HDCR: Cross-lingual Medical Misinformation Detection through Contrastive Claim-Evidence Reasoning

Chaoyuan Zuo*
School of Information and Communication
Nankai University
Tianjin, China
zuocy@nankai.edu.cn

Ritwik Banerjee
Department of Computer Science
Stony Brook University
Stony Brook, USA
rbanerjee@cs.stonybrook.edu

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356375

***Abstract*—The rapid dissemination of health information online enables dangerous distortions that threaten public health. We present HD^{2}CR (health-information distortion detection with contrastive reasoning), a framework for fine-grained detection of medical misinformation. Through systematic analysis of health news patterns, we identify four primary distortion types: over-generalization, exaggeration, under-generalization, and false causality. Our contributions include: a cross-lingual corpus of 72,275 English and Chinese claim-evidence pairs with validated distortion labels; a dual-encoder architecture with contrastive cross-attention that explicitly models semantic divergence between claims and biomedical evidence; and extensive evaluations demonstrating HD^{2}CR's superior performance: 93.1% binary F_1 and 87.3% 5-class accuracy, with robust cross-lingual generalization (only 2.9% degradation between Chinese and English).**

***Index Terms*—Clinical misinformation detection, Cross-lingual NLP, Health informatics, Fact-checking**

I. Introduction

The digital dissemination of medical information poses major risks to public health and clinical decision-making. Research findings often undergo multiple interpretations, from peer-reviewed articles to news outlets and social media for wide public consumption (Figure 1) — each stage introducing potential distortions, with clinical consequences. Leading outlets such as *The New York Times*, *CNN*, *MedPage Today*, and *STAT News* often serve as primary interpreters of research findings, bridging the gap between domain-specific language and public audiences. Even professional health journalism is vulnerable, however: systematic evaluations reveal that only 20% of reports preserve complete clinical accuracy, while nearly half omit crucial details such as study limitations, population constraints, or confidence intervals [1]. The COVID-19 pandemic starkly illustrated the dangers posed by these phenomena. Around 18% of health content on social media contained false information, fueling unsafe self-medication, vaccine hesitancy, and delays in care [2]–[4]. These challenges highlight the urgent need for automated systems to detect and mitigate medical misinformation.

*Corresponding author

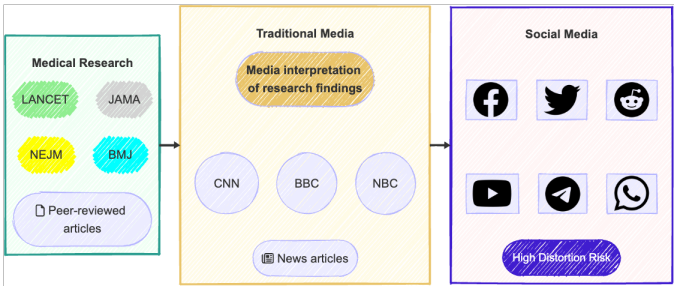

Fig. 1. Flow (→) of medical research information from peer-reviewed articles, through the interpretations of traditional media, to social media.

Crucially, inaccuracies rarely manifest as outright falsehoods. Instead, they appear as a variety of distortions and misinterpretations ranging from overgeneralization and exaggeration to disease-mongering and the omission of clinically relevant context [3], [5]–[7]. Our analysis of health news reporting reveals four predominant distortion types: (i) *overgeneralization*, (ii) *undergeneralization* or improper restriction, (iii) *exaggeration* or *amplification*, and (iv) *false causality*. Exemplars in Table I illustrate how scientific findings are systematically transformed through these mechanisms. The challenge intensifies in cross-lingual contexts. Because most medical research is published in English, reporting in other languages–such as Chinese–introduces additional risks through translation inaccuracies and semantic drift in specialized terminology. This linguistic and interpretive complexity underscores the need for sophisticated detection methods to recognize diverse and varied distortion phenomena.

Current solutions to medical misinformation are limited to fact-checking systems that offer binary `TRUE|FALSE` labels or indicate the 'severity' or 'degree' of falsehoods [8], [9]. These approaches are not equipped to detect the subtle yet clinically significant semantic distortions that healthcare providers need to understand and address (Table I)—an increasingly recognized gap [7], [10], [11]. Thus, we argue that maintaining the integrity of medical information requires a

TABLE I
TAXONOMY OF MEDICAL MISINFORMATION DISTORTIONS WITH REPRESENTATIVE EXAMPLES FROM HEALTH NEWS REPORTING

Case Study: Low-Fat Diet and Breast Cancer Risk		
Original article:	Study hints at potential risk between unhealthy low-fat diets and postmenopausal breast cancer.	
Source:	Health News Daily, February 2024.	
Research reference:	Associations between overall, healthful, and unhealthful low-fat dietary patterns and breast cancer risk in a Mediterranean cohort: The SUN project (DOI: 10.1016/j.nut.2022.111967)	
Exemplars and types of medical information distortion		
Overgeneralization	Inappropriately extending limited research findings to broader populations or situations beyond the validated scope	Low-Fat Diets Increase Breast Cancer Risk in All Women
Undergeneralization (*improper restriction*)	Inappropriately narrowing the applicability of well-established medical evidence to specific populations without scientific justification	Breast Cancer Diet Guidelines Only Apply to Women Over 65
Exaggeration (*amplification bias*)	Inappropriately amplifying treatment effects, risk levels, or statistical significance beyond what evidence supports	Shocking Study: Low-Fat Diets Dramatically Increase Breast Cancer Risk
False causality (*spurious causation*)	Incorrectly interpreting correlation or temporal association as a causal relationship without sufficient evidence	Scientists Prove Low-Fat Diets Cause Breast Cancer

deeper understanding of the semantic distortions that enter the information flow depicted in Fig. 1: from scientific evidence to their traditional and social media interpretations. To address this gap, we make three key contributions:

1. We introduce the **taxonomy of semantic distortion types grounded in clinical communication needs** (Table I), and construct a **large-scale cross-lingual corpus** of 72,000+ claim-evidence pairs linking health claims to peer-reviewed biomedical research literature. This represents the first large-scale resource specifically for evidence-based fine-grained identification of medical information distortion.
2. We propose HD^{2}CR (**h**ealth-information **d**istortion **d**etection with **c**ontrastive **r**easoning), a **novel clinical evidence reasoning architecture** employing dual encoders with cross-attention that enables integration with clinical decision support systems by providing interpretable detection of specific types of information distortion.
3. We provide **comprehensive benchmarking on two tasks: (i) distortion detection, and (ii) distortion type detection**. Our experiments with multiple competitive models find that HD^{2}CRachieves 0.931 F_1-score on the former, and 87.4% accuracy on the latter. Furthermore, we observe only a mild degradation of 2.9% between English and Chinese, underscoring the robustness of our model across languages and its potential for transnational health applications.

We thus identify information distortion as a critical impediment to healthcare, and provide tools to address this problem through the identification of peer-reviewed research evidence.

II. DATASET

In three stages, we construct a large cross-lingual corpus to detect fine-grained information distortion:[1] data collection, where health news articles are linked to peer-reviewed biomedical research; data validation, where claim-evidence pairs are verified for clinical relevance; and generation of medically plausible distortions across the four types described in our taxonomy. The final corpus comprises 72,275 samples labeled for distortion detection and classification of distortion types.

A. Medical News Data Collection

We collect health news articles explicitly linked to peer-reviewed biomedical research, ensuring verifiability of each claim against authoritative clinical evidence.

Our **English-language medical news** collection expands the validated dataset from our earlier work [12], extending through 2025 to yield 15,108 articles (over 200% increase in articles with explicit ground-truth links). The corpus is sourced from Reuters Health, CNN Health, and specialized medical journalism platforms, spanning major domains including oncology, infectious diseases, cardiovascular health, and neurology. For **Chinese-language medical news**, we use the health-claim verification corpus by Zuo et al. [13]: from 13,748 articles collected from Chinese medical news platforms (2017-2023), we retain 1,439 after filtering for verifiable scientific citations and substantial medical content. This collection of 16,547 news articles, each linked to corresponding peer-reviewed publications, offers real-world medical information verification data in monolingual and cross-lingual settings.

B. Clinical Validation of Claims and Evidence

News headlines constitute our primary health claims, given their critical role in shaping public health understanding. For each article, supporting evidence derives from explicitly linked peer-reviewed research. We extract their titles and abstracts to provide sufficient clinical detail for verification.

Each claim is required to (a) contain specific medical information amenable to scientific verification, and (b) demonstrate substantive correspondence with the underlying evidence. We employ GPT-4 for scalable validation [14]. A biomedical expert validated 100 randomly sampled pairs, reporting 98% agreement with GPT-4 assessments—substantially higher than earlier human-LLM agreement evaluations [15]. This process retains 14,455 (87.4%) claim-evidence pairs (13,321 English and 1,134 Chinese) as ground truth.

[1]Available at https://zenodo.org/records/17486207

TABLE II
PRETRAINED LANGUAGE MODELS USED AS BASELINE CLASSIFIERS

General-purpose multilingual encoders:
1. `XLM-RoBERTa` [17], trained on 2.5TB of multilingual data across 100 languages, including medical terminology.
2. `mBERT` [18], a BERT variant pretrained on Wikipedia's medical content in 104 languages.

Multilingual embeddings for semantic similarity:
1. `BGE-M3` [19], state-of-the-art multilingual embeddings supporting medical terminology across 100+ languages with unified representations.
2. `Multilingual-E5` [20], offering contrastive pretraining on 1 billion multilingual text pairs

Domain-specific models:
1. `BioBERT` [21], pretrained on 4.5B words from PubMed abstracts and 13.5B words from PubMed Central articles.
2. `PubMedBERT` [22], pretrained exclusively on biomedical literature without general domain data.

C. Systematic Generation of Plausible Distortions

To mirror real-world medical information propagation, where distortions emerge from interpreting health information without access to original research, our system operates exclusively on health claims deliberately isolated from their research context. For each validated claim, we use GPT-4 to generate distortions of all four types in our taxonomy (Table I).

We design category-specific prompts incorporating linguistic markers from real medical misinformation cases, with 5 variations per category. Clinical plausibility is ensured using Sentence-BERT [16]: we retain variations with 0.4-0.9 cosine similarity to the original claim, a range that yields itself to topical relevance while preventing trivial variations.

Expert validation was conducted where two biomedical scholars independently annotated 100 randomly selected claim-evidence pairs across all distortion types, with no knowledge of the generation process. This revealed near-perfect agreement among all three annotators (two human and GPT-4): 91% complete agreement with Fleiss' Kappa $\kappa = 0.921$. Our final corpus comprises 72,275 $\{c, r, l\}$ triplets, representing the claim c (original or distorted), corresponding evidence r (title and abstract), and distortion type l. The collection is partitioned using stratified sampling into 60% training, 30% test, and 10% development data.

A Brief Descriptive Overview of the Dataset

Our corpus exhibits systematic linguistic patterns: distorted claims average 41% more words than accurate ones (13.7 vs 9.7 words in Chinese; 14.4 vs 10.2 in English), suggesting deliberate elaboration in misinformation. The compression from research abstracts (268 words English, 225 Chinese) to claims creates an approx. 26:1 ratio, risking loss of critical clinical qualifiers. Language-specific framing and cultural medical frameworks are revealed as well: English sources prioritize cardiovascular terms and sustained COVID-19 coverage, while Chinese sources emphasize holistic systemic terms ("blood", "immune", "therapy"). We also find that misinformation often target domains of high patient anxiety (respiratory and cardiovascular), as these areas show higher distortion rates.

III. METHODOLOGY

Problem formulation: Given a public-facing health claim c in English or Mandarin Chinese, and corresponding medical evidence r (in English, from peer-reviewed research), our task is to classify c as either *accurate* ($\checkmark$) or as one of four distortion types: *overgeneralization* ($\Uparrow$), *under-generalization* ($\Downarrow$), *exaggeration* ($\ggg$), or *false causality* ($\nRightarrow$). Formally, we learn a function $f : (c, r) \rightarrow y,$ where $y \in \{\checkmark, \Uparrow, \Downarrow, \ggg, \nRightarrow\}$.

Our approach is to first establish baseline methods ranging from feature-based supervised learning to domain-specific language models, and then introduce our novel HD^{2}CR framework, which explicitly models claim-evidence relationships to identify clinically significant distortions.

A. Baseline Methods: Feature-based Classifier

The first baseline uses `tf-idf` of unigram and bigram features with SVM for classification, an approach with demonstrated success in medical text classification [23], [24]. For each (c, r) pair, we concatenate the texts, apply `tf-idf` encoding with sublinear `tf` scaling, and consider unigrams and bigrams to retain a maximum of 10*K* features. For Chinese-language claims, we use professional translations to English before vectorization (this is also done for our subsequent experiments described in Sections III-B, III-C, and III-D). With these feature vectors, we use a linear SVM with L2 regularization. Despite its simplicity, we find this to be a competitive computationally efficient baseline, especially suitable for low-resource clinical deployment (§IV).

B. Baseline Methods: Pretrained Language Models

We evaluate three categories of pretrained language models (Table II): general-purpose multilingual encoders, which are masked language models that produce contextualized representations and need task-specific heads; multilingual embeddings specifically trained for semantic similarity and retrieval tasks, which are optimized to produce high-quality fixed embeddings for similarity comparison; and domain-specific models.

For the general-purpose encoders, we adopt a standard practice in biomedical text processing, with the input formatted as `[CLS] <claim> [SEP] <evidence> [SEP]`, followed by a classification head with dropout ($p = 0.1$) to prevent overfitting on domain-specific terminology [22]. For the embeddings pretrained specifically for semantic similarity, we use separate representations for the claim and the the evidence, and then concatenate the two to produce the classifier input. This approach preserves their respective medical contexts.

C. Baseline Methods: Large Language Models

We select GPT-4 as a representative baseline for state-of-the-art LLMs, acknowledging that comprehensive comparison across multiple LLMs is beyond the scope of this work. To this end, we use two clinically-informed prompting strategies:
1. the model receives claim-evidence pairs with the definitions of each distortion type (**zero-shot classification**).
2. the task is structured as a clinical decision-making process (**chain-of-thought prompting** [25]): (i) extract key medical

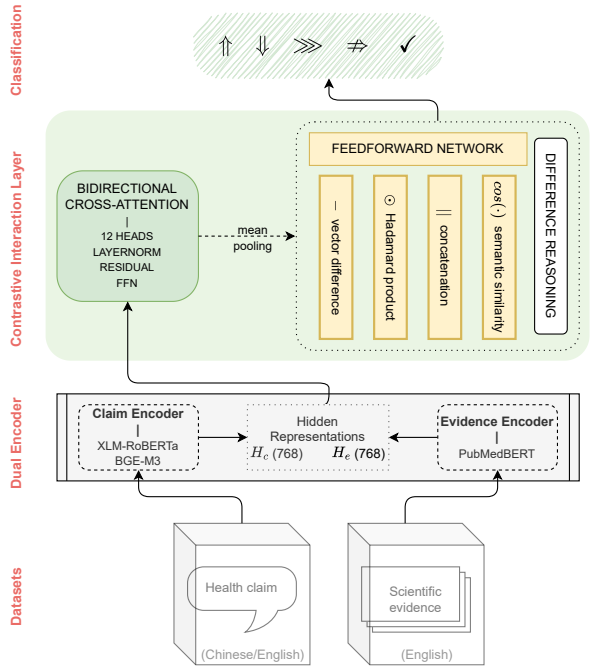

Fig. 2. HD^{2}CR architecture for medical misinformation detection, featuring *dual encoders* for claim and evidence processing, a *contrastive interaction layer* for semantic alignment, a *difference reasoning module* for distortion identification, and a *classification head* for clinical categorization.

facts from evidence, (ii) identify clinical claims, (iii) compare population scope and effect sizes, and (iv) determine the type of information distortion.

D. The HD^{2}CR Clinical Evidence Reasoning Framework

While baseline methods achieve reasonable performance, they lack the ability to capture claim-evidence relationships critical for clinical understanding, and treat the distortion taxonomy as generic categories. HD^{2}CR addresses this with architectural innovations shown in Figure 2, comprising *dual encoders* to separately process claims and medical evidence, a *contrastive interaction layer* to identify semantic divergence, and finally, a *classification head* to identify the distortion type.

Separate encoders handle health claims and scientific evidence independently, as claims use simplified lay language while evidence contains technical medical terminology. We evaluate two encoder configurations to represent the claims — BGE-M3 and XLM-RoBERTa — while always using PubMedBERT to represent the English-language research evidence. Each encoder produces contextual representations:

$$\mathbf{H}_c \in \mathbb{R}^{L_c \times d_c}; \qquad \mathbf{H}_e \in \mathbb{R}^{L_e \times d_e}$$

where L_c and L_e denote sequence lengths for claims and evidence respectively, with d_c and d_e as hidden dimensions (1024 for BGE-M3, 768 for medical BERT variants).

Our core innovation is modeling **contrastive interaction** with *bidirectional* **cross-attention** (CA), to capture how claims relate to scientific evidence. We employ with 12 attention heads to capture complex medical relationships, and identify *how* a claim diverges from (or aligns with) the evidence.

$$\mathbf{H}'_c = \mathrm{CA}_{c\to e}(\mathbf{H}_c, \mathbf{H}_e, \mathbf{H}_e); \qquad \mathbf{H}'_e = \mathrm{CA}_{e\to c}(\mathbf{H}_e, \mathbf{H}_c, \mathbf{H}_c)$$

Modeling contrast in this manner aids the identification of clinically relevant nuances such as omitted population constraints, exaggerations, or misrepresented causal relationships.

Next, our **difference reasoning** module takes four fixed-size representations ($\mathbf{h}_c, \mathbf{h}'_c, \mathbf{h}_e, \mathbf{h}'_e \in \mathbb{R}^d$) obtained through mean pooling), and computes clinically-relevant features including: $[\mathbf{h}'_c - \mathbf{h}_c; \mathbf{h}'_e - \mathbf{h}_e]$ to capture semantic shift, Hadamard product for element-wise differences, and cosine similarity as a coarse measure of semantic alignment. They pass through a domain-optimized feedforward network to produce the representation to be classified into one of four type of clinical information distortion (see Table I) or as accurate. Our classifier is optimized using a class-weighted cross-entropy loss with label smoothing ($\alpha = 0.1$): $\mathcal{L}_{\text{CE}}^{\text{smooth}}(\mathbf{y}, \hat{\mathbf{y}}, \alpha, \mathbf{w})$, where $\mathbf{w}$ denotes inverse class frequency weights.

TABLE III
PERFORMANCE ON HEALTH MISINFORMATION DETECTION. BEST RESULTS IN BOLD, SECOND-BEST UNDERLINED.

Model	**Classification Task**							
	Binary				**4-class**		**5-class**	
	P	R	F_1	MCC	F_1	Acc.	F_1	Acc.
Baseline methods: feature-based classifiers								
SVM	81.1	81.0	81.1	0.055	70.5	66.2	57.9	57.9
Baseline methods: pretrained language models								
mBERT	92.2	92.5	92.4	0.615	93.8	90.7	86.3	86.3
XLM-RoBERTa	91.7	93.2	92.4	0.610	93.8	91.0	86.0	86.1
BGE-M3	92.5	93.0	92.7	0.633	94.4	91.5	87.1	87.1
Multilingual E5	91.2	93.5	92.3	0.598	94.5	91.8	86.1	86.2
BioBERT	92.0	93.3	92.6	0.622	94.4	91.6	86.8	86.8
PubMedBERT	92.3	93.2	92.7	0.629	94.5	91.7	87.1	87.1
Baseline methods: large language models								
GPT-4 (zero shot)	82.9	90.9	86.7	0.198	59.8	57.0	51.3	50.6
GPT-4 (CoT)	86.8	82.9	84.8	0.308	57.4	53.5	53.0	52.8
HD^{2}CR (*with PubMedBERT for evidence encoding*)								
BGE-M3	**93.3**	92.1	92.7	**0.643**	93.9	90.7	87.2	87.2
XLM-RoBERTa	91.8	**94.5**	**93.1**	0.636	**95.1**	**92.8**	**87.3**	**87.4**

IV. EXPERIMENTS AND RESULTS

We evaluate our framework on the cross-lingual biomedical dataset (Section II) of 72,275 claim-evidence pairs with clinically-validated distortion labels. All experiments use a NVIDIA RTX 4090 GPU. The hyperparameters are identical across neural models: batch size 16, learning rate 2e-5 with AdamW optimizer, and 2 epochs with early stopping on the development set. We assess performance for binary ('accurate' vs. 'distorted'), 4-class (distortion types only), and 5-class (all distortion types, plus 'accurate') classification. Given the class imbalance for binary classification (4:1 distorted-to-

TABLE IV
CROSS-LINGUAL F_1 PERFORMANCE. Δ DENOTES AVERAGE ABSOLUTE PERFORMANCE GAP BETWEEN LANGUAGES (LOWER IS BETTER).

Model	English		Chinese		Δ
	Binary	5-class	Binary	5-class	
SVM	81.4	58.5	77.0	50.1	8.4
XLM-RoBERTa	**93.2**	87.6	**92.2**	83.8	3.8
PubMedBERT	92.9	**87.6**	90.0	80.6	7.0
BGE-M3	92.9	87.5	91.0	83.2	4.3
GPT-4 (0-shot)	86.7	51.6	86.9	47.7	3.9
HD^{2}CR (with BGE-M3)	92.7	87.5	91.8	**84.5**	**2.9**

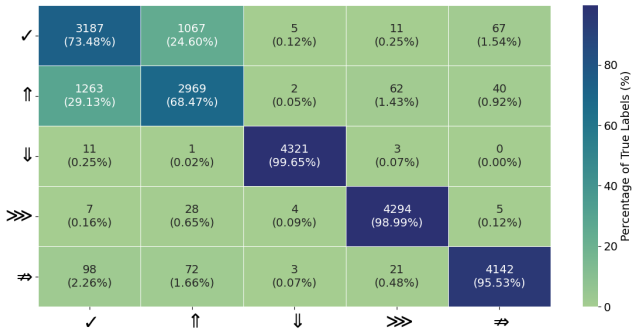

Fig. 3. 5-class confusion matrix for HD^{2}CR (with BGE-M3).

accurate ratio), we report precision, recall, F_1, and Matthews Correlation Coefficient (MCC).

Table III offers several critical insights. The feature-engineering approach with SVM achieves a misleading F_1 of 81.1%, while its MCC is extremely low (0.055). This exposes a fundamental flaw, that the model exploits class imbalance rather than learning meaningful distinctions, and highlights why understanding medical information distortion requires a distillation of nuanced semantic differences.

GPT-4's performance reveals a striking limitation of prompting-based approaches. While zero-shot prompting achieves high recall (90.9%), it comes at the cost of precision, yielding an MCC of only 0.198. More critically, performance collapses dramatically on fine-grained distortion categorization (4-class F1: 59.8%, and clinical chain-of-thought prompting providing a 5-class F1 of 53.0%) This suggests that even sophisticated reasoning capabilities fail to identify subtle clinically important distortions, underscoring the need for architectures designed specifically for this task.

Across diverse pretrained architectures — multilingual (mBERT, XLM-RoBERTa, BGE-M3), general English (Multilingual E5), and biomedical (BioBERT, PubMedBERT) — performance converges remarkably tightly (binary F_1: 92.3-92.7%, 5-class accuracy: 86.1-87.1%). This plateau persists despite architectural differences and pretraining objectives, indicating that standard single-encoder architectures approach a ceiling when treating information distortion as generic text classification. Multilingual embedding models (BGE-M3, Multilingual E5) marginally outperform transformer encoders, which suggest that dense semantic similarity representations are slightly better at capturing claim-evidence relations.

HD^{2}CR breaks the plateau through architectural innovation: the dual-encoder design with contrastive interaction achieves consistent gains where model scale and pretraining alone fail. XLM-RoBERTa reaches 93.1% binary F_1 and 95.1% 4-class F_1. Crucially, the improvement is most pronounced on 4-class categorization (+0.6 F_1), precisely where semantic precision matters most. This validates our hypothesis that explicitly modeling claim-evidence divergence through bidirectional cross-attention captures critical reasoning patterns that remain latent in standard encoders. HD^{2}CR with BGE-M3 achieves the highest binary MCC (0.643) and precision (93.3%), while XLM-RoBERTa excels at recall (94.5%) and fine-grained classification. This configuration-dependent performance suggests that different encoder combinations capture complementary aspects of medical misinformation: BGE-M3's dense retrieval pretraining enhances precision in identifying distortions, while XLM-RoBERTa's cross-lingual capacity improves generalization across claim phrasings.

Biomedical models (BioBERT, PubMedBERT) struggle with machine-translated Chinese claims despite competitive overall performance, as their English-centric pretraining creates sensitivity to translated clinical terminology. In contrast, HD^{2}CR maintains robust cross-lingual performance (only 2.9% degradation) by explicitly comparing claims against evidence rather than analyzing them in isolation, learning distortion patterns that persist across languages. This is crucial, since research literature is primarily in English while health misinformation disseminates globally in many languages.

Figure 3 reveals clinically significant patterns. HD^{2}CR excels at detecting three categories: *false causality* (95.53%), *under-generalization* (99.65%), and *exaggeration* (98.99%). However, separating accurate claims from over-generalizations remains challenging due to omissions (e.g., missing population qualifiers, dropped contraindications), requiring domain expertise beyond linguistic context.

V. RELATED WORK

Medical misinformation detection requires specialized approaches for clinical safety. Kotonya and Toni [26] introduced PUBHEALTH (11.8K expert-validated claims), establishing explainability for medical fact-checking. During COVID-19, Hossain et al. [27] developed COVIDLies (6,761 clinical annotations) combining misconception retrieval with stance detection. However, recent reviews reveal persistent failures in detecting subtle clinical distortions—effect exaggeration, scope misrepresentation, contraindication omissions—that directly impact health outcomes [10].

Cross-lingual approaches by Gupta and Srikumar [28] (X-Fact, 25 languages) and Kazemi et al. [29] (206K fact-checks across 39 languages) address multilingual verification but overlook healthcare's unique challenge: health information disseminates in local languages while biomedical evidence is primarily available only in English.

Recent advances employ fine-grained categorization: Da San Martino et al. [30] identified 18 propaganda techniques, albeit not specifically for medical information, while Song et al. [31] developed categories specific to COVID-19. Others have expanded this area of work by incorporating knowledge graphs [32], [33]. We extend these by explicitly modeling claim-evidence semantic divergence through contrastive reasoning, designed specifically for clinical distortion patterns.

VI. Conclusion

We present HD^{2}CR, a contrastive reasoning framework for the detection of medical information distortion, along with a task-specific taxonomy. Our contributions include: (1) the first large-scale cross-lingual dataset with 72,275 clinically-validated claim-evidence pairs; (2) a dual-encoder architecture modeling semantic divergence between health claims and biomedical evidence, achieving 93.1% F_1 for distortion detection; and (3) robust cross-lingual performance with only 2.9% degradation. HD^{2}CR excels at detecting high-risk distortions like false causality and exaggeration, which are critical for preventing dangerous self-medication and treatment abandonment. As medical misinformation increasingly threatens global health outcomes, HD^{2}CR provides essential infrastructure for clinical decision support systems helping healthcare providers address public misconceptions. Our dataset and models are made available for further research and wider applications of evidence-based healthcare communication.

Acknowledgment

Chaoyuan Zuo was supported by the National Natural Science Foundation of China (Grant No. 62406150). Ritwik Banerjee were supported in part by the U.S. National Science Foundation under the award CNS-2335686.

References

[1] A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud, "Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study," *PLOS Medicine*, vol. 9, no. 9, pp. 1–11, 2012.

[2] R. Kouzy, J. A. Jaoude, A. Kraitem, M. B. E. Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter," *Cureus*, vol. 12, no. 3, p. e7255, 2020.

[3] I. J. B. Do Nascimento, A. B. Pizarro, J. M. Almeida, N. Azzopardi-Muscat, M. A. Gonçalves, M. Björklund, and D. Novillo-Ortiz, "Infodemics and health misinformation: a systematic review of reviews," *Bull. World Health Organ.*, vol. 100, no. 9, p. 544, 2022.

[4] K. Devi and P. Singh, "Truth in the Age of Clickbait: A Review of Social Media Misinformation Through Case Studies," *J. Commun. Manag.*, vol. 4, no. 02, pp. 32–40, 2025.

[5] R. Moynihan, L. Bero, D. Ross-Degnan, D. Henry, K. Lee, J. Watkins, C. Mah, and S. B. Soumerai, "Coverage by the news media of the benefits and risks of medications," *N. Engl. J. Med.*, vol. 342, no. 22, pp. 1645–1650, 2000.

[6] D. Wright and I. Augenstein, "Semi-supervised exaggeration detection of health science press releases," in *EMNLP*, 2021, pp. 10 824–10 836.

[7] C. Zuo, Q. Zhang, and R. Banerjee, "An Empirical Assessment of the Qualitative Aspects of Misinformation in Health News," in *Fourth Workshop on NLP for Internet Freedom*, 2021, pp. 76–81.

[8] L. Cui and D. Lee, "CoAID: COVID-19 Healthcare Misinformation Dataset," 2020, arXiv:2006.00885.

[9] A. Dharawat, I. Lourentzou, A. Morales, and C. Zhai, "Drink Bleach or Do What Now? COVID-HeRA: A Study of Risk-Informed Health Decision Making in the Presence of COVID-19 Misinformation," *ICWSM*, vol. 16, no. 1, pp. 1218–1227, 2022.

[10] V. Papanikou, P. Papadakos, T. Karamanidou, T. G. Stavropoulos, E. Pitoura, and P. Tsaparas, "Health Misinformation in Social Networks: A Survey of Information Technology Approaches," *Future Internet*, vol. 17, no. 3, p. 129, 2025.

[11] O. Vinhas and M. Bastos, "Fact-Checking Misinformation: Eight Notes on Consensus Reality," *Journal. Stud.*, vol. 23, no. 4, pp. 448–468, 2022.

[12] C. Zuo, N. Acharya, and R. Banerjee, "Querying across genres for medical claims in news," in *EMNLP*, 2020, pp. 1783–1789.

[13] C. Zuo, Y. Liu, C. Wang, and R. Banerjee, "From Claim to Evidence: Verifying Chinese Health Claims with Medical Literature," in *Natural Language Processing and Chinese Computing*, 2024, pp. 171–183.

[14] OpenAI, "GPT-4 Technical Report," 2024, arXiv:2303.08774.

[15] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. Do, Y. Xu, and P. Fung, "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity," in *IJCNLP-AACL*, 2023, pp. 675–718.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using siamese BERT-networks," in *EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *ACL*, 2020, pp. 8440–8451.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[19] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, Multi-functionality, Multi-granularity Text Embeddings Through Self-Knowledge Distillation," in *Findings of ACL*, 2024, pp. 2318–2335.

[20] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report ," 2024, arXiv:2402.05672.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019.

[22] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, 2021.

[23] P. Cichosz, "Bag of Words and Embedding Text Representation Methods for Medical Article Classification," *Int. J. Appl. Math. Comput. Sci.*, vol. 33, no. 4, pp. 603–621, 2023.

[24] G. Rabby and P. Berka, "Multi-class classification of COVID-19 documents using machine learning algorithms," *J. Intell. Inf. Syst.*, vol. 60, pp. 571–591, 2023.

[25] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *NeurIPS*, vol. 35, 2022, pp. 24 824–24 837.

[26] N. Kotonya and F. Toni, "Explainable automated fact-checking for public health claims," in *EMNLP*, 2020, pp. 7740–7754.

[27] T. Hossain and et al., "COVIDLies: detecting COVID-19 misinformation on social media," in *NLP COVID-19 Workshop at EMNLP*, 2020.

[28] A. Gupta and V. Srikumar, "X-Fact: A New Benchmark Dataset for Multilingual Fact Checking," in *ACL-IJCNLP*, 2021, pp. 675–682.

[29] M. Pikuliak and et al., "Multilingual previously fact-checked claim retrieval," in *EMNLP*, 2023, pp. 16 477–16 500.

[30] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news articles," in *EMNLP-IJCNLP*, 2019, pp. 5636–5646.

[31] X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, and K. Bontcheva, "Classification aware neural topic model for COVID-19 disinformation categorisation," *PLoS One*, vol. 16, no. 2, p. e0247086, 2021.

[32] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, and E. Choi, "FactKG: fact verification via reasoning on knowledge graphs," in *ACL*, 2023, pp. 16 190–16 206.

[33] Z. Yue, H. Zeng, L. Shang, Y. Liu, Y. Zhang, and D. Wang, "Retrieval augmented fact verification by synthesizing contrastive arguments," in *ACL*, 2024, pp. 10 331–10 343.