cụ cần trả lời được: verdict dựa trên bằng chứng nào, truy cập khi nào, với mức tin cậy bao nhiêu, và do phiên bản engine nào sinh ra. Đầu ra dạng văn bản tự do không trả lời được các câu hỏi đó một cách hệ thống.

Ngoài ba khoảng trống trên, bối cảnh tiếng Việt đặt thêm hai ràng buộc thực tiễn: chất lượng nguồn trên web tiếng Việt phân hóa mạnh và chịu nhiễu SEO nặng, đòi hỏi cơ chế phân bậc độ tin cậy nguồn; và hệ thống văn bản pháp luật Việt Nam (luật, nghị định, thông tư, cùng trạng thái hiệu lực thay đổi theo thời gian) đòi hỏi chế độ phân tích riêng thay vì áp thang đúng sai của tin tức.

2. Nguyên tắc thiết kế

Ba nguyên tắc xuyên suốt định hình mọi quyết định phương pháp của Veribias:

Minh bạch về cấu trúc lập luận. Hệ thống phải cho người dùng thấy văn bản đang lập luận như thế nào (claim nào là nền, claim nào xây trên claim nào), không chỉ câu nào đúng câu nào sai.

Trung thực về độ bất định. Mọi kết luận đi kèm mức tin cậy có căn cứ; khi bằng chứng không đủ, hệ thống nói rõ điều đó thay vì đoán. Từ chối kết luận được coi là một tính năng, không phải thất bại.

Truy vết được đến từng bằng chứng. Mỗi báo cáo, mỗi claim, mỗi nguồn đều có định danh và dấu thời gian, cho phép tái kiểm tra và kiểm toán về sau.

> **Nguyên tắc vận hành:** Veribias đánh giá trạng thái bằng chứng, không quy kết ý định. Đầu ra của hệ thống mô tả claim "không có nguồn độc lập xác nhận" chứ không kết luận tác giả "bịa đặt". Ranh giới này vừa là chuẩn mực học thuật, vừa là yêu cầu trách nhiệm khi công cụ được dùng trong môi trường có hệ quả pháp lý về danh dự.

3. Kiến trúc năm chặng

Veribias tổ chức quy trình kiểm chứng thành năm chặng, tách bạch phần xử lý tất định (deterministic) khỏi phần suy luận bằng mô hình ngôn ngữ. Việc tách bạch này phục vụ hai mục đích: các chặng tất định không được phép sai và không cần tốn chi phí suy luận; các chặng suy luận được thu hẹp thành nhiệm vụ đơn lẻ với đầu ra có cấu trúc, giảm phương sai giữa các lần chạy.

Chặng	Nhiệm vụ	Bản chất
1. Ingest	Nhận đầu vào (văn bản, URL, tệp, ảnh), nhận dạng ký tự nếu cần, chuẩn hóa, tính metadata và định danh đầu vào	Tất định
2. Trích xuất và phân loại claim	Liệt kê các đơn vị phân tích, gán nhãn taxonomy hai chiều, xác định chuỗi phụ thuộc	Suy luận, đầu ra có cấu trúc
3. Vòng kiểm chứng	Với từng claim: tra bộ nhớ claim đã kiểm, tra registry nguồn, tìm kiếm hai chiều (xác nhận và phản bác), chấm điểm và gán mức tin cậy	Suy luận kết hợp truy xuất
4. Tổng hợp và abstention	Xác định verdict tổng theo trọng số vai trò, áp quy tắc abstention, sinh nhận định và khuyến nghị	Suy luận trên kết quả có cấu trúc của chặng 3
5. Render	Sinh báo cáo cho người đọc từ dữ liệu có cấu trúc; xuất các định dạng (văn bản, DOCX, JSON)	Tất định

Nguồn sự thật duy nhất của một báo cáo là bản ghi dữ liệu có cấu trúc (mục 5); mọi định dạng trình bày đều được render từ bản ghi đó, bảo đảm báo cáo người đọc và dữ liệu máy đọc không bao giờ lệch nhau.

4. Ba cơ chế cốt lõi

4.1. Taxonomy hai chiều

Mỗi claim nhận hai nhãn độc lập ngay tại bước trích xuất:

Chiều	Giá trị	Định nghĩa	Ảnh hưởng xử lý
Vai trò	CT (trung tâm)	Tiền đề cốt lõi: nếu sai, luận điểm chính của văn bản sụp đổ	Trọng số lớn nhất trong verdict; sai lệch ở đây có quyền phủ quyết verdict tổng
	CH (hỗ trợ)	Củng cố hoặc minh họa cho claim trung tâm	Trọng số trung bình
	CP (phụ)	Bối cảnh, ví dụ, số liệu màu sắc	Trọng số thấp
Bản chất	V (kiểm chứng được)	Số liệu, sự kiện, trích dẫn, quan hệ nhân quả có thể truy nguồn công khai	Chấm điểm 0–1 theo bằng chứng, kèm mức tin cậy

Chiều	Giá trị	Định nghĩa	Ảnh hưởng xử lý
	I (diễn giải)	Nhận định, dự báo, đánh giá định tính	Không chấm điểm 0–1; đánh giá chất lượng lập luận: căn cứ có đủ không, suy luận có vượt bằng chứng không

Ví dụ minh họa — cùng một chủ đề: "Xuất khẩu dệt may Việt Nam đạt 44 tỷ USD năm 2024" là claim CT–V (truy nguồn được, chấm điểm theo bằng chứng); "Doanh nghiệp dệt may sẽ mất năng lực cạnh tranh nếu không chuyển đổi số" là claim CT–I (nhận định, được đánh giá theo độ chặt của lập luận, không bị ép vào thang đúng sai).

Sự tách biệt hai chiều ngăn hai lỗi thường gặp: ép điểm số cho nhận định diễn giải (làm verdict trông định lượng hơn bản chất của nó), và để một chi tiết phụ sai kéo tụt verdict của một bài có lập luận trung tâm vững.

4.2. Chuỗi phụ thuộc logic

Tại bước trích xuất, hệ thống xác định quan hệ phụ thuộc giữa các claim (claim C3 xây trên C2, C2 xây trên C1). Khi một claim trung tâm bị bác bỏ hoặc có điểm rất thấp, các claim phụ thuộc vẫn được chấm theo bằng chứng riêng của chúng, nhưng verdict tổng phải ghi rõ: chuỗi lập luận mất căn cứ từ gốc. Điểm của từng mắt xích phản ánh logic nội tại; sự sụp đổ của tiền đề phản ánh vào verdict tổng thể. Cơ chế này đưa vào hệ thống tự động điều mà biên tập viên giàu kinh nghiệm làm theo trực giác: đọc bài như một cấu trúc, không phải một danh sách câu.

4.3. Mức tin cậy gắn với phân bậc nguồn

Mỗi claim kiểm chứng được nhận một trong ba mức tin cậy, xác định bởi ba biến: số nguồn độc lập đồng thuận, chất lượng nguồn theo phân bậc (sơ cấp so với thứ cấp), và mức nhất quán giữa các nguồn. Phân bậc nguồn dựa trên một registry được con người thẩm định và cập nhật định kỳ, xếp domain từ nguồn sơ cấp gốc (cơ quan phát hành, công báo, cơ quan thống kê) qua báo chí có quy trình biên tập, đến các trang tổng hợp chỉ dùng dẫn đường và danh sách loại trừ.

> **Quy tắc cứng:** claim chỉ được xác nhận bởi các nguồn thứ cấp cùng dẫn về một gốc duy nhất không bao giờ đạt mức tin cậy cao, bất kể số lượng nguồn, vì đồng thuận bề mặt không phải độc lập thực chất.

4.4. Quy tắc abstention

Veribias áp một ngưỡng bắt buộc ở chặng tổng hợp: khi hơn 40% đơn vị phân tích của một văn bản là không kiểm chứng được hoặc chỉ đạt mức tin cậy thấp, verdict tổng buộc phải là "không đủ bằng chứng

để kết luận", kèm liệt kê cụ thể cần bổ sung nguồn gì. Hệ thống không được phép chọn một nhãn tốt hoặc xấu trong điều kiện đó.

Ngưỡng 40% là tham số thiết kế, được hiệu chỉnh qua bộ mẫu chuẩn (mục 6) thay vì cố định vĩnh viễn. Điều cố định là nguyên tắc: trong kiểm chứng, một kết luận sai đắt hơn một lời từ chối kết luận, và hệ thống phải được thiết kế để phản ánh cấu trúc chi phí đó.

5. Truy vết và chuẩn dữ liệu

Mọi báo cáo Veribias được sinh kèm một bản ghi dữ liệu có cấu trúc theo Veribias Report Schema (VRS), bao gồm: mã định danh báo cáo duy nhất, phiên bản engine, chế độ và độ sâu phân tích, danh sách claim với đầy đủ nhãn taxonomy, điểm, trạng thái, mức tin cậy và chuỗi phụ thuộc, danh sách bằng chứng cho từng claim với loại nguồn và ngày truy cập, verdict tổng kèm cờ abstention, và thống kê vận hành (số lượt tìm kiếm, số claim kế thừa từ bộ nhớ).

Ba hệ quả thực tiễn của chuẩn hóa này:

Kiểm toán được. Khi verdict bị chất vấn, tổ chức trình được toàn bộ chuỗi: claim nào, bằng chứng nào, truy cập ngày nào, engine phiên bản nào. Ngày truy cập là bắt buộc cho từng nguồn vì bằng chứng web có thể thay đổi hoặc biến mất sau thời điểm kiểm.

Tích hợp được. VRS là hợp đồng dữ liệu cho mọi tích hợp hạ nguồn (hệ quản trị nội dung, dashboard thống kê, giao diện lập trình ứng dụng), tách phương pháp luận khỏi hình thức trình bày.

Tích lũy được. Mỗi claim đã kiểm chứng, cùng bằng chứng của nó, gia nhập một bộ nhớ có thể tái sử dụng: claim tương đương xuất hiện trong tài liệu sau được kế thừa kết quả kèm tham chiếu về báo cáo gốc, thay vì kiểm lại từ đầu. Theo thời gian, bộ nhớ này cùng registry nguồn trở thành hai tài sản tri thức tăng giá trị theo mỗi báo cáo chạy qua hệ thống.

6. Phương pháp đánh giá chất lượng

Một hệ thống kiểm chứng phải tự chịu được sự kiểm chứng. Veribias đánh giá chất lượng bằng một bộ mẫu chuẩn (golden set) do người thẩm định gán nhãn, gồm các tài liệu thật phủ đủ các chế độ phân tích và các tình huống bẫy được thiết kế chủ đích: bài viết thuyết phục nhưng sai tiền đề trung tâm, bài đúng nhưng mọi nguồn cùng một gốc, bài thuần nhận định không có claim kiểm chứng được, văn bản pháp lý đã hết hiệu lực, văn bản người viết thật nhưng văn phong dễ bị nghi là máy sinh, và tài liệu có tỷ lệ claim không kiểm chứng được vượt ngưỡng abstention.

Với mỗi mẫu, kết quả kỳ vọng không phải toàn văn báo cáo mà là tập trường so sánh được: verdict kỳ vọng (hoặc dải chấp nhận), các claim bắt buộc phải tìm thấy kèm trạng thái kỳ vọng, và danh mục lỗi cấm.

Một phần bộ mẫu được gán nhãn độc lập bởi người thứ hai để đo mức đồng thuận giữa người gán nhãn; bất đồng ở tầng người là tín hiệu rubric còn mơ hồ và được xử lý trước khi quy trách nhiệm cho mô hình.

Mỗi cấu hình engine được chạy lặp nhiều lần trên toàn bộ mẫu và đo theo bốn chỉ số, xếp theo mức nghiêm trọng:

Chỉ số	Định nghĩa	Ngưỡng chấp nhận
Tỷ lệ lỗi cấm	Tần suất phạm bất kỳ lỗi nào trong danh mục lỗi cấm của mẫu (ví dụ: xác nhận một claim mà người gán nhãn xác định là không kiểm chứng được)	0%, không nhân nhượng
Đồng thuận verdict	Tỷ lệ lần chạy cho verdict trong dải kỳ vọng	Tối thiểu 80% toàn bộ; 100% với mẫu được đánh dấu rõ ràng
Độ phủ claim	Tỷ lệ claim bắt buộc được trích xuất, và độ trùng lặp danh sách claim giữa các lần chạy	Tối thiểu 85%; hệ số trùng lặp tối thiểu 0,7
Độ lệch điểm	Độ lệch chuẩn của điểm từng claim qua các lần chạy	Tối đa 0,15 trên thang 0–1

Bộ mẫu chuẩn đồng thời đóng vai trò kiểm định hồi quy: mọi thay đổi rubric, mọi lần nâng cấp mô hình nền đều phải chạy lại toàn bộ bộ mẫu và giữ được các ngưỡng trên trước khi đưa vào vận hành. Đây là điểm phân biệt giữa một tập prompt và một engine: engine có cam kết chất lượng đo được và bảo vệ được qua thời gian.

7. Triển khai trong quy trình biên tập

Ngoài chế độ kiểm chứng tài liệu, phương pháp luận Veribias được triển khai thành một giao thức tiền xuất bản (Editorial Protocol) ba tầng cho các kênh nội dung:

Tầng Gate đặt bốn câu hỏi bắt buộc trước khi viết hoặc đăng: nguồn gốc thông tin (sơ cấp, thứ cấp, hay không rõ), mục đích bài (biên tập, quảng cáo trá hình, hay nội dung thương hiệu), sự hiện diện của claim rủi ro cao (số liệu định lượng, tuyên bố y tế hoặc pháp lý), và tính khả thi của việc kiểm chứng luận điểm chính. Các tổ hợp trả lời nhất định kích hoạt yêu cầu bắt buộc (ví dụ: có claim rủi ro cao thì bắt buộc quét đầy đủ; không kiểm chứng được thì bài phải gán nhãn quan điểm).

Tầng Scan chạy checklist rủi ro đặc thù của từng kênh, được định nghĩa trong hồ sơ kênh (channel profile) tách khỏi engine lõi: kênh mới gia nhập bằng cách khai báo hồ sơ, không sửa phương pháp.

Tầng Stamp trả về một trong bốn trạng thái xuất bản (đạt, đạt kèm điều kiện, chờ bổ sung, từ chối) cùng danh sách hành động cụ thể.

Trong toàn bộ thiết kế, verdict máy là draft verdict. Vai trò của con người được định nghĩa theo nguyên tắc duyệt ngoại lệ: người thẩm định tập trung vào các báo cáo có cờ cảnh báo, mức tin cậy thấp, verdict xấu, hoặc claim thuộc nhóm rủi ro cao, thay vì đọc lại mọi thứ. Cách phân công này vừa là kiểm soát chất lượng, vừa là cơ chế trách nhiệm: quyết định cuối ở các tình huống hệ trọng luôn thuộc về con người.

8. Sử dụng có trách nhiệm

Ba quy tắc trách nhiệm được mã hóa thành ràng buộc bắt buộc của hệ thống, không phải khuyến nghị:

Đánh giá bằng chứng, không quy kết ý định. Đầu ra mô tả trạng thái bằng chứng của claim; hệ thống không sinh ngôn ngữ quy kết sự gian dối hay động cơ của tác giả. Với claim liên quan đến cá nhân hoặc doanh nghiệp cụ thể, hệ thống chỉ trình bày điều nguồn nói kèm dẫn nguồn, không suy diễn thêm.

Phát hiện văn bản máy sinh là chỉ báo xác suất, không phải phán quyết. Chế độ phân tích dấu hiệu văn bản do máy sinh luôn trả kết quả dưới dạng xác suất kèm dấu hiệu cụ thể, kèm tuyên bố bắt buộc rằng kết quả không được dùng làm căn cứ duy nhất cho các quyết định nhân sự hoặc học thuật. Nghiên cứu và thực tiễn đều cho thấy các bộ phát hiện loại này có tỷ lệ dương tính giả đáng kể; một cáo buộc sai nhắm vào người viết thật gây tổn hại không khắc phục được, và thiết kế hệ thống phải phản ánh rủi ro đó.

Minh bạch về phương pháp trong chính đầu ra. Mỗi báo cáo kết thúc bằng ghi chú phương pháp nêu rõ verdict là kết quả diễn giải bằng chứng công khai bởi mô hình ngôn ngữ, không phải thẩm định của chuyên gia độc lập, và khuyến nghị kiểm tra chéo cho các claim hệ trọng về tài chính, pháp lý, y tế.

9. Vị trí trong bức tranh chung

Bảng dưới định vị các lựa chọn thiết kế của Veribias so với hai hình mẫu phổ biến trong tài liệu nghiên cứu và thực tiễn: pipeline AFC cổ điển tập trung vào phát hiện claim và dự đoán tính xác thực, và các hệ thống kiểm chứng dựa trực tiếp trên mô hình ngôn ngữ lớn với truy xuất web. Cần nói rõ đây là so sánh về lựa chọn thiết kế, không phải kết quả đối chứng thực nghiệm; đánh giá định lượng so sánh trực tiếp nằm trong kế hoạch công bố sau khi bộ mẫu chuẩn đạt quy mô đủ lớn.

Khía cạnh thiết kế	Pipeline AFC cổ điển	Hệ LLM kèm truy xuất	Veribias
Đơn vị phân tích	Claim rời rạc	Claim rời rạc hoặc toàn văn	Claim trong cấu trúc lập luận (taxonomy hai chiều, chuỗi phụ thuộc)
Claim diễn giải	Thường nằm ngoài phạm vi	Thường bị ép vào thang xác thực	Đánh giá chất lượng lập luận, tách khỏi thang điểm

Khía cạnh thiết kế	Pipeline AFC cổ điển	Hệ LLM kèm truy xuất	Veribias
Khi bằng chứng mỏng	Trả nhãn với độ tin cậy thấp	Có xu hướng vẫn kết luận	Quy tắc abstention bắt buộc theo ngưỡng
Truy vết	Tùy hệ thống	Thường yếu	Schema chuẩn hóa, mã báo cáo, ngày truy cập từng nguồn
Vai trò con người	Thường tách rời	Thường tách rời	Tích hợp từ thiết kế (draft verdict, duyệt ngoại lệ, giao thức tiền xuất bản)
Bối cảnh ngôn ngữ	Chủ yếu tiếng Anh	Đa ngôn ngữ nhưng không chuyên biệt	Tối ưu tiếng Việt: registry nguồn Việt, chế độ riêng cho văn bản pháp luật Việt Nam

10. Giới hạn hiện tại

Phương pháp luận Veribias có bốn giới hạn cần nêu thẳng.

Một. chi phí tính toán cao hơn các hệ thống một bước do pipeline nhiều chặng và tìm kiếm hai chiều; chi phí này được kiểm soát bằng bộ nhớ claim, phân tuyến mô hình theo độ khó nhiệm vụ, và ba mức độ sâu phân tích, nhưng không triệt tiêu được.

Hai. chất lượng kiểm chứng bị chặn trên bởi chất lượng nguồn truy cập được: claim không có dấu vết trên nguồn công khai và không có trong tài liệu nội bộ được cung cấp sẽ dừng ở trạng thái không kiểm chứng được, đúng theo thiết kế.

Ba. verdict là kết quả diễn giải của mô hình ngôn ngữ trên bằng chứng; các quyết định rủi ro cao luôn cần thẩm định con người, và hệ thống được thiết kế để làm rõ ranh giới này thay vì xóa nhòa nó.

Bốn. phạm vi hiện tại tập trung vào tiếng Việt và bối cảnh Việt Nam; các tuyên bố về hiệu quả chưa nên ngoại suy sang ngôn ngữ và hệ thống pháp lý khác khi chưa có bộ mẫu chuẩn tương ứng.

11. Veribias hướng đến điều gì

Veribias không tuyên bố dự đoán đúng sai chính xác hơn mọi hệ thống khác. Đóng góp của nó nằm ở chỗ khác: biến quá trình kiểm chứng thành một quy trình có cấu trúc, nơi mọi kết luận đều chỉ ra được nó đứng trên bằng chứng nào, chắc chắn đến đâu, và trung thực thừa nhận khi chưa đủ căn cứ để kết luận.

Ba hướng phát triển đã nằm trong lộ trình: mở rộng bộ mẫu chuẩn để công bố kết quả đánh giá định lượng có thể tái lập; phát triển registry nguồn tiếng Việt thành tài sản dữ liệu được duy trì liên tục; và mở

rộng dần sang tiếng Anh cùng các lĩnh vực chuyên sâu, với điều kiện mỗi phạm vi mới có bộ mẫu chuẩn riêng trước khi đưa vào vận hành.

> ***Trong môi trường thông tin nơi chi phí tạo nội dung tiến về không, giá trị dịch chuyển về phía những gì kiểm chứng được. Veribias được xây để phục vụ đúng sự dịch chuyển đó: không thay người kiểm chứng, mà làm cho công việc kiểm chứng nhanh hơn, minh bạch hơn, và chịu được sự chất vấn.***

Tài liệu tham khảo

1. Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A Survey on Automated Fact-Checking. Transactions of the Association for Computational Linguistics, 10, 178–206.

2. Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. Proceedings of KDD 2017.

3. Thorne, J., & Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. Proceedings of COLING 2018.

4. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. Proceedings of NAACL 2018.

5. Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., & Da San Martino, G. (2021). Automated Fact-Checking for Assisting Human Fact-Checkers. Proceedings of IJCAI 2021.

6. Schlichtkrull, M., Guo, Z., & Vlachos, A. (2023). AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. Proceedings of NeurIPS 2023.

7. Kamath, A., Jia, R., & Liang, P. (2020). Selective Question Answering under Domain Shift. Proceedings of ACL 2020.

8. Veribias Internal Methodology Documentation v3.0 (2026). Tài liệu nội bộ.

Ghi chú trích dẫn: tên các công trình được giữ nguyên văn tiếng Anh theo bản gốc để bảo đảm truy nguyên chính xác.

Liên hệ

Zen (Nguyễn Hiếu Thượng)
Veribias — Verify the truth. Detect the bias.
Email: zen@veribias.com · Website: www.veribias.com