ChatGPT

So sánh các bộ dữ liệu Fact-checking và Suy luận ngôn ngữ tự nhiên tiếng Việt

Bảng so sánh chi tiết các bộ dữ liệu

Bảng dưới đây tóm tắt đặc trưng của các bộ dữ liệu: **ViAdverNLI** (bao gồm 3 vòng R1, R2, R3), **ViNLI**, **ViWikiFC**, **ViFactCheck** và **ISE-DSC01**, dựa trên các tiêu chí đã liệt kê:

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
Nguồn dữ liệu	Các **câu premise** trích từ báo điện tử VnExpress (700+ tin tức đa chủ đề) [1] [2]. **Hypothesis** do người chú thích viết để đánh lừa mô hình NLI.	Câu **premise** lấy từ >800 bài báo trực tuyến (13 chủ đề: thời sự, giải trí, thể thao, v.v.) [3]. **Hypothesis** do người tạo theo hướng dẫn NLI tiêu chuẩn.	**Câu bằng chứng** (evidence) lấy từ **Wikipedia tiếng Việt** – trích các phát biểu thực tế từ bài wiki rồi chuyển thành *claim* cần kiểm chứng [4].	**Claim** (phát biểu) và **đoạn evidence** (bằng chứng) trích từ **bài báo tin tức Việt Nam** uy tín (Thanh Niên, Tuổi Trẻ, VNExpress, ...) thuộc 12 lĩnh vực khác nhau [5]. Mỗi claim gắn với bối cảnh bài báo thực tế.	**Claim** và **đoạn văn ngữ cảnh** lấy từ **nhiều bài báo tiếng Việt** đa lĩnh vực (thời sự, kinh tế, khoa học, v.v.) [6] [7]. Nguồn tin chọn từ báo phổ biến, đáng tin cậy, đảm bảo tính đa dạng và thực tế.

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
Số lượng mẫu (train/dev/test)	~**10.0k** cặp premise–hypothesis (tổng hợp 3 vòng) [2]. Chia 8:1:1 thành 8,012 train / 1,000 dev / 1,000 test [8]. *(Mỗi vòng ~3.3k cặp: R1 ~2.6k train, 330 dev, 330 test; R2 ~2.7k train, 330 dev, 330 test; R3 ~2.7k train, 340 dev, 340 test)* [9].	>**30k** cặp premise–hypothesis có nhãn (khoảng 24k train, 3k dev, 3k test – chia 80/10/10%) [10] [11]. Mỗi premise đi kèm 8 hypothesis (2 cho mỗi nhãn) nhằm cân bằng dữ liệu [12].	>**20k** claim được gán nhãn kèm bằng chứng (≈16k train / 2k dev / 2k test) [13]. Phân bố nhãn **cân bằng** giữa 3 lớp [14]. Mỗi claim có một hoặc vài câu bằng chứng tương ứng từ wiki.	**7,232** cặp claim–evidence (5,126 train / 1,026 dev / 1,080 test – ~70/10/20%) [15].	~**49.7k** cặp claim–đoạn văn (38,684 train / 4,793 dev / 5,396 test) [7] – một trong những bộ dữ liệu fact-checking lớn nhất cho tiếng Việt. Nhãn phân bố gần như **cân bằng** giữa các lớp [16].

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
Định dạng dữ liệu	Mỗi mẫu gồm: một **premise** (câu gốc từ tin tức) + một **hypothesis** (câu do annotator viết) + một **nhãn suy luận** (Entail/ Contr/ Neutral). Được xây dựng qua nhiều vòng điều chỉnh để tăng độ khó cho mô hình [17] [18].	Mỗi mẫu gồm: một câu **premise** và một câu **hypothesis**, kèm **nhãn suy luận logic** giữa chúng (kéo theo, mâu thuẫn hoặc trung tính) [19] [20]. Định dạng tuân theo chuẩn NLI (giống SNLI/ MultiNLI).	Mỗi mẫu gồm: một **claim** (câu phát biểu cần kiểm chứng) + **bộ bằng chứng** (1 hoặc nhiều câu Wikipedia liên quan) + **nhãn** đánh giá độ đúng/sai của claim dựa trên bằng chứng [21] [22]. Định dạng tương tự FEVER cho tiếng Anh.	Mỗi mẫu gồm: một **claim** và một **đoạn evidence** (trích từ bài báo liên quan) + **nhãn** xác định tính đúng/sai của claim dựa trên evidence đó [23] [24]. Định dạng tương tự bài toán kiểm chứng claim FEVER nhưng áp dụng cho ngữ liệu báo chí Việt [24].	Mỗi mẫu gồm: một **claim** và một **đoạn văn ngữ cảnh** liên quan + **nhãn** xác định claim được hỗ trợ, bác bỏ hay không đủ thông tin từ đoạn văn [25]. (*Trong tập huấn luyện, đoạn văn thường chứa bằng chứng trực tiếp/gián tiếp; trong thi đấu, mô hình phải tự tìm câu chứng cứ từ văn bản dài*) [26] [27].

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
Bộ nhãn & phân bố	**3 nhãn NLI**: Entailment (kéo theo), Contradiction (mâu thuẫn), Neutral (trung tính) – tương ứng Hỗ trợ, Phản bác, Không đủ thông tin (theo ngữ cảnh) [28]. *Phân bố:* gần **cân bằng** tổng thể (mỗi nhãn ~33%) nhưng thay đổi theo vòng: ví dụ vòng 1 ít NEI hơn (chủ yếu Entail/Contr) so với vòng 3 (NEI tăng) – phản ánh chiến lược tạo câu ngày càng đa dạng.	**3 nhãn NLI tiêu chuẩn**: Entailment (kéo theo), Contradiction (mâu thuẫn), Neutral (trung tính) [20]. Nhờ thiết kế mỗi premise sinh 2 hypothesis mỗi loại, tập dữ liệu được cân bằng 3 nhãn (xấp xỉ 1/3 mỗi lớp).	**3 nhãn**: *Supports* (có căn cứ), *Refutes* (bác bỏ), *Not Enough Information* (thiếu thông tin) – theo schema FEVER [29] [30]. Phân bố nhãn đã được **cân bằng** giữa các lớp (~33% mỗi loại) trên toàn bộ tập [14].	**3 nhãn**: *Support* (Hỗ trợ), *Refute* (Bác bỏ), *Not Enough Information (NEI)* (Không đủ thông tin) [31] [32]. Mỗi bài báo gốc tạo đúng 2 claim mỗi nhãn, nên tập dữ liệu **cân bằng hoàn hảo** (mỗi lớp ~1/3 tổng số claim) [33] [34]. (Kappa đồng thuận giữa các annotator = 0,83 – rất cao) [35].	**3 nhãn**: *Supported* (Được hỗ trợ), *Refuted* (Bị bác bỏ), *NEI* (Không đủ thông tin) [36] [37]. Nhãn được xác định qua đối chiếu claim với nội dung bài báo: nếu bài báo khẳng định claim ⇒ Supported; mâu thuẫn claim ⇒ Refuted; không đề cập đủ ⇒ NEI [38] [39]. Phân bố nhãn ~cân bằng (vd. train: 12.8k Sup, 12.6k Ref, 13.3k NEI) [40].

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
Độ dài TB (claim & evidence)	**Premise (tin tức)**: ~24 từ (tương tự ViNLI) [41]. **Hypothesis**: ngắn, trung bình *~12-15 từ* (ngắn hơn hẳn so với ViNLI ~18 từ) [42]. Các câu hypothesis thường súc tích, tập trung thay đổi nhỏ để đánh lừa mô hình.	**Premise**: ~24,5 từ; **Hypothesis**: ~18,1 từ (trung bình) [11]. Hypothesis thường ngắn hơn premise, nhưng vẫn đủ rõ nghĩa cho suy luận [43]. (Ngắn nhất 4 từ, dài nhất 68 từ; đa số 10–23 từ) [44].	**Claim**: thường là một câu ngắn gọn (*~15–20 từ*). **Bằng chứng**: mỗi câu bằng chứng Wikipedia *~20 từ*; một claim có thể kèm 1–2 câu bằng chứng nên tổng độ dài ngữ cảnh ~20–40 từ. *(Suy ra từ độ dài câu wiki trung bình, tương tự FEVER).*	**Claim**: ngắn, thường 1 câu (*~12–15 từ*). **Evidence**: đoạn trích báo ~1–3 câu liên quan (có thể *~30–50 từ*). Do mỗi claim chỉ dựa vào một đoạn chứng cứ cụ thể nên ngữ cảnh khá gọn. *(Theo mô tả, evidence là "đoạn trích nội dung báo liên quan", thường chỉ vài dòng)* [23].	**Claim**: thường 1 câu (*~10–20 từ*). **Đoạn ngữ cảnh**: trung bình dài hơn evidence của ViFactCheck, có thể là một đoạn văn (*~50–100 từ*). Trong dataset train, đoạn văn đã được cắt để chứa thông tin chính yếu (tránh quá dài) [26]. Trong chế độ thi, mô hình phải xử lý cả bài báo dài nên độ dài ngữ cảnh thực tế có thể lớn.

Quy mô từ vựng & đa dạng ngôn ngữ	Tương đối **đa dạng, nhiều từ mới**: Annotator cố ý **không lặp từ** của premise trong hypothesis – tỷ lệ trùng từ thấp và thường dùng từ đồng nghĩa, cách diễn đạt khác [45]. Mỗi vòng bổ sung từ vựng mới (đặc biệt là danh từ, động từ) để mô hình khó suy luận dựa trên từ khóa bề mặt [46] [47]. Chủ đề tin tức phong phú (thời sự, giải trí, khoa học,...).	Quy mô **lớn, phủ nhiều chủ đề** (13 lĩnh vực tin tức) nên từ vựng khá phong phú [3]. Hướng dẫn sinh dữ liệu yêu cầu **không sao chép y nguyên** premise mà dùng từ ngữ của riêng mình [48] – do đó corpus chứa nhiều cách diễn đạt lại, bao gồm từ đồng nghĩa, câu chủ–bị động, thay đổi trạng từ, v.v. (theo các luật tạo dữ liệu) [49] [50].	**Đa dạng kiến thức bách khoa**: bao gồm nhiều tên riêng, thuật ngữ về địa danh, nhân vật, sinh vật, lịch sử... do lấy từ Wikipedia. Văn phong bách khoa chung, nhưng nhờ cơ chế tạo claim (đổi chi tiết hoặc thêm thông tin ngoài phạm vi) nên có xuất hiện cả câu đúng lẫn sai, thông tin hư cấu. Từ vựng bao trùm các lĩnh vực có trên Wikipedia tiếng Việt (khá rộng, nhưng bị giới hạn ở phạm vi tri thức wiki).	**Đa miền ngôn ngữ báo chí**: bao quát 12 lĩnh vực từ chính trị, y tế đến giải trí [5] nên tập hợp từ vựng rất **phong phú**. Các claim và evidence chứa nhiều số liệu, tên riêng, thuật ngữ chuyên ngành (phù hợp từng lĩnh vực). Tuy quy mô mẫu vừa phải (~7k) nhưng do tạo 6 claim mỗi bài báo theo kịch bản khác nhau [33], ngôn ngữ biểu đạt khá đa dạng (bao gồm diễn đạt lại sự kiện thực tế, cố ý sửa chi tiết tạo thông tin sai, và đặt câu hỏi ngoài phạm vi bài).	**Rất lớn và phong phú**: gần 50k mẫu từ hàng nghìn bài báo trải rộng nhiều chuyên mục, do đó **quy mô từ vựng lớn** nhất trong các bộ so sánh. Văn bản ngữ cảnh dài hơn, chứa câu phức và thông tin nền. Đa số từ vựng là ngôn ngữ báo chí phổ thông, nhưng nhờ lượng dữ liệu lớn, mô hình học được nhiều cách diễn đạt khác nhau của cùng một nội dung. Tập dữ liệu cũng bao gồm nhiều **tên riêng, thuật ngữ chuyên ngành** từ các mảng khác nhau (kinh tế, khoa học, thể thao,...).

Tính chất *adversarial*	**Có** – Bộ dữ liệu *adversarial* duy nhất: Được xây dựng theo quy trình **human-and-model-in-the-loop** qua 3 vòng [17]. Mỗi vòng, annotator tìm cách viết hypothesis gây **hiểu lầm cho mô hình** NLI hiện tại, sau đó lọc những mẫu mô hình dự đoán sai (xác nhận bởi người) làm dữ liệu vòng tiếp theo [51] [52]. Qua từng vòng, mô hình ngày càng mạnh nhưng vẫn bị đánh lừa bởi câu khó hơn. Kết quả, ViAdverNLI chứa nhiều mẫu hóc búa mà **mô hình SOTA chỉ đạt ~48%** độ chính xác trên test [53] – thấp hơn hẳn các dataset khác.	**Không (chuẩn)** – Được xây dựng theo quy trình truyền thống, chú trọng chất lượng hơn là đánh lừa mô hình. Các cặp câu được viết và gán nhãn cẩn thận, tránh cặp quá đơn giản hay mơ hồ [54] [55], nhưng **không có yếu tố adversarial chủ đích**. Mục tiêu là phản ánh suy luận thực tế, không tập trung khai thác lỗ hổng của mô hình.	**Không (theo chuẩn FEVER)** – Claim được tạo thủ công từ câu wiki nhưng không dựa trên phản hồi của mô hình. Tuy nhiều claim đòi hỏi suy luận phức tạp (cần kết hợp kiến thức hoặc phát hiện thông tin bị đảo), bộ dữ liệu **không sử dụng mô hình trong vòng tạo dữ liệu**. Vì thế, tính adversarial chỉ ở mức thay đổi thông tin để đánh đố người/máy theo kiểu FEVER, chưa phải adversarial multi-round.	**Không hẳn (nhưng có yếu tố sáng tạo)** – Dù không có mô hình trong vòng tạo dữ liệu, nhóm tác giả **cố tình tạo claim sai bằng cách chỉnh sửa chi tiết** và claim NEI bằng cách đưa thông tin ngoài bài [34] [56]. Do đó, dataset có những mẫu "gây nhiễu" giống thật (adversarial đối với người đọc). Tuy nhiên, các claim này được thiết kế nhằm đảm bảo đủ thông tin để con người nhận biết (dựa vào bài báo), chứ không nhằm đánh bại mô hình cụ thể nào.	**Không (bán tự động)** – Dữ liệu được tạo ra bằng **đối chiếu tự động + kiểm duyệt thủ công**, không phải qua nhiều vòng tương tác người-máy. Mục tiêu chính là quy mô lớn và đủ độ khó tổng quan, không nhắm đến việc tìm lỗi mô hình cụ thể. Tuy nhiên, do có khâu tự động, có thể chứa một số trường hợp nhiễu tự nhiên (mô hình dễ nhầm) – nhưng đây là tác dụng phụ hơn là thiết kế có chủ đích.

Phương pháp gán nhãn	**Kết hợp con người & mô hình**: Người viết hypothesis, mô hình dự đoán, sau đó người xác nhận và lọc. Cụ thể, vòng 1 dùng XLM-R gán nhãn tạm cho mẫu annotator viết; vòng 2–3 dùng InfoXLM (huấn luyện trên dữ liệu mở rộng) đánh giá mẫu mới [1] [57]. Chỉ giữ lại những cặp mà mô hình dự đoán *sai* nhưng được người xác minh là hợp lệ để đưa vào tập dữ liệu [51] [52]. Cuối cùng, toàn bộ ~10k cặp được gán nhãn vàng bởi con người (theo đa số phiếu nếu cần).	**Gán nhãn thủ công chất lượng cao**: Các cặp premise-hypothesis do người viết theo hướng dẫn chi tiết (có bộ quy tắc tạo câu cho entailment và contradiction) [58] [49]. Mọi mẫu (nhất là dev/test) được **5 annotator gán nhãn độc lập**, chọn nhãn vàng bằng bỏ phiếu đa số [59] [60]. Kết quả 99,4% cặp được ≥3/5 phiếu trùng, chất lượng rất cao [61].	**Gán nhãn thủ công theo FEVER**: Annotator được huấn luyện viết claim dựa trên câu wiki gốc rồi tự gán nhãn cho claim đó luôn (dựa trên bằng chứng wiki) [62] [63]. Mỗi claim sau đó đều được kiểm tra chéo với câu nguồn để đảm bảo nhãn đúng và câu claim hợp lệ. Bộ dữ liệu được nhóm tác giả soát xét nhằm loại bỏ lỗi và cân bằng tập. (Hiện tại dữ liệu được công bố có thể qua liên hệ nhóm tác giả) [64].	**Gán nhãn thủ công nghiêm ngặt, có giám sát chuyên gia**: Chọn các bài báo, mỗi bài giao cho annotator tạo **6 cặp claim–evidence** (2 Hỗ trợ, 2 Bác bỏ, 2 NEI) dựa trên nội dung bài [33] [56]. Cụ thể: claim *Support* lấy sự kiện **có thật** trong bài, viết lại thành câu đúng; claim *Refute* **sửa một chi tiết** trong bài (số liệu, tên, thời gian…) tạo câu sai; claim *NEI* nêu thông tin **liên quan chủ đề nhưng không có trong bài** [34] [56]. Mỗi claim được đối chiếu lại với bài và gắn nhãn + chỉ rõ đoạn bằng chứng. Nhờ quy trình kỹ lưỡng và	**Bán tự động + hiệu chỉnh thủ công**: Với quy mô lớn, ban đầu dùng kĩ thuật **cross-check tự động**: đối chiếu claim với bài báo gốc, tự động gán nhãn Supported nếu bài chứa thông tin khẳng định claim, Refuted nếu mâu thuẫn, NEI nếu không đề cập đủ [38] [39]. Sau đó, một phần dữ liệu được **kiểm tra thủ công** để sửa lỗi và đảm bảo cân bằng nhãn [65] [66]. Kết quả cuối đạt bộ dữ liệu ~50k với nhãn phân bố gần đều (vd. train mỗi nhãn ~12-13k) [40].

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
				huấn luyện thống nhất, dữ liệu đạt chất lượng cao (Fleiss' Kappa = 0,83) [35] .	

Mục tiêu ứng dụng	Cung cấp **benchmark NLI adversarial** đầu tiên cho tiếng Việt, nhằm thử thách mô hình và cải thiện độ **robust**. Các mô hình SOTA khi đánh giá trên ViAdverNLI cho kết quả rất thấp (mô hình mạnh nhất chỉ ~48% trên test) [53], cho thấy những điểm yếu cần khắc phục. Ngoài ra, khi huấn luyện trên ViAdverNLI, mô hình **cải thiện đáng kể hiệu quả trên các dataset NLI khác** [53] – chứng tỏ tập dữ liệu này giúp nâng cao khả năng khái quát và chống "bẫy" của mô hình. ViAdverNLI đặt nền móng cho hướng nghiên cứu NLI chống lại mẫu nhiễu,	Thiết lập **benchmark NLI tiếng Việt đầu tiên** (COLING 2022) [67], làm chuẩn đánh giá các mô hình suy luận tiếng Việt. ViNLI đã được sử dụng rộng rãi để so sánh hiệu quả mô hình và thúc đẩy nghiên cứu chuyên biệt (ví dụ, phát sinh ViHealthNLI cho y tế) [68]. Nhờ ViNLI, cộng đồng có điểm bắt đầu để phát triển các mô hình NLI cho tiếng Việt, khi trước đó hầu như chưa có dữ liệu suy luận công khai nào [69].	Cung cấp **benchmark fact-checking trên tri thức Wikipedia** đầu tiên cho tiếng Việt [70]. Hỗ trợ nghiên cứu các hệ thống kiểm chứng tự động, đặc biệt về **truy hồi bằng chứng** và suy luận phân loại đúng/sai. ViWikiFC được dùng làm bài toán thách thức: hệ thống SemViQA (thắng UIT Workshop) đạt 80,82% độ chính xác nghiêm ngặt, còn mô hình baseline InfoXLM+BM25 chỉ ~67% [71] [72] – cho thấy yêu cầu mô hình phải hiểu ngữ nghĩa sâu và suy luận phức tạp [73]. Dataset này dự kiến công bố rộng rãi, kỳ vọng thúc đẩy nghiên cứu fact-checking dùng tri thức mở (Wikipedia) cho các ngôn ngữ ít tài nguyên.	Đưa ra **benchmark kiểm chứng tin tức đa lĩnh vực** đầu tiên cho tiếng Việt (AAAI 2025) [74] [75]. Mục tiêu đáp ứng nhu cầu cấp bách về công cụ kiểm chứng tự động trong bối cảnh tin giả lan nhanh [76], nâng độ chính xác của hệ thống kiểm tin tiếng Việt [77]. Nhóm tác giả đã thử nghiệm nhiều mô hình SOTA (kể cả model ngôn ngữ lớn) trên ViFactCheck; đáng chú ý mô hình Gemma đạt F1 macro 89,9%, thiết lập SOTA mới [78] [79]. Dataset kỳ vọng được dùng rộng rãi để huấn luyện & đánh giá hệ thống phát hiện tin sai, và có thể	Làm **tập huấn luyện & đánh giá chính** cho cuộc thi **UIT Data Science Challenge 2023 (task Fact-Checking)** [83] [84], nơi nhiều đội phát triển mô hình kiểm chứng tự động. Kết quả cuộc thi cho thấy dataset này đã **đẩy mạnh hiệu quả mô hình**: hệ thống thắng cuộc (SemViQA) đạt 78,97% accuracy nghiêm ngặt trên test – vượt xa baseline ban đầu [85]. Sau cuộc thi, ISE-DSC01 tiếp tục được sử dụng trong nghiên cứu học thuật về fact-checking tiếng Việt (huấn luyện BERT/ RoBERTa, thử nghiệm

Tiêu chí	ViAdverNLI (R1–R3)	ViNLI	ViWikiFC	ViFactCheck	ISE-DSC01
	nâng cao độ tin cậy của hệ thống suy luận tự động.			làm nền tảng cho các cuộc thi, ứng dụng thực tiễn trong tương lai [80] [81]. (Dữ liệu và mã nguồn đã công khai trên GitHub [82]).	phương pháp truy hồi chứng cứ mới) [86]. Nhìn chung, ISE-DSC01 đánh dấu bước tiến lớn, tạo nền tảng dữ liệu lớn thực tế để phát triển & benchmark hệ thống kiểm chứng tự động tiếng Việt [87].

(Nguồn: tổng hợp thông tin từ các công bố ViNLI [3] [19] [20], ViWikiFC [13] [62], ViFactCheck [15] [56], ISE-DSC01 [7] [40] và ViAdverNLI [2] [18].)

Nhận xét so sánh và điểm mạnh của ViAdverNLI

Qua bảng trên, có thể thấy **ViAdverNLI** sở hữu nhiều điểm khác biệt nổi bật so với các bộ dữ liệu trước đây:

- **Độ khó và tính thách thức cao:** ViAdverNLI được thiết kế có chủ đích để *làm khó mô hình*. Điều này thể hiện ở việc mô hình mạnh nhất chỉ đạt ~48% chính xác trên dữ liệu này [53] – thấp hơn đáng kể so với độ chính xác ~79–90% trên ViFactCheck hay ISE-DSC01 (và ~80%+ trên ViWikiFC, ViNLI). ViAdverNLI liên tục *phơi bày điểm yếu* của mô hình qua các vòng, buộc mô hình phải cải thiện khả năng hiểu ngữ cảnh và tránh suy luận dựa trên mẹo. Đây là ưu điểm quan trọng: dataset càng khó sẽ càng thúc đẩy nghiên cứu mô hình NLI **robust** hơn trước những trường hợp đánh lừa.

- **Chiến lược *adversarial* nhiều vòng độc nhất:** Khác với các bộ dữ liệu còn lại xây dựng một lần, ViAdverNLI áp dụng quy trình *human-in-the-loop* qua 3 vòng (R1→R2→R3). Ở mỗi vòng, annotator tìm cách tạo câu hypothesis mà mô hình hiện tại dự đoán sai; những mẫu "đánh bại" được mô hình xác nhận và thêm vào dữ liệu [51]. Chiến lược này giúp ViAdverNLI thu thập được nhiều hiện tượng suy luận **hiểm hóc** (từ việc chơi chữ, tráo đổi từ đồng nghĩa đến tạo bẫy ngữ nghĩa) – những thứ không xuất hiện nhiều trong các dataset chuẩn. Nhờ đó, ViAdverNLI có độ đa dạng cao về cách "bẫy" mô hình, làm cho nó trở thành thước đo tốt để đánh giá khả năng **tổng quát hóa** và chống lại mẫu adversarial của mô hình NLI.

- **Tính đa dạng ngôn ngữ và hiện tượng suy luận:** ViAdverNLI khuyến khích annotator **diễn đạt lại** premise bằng nhiều cách mới (giữ nghĩa hoặc đảo nghĩa) thay vì sao chép từ ngữ. Điều này tạo nên tập dữ liệu với tỷ lệ trùng từ rất thấp và nhiều từ vựng mới trong hypothesis [45]. Các mẫu trong ViAdverNLI bao phủ nhiều hiện tượng: từ phủ định gián tiếp, ẩn dụ, đến thay đổi chi tiết nhỏ (như "30 phút" vs "nửa giờ") khiến mô hình dễ nhầm lẫn [88] [89]. So với ViNLI hay ViFactCheck – nơi câu hypothesis/claim thường dùng từ ngữ khá sát với premise/bằng chứng – ViAdverNLI *đa dạng và khó đoán* hơn, buộc mô hình học cách **hiểu ngữ nghĩa thực sự** thay vì dựa vào từ khóa đơn thuần.

- **Hiệu quả trong cải thiện mô hình:** Một lợi thế đáng chú ý là khi huấn luyện mô hình trên ViAdverNLI, hiệu suất trên các bộ dữ liệu khác được cải thiện rõ rệt [53]. Điều này gợi ý rằng ViAdverNLI không chỉ đóng vai trò kiểm tra độ bền của mô hình, mà còn như một bộ dữ liệu huấn luyện giúp mô hình **tổng quát tốt hơn** (có lẽ do học được cách xử lý các bẫy ngôn ngữ). Ngược lại, các dataset khác tuy hữu ích để đánh giá thông thường, nhưng có thể chưa đủ phong phú để nâng cao độ bền cho mô hình trước các ví dụ "hiểm".

- **Bổ sung cho khoảng trống tài nguyên:** Trước ViAdverNLI, các benchmark NLI tiếng Việt (như ViNLI, VnNewsNLI) đều tập trung vào dữ liệu "thẳng" và chất lượng, chưa chú trọng khía cạnh adversarial [90]. ViAdverNLI ra đời đã **bổ khuyết khoảng trống** này, đưa tiếng Việt vào xu hướng nghiên cứu NLI tiên tiến giống như tiếng Anh (ANLI) hay tiếng Trung [91]. Trong bối cảnh phát hiện sai lệch thông tin ngày càng quan trọng, dataset này đóng góp cách đánh giá mới, khuyến khích phát triển mô hình NLI và fact-checking có khả năng chống chịu tốt hơn với thông tin gây nhiễu.

Tóm lại, **ViAdverNLI** nổi trội nhờ cách xây dựng sáng tạo và độ phức tạp cao, giúp kiểm tra và huấn luyện mô hình ở một mức độ **khó** hơn so với các bộ dữ liệu hiện có. Trong khi ViNLI cung cấp nền tảng cơ bản về suy luận, ViWikiFC mở rộng sang tri thức Wikipedia, ISE-DSC01 và ViFactCheck đem lại khối lượng dữ liệu lớn và đa miền cho fact-checking, thì ViAdverNLI tập trung vào khía cạnh *thử thách mô hình*, khiến nó trở thành **benchmark đặc biệt giá trị** để thúc đẩy các nghiên cứu NLI/fact-checking tiếng Việt lên tầm cao mới. Các điểm mạnh này của ViAdverNLI bổ sung cho những benchmark sẵn có, cùng nhau tạo nên một bộ dữ liệu đa dạng giúp phát triển các hệ thống kiểm chứng thông tin và suy luận ngôn ngữ tự nhiên ngày càng toàn diện và đáng tin cậy hơn.

Nguồn tài liệu tham khảo: ViNLI [3] [20]; ViWikiFC [13] [71]; ViFactCheck [15] [56]; ISE-DSC01 [7] [40]; ViAdverNLI [18] [53].

---

[1] [2] [8] [9] [17] [18] [41] [42] [45] [46] [47] [51] [52] [53] [57] [88] [89] [90] [91] ViANLI: Adversarial Natural Language Inference for Vietnamese
https://arxiv.org/html/2406.17716v2

[3] [4] [5] [6] [7] [13] [14] [15] [16] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [54] [55] [56] [62] [63] [64] [65] [66] [67] [68] [69] [70] [71] [72] [73] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83] [84] [85] [86] [87] VNese Fact-Checking Data Collection.pdf
file://file-UxTqzMGcQumdBMqzfR2UFu

[10] [11] [12] [43] [44] [48] [49] [50] [58] [59] [60] [61] ViNLI: A Vietnamese Corpus for Studies on Open-Domain Natural Language Inference
https://aclanthology.org/2022.coling-1.339.pdf