Evaluating the Generalization and Robustness of Language Models for Consistency Classification under Domain Shift and Adversarial Conditions

24149686, 21081625, 24055520,24147724, 24118326, 24073584,23070164

Abstract

Consistency is essential for reliable language model performance, yet remains underexplored in lightweight models. This study evaluates four compact models-DistilBERT, MiniLM, Electra, and T5-small-on binary consistency classification tasks across semantic, logical, and factual dimensions, benchmarking on six diverse datasets, including synthetic, adversarial, and real-world sources. Our experiments use zero-shot testing, large-data fine-tuning, and small-data training under in-domain, cross-domain, and adversarial conditions.

Results show that while lightweight models perform well on matched-domain data, they generalize poorly to unseen domains. MiniLM is more robust in adversarial settings, DistilBERT performs best in-domain, and T5-small benefits from domain-augmented training. However, all models exhibit a tendency to memorize dataset-specific patterns rather than demonstrate true reasoning. These findings highlight the need for improved model architectures and training strategies to enhance generalization and consistency understanding in lightweight models.

1 Introduction

As language models are increasingly deployed in real-world applications-such as fact-checking systems, virtual assistants, and scientific summarization-their ability to maintain logical, semantic, and factual consistency has become a critical requirement(Saxena et al., 2024). However, recent research has shown that even state-of-the-art pre-trained models often generate or misclassify outputs that contradict earlier context, common sense, or real-world knowledge(Huang et al., 2025). This inconsistency undermines model reliability and poses risks in diverse domains.

While most existing work focuses on large models, limited research has explored the behavior of lightweight models under such challenging conditions(Wang et al., 2024). Given the growing demand for resource-efficient solutions, it is crucial to understand how lightweight language models perform in consistency-sensitive tasks across diverse domains and data sources, particularly their robustness under domain shift, adversarial inputs, and limited data scenarios remains questionable.

This work addresses that gap by evaluating the generalization and robustness of four lightweight models-DistilBERT, MiniLM, Electra, and T5-small-in the context of logical consistency classification. We define consistency as a binary classification task: determining whether a sentence is logically, semantically, or factually coherent. To this end, we construct a diverse benchmark of six datasets, covering synthetic model-generated data, real-world fact-checked corpora, adversarially crafted examples, and NLI-based logical consistency tasks. These datasets vary in linguistic complexity, consistency type, and topical domain, allowing us to examine model behavior under in-domain, cross-domain, and adversarial conditions.

Our evaluation comprises three experimental phases: (1) zero-shot testing to examine inductive biases; (2) large data fine-tuning on in-domain and cross-domain performance; and (3) small-data experiments to explore pattern learning versus genuine reasoning. Through this setup, we aim to answer the following research question:

To what extent can lightweight language models generalize and remain robust in consistency classification under domain shifts, adversarial perturbations, and varying training conditions?

Our results reveal that while lightweight models achieve strong performance on in-domain data, their generalization to unseen domains is limited. Models like MiniLM show improved robustness with adversarial conditions, but all tend to rely on dataset-specific patterns over genuine reasoning.

2 Related Work

We review prior work on consistency evaluation, domain robustness, and lightweight Pre-trained Language Models (PLMs), focusing on how compact models behave consistently across tasks and domains.

2.1 Consistency Evaluation

NLI datasets such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) are standard for testing semantic reasoning, and we repurpose them to assess model consistency under semantic, factual, and logical perturbations. However, their simplicity limits their applicability in the real world.

BECEL (Ravichander et al., 2022) provides a more comprehensive view, evaluating five types of consistency in diverse tasks beyond NLI. While prior work centers on large models (e.g., BERT, GPT-3), we focus on compact models in realistic and adversarial settings, addressing an underexplored area.

2.2 Domain Generalization and Robustness

Models often fail to generalize beyond training data due to domain change. This problem has been observed in various NLP tasks, including sentiment classification (Blitzer et al., 2007), QA (Fisch et al., 2019), and machine translation (Koehn and Knowles, 2017). Domain-adversarial training (Ganin et al., 2016), data augmentation (Wei and Zou, 2019), and other techniques have been proposed, but their application to consistency classification is limited. Our study includes domain generalization and adversarial robustness as core evaluation criteria.

2.3 Lightweight Pretrained Language Models

Several lightweight PLMs have been introduced to reduce computational costs. DistilBERT (Sanh et al., 2019) compresses BERT while retaining most of its performance. MiniLM (Wang et al., 2020) uses attention distillation, Electra-small (Clark et al., 2020) adopts a replaced token detection approach, and T5-small (Raffel et al., 2020) generalizes through a text-to-text paradigm. Although these models are commonly evaluated on GLUE, they have not been systematically assessed for consistency under domain-shift or adversarial conditions. Our work directly benchmarks these models on consistency tasks in realistic settings.

2.4 Our Work

Existing consistency studies often focus on clean, in-domain data with balanced labels. Moreover, we are most relying on challenges in domain shift, adversarial inputs, and low-resource settings in large-language models, with limited exploration of lightweight alternatives. We propose a unified, realistic framework to evaluate consistency in NLP. Using four compact models, DistilBERT, MiniLM, Electra-small, and T5-small, we test performance across in-domain, cross-domain, data-augmented, and adversarial setups, under both high- and low-resource conditions. This enables a thorough analysis of their generalization, robustness, and scalability.

3 Methods

3.1 Problem Formulation and hypothesis

Collecting, annotating, and validating real-world data resources is an extremely costly and time-consuming task, often constrained by ethical considerations, privacy concerns, or high acquisition costs. (Nowruzi et al., 2019) In today's era of widespread use of large language models (LLMs), generating synthetic data offers significant flexibility and controllability, making it a promising alternative to real data. Existing LLMs, such as GPT-4 and Cohere, already possess the ability to generate high-quality, coherent text, and can produce data that meets specific requirements when provided with appropriate prompts. (Nadas et al., 2025) However, whether synthetic data can fully replace real data remains uncertain. (Li et al., 2023)

In addition, recent studies have made significant progress in diagnosing logical fallacies using machine learning, including classification and the development of custom models. Building upon these efforts, we propose two hypotheses:

- Lightweight transformer models perform well in achieving reliable binary classification of consistency of sentences .
- Models trained on synthetic datasets can perform well in consistency detection tasks given the scarcity and diverse of real-world data.

3.2 General Progress

Our objective is to investigate the ability of lightweight models evaluation of generalization and robustness on synthetic or real data to assess consistency in sentences.

We selected four lightweight transformer models due to their distinct designs and pretraining strategies. This diversity allows us to evaluate how different learning mechanisms-ranging from masked language modeling and attention-based distillation to discriminative token replacement and text-to-text generation-affect consistency classification.

Furthermore, we selected six distinct datasets covering three types of consistency with label 1 as consistent and label 0 as inconsistent: semantic, logical, and factual. These datasets include both synthetic and real-world samples, enabling a robust investigation of textual consistency while evaluating model performance under in-domain, cross-domain, and adversarial conditions.

After selecting models and datasets, we plan to conduct a series of experiments to evaluate model performance under various conditions. We will begin with zero-shot evaluations to assess the models' inherent reasoning capabilities without any fine-tuning. This will be followed by fine-tuning on synthetic datasets, and subsequently testing on both synthetic and real-world datasets to examine cross-domain generalization.

3.3 Model Architectures

The four lightweight transformer-based models are **Distilbert** (Sanh et al., 2019), **MiniLM** (Wang et al., 2020), **T5-small** (Colin, 2020), and **Electra** (Clark et al., 2020). These models provide an effective trade-off between model size and performance, offering faster inference times while maintaining competitive accuracy, making them suitable for resource-constrained scenarios. Each model uses its native HuggingFace tokenizer. DistilBERT, MiniLM, and Electra utilize WordPiece tokenization, while T5-small uses SentencePiece. All input texts were tokenized using a maximum sequence length of 128, with truncation and padding applied for consistent dimensions across models. The classification loss function was *CrossEntropyLoss*, with T5-small using the corresponding sequence-to sequence objective.

A unified training configuration was adopted across models. Optimization was performed using the AdamW algorithm, with batch sizes adjusted according to model size and memory limitations (e.g., 16/32 for DistilBERT, 4 for T5-small). Early stopping was not applied.

Model	**Params**	**Architecture**
DistilBERT	∼66M	BERT-based
MiniLM	∼33M	BERT-style (light)
T5-small	∼60M	Seq2Seq Transformer
Electra	∼42M	Discriminator-based

Table 1: Model size and architecture type.

Model	**Pretraining**	**Tokenizer**
DistilBERT	Distilled MLM	WordPiece
MiniLM	Contrastive	WordPiece
T5-small	Denoising Autoencoder	SentencePiece
Electra	Replaced Token Detection	WordPiece

Table 2: Pretraining objectives and tokenization methods.

3.4 Evaluation Metrics

To comprehensively evaluate model performance in consistency classification, we report two core metrics: **Accuracy** and **F1-score**. Each metric captures a different aspect of model behavior and generalization. All metrics were computed using the `scikit-learn` library (Pedregosa et al., 2011). For the T5-small model, text-based predictions were decoded and post-processed into binary labels prior to metric computation.

Accuracy reflects overall correctness and serves as a straightforward measure of classification performance across the entire test set. However, under conditions of domain shift or class imbalance, accuracy can be misleading as it may overestimate model robustness (Goutte and Gaussier, 2005).

F1-score, computed using the macro average (Schütze et al., 2008), provides a more balanced view by considering both precision and recall across classes. It is particularly informative in scenarios with non-uniform class distributions, and thus serves as a more sensitive indicator of robustness under distributional shifts.

In our evaluation, F1 score serves as the primary metric for assessing zero-shot performance. This is because F1 provides a balanced measure of a model's ability to correctly identify both consistent and inconsistent samples, particularly important in cases where the model may be biased toward a single class (Powers, 2020). While accuracy is reported for completeness, it can be misleading in the presence of asymmetric model behavior or imbalanced predictions, which are common in zero-shot scenarios. Thus, F1 better reflects the actual reasoning ability of the models under test. (Chicco and Jurman, 2020).

4 Experiments

4.1 Datasets

To evaluate the consistency capabilities of lightweight language models, we construct a comprehensive benchmark consisting of six diverse datasets. These datasets are carefully selected to cover three major types of consistency - **semantic, logical, and factual** (Examples are detailed in the Appendix) - and span a wide range of linguistic structures, topical domains, task types, and data sources, ensuring a thorough and robust evaluation across a wide range of natural language scenarios. Specifically, our dataset selection enables:

- **Multi-perspective training and evaluation**, including synthetic texts from large language models, real-world annotated corpora, and adversarial examples, incorporating balanced and imbalanced datasets;
- **Stylistic diversity**, ranging from simple, well-structured sentences to long-form passages and ambiguous, philosophically nuanced claims;
- **Cross-domain representation**, incorporating content from diverse domains including science, history, politics, law, health, and everyday knowledge.

These datasets allow us to systematically assess both generalization and robustness of models in consistency classification.

Cohere Dataset: We generated 55,000 synthetic samples using the Cohere Command model, evenly split between consistent and inconsistent statements. Prompts targeted seven domains: *History & Culture, Technology & AI, Science & Environment, Education & Learning, Law & Ethics, Sports & Entertainment, and Business & Marketing*. The outputs are grammatically correct and diverse, making this dataset ideal for in-domain training and baseline evaluation.

ChatGPT General Dataset: This dataset includes 16,500 balanced samples produced by ChatGPT, containing factual and logical claims across open domains. It is used to test whether models trained on Cohere data can generalize to a distinct generative style and semantic structure.

Adversarial Dataset: Comprising 4,950 high-complexity samples, this dataset was created via ChatGPT to probe model robustness. Samples contain subtle inconsistencies and abstract, ambiguous expressions (e.g., "It is debated whether…"), covering topics like language, cognition, and philosophy. It evaluates models' behavior under uncertainty and distributional edge cases.

Dataset	Description	Focus
Cohere	Synthetic from Cohere	Multi-domain semantic/logical/factual consistency
ChatGPT	Synthetic from ChatGPT	General-domain consistency
Adversarial	Synthetic from ChatGPT	Robustness under ambiguity and borderline cases
SNLI	SNLI from Stanford NLI Corpus	Entailment vs. contradiction-based logical consistency
Merged	Mixed from WikiText + Logical Fallacy	Wikipedia facts vs. formal reasoning errors
Real-life	Real-world from LIAR + FEVER + PubHealth	Factual consistency of real-world knowledge claims

Table 3: Description and focus of datasets used in our evaluation.

SNLI Dataset: The SNLI corpus is a standard benchmark for logical inference. We converted each *premise-hypothesis* pair into a single input and binarized the labels: *entailment* → 1 (consistent), *contradiction* → 0 (inconsistent), and excluded *neutral*. The final dataset contains 6,780 samples (3,368 consistent, 3,412 inconsistent) and supports fine-grained logical consistency testing, including negation and transitivity (Jang et al., 2022).

Merged Dataset: This dataset simulates mixed-domain input by combining two sources. Consistent samples (label 1) come from the WikiText corpus, composed of high-quality Wikipedia articles ("Good" or "Featured"), representing long-form factual text. Inconsistent samples (label 0) are drawn from a logical fallacy classification dataset, covering diverse reasoning errors such as false causality and circular reasoning. The combined dataset contains 8,590 sentences (5,380 consistent, 3,210 inconsistent) and is designed to test model robustness to varied linguistic styles and formal logic violations.

Real-Life Dataset: To evaluate factual consistency in real-world contexts, we merged three publicly available datasets: LIAR (political claims), FEVER (Wikipedia-based fact-checking), and PubHealth (health-related verification). Labels were binarized: true/mostly-true → 1; false/pants-fire → 0. After preprocessing, the corpus includes 112,544 samples (78,114 consistent, 34,430 inconsistent), serving as the largest and most diverse real-world benchmark in our evaluation, particularly valuable for testing model performance on factual consistency tasks.

4.2 Experiment Design

4.2.1 From Zero-Shot to Fine-Tuned

We trained four models on Cohere datasets and evaluated their performances on both synthetic and real-world datasets using a 70% training and 30% testing split.

Before fine-tuning, we assessed the models' zero-shot capabilities. DistilBERT performed best, achieving ≈ 0.5 accuracy and ≈ 0.6 F1 score. In contrast, MiniLM and T5-small failed to grasp the task, with zero F1 score across most datasets.

Training details: At fine-tuning, we used AdamW optimiser. All models are trained for 5 epochs across. The batch size and learning rate are different across model size and tasks.

After training on Cohere's synthetic data, all models achieved near-perfect performance on the corresponding test sets (accuracy and F1 > 0.99). However, they failed to generalize to other datasets. MiniLM and T5-small showed slight F1 improvements on synthesised datasets only($\approx$ 0.5), but minimal gains in accuracy.

Retraining on Cohere + ChatGPT data preserved performance on corresponding synthetic test sets (accuracy and F1 > 0.99). Generalization remained limited, but MiniLM and T5-small showed marked improvements on Real-life and Adversarial data. MiniLM's Real-life accuracy rose from 0.5032 to 0.6818 and F1 from 0 to 0.7727. T5-small's performance on Adversarial data improved in accuracy from 0.5 to 0.7079 and F1 from 0 to 0.7543; on Real-life data, accuracy rose from 0.3059 to 0.6762 and F1 from 0 to 0.7633.

These results suggest models perform well on matched train-test pairs but generalize poorly, likely leveraging dataset-specific patterns rather than task understanding.

4.2.2 Small-Scale Training and Generalization

To test for pattern memorization, we trained DistilBERT and MiniLM on 10% subsets of each dataset and evaluated cross-dataset performance.

DistilBERT continued the pattern observed in previous experiments: it achieved near-perfect scores (accuracy and F1 > 0.97) only on test sets corresponding to its training data. An exception was observed when training on ChatGPT data, where DistilBERT achieved 0.7324 accuracy and 0.8366 on Real-life dataset. Also, when trained on the Real-life dataset, it generalized moderately well to three other datasets, but training on SNLI yielded no notable improvements.

MiniLM displayed slightly better generalization. Training on Cohere led to strong performance on both Cohere and ChatGPT test sets (accuracy and F1 > 0.97), and decent results on the merged dataset (accuracy ≈ 0.84, F1 ≈ 0.83). Notably, training on ChatGPT and Adversarial data enabled broader generalization (accuracy ≈ 0.7, F1 ≈ 0.77 across most sets), though performance dipped when testing on Cohere after training on ChatGPT. An exception was observed when training on Real-life data, where MiniLM achieved solid results on SNLI (accuracy ≈ 0.73, F1 ≈ 0.75).

These findings reinforce the diagonal pattern—models perform well when training and test domains align—and highlight limited generalization, suggesting potential reliance on dataset-specific cues rather than true task comprehension.

5 Results and Discussion

5.1 Zero-Shot Performance

In our zero-shot evaluation of light-weighted models, we observe direct model performance (Table 4& Table 7) without fine-tuning. According section 3.4, we choose F1 as our primary evaluation metric. DistilBERT consistently achieves the highest F1 scores on two thrids of the datasets and maintains stable accuracy overall, despite underperforming on GPT. Electra exhibits greater variability across models. MiniLM and T5-small performs poorly across almost all zero-shot with near zero F1 score. Overall, DistilBERT is most reliable across a wider range of data.

DistilBERT well performance may be largely attributed to its masked language modelling pretraining. (Devlin et al., 2019)While underformance on GPT may be caused by domain mismatch between the GPT-generated and the kind of data DistilBERT was trained on. MiniLM performance suggests that it maybe sensitive to nuanced logical patterns in specific contexts. (Wang et al., 2020) On the contrary, indicating that T5-small is defaulted to random or majority class predictions without fine-tuning. (Zhou et al., 2021) This distinct contrast

Dataset	DistilBERT	Electra	MiniLM	T5-small
Cohere	0.6623	0.5780	0.0000	0.0000
GPT	0.0633	0.6043	0.0000	0.0000
Adversarial	0.6192	0.0737	0.6667	0.0000
Merged	0.6592	0.1286	0.0000	0.0000
Real-life	0.6107	0.4319	0.0000	0.0000
SNLI	0.5646	0.3576	0.0000	0.0000

Table 4: Zero-shot F1 scores of lightweight models across datasets.

suggests that while some lightweight models possess initial inductive biases suitable for consistency reasoning, most require targeted training to function effectively in identifying consistency. (Brown et al., 2020)

5.2 Performance with Large Training Datasets

Based on the fine-tuned results in Tables 8–11, key observations have been made regarding generalisation across datasets. DistillBERT is overall the most robust, while Electra shows similar trends—performing slightly better on SNLI but less effectively on Merged and Real-life datasets. MiniLM demonstrates balanced generalization, notably improving on adversarial and real-life data when trained on both Cohere and GPT. T5-small excels on GPT and adversarial sets but generalizes inconsistently on others.

To better understand these inconsistencies and the persistently high accuracy scores on the Cohere test set, we further explore the underlying reasons as follow:

5.2.1 Variability of Model Architecture Difference

The differences in lightweight model performance are closely tied to their architectural designs and pretraining approaches. DistilBERT, a distilled version of BERT trained with masked language modeling, generally excels in tasks with clear syntactic structures (Sanh et al., 2019). MiniLM employs contrastive distillation to preserve attention patterns, which may account for its strong performance on semantically or logically challenging datasets (Wang et al., 2020). Electra, which is pretrained to discriminate between real and replaced tokens, is highly efficient but might be less flexible when dealing with real-world inconsistencies (Clark et al., 2020), leading to poorer performance. Although earlier evaluations of T5-small suggested weak generalization, results from training on the combined Cohere + GPT dataset paint a more nuanced picture. T5-small achieved high performance on Adversarial(F1: 0.7542) and Real-life datasets (F1: 0.7633), significantly outperforming its previous zero-shot and single-source training baselines. This improvement may be attributed to the nature of the dataset generated - broader domain coverage and stylistic variety introduced - by combining Cohere and ChatGPT data, which enhanced its ability to align generation-based predictions with classification objectives (Colin, 2020)

Overall, while these models are well-suited for fine-tuning on large-scale datasets, their ability to generalize depends still heavily on their internal mechanisms. This highlights the need to choose architectures that align with the consistency task's logical and semantic demands.

5.2.2 Variability of Datasets Patterns

To investigate dataset formation, we consider three perspectives:

First, synthetic data generated from the Cohere model consistently produced near-perfect performance across all lightweight models. This can be explained by the low variance and high regularity of language used in the Cohere dataset, which often recycles predictable topical domains (e.g., ethics, human rights, climate change) and associated expressions. (Brown et al., 2020) These repeated linguistic structures reduce ambiguity and provide strong surface-level cues that models can exploit. Furthermore, as revealed in the N-gram top frequent analysis(Figure 2) , Cohere samples showed dense clusters of repeated n-gram phrases, especially in consistent examples, making it easier for models with recurring syntactic patterns. (Manning and Schütze, 1999) This predictability reduces the models' need for deeper reasoning and encourages memorization, explaining why even models with limited reasoning capacity achieved nearly perfect scores in this setting. (Bowman et al., 2015)

Second, a Part-of-Speech (POS) analysis (Figure 3) shows that Cohere exhibits a tightly clustered POS pattern (dominated by determiners, common nouns, and modal verbs), which simplifies learning, (Toutanova et al., 2003; Jurafsky and Martin, 2023), while the GPT dataset presents a broader, more dispersed distribution, demanding more flexible generalization.

Lastly, we noticed that Cohere primarily uses linear logical constructions (e.g., cause-effect, direct contrast), which are easier for models to track, whereas the Real-life and SNLI datasets involve more abstract or implicit logic forms that require multi-step, context-dependent inference—highlighting why models excel on Cohere but struggle with less explicit logical patterns. (Williams et al., 2018; Camburu et al., 2018)

5.2.3 Further Verification Experiment

To further investigate model generalization and pattern reliance, we selected the Merged dataset as a base training datasets due to its inclusion of both structured and logically inconsistent content (see Section 4.1). This allows us to explore for a upper bound evaluation of model's performance on real consistency cases since sythetic data may have similar contexts.

DistilBERT, Electra, and MiniLM were trained on this dataset and then evaluated on three test sets: Merged (in-domain), Real-life, and SNLI. In-domain performance was high for all models (e.g., DistilBERT achieved 0.9640 accuracy and 0.9713 F1), but cross-domain generalization varied. On the Real-life set, DistilBERT maintained relatively strong performance (F1:0.7292), MiniLM was moderate (F1: 0.5585), and Electra performed poorly (F1:0.0632). On SNLI, performance dropped for all, with MiniLM slightly leading (F1:0.5450 compared to 0.1274 for DistilBERT and 0.2051 for Electra). [Tables 12].

These findings reinforce the need to evaluate models beyond in-domain tests, as high performance may result from memorizing superficial stylistic or lexical patterns rather than genuine understanding, echoing earlier work on shortcut learning and spurious correlations. (McCoy et al., 2019; Gururangan et al., 2018)

5.3 Performance with Small Training Datasets

Above results may suggest overfitting in the datasets. In order to avoid this, we tried to train with smaller datasets.

Based on the data in Fine-tuned Cross-Dataset Evaluation, we hypothesize that there may be certain patterns present in different datasets (especially real-world datasets) that allow the model to learn these patterns and make logical judgments even with a small amount of training data.

We used the **t-SNE** (Van der Maaten and Hinton, 2008) method to independently visualize each dataset, and we can observe the distribution patterns of the embeddings of different datasets in space, as shown in Figure 1. We found that the Cohere and SNLI datasets have compact embeddings, indicating the presence of clear patterns that allow the model to make easier judgments. In contrast, Adversarial or Real-life datasets are more diverse or complex, leading to more scattered embeddings. After training on these datasets, the model might rely solely on superficial features, unable to perform true logical reasoning.

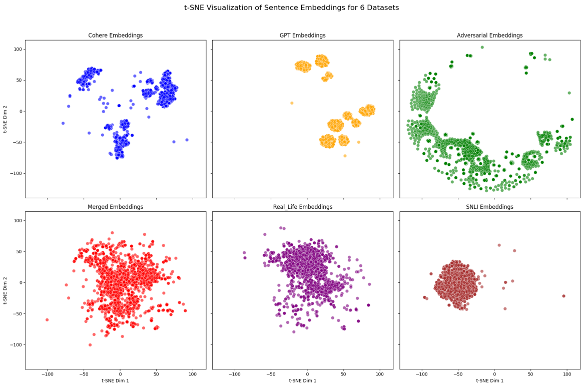

Figure 1: t-SNE Visualization of Sentence Embeddings for 6 Datasets

Taking the DistillBERT model trained on the Cohere dataset as an example (see Table 5), DistillBERT achieves F1 = 0.9767 on this dataset, which is quite excellent. This high accuracy and F1 score indicate that DistillBERT can effectively capture patterns in the data and make logical judgements. However, when the model is tested on SNLI data judgements, performance drops dramatically, with F1 = 0.0135. This result suggests that the pattern in the SNLI dataset differs significantly from that in the training dataset, causing the model to fail to adapt to the new pattern, and therefore the prediction performance suffers.

Dataset	ACC	F1 Score
Cohere	0.9766	0.9767
GPT	0.5310	0.2430
Merged	0.3800	0.1063
Real-Life	0.3387	0.1233
SNLI	0.5046	0.0135

Table 5: DistillBERT trained on Cohere dataset

Compared to DistillBERT, the performance of MiniLM is more balanced, as shown in Table 6, especially across multiple datasets. Although it still performs best on the Cohere dataset, it remains relatively stable when faced with other datasets such as SNLI or Adversarial. While this stability is notable, it can also be inferred that the MiniLM model rely on a certain local pattern rather than a deep understanding of logic itself.

Dataset	ACC	F1 Score
Cohere	0.9783	0.9737
GPT	0.9738	0.9737
Merged	0.8412	0.8314
Real-Life	0.3387	0.1233
SNLI	0.5046	0.0135

Table 6: MiniLM trained on Cohere dataset

This situation aligns with our hypothesis that the model has learned the local patterns in the data, rather than universal logical reasoning rules. In the experiment, both DistillBERT and MiniLM rely on specific patterns in the datasets for reasoning.

Thus, even though data augmentation improves the model's adaptability to similar datasets, it does not significantly enhance the model's generalization ability to different data patterns because it does not enable the model to truly understand logic. According to the experimental results, even with enhanced diversity in the training set, the model still shows significant variation when faced with datasets with different patterns. This further confirms the model's reliance on pattern learning rather than mastering universal logical reasoning rules.

6 Conclusion

Our results show that lightweight models perform well in matched-domain settings but lose accuracy in cross-domain and adversarial scenarios. DistilBERT excels in domain-specific tasks thanks to its masked language modeling pretraining, while MiniLM is more resilient to adversarial conditions due to its attention-focused distillation. T5-small improves significantly with diverse, domain-augmented training data, underscoring the importance of dataset diversity. In contrast, Electra struggles with distribution shifts, likely due to its discriminative pretraining that emphasizes local token patterns over global logic.

However, our additional analyses further supported our hypothesis, revealing that most models predominantly rely on memorizing superficial dataset-specific patterns rather than genuinely learning underlying logical reasoning. This reliance was especially evident in our small-data experiments, where the models exhibited high accuracy only when training and testing data patterns closely aligned, indicating limited true reasoning capability.

Despite the rigor of our approach, several limitations highlight promising directions for future research:

Scale and Coverage of Data:

Our limited synthetic and real-world datasets restrict the full spectrum of consistency classification scenarios. Future work should expand the data to cover more varied domains and a broader range of consistency types.

Richness of Consistency Types:

Currently treating consistency as a binary task oversimplifies the problem. Future studies could employ multi-label or hierarchical annotations to distinguish between semantic, logical, and factual inconsistencies for more nuanced evaluations.

Architectural Innovations:

Exploring hybrid lightweight models that integrate structured logic components with pre-trained transformers may improve logical reasoning while enhancing interpretability. Such designs could offer insights into the models' internal decision-making processes and reduce reliance on superficial cues.

Robust Evaluation under Distribution Shifts: Our current assessment relies on F1 and Accuracy which may not fully capture how our classifier performs under varying real-world conditions. It may benefit from multi-ID effective robustness framework (Shi et al., 2023) for evaluating model performance across diverse in-distribution scenarios, obtaining a more comprehensive understanding of data shifts.

In conclusion, our study reveals that while lightweight language models can excel in controlled settings, they struggle to generalize across domains and handle adversarial input. Data augmentation and light-weighted model architecture play critical roles in improving performance, but pattern memorization remains a central challenge. Addressing these limitations through richer data, finer-grained tasks, and more interpretable modeling approaches will be key to advancing consistency classification in real-world NLP applications.

A Appendix

Data Availability Statement:

*All datasets and codes are available in the uploaded zip file. The 6 datasets used in our experiment are named accordingly in the folder.

Experiment Results Table & Figures:

Dataset	DistilBERT	Electra	MiniLM	T5-small
Cohere	0.5313	0.5348	0.5013	0.5013
GPT	0.4747	0.5605	0.4950	0.4972
Adversarial	0.4586	0.4696	0.5000	0.5000
Merged	0.5670	0.3907	0.5031	0.3736
Real-life	0.5088	0.4246	0.5032	0.3059
SNLI	0.4918	0.5136	0.5030	0.5033

Table 7: Zero-shot accuracy of lightweight models across datasets.

Dataset	ACC	F1	ACC	F1
	Cohere Train		Cohere + GPT Train	
Cohere_Test	0.9990	0.9990	0.9967	0.9967
GPT	0.6175	0.4679	0.9998	0.9998
Adversarial	–	–	0.6489	0.6195
Merged	0.3892	0.2080	0.3927	0.2012
Real-life	0.3820	0.2668	0.4903	0.4875
SNLI	0.5046	0.0135	0.5038	0.0059

Table 8: DistilBERT performance across datasets when trained on Cohere vs. Cohere + GPT.

Dataset	Cohere Train		Cohere + GPT Train	
	ACC	F1	ACC	F1
Cohere_Test	0.9967	0.9967	0.9966	0.9966
GPT	0.5554	0.3262	0.9998	0.9998
Adversarial	–	–	0.6592	0.5170
Merged	0.3051	0.2750	0.3465	0.1083
Real-life	0.4036	0.3765	0.3827	0.2709
SNLI	0.5058	0.1528	0.5057	0.0271

Table 9: Electra performance across datasets when trained on Cohere vs. Cohere + GPT.

Dataset	Cohere Train		Cohere + GPT Train	
	ACC	F1	ACC	F1
Cohere_Test	0.9945	0.9945	0.9955	0.9955
GPT	0.5710	0.4196	0.9992	0.9992
Adversarial	0.5273	0.6378	0.6487	0.6136
Merged	0.3144	0.1412	0.3535	0.3113
Real-life	0.3526	0.1903	0.6818	0.7727
SNLI	0.5055	0.0296	0.5063	0.1014

Table 10: MiniLM performance across datasets when trained on Cohere vs. Cohere + GPT.

Dataset	Cohere Train		Cohere + GPT Train	
	ACC	F1	ACC	F1
Cohere_Test	0.9957	0.9957	0.9962	0.9962
GPT	0.4830	0.5373	1.0000	1.0000
Adversarial	0.6451	0.7072	0.7079	0.7542
Merged	0.3599	0.2546	0.3929	0.2551
Real-life	0.4491	0.4486	0.6762	0.7633
SNLI	0.5141	0.1340	0.5101	0.0731

Table 11: T5-small performance across datasets when trained on Cohere vs. Cohere + GPT.

Train	Test	DistilBERT	Electra	MiniLM
Accuracy				
Merged	Merged	0.9640	0.9558	0.9474
Merged	Real-life	0.6328	0.3204	0.5255
Merged	SNLI	0.5029	0.4937	0.5244
F1 Score				
Merged	Merged	0.9713	0.9645	0.9567
Merged	Real-life	0.7292	0.0632	0.5585
Merged	SNLI	0.1274	0.2051	0.5450

Table 12: Model performance on cross-domain evaluations using Merged-based training data.

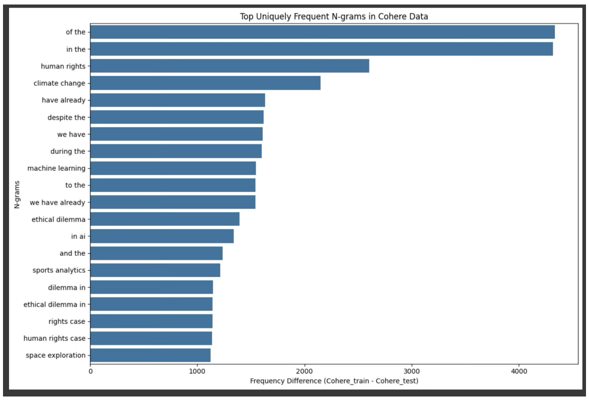

Figure 2: Top Frequent N-grams in Cohere Data

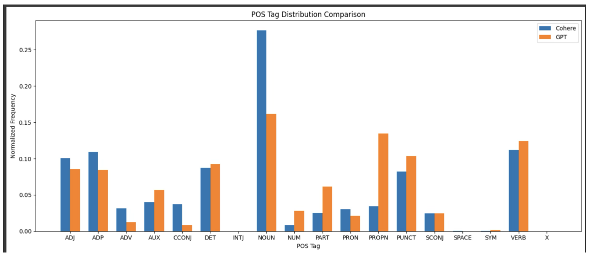

Figure 3: Part of Sentence Tag Distribution Comparison between Cohere and GPT

B References

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 440–447.

Samuel R 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 (EMNLP)*, pages 632–642.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Raffel Colin. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adam Fisch et al. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: Benchmark for Consistency Evaluation of Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3837–3852. International Committee on Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing (3rd ed. draft)*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *arXiv preprint arXiv:2503.14023*.

Farzan Erlik Nowruzi, Prince Kapoor, Dhanvin Kolhatkar, Fahed Al Hassanat, Robert Laganiere, and Julien Rebut. 2019. How much real data do we actually need: Analyzing object detection performance using synthetic and real data. *arXiv preprint arXiv:1907.07061*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

David M W Powers. 2020. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhilasha Ravichander, Nicholas Lourie, Matt Gardner, and Eduard Hovy. 2022. Probing the consistency of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3716–3732.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yash Saxena, Sarthak Chopra, and Arunendra Mani Tripathi. 2024. Evaluating consistency and reasoning capabilities of large language models. In *2024 Second International Conference on Data Science and Information System (ICDSIS)*, pages 1–5. IEEE.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, and Yao Qin. 2023. Effective robustness against natural distribution shifts for models with different training data. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*. https://shizhouxing.github.io/effective-robustness.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP (EMNLP Workshop)*, pages 638–644.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*, pages 1112–1122.

Da Zhou, Pengcheng He, Xueyun Yuan, and Weizhu Liu. 2021. Rethinking pre-training and self-training. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.