GraDuLa: Graph-based Dual-branch Language-agnostic Modeling for Misinformation Detection with Cross-lingual Evidences

Atul Kumar Singh
Fujitsu Research of India Pvt. Ltd.
Bengaluru, India
atul.singh@fujitsu.com

Rahul Mishra
Fujitsu Research of India Pvt. Ltd.
Bengaluru, India
Rahul.N.Mishra@fujitsu.com

Abstract

Misinformation detection has predominantly been studied in monolingual or language-aligned settings, where claims and evidences are assumed to be written in the same language. However, this assumption does not reflect real-world information ecosystems, in which misinformation frequently spans multiple languages and verification requires cross-lingual reasoning over heterogeneous evidence sources. As a result, cross-lingual claim-evidence reasoning remains largely unexplored and has not been systematically studied, and existing methods are not explicitly designed to support such scenarios. To address this gap, we introduce a novel setting for multilingual misinformation detection with cross-lingual evidences and propose GraDuLa, a graph-based dual-branch language-agnostic modeling for misinformation detection. GraDuLa projects evidential information into a language-agnostic subspace, disentangling semantic representations from language-specific associations through dual-branch graph modeling. For evaluation, we construct and release a benchmark *GraDuLaEval* and compare our approach against strong baseline methods. We further plan to release the dual-branch graph representations as an additional resource for future research.

1 Introduction

Current information ecosystems remain heavily biased toward resource-rich languages such as English (Yang et al., 2024)(Vykopal et al., 2025). Modern information processing and search systems predominantly rely on high-resource languages for information consumption, which limits their usability across many multilingual real-world scenarios(Goworek et al., 2025). While misinformation detection is a well-established research task and is increasingly seeing real-world deployment and productization, multilingual misinformation detection remains largely underexplored. Most existing studies focus on monolingual settings, where both the claim and the supporting evidence are expressed in the same language(Thorne et al., 2018)(Wang, 2017)(Gupta and Srikumar, 2021)(Augenstein et al., 2019). However, this assumption rarely holds, as claims and evidence may appear in different languages depending on factors such as the geographic location of events, media coverage, and the popularity of the events or involved entities (Dementieva and Panchenko, 2021)(Huang et al., 2022). This cross-lingual claim-evidence setting is largely unexplored and poses a significant challenge for existing misinformation detection systems, which are not designed to handle such scenarios.

Specifically, in our study, we observe that for many popular large language models (LLMs), veracity prediction performance drops by more than 10% when moving from same-language to cross-lingual claimevidence pairs (Table 5), consistent with prior findings on multilingual retrieval and cross-lingual reasoning degradation (Conneau et al., 2020)(Park and Lee, 2025). We attribute this drop in performance to language-specific clusters created within parametric memory of LLMs due to pre-training and post-training strategies, where concepts expressed in different languages are not uniformly aligned in the embedding space(Amiraz et al., 2025). As a result, semantically equivalent facts written in different languages may not be treated as similar by the model. Furthermore, we observe that currently available misinformation datasets do not provide sufficiently large cross-lingual claimevidence pairs to systematically study this challenge(Nakov et al., 2021).

Motivated by these observations and research gaps, this paper investigates cross-lingual claim-based misinformation detection along three primary directions. First, we systematically analyze the challenges of cross-lingual claim verification and demonstrate how language mismatch de-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33112–33131
October 24-29, 2026 ©2026 Association for Computational Linguistics

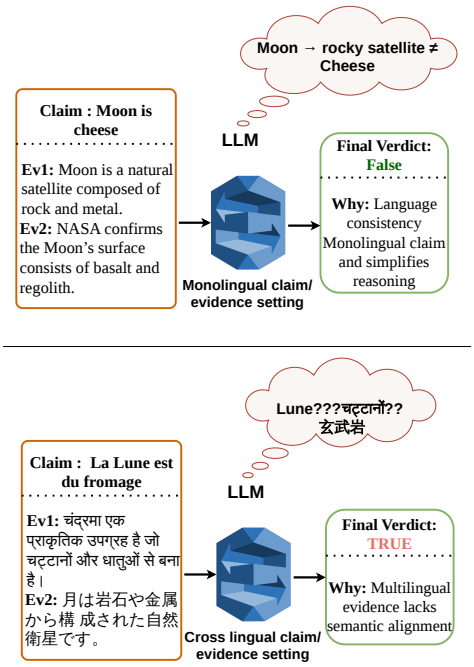


Figure 1: Illustration of cross-lingual claimevidence verification challenge. When claims and evidences appear in different languages, existing models often fail to align semantically equivalent facts, leading to incorrect veracity predictions.

grades the performance of existing models. Second, we propose a framework that mitigates language-specific biases in LLM-derived representations by learning language-agnostic representations of factual information. The proposed model, GraDuLa, constructs a multilingual evidence graph and employs a dual-branch graph neural network to disentangle semantic content from language-specific signals. Third, we introduce GraDuLaEval, a cross-lingual evidence-based benchmark constructed from existing claim verification datasets, enabling systematic evaluation of models in controlled synthetic cross-lingual stress settings.

The key contributions of this work are summarized as follows: 1) We propose GraDuLa, a graph-based framework that learns language-agnostic representations of factual information from multilingual evidence, enabling cross-lingual claim verification without relying on machine translation during inference or training and only during benchmark construction. 2) We propose a novel dual-branch graph modeling approach that leverages contrastive and orthogonality losses to disentangle semantic representations from language-specific representations of underlying concepts. 3) We construct two large-scale cross-lingual, evidence-based claim veracity prediction benchmarks designed as controlled stress-test environments for evaluating language-robust reasoning under uniform cross-lingual evidences availability, on which GraDuLa consistently outperforms all suitable baseline methods. 4) We perform extensive ablation studies, error analyses, and insightful qualitative experiments to analyze decision-making behavior, validate the datasets, and highlight key empirical observations.

2 Related Work

Misinformation Detection. Early fact-checking benchmarks such as FEVER (Thorne et al., 2018) and LIAR (Wang, 2017) focus primarily on monolingual English settings (Mishra and Setty, 2019). Later datasets including MultiFC (Augenstein et al., 2019) and XFact (Gupta and Srikumar, 2021) extend fact-checking to multiple languages but typically assume that claims and evidence are written in the same language. In practice, misinformation often requires cross-lingual verification, where claims and evidence appear in different languages (Goworek et al., 2025). Prior work shows that cross-lingual evidence retrieved from web can improve fake news detection, indicating relevant evidence frequently exists across languages (Dementieva and Panchenko, 2021; Huang et al., 2022).

Graph-based Fact Verification. Graph neural networks are widely used for misinformation detection because they model relationships between claims, evidence, and entities. Approaches such as Kernel Graph Attention Network (KGAT) (Liu et al., 2020) and evidence-aware graph models (Xu et al., 2022) construct evidence graphs to capture claimdocument interactions. Other methods leverage knowledge graphs or heterogeneous evidence graphs to support multi-hop reasoning over factual relations (Kim et al., 2023). However, most graph-based fact-checking models assume monolingual settings and do not explicitly address cross-lingual reasoning.

Cross-lingual Misinformation Detection. Prior work explores cross-lingual misinformation detection using multilingual models and zero-shot transfer, demonstrating that models trained on high-resource languages can generalize to others (Conneau et al., 2020; Dementieva and Panchenko, 2021). Shared tasks such as CLEF CheckThat! and studies on COVID-19 misinformation highlight the challenges of verifying

claims across languages (Nakov et al., 2021; Shahi et al., 2021). However, most approaches rely on translated/monolingual evidence, whereas our work focuses on settings where evidences occur in different languages.

Language-Agnostic Representation Learning. Multilingual representation learning aims to build shared cross-lingual embedding spaces through unsupervised or weakly supervised objectives (Conneau et al., 2020). However, sentence-level encoders such as Sentence-BERT often retain language-specific biases, causing embeddings to cluster by language rather than semantic meaning (Reimers and Gurevych, 2019; Yang et al., 2024). Such biases also appear in large language models and multilingual retrieval systems, degrading performance in cross-lingual queryevidence scenarios (Vykopal et al., 2025; Goworek et al., 2025).

3 Methodology

Given a claim c expressed in language ℓ_c and a set of evidence documents $E = \{e_1, \ldots, e_m\}$ written in different languages $\{\ell_1, \ldots, \ell_m\}$, the task of *cross-lingual fact verification* is to predict veracity label $y \in \{\text{Real}, \text{Fake}\}$ for the claim. The model must perform reasoning across languages where the claim and supporting evidences may not share the same linguistic representation.

3.1 Multilingual Dataset Construction

In real-world fact-checking, claims and supporting evidence often appear in different languages. To simulate this cross-lingual challenge, we create **GraDuLaEval**, a dataset where claims and evidence are intentionally distributed across languages. The benchmark assumes an ideal setting where relevant evidence in all languages is equally available online and the retriever has no language bias, giving evidence from different languages an equal chance of being retrieved and forming uniform cross-lingual claimevidence pairs.

3.1.1 Cross Lingual translation

Translation Setup: We translate all claims and their supporting evidences using NLLB-200 model (Costa-Jussà et al., 2022). We select six languages spanning diverse language families and writing families: English, Hindi, Japanese, Korean, French, and German. Each claim is translated into a randomly selected target language. The corresponding evidence is translated into a different randomly selected language with uniform probability, ensuring systematic cross-lingual mismatch.

Translation Quality Control. To reduce noise from automatic translation, particularly for long evidence documents, we apply a back-translation consistency check. Each sentence is translated into the target language and then back into English using the same model, and we use ChrF(Popović, 2015)-based back-translation consistency as a heuristic to identify severely degraded translations rather than as a full measure of translation correctness. Because evidence sentences in PolitiFact(Popat et al., 2018) and Snopes(Vo and Lee, 2020) are often long, even correct translations can receive relatively low ChrF scores. We therefore avoid using a fixed global threshold. Instead, we apply a language-wise, document-level percentile filtering strategy, reverting only the bottom 25% of sentences with the lowest ChrF scores within each language back to english. While back-translation uses the same model for scalability, we complement it with human validation to mitigate bias.

3.1.2 Structured Knowledge Extraction

To reduce context length and isolate factual signal, we extract subjectrelationobject triplets, which provide a compact representation of factual information while removing irrelevant content. We construct a structured representation from translated evidence by extracting (1) factual subjectrelationobject triplets and (2) concise sentences that directly support or contradict the claim. Both are extracted using the multilingual LLaMA-3.2 8B-Instruct model (Dubey et al., 2024).

Triplet Extraction. Evidence documents are often long and entity-dense. We apply language-specific few-shot prompts to extract meaningful subjectrelationobject triplets that represent structured facts. Post-processing removes malformed triplets, duplicates, and overly generic relations.

Sentence Extraction. While triplets capture core factual relations, they may omit context and supporting details. To preserve contextual evidence, we extract 35 concise factual sentences per document using a few-shot prompt. During post-processing, sentences below a minimum token threshold are removed and near-duplicates are filtered using embedding similarity.

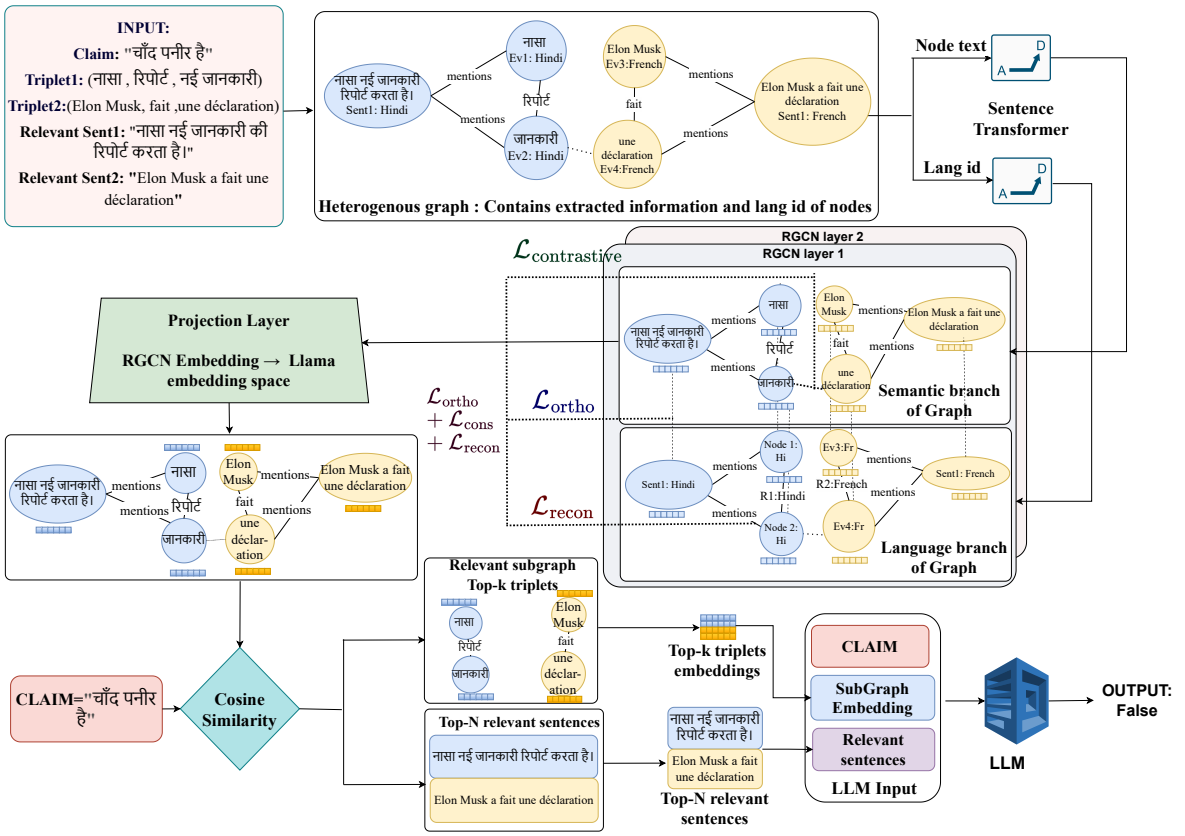


Figure 2: Overview of GraDuLa. Multilingual evidence is converted into a heterogeneous graph, encoded using a dual-branch R-GCN to learn language-agnostic representations, and injected into an LLM for cross-lingual claim verification.

Triplets capture structured facts, while sentences capture contextual evidence. Combining structured triplets with natural-language evidence enables the model to reason over both relational and contextual information.

3.2 Heterogeneous Graph Construction

For each claim, we construct a compact heterogeneous knowledge graph that integrates extracted triplets and supporting sentences. Formally, the graph is defined as $G = (V, E, R)$, where V denotes nodes, E denotes edges, and R denotes relation types. The node set consists of two types: $V = V_{\text{entity}} \cup V_{\text{sentence}}$. V_{entity} contains entity nodes derived from subjects and objects of extracted triplets, while V_{sentence} contains nodes corresponding to factual evidence sentences. We define two edge types. *Relation edges* connect entity pairs according to triplets $(subject, relation, object)$. *Mention edges* link an entity node to a sentence node when the entity appears in that sentence, forming edges of the form $(entity, sentence)$. Each node $v \in V$ is initialized with a textual embedding $h_v^{(0)} = \text{Encoder}(\text{text}_v)$, where encoder is implemented using a SentenceTransformer model.

3.3 Dual-Branch Graph Encoder

Sentence-transformer embeddings retain language biases and cannot fully capture relational structures like triplets, causing GNN message passing to cluster nodes by language rather than meaning. To address this, we introduce a dual-branch relational graph encoder that separates semantic content from language identity. Built on a Relational Graph Convolutional Network (R-GCN), it leverages graph relations to mitigate language bias while enriching semantic representations. The semantic and language branches share the same graph structure but use different input features.

The encoder has two parallel branches: (1) a *semantic branch* capturing factual meaning and relational information, and (2) a *language branch* encoding each node's language. Each node v has two feature vectors. The semantic feature is obtained by encoding node text with a multilingual SentenceTransformer. The language fea-

ture is obtained from fixed reference sentences of the form `sent_id/entity_id: "This is an English sentence."`, translated into target language as needed, then encoded using same transformer. These embeddings capture language identity rather than factual meaning, allowing semantic branch to focus on facts and language branch to model language identity explicitly.

The dual-branch R-GCN is trained with three self-supervised losses.

3.3.1 Graph Reconstruction Loss (Edge Prediction)

Graph Reconstruction Loss (BCE) To ensure that semantic embeddings preserve the relational structure of the graph, we train the model to distinguish true graph edges from randomly sampled non-edges. For each real edge (u, v) in the graph, we compute a positive score using the dot product of semantic node embeddings, and negative scores by pairing u with randomly selected unconnected nodes n:

$$s_{uv}^{+} = \langle z_u^{\text{sem}}, z_v^{\text{sem}} \rangle, \quad s_{un}^{-} = \langle z_u^{\text{sem}}, z_n^{\text{sem}} \rangle \tag{1}$$

Here, z_u^{sem} and z_v^{sem} denote the semantic embeddings of nodes u and v, respectively. The dot product $\langle \cdot, \cdot \rangle$ measures the similarity between two node embeddings. Binary cross-entropy is applied to all scores:

$$\mathcal{L}_{\text{recon}} = \text{BCE}(s^{+}, 1) + \text{BCE}(s^{-}, 0) \tag{2}$$

3.3.2 Cross-Lingual Contrastive Loss (InfoNCE)

The graph reconstruction loss helps the model learn which nodes are connected, but it does not ensure that similar fact written in different languages is represented similarly. As a result, node embeddings may still group by language instead of by meaning. To address this, we apply a cross-lingual contrastive loss where for each node i, we identify its positive example as the node j from a different language with the highest semantic similarity within the claim-specific evidence graph. This selection is based on semantic similarity rather than direct translation alignment. Negative examples are sampled from semantically unrelated nodes belonging to different evidence documents and languages. The contrastive loss is defined as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_i^{\text{sem}}, z_j^{\text{sem}}))}{\sum_{n \in \text{neg}} \exp(\text{sim}(z_i^{\text{sem}}, z_n^{\text{sem}}))} \tag{3}$$

Here, $\text{sim}(\cdot)$ denotes cosine similarity. This objective encourages nodes with the similar meaning to be close to each other, even when written in different languages, and prevent the embeddings from forming language-specific clusters.

3.3.3 Semantic-Language Orthogonality Loss

The language branch explicitly encodes the language of each node. However, semantic embeddings inherited from multilingual encoders often contain residual language-specific information, leading to overlap between semantic and language representations. To mitigate this bias, we encourage the two representations to be orthogonal by:

$$\mathcal{L}_{\text{ortho}} = \left\| \hat{z}^{\text{sem}} \cdot \hat{z}^{\text{lang}} \right\|_2 \tag{4}$$

Here, $\hat{z}^{\text{sem}}$ and $\hat{z}^{\text{lang}}$ are the normalized semantic and language embeddings, respectively. This loss forces the semantic branch to focus on factual content while relegating language identity to the language branch, leading to cleaner disentanglement.

3.3.4 Final Objective

The full training objective is a weighted combination of the three losses:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{recon}} + (1 - \alpha) \mathcal{L}_{\text{contrast}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} \tag{5}$$

After training is finished, only the semantic branch is retained after training which represents language agnostic node embedding. We analyze GraDuLa's sensitivity to the loss weights α and λ in Table 13 and use best-performing setting for all experiments.

3.4 Aligning graph and LLM representation

The representations learned by the graph encoder $\mathbf{z}^{\text{graph}}$ and LLM token embeddings $\mathbf{z}^{\text{LLM}}$ exist in different embedding spaces and cannot be used together directly. To make them compatible, we train a small projection network f_θ which is implemented as a two-layer MLP with GELU activation that maps graph embeddings into the LLM embedding space.

For each node i:

$$\mathbf{z}_i^{\text{proj}} = f_\theta(\mathbf{z}_i^{\text{graph}}), \quad \mathbf{z}_i^{\text{LLM}} = \text{LLM}(x_i) \tag{6}$$

The alignment is learned using a cosine embedding loss:

$$\mathcal{L}_{\text{proj}} = 1 - \mathbf{z}_i^{\text{proj}} \cdot \mathbf{z}_i^{\text{LLM}} \tag{7}$$

After training, this produces LLM-compatible, language-agnostic graph embeddings that can be directly injected into the model.

3.5 Graph-Augmented LLM Training

We enhance the LLM by injecting structured graph embeddings directly into its input space, rather than converting graph information back into text. Unlike textual serialization, embedding-level injection preserves relational structure and remains language-agnostic, enabling more robust reasoning.

Claim-Specific Subgraph Selection For each claim, we compute R-GCN embeddings for all graph nodes, project them into LLM space, and select the most relevant top-k triplet nodes and top-n sentence nodes based on similarity with claim embedding.

Direct Graph Embedding Injection Instead of converting graph triplets into text, we directly inject the projected graph embeddings into the LLMs input layer as prefix embeddings before the claim tokens. This allows the model to use structured factual information without rewriting it in different languages or surface forms. We treat sentences and triplets differently because they serve different roles. Sentences provide contextual explanations and are naturally suited to text-based processing, while triplets encode compact factual relations. Converting triplets back into text can distort their structure and introduce language-specific noise, whereas embedding-level injection preserves their meaning in a concise, language-agnostic form. By combining text-based sentences with embedding-level triplets, the LLM can reason over both context and structured facts, leading to more robust cross-lingual verification.

Prompt and Final Prediction The final input consists of projected triplet embeddings, relevant evidence sentences, and the claim. Combining structured and textual evidence in a shared embedding space enables effective cross-lingual fact verification.

4 Experimental Setup

4.1 Datasets

We evaluate our approach on **GraDuLaEval**, a cross-lingual, evidence-based misinformation detection benchmark introduced in this work and constructed from PolitiFact (Popat et al., 2018) and Snopes (Vo and Lee, 2020). Each dataset contains claims paired with supporting evidence documents and annotated with veracity labels. PolitiFact originally provides six graded veracity labels (*true*, *mostly true*, *half true*, *mostly false*, *false*, and *pants-on-fire*). Following common practice in prior misinformation detection studies (Popat et al., 2018; Wang, 2017; Vo and Lee, 2021; Singh and Mishra, 2026), we convert these labels into binary classes by merging the first three into *real* and the latter three into *fake*. This binarization simplifies the prediction task and ensures consistency with widely used evaluation settings in the literature. Snopes consists of fact-checked claims collected from the official Snopes fact-checking portal and already provides binary veracity labels. To create a controlled cross-lingual verification benchmark, we translate claims and evidence into six languages and extract structured knowledge from the translated evidence using LLaMA-3.2 8B.

4.2 Baselines

We compare GraDuLa against a diverse set of baselines representing current approaches to multilingual and cross-lingual misinformation detection.

Misinformation detection models We include GET (Xu et al., 2022) and MUSER (Liao et al., 2023), two evidence-aware graph-based methods designed for claim verification using multiple supporting documents. GET and MUSER were originally developed for evidence-aware claim verification rather than multilingual or cross-lingual verification. We include them as architectural baselines to evaluate whether existing graph-based evidence reasoning approaches can generalize to the cross-lingual claimevidence setting.

RAG-based LLM baselines We evaluate retrieval-augmented generation (RAG) using LLaMA and Gemma, where retrieved evidence sentences are appended to the input as textual context. We consider both full-evidence and top-k evidence variants.

Translation-based baseline: We additionally evaluate a translate-to-English pipeline, where multilingual evidence is first translated back into English before being processed by the LLaMA-based RAG system.

Cross-lingual retrieval baselines We compare against LLaMA2Vec (Li et al., 2024), which adapts LLM representations for dense retrieval, and DKM-RAG (Park and Lee, 2025), a translation-based approach that translates cross-lingual evidence into the query language before retrieval and verification.

5 Results and discussion

5.1 Comparison with Diverse State-of-the-art Baselines

Table 1 compares GraDuLa with representative baselines across two datasets. GraDuLa consistently achieves the best overall performance in terms of accuracy and macro-F1 on both PolitiFact and Snopes. In particular, GraDuLa improves macro-recall compared to strong RAG-based baselines, which is especially important for class-imbalanced datasets such as Snopes. This suggests that the model is better able to recover true instances while maintaining balanced predictions across classes. Compared with text-only LLM baselines such as RAG with LLaMA and Gemma, GraDuLa achieves improvements of up to several F1 points across both datasets. These gains indicate that explicitly modeling structured evidence relationships through the graph encoder provides complementary reasoning signals beyond text-only retrieval. GraDuLa performs consistently well with both LLaMA and Gemma backbones. This suggests that observed gains primarily stem from the proposed graph-based language-agnostic representation learning rather than from the choice of underlying language model.

5.2 Ablation study

Role of contrastive and orthogonal loss: Removing either contrastive or orthogonality loss degrades performance, indicating that both cross-lingual semantic alignment and explicit separation of language and semantic information are essential for learning language-agnostic graph representations.

Embedding-level vs. text-level graph injection: Replacing embedding-level graph injection with textual subgraphs results in weaker performance, suggesting direct embedding injection better preserves relational structure than text serialization.

Impact of graph reasoning: Removing the R-GCN leads to a notable performance drop, highlighting the importance of explicit graph reasoning and language-agnostic evidence representation.

Contribution of sentence-level nodes: Excluding sentence nodes causes the largest performance drop, showing that sentence-level evidence is crucial for grounding claims beyond entity-level signals in cross-lingual settings.

Table 2: Ablation study on GraDuLaEval (PolitiFact)

GraDuLaEval (Politifact)				
Variant	**Acc**	**Macro$_P$**	**Macro$_R$**	**Macro$_{F1}$**
GraDuLa	0.7143	0.7208	0.7050	0.7051
GraDuLa w/o Contrastive Loss	0.6828	0.6659	0.7010	0.6607
GraDuLa w/o Orthogonal Loss	0.7003	0.7060	0.7052	0.7002
GraDuLa w/o Embed Injection	0.6751	0.6793	0.6792	0.6751
GraDuLa with Textual Subgraph	0.7003	0.7061	0.6907	0.6903
GraDuLa w/o R-GCN	0.6807	0.6866	0.6857	0.6806
GraDuLa w/o Sentence node	0.6246	0.6357	0.6080	0.5960

5.3 Cross lingual retriever performance

We evaluate the quality of learned graph representations via cross-lingual node retrieval. For each claim, we retrieve the top-K nodes based on cosine similarity between the claim embedding and projected R-GCN node embeddings. Retrieval is considered successful when retrieved nodes belong to languages different from the claim language, indicating language-agnostic semantic alignment. Table 3 reports the impact of different training objectives on retrieval performance.

Table 3: Retrieval performance comparison across different model configurations.

Model	**R@5**	**R@10**
No Loss	0.7284	0.8083
Contrastive Learning Only	0.7571	0.8109
Contrastive + Orthogonality	0.7766	0.8297

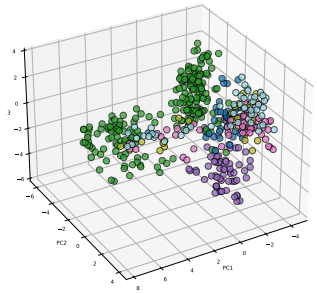
Adding contrastive learning substantially improves Recall@5 and Recall@10, demonstrating better cross-lingual semantic alignment. Incorporating the orthogonality loss further boosts performance, confirming that explicitly disentangling language and semantic information yields more discriminative embeddings.

5.4 Visualization of Cross-Lingual Graph representation

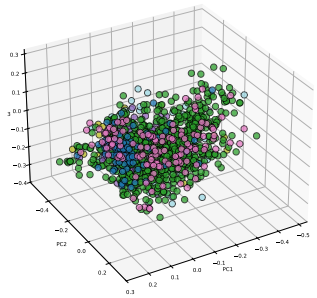
Figure 3(a) shows node embeddings produced using only a SentenceTransformer, without any graph modeling. In this case, nodes from the same language form clear clusters. This means the embeddings are grouped mainly by language, even when they describe the same fact. Figure 3(b) shows embeddings learned by GraDuLa using the dual-branch R-GCN. Here, nodes from different languages are mixed together instead of forming separate language clusters. This indicates that GraDuLa focuses on meaning rather than language, producing more language-agnostic representations.

Table 1: Performance comparison between GraDuLa and representative baselines. Improvements over DKM-RAG are statistically significant according to a two-tailed Student's t-test ($p < 0.05$).

Model	GraDuLaEval(Politifact)				GraDuLaEval(Snopes)			
	Accuracy	**Macro$_P$**	**Macro$_R$**	**Macro$_{F1}$**	**Accuracy**	**Macro$_P$**	**Macro$_R$**	**Macro$_{F1}$**
GET	0.4958	0.6137	0.5287	0.4026	0.7172	0.6343	0.5844	0.5874
MUSER model	0.5126	0.5865	0.5405	0.5126	0.7379	0.6731	0.6443	0.6532
LLaMA RAG(all evidence)	0.6611	0.6960	0.6419	0.6269	0.7977	0.8398	0.6599	0.6811
Gemma RAG(all evidence)	0.6078	0.6123	0.6120	0.6078	0.8092	0.8501	0.6799	0.7057
LLaMA RAG (top-10)	0.6695	0.6924	0.6804	0.6666	0.7816	0.7382	0.6940	0.7084
Llama2vec	0.6695	0.6869	0.6544	0.6476	0.7724	0.7215	0.7138	0.7173
Translate-to-English(LLama)	0.6919	0.7072	0.6786	0.6752	0.7227	0.7211	0.7205	0.7207
DKM-RAG	0.6471	0.6950	0.6250	0.6013	0.7816	0.7412	0.6868	0.7028
GraDuLa (LLama)	0.7143	0.7208	0.7050	0.7051	0.8460	0.8302	0.7774	0.7968
GraDuLa (Gemma)	0.7087	0.7445	0.6921	0.6855	0.8276	0.8019	0.7573	0.7739



(a) SentenceTransformer only (no graph)



(b) GraDuLa (full model)

Figure 3: PCA visualization of node embeddings across languages.

5.5 Human and LLM Evaluation of ChrF-Based Translation Filtering

To check whether our ChrF-based filtering reflects translation quality, we compare ChrF scores with judgments from both humans and an independent LLM. We randomly select 200 translated evidence sentences and divide them into three groups based on their ChrF scores: 025, 2550, and 50100. Table 4 shows the average quality scores assigned by humans and the LLM. Sentences with low ChrF scores (025) receive poor ratings, indicating their meanings are often incorrect. In contrast, sentences with medium and high ChrF scores (2550 and 50100) are judged to preserve the original meaning well. The close agreement between human and LLM evaluations suggests ChrF is reliable signal for identifying low-quality translations and supports its use for filtering translated evidence.

Table 4: Alignment between ChrF score ranges and semantic adequacy as judged by humans and LLM.

ChrF Range	Human Score	LLM Score
0-25	0.52	0.74
25-50	1.31	1.36
50-100	1.63	1.78

5.6 Effect of Translation and Graph-based Reasoning

Large language models perform notably worse on long, multilingual evidence than on monolingual English inputs, where translation noise and language switching hinder reasoning. To analyze this effect, we compare three settings: (i) English-only evidence, (ii) translated cross-lingual evidence directly provided to the LLM, and (iii) GraDuLa, which injects language-disentangled graph representations into the LLM.

Table 5: Comparison of performance between the original English-only data, translated cross-lingual evidence and GraDuLa.

PolitiFact				
Setting	**Acc.**	**Prec.**	**Rec.**	**F1**
LLama 8B(English-only)	0.7283	0.7343	0.7197	0.7204
LLama 8B(Cross-lingual)	0.6611	0.6960	0.6419	0.6269
GraDuLa(Cross-lingual)	0.7143	0.7208	0.7050	0.7051
Snopes				
Setting	**Acc.**	**Prec.**	**Rec.**	**F1**
LLama 8B(English-only)	0.9034	0.8834	0.8797	0.8816
LLama 8B(Cross-lingual)	0.7977	0.8398	0.6599	0.6811
GraDuLa(Cross-lingual)	0.8460	0.8302	0.7774	0.7968

As shown in Table 5, direct use of translated cross-lingual evidence leads to a substantial performance drop, particularly in recall. In contrast, GraDuLa recovers much of this degradation, improving F1 by 7.8 and 11.6 points on PolitiFact and Snopes, respectively, demonstrating the benefit of structured, language-disentangled representations for cross-lingual reasoning.

5.7 Zero-shot Out-of-Language Evaluation

To evaluate the cross-lingual generalization of GraDuLa beyond the languages observed during training, we conduct a zero-shot out-of-language evaluation. Both GraDuLa and the LLaMA baseline are trained under identical settings using only the training languages, with the same data splits, supervision signals, and fine-tuning configuration. For the LLaMA baseline, all available evidence documents are directly provided to the model during both training and inference. We evaluate the models on three unseen languages: Spanish, Gujarati, and Chinese without additional fine-tuning or translation-time adaptation.

Model	Language	Acc	Prec.	Recall	Macro-F1
LLaMA	Spanish	0.6190	0.6444	0.6339	0.6156
	Gujarati	0.6744	0.7190	0.7044	0.6728
	Chinese	0.5870	0.5870	0.5885	0.5852
	Avg	0.6534	0.6738	0.6655	0.6516
GraDuLa	Spanish	0.6667	0.6635	0.6590	0.6597
	Gujarati	0.7209	0.7133	0.7133	0.7133
	Chinese	0.7174	0.7331	0.6923	0.6929
	Avg	**0.7092**	**0.7115**	**0.6990**	**0.6999**

Table 6: Zero-shot out-of-language performance on unseen languages. Models are trained without exposure to Spanish, Gujarati, or Chinese.

As shown in Table 6, GraDuLa consistently outperforms the LLaMA baseline across all three languages. The largest improvement is observed for Chinese, where GraDuLa achieves a Macro-F1 score of 0.6929 compared with 0.5852 for LLaMA. GraDuLa also improves performance on Spanish and Gujarati, achieving Macro-F1 scores of 0.6597 and 0.7133, respectively. On average, GraDuLa achieves an accuracy of 0.7092 and a Macro-F1 score of 0.6999, compared with 0.6534 accuracy and 0.6516 Macro-F1 for the LLaMA baseline. This setting indicates that the learned representations can transfer factual reasoning capabilities to previously unseen linguistic contexts.

5.8 Language-wise Evaluation Metrics

Table 7 presents the performance of GraDuLa broken down by the language of the claim. Overall, the results show relatively consistent performance across the six evaluated languages, with F1 scores ranging from 0.6619 to 0.7939. German achieves the strongest performance, while French and Korean show comparatively lower F1 scores.

Table 7: Table 7 reports performance broken down by the language of the claim.

Language	Acc	Precision	Recall	F1
German (de)	0.8000	0.7976	0.7917	0.7939
English (en)	0.7290	0.7348	0.7244	0.7243
French (fr)	0.6792	0.6762	0.6609	0.6619
Hindi (hi)	0.6852	0.6865	0.6827	0.6825
Japanese (ja)	0.7347	0.7680	0.7024	0.7030
Korean (ko)	0.6780	0.6955	0.6753	0.6684
Overall	**0.7143**	**0.7208**	**0.7050**	**0.7051**

Importantly, the model maintains competitive performance across languages with different scripts and linguistic characteristics. In particular, the results for Hindi, Japanese, and Korean demonstrate that the framework can effectively process non-Latin scripts, with performance remaining broadly comparable to that observed for English and the European languages. This consistency suggests that the learned multilingual representations are not strongly dependent on a single language or script. Overall, these results support the effectiveness of GraDuLa for cross-lingual fact verification across diverse linguistic settings.

6 Conclusions

Multilingual misinformation detection using cross-lingual claimevidence pairs poses significant challenges for existing baseline methods. We introduced GraDuLa, a novel dual-branch graph modeling framework that employs contrastive and orthogonality losses to effectively disentangle language-invariant semantic representations from language-specific features. In addition, we constructed two large-scale cross-lingual, evidence-based claim veracity prediction datasets to support comprehensive evaluation. Experimental results demonstrate that GraDuLa consistently outperforms all suitable baseline approaches across both datasets, highlighting its effectiveness and robustness for multilingual misinformation detection.

7 Limitations

While GraDuLa demonstrates strong performance for cross-lingual misinformation detection, it has few limitations. Evaluation is restricted to a limited set of languages selected to cover diverse scripts and language families; extending the framework to a broader range of low-resource and typologically distant languages remains future work. Experiments are conducted on general-domain datasets (PolitiFact and Snopes). More specialized domains such as scientific, environmental, or medical misinformation introduce substantial cross-lingual jargon variation, which may require domain-adapted extraction and alignment strategies. Due to computational resource constraints, we evaluate GraDuLa using mid-sized LLM backbones and do not explore larger models (e.g., LLaMA-70B), which may further improve performance. Our benchmark relies on automatic translation, which may not fully capture language-specific nuances or cultural grounding.

References

Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan. 2025. The cross-lingual cost: Retrieval biases in rag over arabic-english corpora. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 69–83.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Roksana Goworek, Olivia Macmillan-Scott, and Eda B Özyiğit. 2025. Bridging language gaps: Advances in cross-lingual information retrieval with multilingual llms. *arXiv preprint arXiv:2510.00908*.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. Concrete: Improving cross-lingual fact-checking with cross-lingual retrieval. *arXiv preprint arXiv:2209.02071*.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206.

Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4461–4472.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7342–7351.

Rahul Mishra and Vinay Setty. 2019. Sadhan: Hierarchical attention networks to learn latent aspect embeddings for fake news detection. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 197204, New York, NY, USA. Association for Computing Machinery.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeno, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, and 1 others. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *European Conference on Information Retrieval*, pages 639–649. Springer.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual rag systems. *arXiv preprint arXiv:2502.11175*.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22:100104.

Atul Kumar Singh and Rahul Mishra. 2026. Iconmoe: Modeling intents of misinformation using concept activation vector-based mixture of experts. In *Proceedings of the ACM Web Conference 2026*, pages 1684–1695.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. *arXiv preprint arXiv:2010.03159*.

Nguyen Vo and Kyumin Lee. 2021. Hierarchical multi-head attentive network for evidence-aware fake news detection. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 965–975.

Ivan Vykopal, Antonia Karamolegkou, Jaroslav Kopčan, Qiwei Peng, Tomáš Javrek, Michal Gregor, and Marián Šimko. 2025. Investigating language and retrieval bias in multilingual previously fact-checked claim detection. *arXiv preprint arXiv:2509.25138*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 2501–2510.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2024. Language bias in multilingual information retrieval: The nature of the beast and mitigation methods. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 280–292.

A Appendix

A.1 Dataset Creation and Stats

In this section, we describe the dataset diagram creation flow and dataset statistics for Politifact and Snopes along with per-claim statistics.

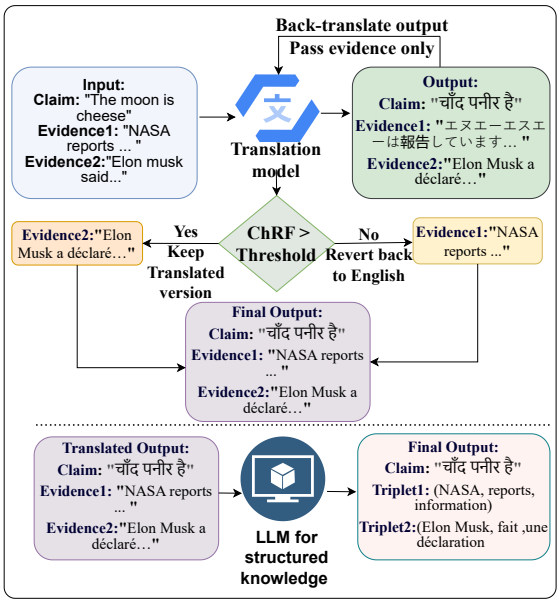


Figure 4: Overview of the translation quality filtering and structured knowledge extraction process. Evidence sentences are back-translated and evaluated using a ChrF-based check; low-quality translations are reverted, while high-quality ones are retained. The resulting multilingual inputs are then used to extract structured factual representations for downstream verification.

Table 8: Dataset statistics for Politifact and Snopes

Dataset	Train		Validation		Test	
	Real	Fake	Real	Fake	Real	Fake
Politifact	1074	939	136	114	134	118
Snopes	2343	888	269	131	293	107

Table 9: Per-claim statistics for Politifact and Snopes

Dataset	Docs / Claim	Triplets / Claim	Sents / Claim
Politifact	8.34	56.23	29.67
Snopes	6.73	49.99	24.90

A.2 Training Configuration and Evaluation

We first pretrain a dual-branch relational graph convolutional network (RGCN) to obtain language-agnostic graph representations. All nodes are initialized using paraphrase-multilingual-mpnet-base-v2 (768-d)(Reimers and Gurevych, 2019). RGCN has 2 layers, hidden size 256 and basis decomposition for up to 256 relation types. We train the RGCN for 3 epochs using the Adam optimizer with a learning rate of 4×10^{-4}, batch-free graph updates, gradient clipping at 1.0, and a temperature of 0.07 for contrastive learning. The projection network is a two-layer multilayer perceptron (MLP) with dimensions $768 \rightarrow 1024 \rightarrow d_{\text{LLM}}$, where d_{LLM} denotes the hidden size of the target large language model. The network uses GELU activation, dropout, and layer normalization. Training is performed using cosine embedding loss with positive pairs only, optimized with Adam (learning rate 5×10^{-3}, weight decay 10^{-4}), a batch size of 64, and for 20 epochs. For each claim, we extract a relevance-filtered subgraph (up to 20 nodes) using cosine similarity between claim and projected rgcn embedding in LLM space. Projected graph embeddings are prepended as a soft prefix to the token embeddings, with appropriate attention masking. We fine-tuned LLMs using QLoRA with 4-bit NF4 quantization.

A.3 Translation data analysis

Here, we summarize translation quality statistics across datasets and languages. Table 10 reports the distribution of ChrF scores before filtering, showing consistent average scores across languages and splits.

A.4 Stability and Statistical Significance Analysis

To assess the robustness of our results, we evaluate GraDuLa across five independent random seeds and report the mean and standard deviation. On PolitiFact, GraDuLa achieves an accuracy of 71.43 ± 0.26, while on Snopes it achieves 84.60 ± 0.51. The corresponding coefficients of variation are 0.36% and 0.60%, respectively, indicating low run-to-run variability and consistent performance across random initializations. The approximate 95% confidence intervals are 71.43 ± 0.32 for PolitiFact and 84.60 ± 0.63 for Snopes. Furthermore, a two-tailed Student's t-test shows that the improvements over the strongest baseline, DKM-RAG, are statistically significant ($p < 0.01$).

A.5 Cross-lingual degradation on Gemma

We also evaluated cross-lingual degradation on Gemma for Politifact Dataset and observed consistent trends across both LLaMA and Gemma backbones. Below is the Gemma table:

Table 11: Comparison of performance between the original English-only data, translated cross-lingual evidence and GraDuLa.

PolitiFact				
Setting	**Acc.**	**Prec.**	**Rec.**	**F1**
Gemma 8B(English-only)	0.6975	0.6965	0.6974	0.6966
Gemma 8B(Cross-lingual)	0.6078	0.6123	0.6120	0.6078
GraDuLa(Cross-lingual)	0.7087	0.7445	0.6921	0.6855

A.6 Human Evaluation of Triplet Relevance

We conduct a post-hoc human evaluation to assess the relevance of triplets selected from claim-specific subgraphs, independent of model predictions. We randomly sample 50 claims from GraDuLaEval (PolitiFact) and present annotators with the top-k triplets ($k = 10$) selected for each claim. As a baseline, we use a sentence-transformer-based method that ranks triplets solely by embedding similarity to the claim.

Three graduate-level annotators with prior experience in NLP and misinformation analysis rate each triplet based on its usefulness for verifying the claim, considering whether it supports, contradicts, or provides essential factual context. Relevance is scored on a 1–10 numerical scale, with scores averaged across triplets and claims. Annotators are provided no ground-truth triplet labels and make judgments solely from the claim and triplet content.

Table 12: Human evaluation of triplet relevance on GraDuLaEval (PolitiFact). Scores indicate the average relevance over the top-10 selected triplets per claim.

Method	**Avg. Relevance Score**
SentenceTransformer Selection	4.2
GraDuLa Subgraph Selection	**7.3**

GraDuLa achieves a substantially higher relevance score (7.3 vs. 4.2), indicating that graph-based, language-agnostic subgraph selection more effectively prioritizes factually meaningful relations for claim verification.

Table 10: Evidence Translation Quality Statistics for PolitiFact and Snopes. For each dataset split and language, we report the total number of evidence sentences, the number rejected by ChrF-based filtering, and the minimum, average, and maximum ChrF scores computed via back-translation.

		PolitiFact					Snopes				
Split	**Lang**	**Total**	**Rej.**	**Min**	**Avg**	**Max**	**Total**	**Rej.**	**Min**	**Avg**	**Max**
Train	Hindi	4480	820	0.22	51.79	92.4	4544	770	1.00	52.98	97.1
	Japanese	4415	839	0.26	27.84	72.8	4398	780	0.36	30.27	89.7
	Korean	4444	739	0.00	28.86	76.8	4519	785	1.47	29.80	86.0
	French	4400	290	2.73	59.46	100.0	4481	261	2.28	60.53	100.0
	German	4437	504	1.05	54.76	100.0	4444	349	1.10	58.12	95.1
Val	Hindi	561	92	0.21	52.39	96.9	592	99	5.11	52.85	79.7
	Japanese	547	97	2.07	28.16	71.6	601	115	3.54	29.36	77.9
	Korean	576	98	1.09	29.10	76.4	571	90	2.44	30.49	71.2
	French	570	33	5.49	59.87	92.3	590	28	3.17	59.96	96.0
	German	571	58	3.57	55.68	94.9	551	35	1.89	58.72	92.8
Test	Hindi	505	90	2.19	51.91	90.0	518	78	1.78	52.91	83.5
	Japanese	514	102	0.36	27.74	65.9	553	114	4.11	29.82	72.3
	Korean	520	93	0.90	28.73	68.0	554	89	1.50	29.32	81.8
	French	524	30	0.87	60.03	92.0	563	32	7.31	61.85	94.5
	German	506	57	2.17	54.86	92.3	533	42	1.40	57.96	90.0

A.7 Hyperparameter tuning

Table 13: Ablation study on PolitiFact showing the effect of varying the loss weights α and λ. α controls the balance between graph reconstruction and contrastive alignment, while λ controls the strength of the orthogonality loss.

Setting	**Acc**	**Prec.**	**Rec.**	**Macro-F1**
Varying α (with fixed λ)				
$\alpha = 0.0$	0.6611	0.6745	0.6465	0.6402
$\alpha = 0.2$	0.6639	0.7138	0.6428	0.6234
$\alpha = 0.4$	0.6975	0.7231	0.6821	0.6765
$\alpha = 0.6$	0.7143	0.7208	0.7050	0.7051
$\alpha = 0.8$	0.6751	0.6937	0.6600	0.6535
$\alpha = 1.0$	0.6828	0.6659	0.7010	0.6607
Varying λ (with fixed α)				
$\lambda = 0.0$	0.7003	0.7060	0.7052	0.7002
$\lambda = 0.2$	0.7031	0.7060	0.7065	0.7031
$\lambda = 0.4$	0.7003	0.7043	0.7043	0.7003
$\lambda = 0.6$	0.6919	0.6971	0.6965	0.6919
$\lambda = 0.8$	0.7143	0.7208	0.7050	0.7051
$\lambda = 1.0$	0.7003	0.7061	0.6907	0.6903

Table 13 analyzes the sensitivity of GraDuLa to the loss weights α and λ on PolitiFact. When varying α, performance improves as contrastive alignment is gradually introduced, peaking at $\alpha = 0.6$, which balances graph reconstruction and cross-lingual semantic alignment. Very small or large α values degrade performance, indicating that relying exclusively on either reconstruction or contrastive learning is suboptimal. Similarly, varying λ shows that moderate orthogonality regularization is beneficial, with best results obtained at $\lambda = 0.8$. Lower values insufficiently disentangle language and semantic representations, while overly strong regularization degrades performance. Based on these results, we use $\alpha = 0.6$ and $\lambda = 0.8$ for all experiments.

A.8 Prompt Used for Knowledge Graph Triplet Extraction

To extract high-quality factual triplets from text, we employ the following prompt. The prompt instructs the model to extract structured factual relationships in the form of subject–relation–object triplets while enforcing strict constraints on entity specificity and factual validity.

> You are an expert knowledge graph extraction system. Your task is to extract **high-quality factual triplets**. **Strict Rules:**
>
> 1. Extract only factual relationships from the text.
> 2. Each triplet must follow the format: (Subject, Relation, Object).
> 3. Subject and Object must be specific named entities (e.g., people, organi-

zations, places, dates, numbers, or events).

4. Relations must represent meaningful actions or relationships (e.g., announced, located in, occurred on, manufactured by).
5. Preserve the exact language used in the input text.
6. Extract a minimum of 5–10 triplets (more if the text is information-rich).
7. Do not use generic subjects or objects (e.g., "this", "that", "it", "something").
8. Every triplet must contain concrete, verifiable information.

Few-shot examples: `{few_shot}` **Extraction target text:**

`{text}`

Output format: One triplet per line, listed under the heading `TRIPLETS`.

A.9 Error Analysis

GraDuLaEval involves multiple stages, and errors introduced during multilingual processing, knowledge extraction, graph construction, and evidence selection can propagate to the final verification decision.

The analysis shows that translation may distort contextual meaning, while triplet extraction can omit attribution, negation, or quantitative information. In addition, incomplete cross-lingual entity alignment can fragment related evidence, and sentence extraction may omit important contextual details.

At the verification stage, semantically related evidence may be incorrectly treated as factual support. The examples show that the model can confuse attributed statements with factual knowledge or rely on topically relevant evidence that does not directly establish the claim. These results highlight the need for improved modeling of attribution, context, and claim-specific evidence sufficiency.

A.10 Ethical Considerations

This work focuses on improving multilingual misinformation detection and does not involve human subjects, personal user interactions, or the collection of private data. All source datasets used in this study (PolitiFact and Snopes) are publicly available and have been widely used in prior research. The constructed benchmark, **GraDuLaEval**, contains no personally identifiable information beyond what is already present in these public sources and does not introduce or disclose any new personal, sensitive, or private information.

Our dataset construction process relies on automatic translation and structured information extraction, and no new sensitive content is generated beyond the original datasets. Large language models are used strictly for offline preprocessing tasks such as translation and knowledge extraction, and not for generating novel misinformation or deceptive content. Any synthetic content produced during preprocessing is not released independently and is only used to support model training and evaluation.

While our approach aims to improve factual verification across languages, including low-resource settings, automated misinformation detection systems may still produce incorrect predictions. We emphasize that such systems should be used as decision-support tools rather than as replacements for human judgment, particularly in high-stakes domains. Finally, we acknowledge that pretrained models and translation systems may reflect societal or linguistic biases present in their training data. Although our language-agnostic design seeks to reduce language-based bias, it does not eliminate all forms of bias, and future work should include broader bias analysis and human-in-the-loop validation.

A.11 Graph visualization

Figure 5 shows examples of claim-specific heterogeneous graphs, where entities are linked through sentence-level evidence nodes.

A.12 Reference Sentence Initialization for the Language Branch

To explicitly encode language information while keeping it separate from semantic content, we initialize the language branch using fixed reference sentences. For each node in the graph (sentence or entity), we assign a language-specific reference text of the form `sent_id/entity_id: "This is an English sentence."`, where the identifier is unique for each node. The same template sentence is translated into the corresponding target

Error Category	Representative Examples	Observation
Translation errors	**Example 1 (EN→HI):** "Barack Obama wants to unilaterally disarm America" → "बराक ओबामा एकतरफा निरस्त्रीकरण चाहते हैं राष्ट्र chain email का कहना है". **Example 2 (EN→JA):** "employer, you have the choice to decide not to choose Sebelius" → 「選択する選択はできますセベリアスを選択しない」	The first example leaves the phrase "chain email" untranslated and produces an unnatural sentence, resulting in the loss of contextual information. The second omits the subject "employer" and fails to preserve the negation in "choose not to choose," leading to a semantic distortion. Such translation errors introduce noisy multilingual evidence that propagates to subsequent stages of the pipeline.
Triplet extraction errors	**Example 1:** Evidence: "Thomas McInerney shared his opinion that Barack Obama wants to unilaterally disarm America." Extracted triplet: *(Barack Obama, wants, unilaterally disarm America)*. **Example 2:** Evidence: "...at least 63% of hip replacements performed in Canada..." Extracted triplet: *(Hip replacement, est disponible en, Canada et England)*	The first triplet removes the attribution ("Thomas McInerney shared his opinion"), incorrectly transforming a quoted opinion into a factual assertion. The second converts a quantitative statement into a generic availability relation, discarding numerical information essential for claim verification.
Graph construction errors	**Example 1:** The entities *Roy Blunt*, *Mr. Blunt*, *Congressman Roy Blunt*, *Kongressmann*, and श्री ब्लंट are represented as separate nodes. **Example 2:** Triplet 1: *(ディック・チェイニー, 反対している, バラク・オバマ)*. Triplet 2: *(Dick Cheney, est en désaccord avec, Barack Obama)*	Semantically equivalent entities and relations expressed in different languages are not consistently merged into a unified graph representation. This creates duplicate nodes and edges, prevents effective aggregation of multilingual evidence, and reduces graph connectivity.
Sentence extraction errors	**Example 1:** From an article discussing Thomas McInerney's opinion, the extracted evidence includes "Obama is not a sane military leader" but omits the sentence identifying this statement as McInerney's opinion. **Example 2:** Extracted sentence: "당신은 Guantanamo 폐쇄되어야 한다고 생각합니까"	Although relevant documents are retrieved, the extracted sentences often omit crucial contextual information, such as attribution, negation, explanatory context, or supporting evidence. In addition, untranslated named entities and mixed-language sentences reduce the fluency and coherence of the evidence provided to the downstream language model.
Graph representation errors	**Example:** Claim: "Obama administration spent taxpayer dollars on electric cars in Finland and windmills in China." Selected triplets include: *(Paul Ryan, said, taxpayer dollars were spent overseas)*, *(Chine, bénéficiaire, moulins à vent)*, and *(Die Obama-Administration, hat Steuergelder ausgegeben, für Elektroautos)*. The ground truth is **False**, but the model predicts **True**.	The graph captures strong semantic and relational overlap with the claim, including concepts related to taxpayer dollars, electric cars, and windmills, but does not explicitly preserve stance or attribution. Consequently, an attributed political statement may be represented as factual knowledge, causing claim–speech confusion and resulting in an incorrect veracity prediction.
Claim-specific evidence selection and inference errors	**Example:** Claim: "The anti-terrorism presentation describes suspicious activities in the Fresno, California area." Selected evidence includes *(Aktivitäten, wurden beschrieben als, antiterroristisch)*, *(Fresno, befindet sich in, Kalifornien)*, and *(Fresno, fand statt, Präsentation)*. The ground truth is **False**, but the model predicts **True**.	The selected evidence provides generic multilingual context related to an anti-terrorism presentation and the Fresno area but does not establish the specific proposition that suspicious activities occurred there. The model therefore treats topical and semantic similarity as sufficient support, revealing limitations in claim-specific evidence selection and reasoning.

language (Hindi, Japanese, Korean, French, and German) and used to initialize nodes written in that language.

These reference sentences are encoded using the same multilingual SentenceTransformer as the semantic branch, but they are only provided to the language branch. Since all reference sentences are semantically identical and contain no factual information, their embeddings mainly capture language and script characteristics rather than meaning.

This initialization allows the language branch to explicitly model language identity, while the semantic branch focuses on factual content. By separating these two signals from the beginning, the

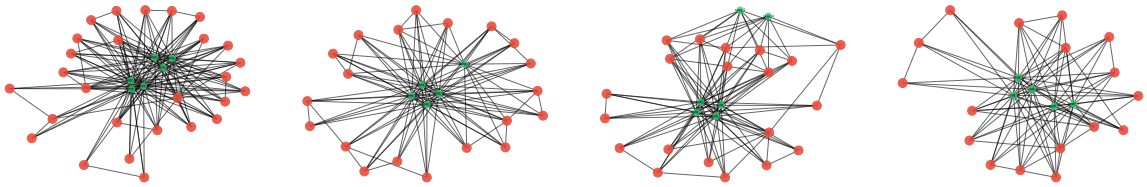


Figure 5: Visualization of heterogeneous claim-specific graphs constructed in GraDuLaEval. Red nodes denote **entity nodes**, and green nodes denote **sentence nodes**. Edges represent semantic and relational connections extracted from evidence documents. The graphs illustrate how multiple entities are linked through shared sentence-level evidence, forming a heterogeneous structure that enables structured cross-lingual reasoning.

model reduces the influence of language-specific patterns on semantic reasoning in the graph.

A.13 Human Annotation Details.

We conduct a post-hoc human evaluation to assess the relevance of triplets selected by the proposed GraDuLa framework. The goal of this annotation is to verify whether the automatically selected triplets are useful for claim verification, independent of model predictions.

We randomly sample 50 claims from GraDuLaEval (PolitiFact). For each claim, we extract the top-10 triplets selected from the claim-specific subgraph. As a baseline, we also extract the top-10 triplets selected using a sentence-transformer-based subgraph selection method, where triplets are ranked solely by embedding similarity to the claim.

The selected triplets are presented to human annotators along with the corresponding claim. Annotators are asked to judge how relevant each triplet is for verifying the claim, based on whether it directly supports, contradicts, or provides essential factual context. Relevance is assessed using a numerical rating scale, and scores are averaged across triplets and claims to obtain a final relevance score for each method.

Three graduate-level annotators volunteered for the study, all with prior experience in NLP and misinformation analysis. No ground-truth triplet labels are provided during annotation; judgments are based purely on annotators understanding of the claim and the factual content expressed by each triplet.

The annotation is conducted using spreadsheet-based interfaces, and no personally identifiable information is collected. Annotators are compensated fairly in accordance with standard academic annotation practices. By participating in the study, annotators consent to the use of their anonymized judgments for research purposes only.

B Datasheet for GraDuLaEval

B.1 Motivation

Q: For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)
A: This dataset aims to advance research in cross-lingual misinformation detection by addressing the shortage of large-scale, multilingual benchmark datasets. With coverage across multiple languages, it facilitates the development and evaluation of models that can effectively detect misinformation in a cross-lingual context, offering valuable resources to the research community.

Q: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
A: The datasets used in this study were constructed by re-purposing two publicly available datasets by authors of this paper.

Q: Who funded the creation of the dataset?
A: It is not disclosed to adhere to the anonymity policy.

Q: Any other comments? **A:** No.

B.2 Composition

Q: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)
A: Each sample in the dataset contains a claim, and corresponding evidences retrieved from the Web. Claims and Evidences can be in various languages such as English, Japanese, Hindi, Korean, French and German.

Q: How many instances are there in total (of each type, if appropriate)?
A: Please refer to Table 8 and Table 9.

Q: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
A: The dataset consists of all instances contained in the original datasets.

Q: Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information but might include, e.g., redacted text.
A: No.

Q: Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.)
A: No

Q: Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
A: Yes. please refer to the Table 8.

Q: Are there any errors, sources of noise, or redundancies in the dataset?
A: We perform rule-based filtration to remove noisy samples present in the dataset; however, it is not feasible to manually inspect all data instances.

Q: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)
A: The dataset is self-contained and can be freely accessed, used, modified, and shared without restrictions.

Q: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
A: No, as all samples in the dataset are publicly available.

Q: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
A: No.

Q: Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)
A: No.

Q: Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
A: No.

Q: Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
A: No.

Q: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
A: No.

Q: Any other comments?
A: No.

B.3 Collection process

Q: How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)
A: The dataset is re-purposed from already existing publicly available datasets Politifact and Snopes.

Q: What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?)
A: Both Politifact and Snopes datasets are available to be downloaded from the Web.

Q: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
A: NA.

Q: Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?
A: The dataset is re-purposed from already existing publicly available datasets Politifact and Snopes.

Q: Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., a recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)
A: Both original Politifact and Snopes datasets are frozen datasets, not evolving with time.

Q: Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
A: No.

Q: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
A: The dataset is re-purposed from already existing publicly available datasets Politifact and Snopes downloaded from the Web.

Q: Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)
A: NA.

Q: Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)
A: NA.

Q: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)
A: N/A.

Q: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)
A: No.

Q: Any other comments?
A: No.

B.4 Preprocessing, cleaning, labeling

Q: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions

in this section.)

A: Yes, please refer Section 3.2 in main content.

Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? (If so, please provide a link or other access point to the "raw" data.)

A: The "raw" data is already publicly available.

Q: Any other comments?

A: No.

B.5 Uses

Q: Has the dataset been used for any tasks already? (If so, please provide a description.)

A: Dataset is used for the task of cross-lingual misinformation detection, in this paper. Please refer to methodology and results sections in the main content.

Q: Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.)

A: No.

Q: What (other) tasks could the dataset be used for?

A: Cross-lingual misinformation detection.

Q: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)

A: No.

Q: Are there tasks for which the dataset should not be used? (If so, please provide a description.)

A: NA.

Q: Any other comments?

A: No.

B.6 Distribution

Q: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? (If so, please provide a description.)

A: Yes, the data will be free to the public to download, use, modify, and re-distribute.

Q: How will the dataset be distributed (e.g., tarball on the website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?)

A: Suitable access will be provided once the paper gets accepted.

Q: When will the dataset be distributed?

A: Suitable access will be provided once the paper gets accepted.

Q: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)

A: Yes, the dataset will be distributed under suitable license.

Q: Have any third parties imposed IP-based or other restrictions on the data associated with the instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.).

A: No.

Q: Do any export controls or other regulatory restrictions apply to the dataset or individual instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)

A: No.

Q: Any other comments?

A: No.

B.7 Maintenance

Q: Who is supporting/hosting/maintaining the dataset?
A: Authors.

Q: How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
A: Author emails.

Q: Is there an erratum? (If so, please provide a link or other access point.)
A: No.

Q: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)
A: As of now, there is no such plan.

Q: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.)
A: NA.

Q: Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)
A: No older version is available.

Q: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)
A: Anyone can freely extent or re-purpose the dataset.

Q: Any other comments?
A: No.