BNLI: A Linguistically-Refined Bengali Dataset for Natural Language Inference

Farah Binta Haque[1], **Md Yasin**[1,*], **Shishir Saha**[1,*],
Md Shoaib Akhter Rafi[2], **Farig Sadeque**[1]
[1]Brac University, [2]Bangladesh University of Engineering and Technology
[1]Kha 224 Pragati Sarani, Merul Badda, Dhaka 1212, Bangladesh,
[2]ECE Building, West Palashi Campus, Dhaka 1205, Bangladesh
farahhaq2023@gmail.com, yasinislam046@gmail.com, shishirsaha830@gmail.com,
1806114@eee.buet.ac.bd, farig.sadeque@bracu.ac.bd
*These authors contributed equally to this work.

Abstract

Despite the growing progress in Natural Language Inference (NLI) research, resources for the Bengali language remain extremely limited. Existing Bengali NLI datasets exhibit several inconsistencies, including annotation errors, ambiguous sentence pairs, and inadequate linguistic diversity, which hinder effective model training and evaluation. To address these limitations, we introduce BNLI, a refined and linguistically curated Bengali NLI dataset designed to support robust language understanding and inference modeling. The dataset was constructed through a rigorous annotation pipeline emphasizing semantic clarity and balance across entailment, contradiction, and neutrality classes. We benchmarked BNLI using a suite of state-of-the-art transformer-based architectures, including multilingual and Bengali-specific models, to assess their ability to capture complex semantic relations in Bengali text. The experimental findings highlight the improved reliability and interpretability achieved with BNLI, establishing it as a strong foundation for advancing research in Bengali and other low-resource language inference tasks. The link to the BNLI dataset: https://github.com/FarahHaque/BNLI-Dataset.git

Keywords: Natural Language Processing, Natural Language Inference, Bengali Language, Deep Learning

1. Introduction

Natural Language Inference (NLI), also referred to as Recognizing Textual Entailment (RTE), represents a core task in natural language understanding that requires determining whether a given hypothesis can be inferred from a corresponding premise. Over the past decade, large-scale English NLI datasets such as SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), and ANLI (Nie and Bansal, 2017) have become foundational benchmarks for advancing deep language models including BERT, RoBERTa, and LLaMA. These resources have substantially contributed to progress in semantic reasoning and contextual understanding across diverse NLP applications (Bowman et al., 2015; Williams et al., 2018; Gururangan et al., 2018). However, comparable resources for low-resource languages—particularly Bengali—remain strikingly underdeveloped.

Although Bengali ranks among the top ten most spoken languages globally, Bengali Natural Language Processing (BNLP) still suffers from a scarcity of high-quality annotated corpora for complex tasks such as entailment recognition (Aggarwal et al., 2022; Bhattacharjee et al., 2022; Kabir et al., 2024). Only a handful of preliminary efforts have been made to construct Bengali NLI datasets; however, most existing corpora suffer from severe shortcomings, including inaccurate translations from English datasets (Aggarwal et al., 2022; Bhattacharjee et al., 2022), inconsistent labeling, syntactic errors, and a lack of semantic and contextual diversity (see Fig. 2). Consequently, models trained on these resources often exhibit limited generalization, biased performance, and poor interpretability when exposed to real-world Bengali text (Kabir et al., 2024; Faria et al., 2024; Mahfuz et al., 2025). This persistent data deficiency has impeded meaningful advancement in Bengali inference modeling and constrained cross-lingual transfer research.

To bridge this critical gap, we present BNLI—a refined and linguistically curated Bengali Natural Language Inference dataset—specifically designed to facilitate robust evaluation and training of NLI models in Bengali. BNLI was developed through a rigorous multi-stage pipeline that combines manual data curation, human validation, and linguistic refinement to ensure both semantic consistency and syntactic authenticity. Unlike earlier datasets, BNLI emphasizes balanced representation across entailment, contradiction, and neutrality classes and incorporates examples reflecting diverse linguistic structures and contextual nuances present in natural Bengali text.

To establish baselines, we evaluated BNLI using several transformer-based architectures, including

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 85–91
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Table 1: List of existing NLI Datasets on diverse languages including Bangla.

Dataset	Language	Year	Sample Size
SNLI (Bowman et al., 2015)	English	2015	∼570k
MultiNLI (Williams et al., 2018)	English	2017	∼433k
e-SNLI (Camburu et al., 2018)	English	2018	∼570k
ANLI (Nie et al., 2020)	English	2018	∼162k
SICK (Marelli et al., 2014)	English	2019-2020	∼10k
SciTail (Khot et al., 2018)	English	2014	∼27k
RTE (Castillo, 2010)	English	2004-2013	Not Found
SCINLI (Sadat and Caragea, 2022)	English	2022	∼107k
FEVER (Thorne et al., 2018)	English	2018	∼185K
QNLI (Swayamdipta et al., 2020)	English	2018	∼100K
ARABIC-XNLI (Abdelali et al., 2024)	Arabic	2018	∼392k
ARBERT / MARBERT benchmarks (Abdul-Mageed et al., 2021)	Arabic	2021	∼20k
Arabic Natural Language Inference Corpus (Obeidat et al., 2025)	Arabic	2019	∼9k
SANA (Sentiment and NLI Arabic) (Abdul-Mageed and Diab, 2014)	Arabic	2020	∼7k
OSIAN (Open Source International Arabic News) (Zeroual et al., 2019)	Arabic	2019	∼1.1M
ViNLI (Van Huynh et al., 2022)	Vietnamese	2021	∼7.5k
ViHealthNLI (Nguyen et al., 2024)	Vietnamese	2024	∼5k
VietNLI (Bui et al., 2025)	Vietnamese	2023	∼25k
XNLI (Cross-lingual NLI) (Conneau et al., 2018)	Multilingual	2018	∼112k
INDICXNLI (Aggarwal et al., 2022)	Multilingual	2022	∼392k
Multilingual-NLI-26lang-2mil7 (He et al., 2020)	Multilingual	2022	∼2.73M
Cross-Lingual NLI for Low-Resource Languages (MT-NLI) (Zhao, 2022)	Multilingual	2023	∼3.5M
SICK-NL (Wijnholds and Moortgat, 2021)	Dutch	2021	∼10k
NLI-TR (Budur et al., 2020)	Turkish	2020	∼393k
farsTail (Amirkhani et al., 2023)	Persian	2020	∼10k
OCNLI (Original Chinese NLI) (Hu et al., 2020)	Chinese	2020	∼56k
Polish NLI (Ziembicki et al., 2024)	Polish	2022	∼10k
Bangla NLI (BanglaBERT benchmarks) (Bhattacharjee et al., 2022)	Bangla	2022	∼381k

multilingual and Bengali-specific models. The experimental results demonstrate that transformer models outperform recurrent approaches by effectively capturing contextual semantics and long-range dependencies in Bengali. These findings underscore the reliability of BNLI as a benchmark and highlight its potential to accelerate progress in Bengali and other low-resource language inference research (Rashad Al Hasan Rony et al., 2024).

In summary, the main contributions of this study are:

- We develop BNLI, a high-quality and linguistically refined Bengali NLI dataset addressing critical flaws in previous resources.
- We provide comprehensive benchmark evaluations using multiple transformer-based architectures to establish reliable baselines for future research.

By making BNLI publicly available, we aim to promote further exploration in Bengali semantic understanding, encourage resource development for other low-resource languages, and foster inclusive multilingual NLI research.

2. Literature Review

Table 1 presents an overview of major Natural Language Inference (NLI) datasets developed across various languages over the past two decades. As illustrated, extensive and well-curated resources exist for high-resource languages such as English, Arabic, and Chinese, as well as for several emerging multilingual benchmarks like XNLI, INDICXNLI, and MT-based multilingual corpora. These datasets—ranging from early collections such as RTE and SNLI to large-scale multilingual resources—have substantially driven progress in semantic understanding and cross-lingual inference modeling.

In contrast, Bengali remains severely underrepresented in this landscape. Only a limited number of Bengali NLI datasets have been introduced, and most of them are either small in scale or derived from machine-translated versions of English benchmarks. Moreover, many samples in these datasets suffer from semantic inconsistencies, mistranslations, and contextually incorrect entailment labeling, thereby limiting their reliability for model training and evaluation. This evident disparity underscores the urgent need for a linguistically consistent and semantically validated Bengali NLI resource. To address this gap, we introduce BNLI, a refined and meticulously curated dataset aimed at ensuring semantic integrity, syntactic correctness, and balanced class representation for robust Bengali inference modeling.

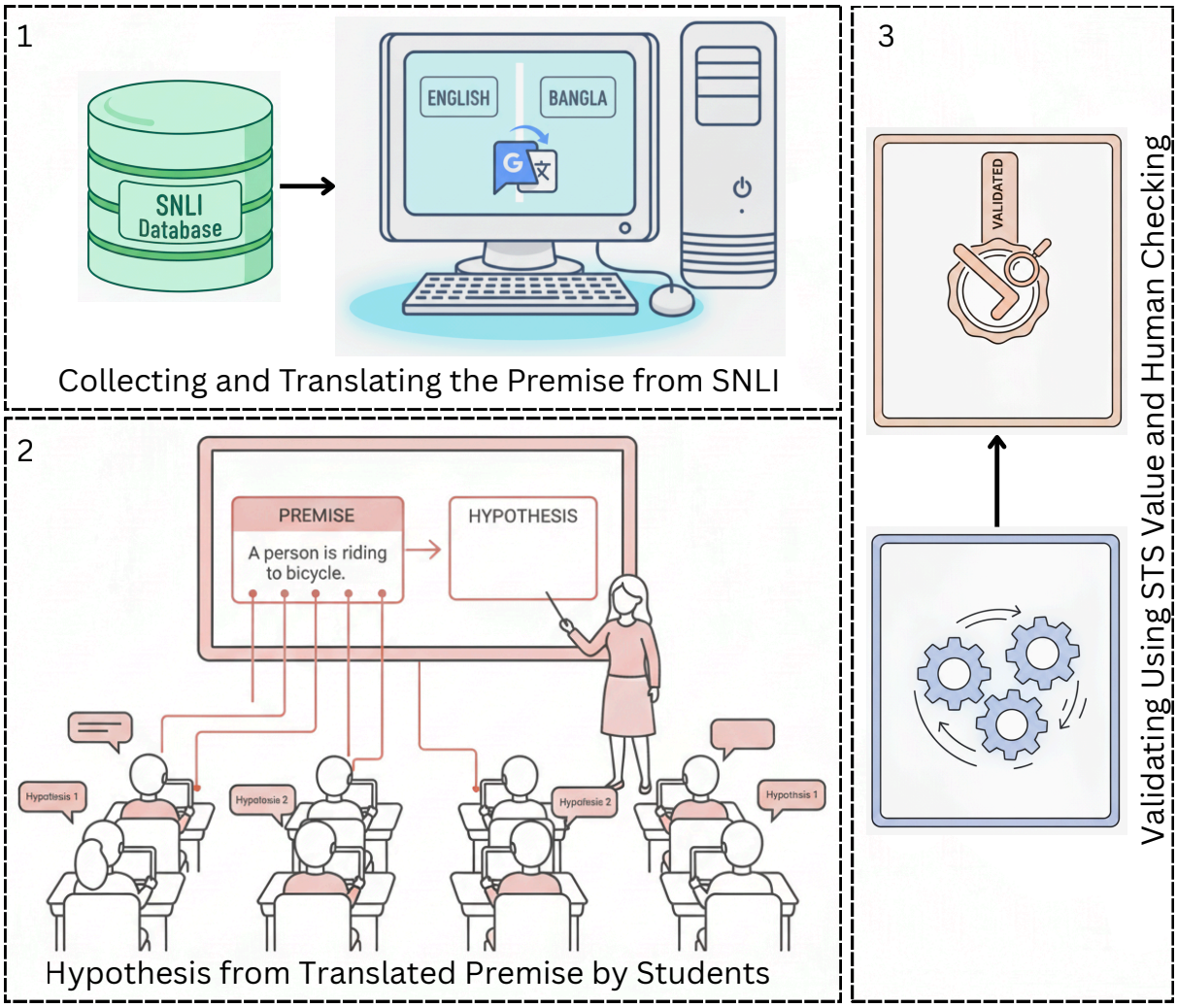

Figure 1: Overall workflow of the proposed multi-stage pipeline adopted for constructing the BNLI dataset, illustrating data collection, linguistic refinement, human validation, and final dataset compilation processes.

Premise	Hypothesis	Label
আমি একটা কেবিন দরজা ভেঙে মাটিতে পড়ে গেলাম-	আমি দরজা ভেঙে পড়ি এবং পড়ে যাই।	entailment
এটা ধীর গতির এটা উহ এই মুহূর্তে বাজারে অনেক ভাল মেশিন আছে	এটাই সবচেয়ে দ্রুতগতির মেশিন, তুমি এর চেয়ে ভালো মেশিন খুঁজে পাবে না।	contradiction
কারন আমি বলতে চাচ্ছি যে শীতের সময় সেখানে কত গরম লাগে সেটা কত	আমি যেখানে থাকি, শীতকালসহ সব সময় গরম থাকে।	neutral

Figure 2: Sample entries from the Bangla NLI dataset, showing grammartically inconsistent premise-hypothesis pairs with entailment, neutral, and contradiction labels.

3. Methodology

The BNLI dataset was constructed through a three-stage data collection and refinement process to ensure linguistic accuracy and semantic reliability. The data collection process is illustrated in Fig. 1.

In the first stage, 8885 sentences were selected as premises from the SNLI corpus (Bowman et al., 2015). These sentences were translated into Bengali with careful attention to preserving grammatical integrity and natural sentence flow. In the second stage, we focused on constructing semantically aligned hypotheses for the collected Bengali premises. To generate corresponding hypotheses, 60 native Bengali-speaking students from a reputed university participated in a structured data collection process. Each participant was provided with premise–scene descriptions and tasked with composing hypothesis sentences through Google

Premise	Hypothesis	Label	Similarity
শিশুরা হাসতে হাসতে ক্যামেরার দিকে হাত নাড়াচ্ছিল ।	শিশুরা হাসিমুখে ক্যামেরার দিকে তাকিয়ে হাত নাড়াচ্ছিল ।	entailment	0.923020
শিশুরা হাসতে হাসতে ক্যামেরার দিকে হাত নাড়াচ্ছিল ।	মেয়েরা গম্ভীর মুখে তার কথা ভিডিও করছিল ।	contradiction	0.110095
শিশুরা হাসতে হাসতে ক্যামেরার দিকে হাত নাড়াচ্ছিল ।	শিশুরা হাসতে হাসতে খেলা করছিল ।	neutral	0.622482
একজন বৃদ্ধ লোক একটি রেস্তরায় বসে কমলার জুস পান করছিল ।	একজন বৃদ্ধ লোক জুস খাচ্ছে ।	entailment	0.896547
একজন বৃদ্ধ লোক একটি রেস্তরায় বসে কমলার জুস পান করছিল ।	দুই মহিলা একটি রেস্তোরাঁয় মদ খাচ্ছেন ।	contradiction	0.182121
একজন বৃদ্ধ লোক একটি রেস্তরায় বসে কমলার জুস পান করছিল ।	একজন লোক একটি দোকানে জুস খাচ্ছেন ।	neutral	0.721758
দুই স্বর্ণকেশী মহিলা একে অপরকে আলিঙ্গন করছে ।	দুইজন মহিলা একে অপরকে আলিঙ্গন করছে ।	entailment	0.927653
দুই স্বর্ণকেশী মহিলা একে অপরকে আলিঙ্গন করছে ।	দুই কিশোর একসাথে খেলছে ।	contradiction	0.153298
দুই স্বর্ণকেশী মহিলা একে অপরকে আলিঙ্গন করছে ।	কিশোরী দুইজন একে অপরকে আপ্যায়ন করছে ।	neutral	0.684532

Figure 3: Sample entries from the BNLI dataset, showing premise-hypothesis pairs with entailment, neutral, and contradiction labels, along with estimated STS scores.

Algorithm 1 STS-Based Filtering

Require: Premise–hypothesis pair (p, h) and label $y \in \{\text{Contr.}, \text{Neut.}, \text{Ent.}\}$
Ensure: Decision $d \in \{\text{Keep, Drop}\}$
Translate (p, h) to English
Compute STS score $s \in [0, 1]$ using *en_stsb_bert_large* of Spacy
Set $d \leftarrow$ Drop
if $(y = \text{entailment}$ **and** $s > 0.70)$ **or** $(y = \text{neutral}$ **and** $0.30 < s < 0.75)$ **or** $(y = \text{contradiction}$ **and** $s < 0.30)$ **then**
    $d \leftarrow$ Keep
end if
return d

Forms. This process resulted in 26655 sentence pairs, evenly distributed across the three NLI labels: entailment, contradiction, and neutral.

In the third stage, all collected sentence pairs underwent a comprehensive validation process combining Semantic Textual Similarity (STS) analysis and human evaluation. Each premise–hypothesis pair was translated into English using the Google Translator Python package, and STS scores were computed to assess semantic alignment with the assigned NLI labels using *en_stsb_bert_large* model of Spacy library to determine STS value within 0 to 1. The premise-hypothesis dropping criteria is as described in Algorithm 1.

Pairs with inconsistent or outlier scores were flagged for manual inspection. Subsequently, five native Bengali linguists reviewed these pairs to verify grammatical correctness, semantic coherence, and label validity, resolving discrepancies through consensus. The final dataset contains total 23067 refined sentence pairs (Entailment: 7682 pairs, Contradiction: 7696 pairs, Neutral: 7661 pairs). This dual validation approach ensured that the BNLI dataset maintained high linguistic fidelity, semantic consistency, and balanced representation across inference categories. Arbitrary samples of the premise and hypothesis pairs from BNLI dataset has been shown in Fig. 3.

4. Result Analysis

The results on the BNLI dataset (Table. 2) show that transformer-based models substantially outperform traditional architectures. The LSTM baseline performs poorly, reflecting its limited ability to capture semantic nuances. BERT Base and RoBERTa deliver strong and balanced results, with RoBERTa slightly ahead due to improved contextual representations. BanglaBERT shows inconsistent performance across classes, suggesting limitations in domain coverage. MultiBERT performs reasonably well, benefiting from multilingual training. The LLaMA-2 model achieves the highest overall F1-score (79.13%), demonstrating superior cross-lingual reasoning and robust generalization compared to all other models.

Table 2: Model-wise performance comparison on BNLI dataset.

Model	#Params(M)	Classwise Performance			Average Performance			
		Contr.	Neut.	Ent.	Acc.	Prec.	Rec.	F1
LSTM	∼10	45.12%	46.38%	41.57%	44.19%	45.06%	44.32%	44.68%
BERT Base	110	76.42%	73.19%	64.87%	70.90%	72.45%	71.66%	71.92%
BERT Large	340	83.18%	68.73%	53.24%	67.5%	68.12%	68.44%	68.27%
BanglaBERT	168	32.46%	11.58%	88.12%	45.4%	49.37%	48.73%	44.26%
RoBERTa	125	79.68%	74.91%	67.43%	69.66%	74.65%	73.92%	74.21%
MultiBERT	185	86.23%	75.62%	48.91%	68.75%	69.58%	70.34%	69.93%
LLaMA-2	7,000	**84.76%**	**80.35%**	**72.41%**	**74.56%**	**79.51%**	**78.92%**	**79.13%**

The strong performance of LLaMA-2, although unexpected for NLI tasks on relatively small datasets, can be attributed to several factors. Its large-scale pretraining on diverse corpora enables implicit learning of semantic and inference patterns, while instruction-following and contextual reasoning capabilities allow it to interpret premise–hypothesis relationships effectively without extensive task-specific fine-tuning. Additionally, our STS-based filtering strategy produces semantically well-aligned and less noisy pairs, which may favor models with strong semantic similarity understanding, helping LLaMA-2 outperform smaller, fine-tuned encoders such as BERT variants. In contrast, BanglaBERT exhibits substantially lower performance, particularly for the neutral class, due to the inherent ambiguity of neutral samples, potential class imbalance, and dataset-specific characteristics that differ from prior benchmarks such as Bhattacharjee et al. This discrepancy further suggests that domain differences and implementation-specific factors (e.g., fine-tuning strategy, preprocessing, or tokenization) can significantly influence results, highlighting that previously reported performance may not directly generalize to the BNLI dataset.

5. Conclusion

In this work, we presented BNLI, a high-quality Bengali NLI dataset addressing the limitations of existing resources, including annotation inconsistencies and linguistic imbalance. Through a careful curation and annotation process, BNLI ensures clear semantic distinctions across entailment, contradiction, and neutral classes, making it suitable for reliable model evaluation and training. Our extensive benchmarking with both multilingual and Bengali-specific transformer models demonstrates the dataset's effectiveness in revealing model strengths and weaknesses, as well as the substantial performance gains achievable with large-scale language models such as LLaMA-2. Overall, Bengali NLI dataset serves as a foundational resource for advancing natural language understanding in a low-resource setting. A key future direction is the development of a Bengali foundation model, leveraging this dataset for large-scale pretraining and downstream reasoning tasks. The dataset supports applications in natural language inference and can further be utilized to train LLMs for meaningful sentence generation in the Bengali language.

6. References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, et al. 2024. Larabench: Benchmarking arabic ai with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520.

Muhammad Abdul-Mageed and Mona T Diab. 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages. *arXiv preprint arXiv:2204.08776*.

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023.

Farstail: A persian natural language inference dataset. *Soft Computing*, pages 1–13.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Emrah Budur, Rıza Özçelik, Tunga Güngör, and Christopher Potts. 2020. Data and representation for turkish natural language inference. *arXiv preprint arXiv:2004.14963*.

Tran Bao Bui, Linh Thi-Thuy Nguyen, and Tin Van Huynh. 2025. Vietx-nli: A cross-lingual natural language inference dataset with vietnamese as the source language. In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Julio J Castillo. 2010. Recognizing textual entailment: experiments with machine learning algorithms and rte corpora. *Special issue: Natural Language Processings and its Applications, Research in Computing Science*, 46:155–164.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2475–2485.

Fatema Tuj Johora Faria, Mukaffi Bin Moin, Asif Iftekher Fahim, Pronay Debnath, and Faisal Muhammad Shah. 2024. Unraveling the dominance of large language models over transformer models for bangla natural language inference: A comprehensive study. In *International Conference on Computing and Communication Networks*, pages 13–24. Springer.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. Benllm-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Tamzeed Mahfuz, Satak Kumar Dey, Ruwad Naswan, Hasnaen Adil, Khondker Salman Sayeed, and Haz Sameen Shahgir. 2025. Too late to train, too early to use? a study on necessity and viability of low-resource bengali llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1183–1200.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Huyen Nguyen, Thanh-Ha Do, Tuan-Anh Hoang, et al. 2024. Vihealthnli: A dataset for vietnamese natural language inference in healthcare. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 404–409.

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4885–4901.

Rasha Obeidat, Yara Al-Harahsheh, Mahmoud Al-Ayyoub, and Maram Gharaibeh. 2025. Arentail: manually-curated arabic natural language inference dataset from news headlines. *Language Resources and Evaluation*, 59(1):509–535.

Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, Rakib Al Hasan, Sumon Kanti Dey, Amzad Hossain Rafi, Amzad Hossain Rafi, Ashraf Hasan Sirajee, and Jens Lehmann. 2024. Banglaquad: A bengali open-domain question answering dataset. *arXiv e-prints*, pages arXiv–2410.

Mobashir Sadat and Cornelia Caragea. 2022. Scinli: A corpus for natural language inference on scientific text. *arXiv preprint arXiv:2203.06728*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

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*.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. Vinli: A vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872.

Gijs Wijnholds and Michael Moortgat. 2021. Sicknl: A dataset for dutch natural language inference. *arXiv preprint arXiv:2101.05716*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 1112–1122.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the fourth arabic natural language processing workshop*, pages 175–182.

Jiawei Zhao. 2022. *Cross-lingual learning in low-resource: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science, School of Natural and Computational Sciences, Massey University, Auckland, New Zealand*. Ph.D. thesis, Massey University.

Daniel Ziembicki, Karolina Seweryn, and Anna Wróblewska. 2024. Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering*, 30(2):385–416.