Improving OOD Robustness via Background-Aware Test-Time-Augmentation in Black-Box and Resource Constrained Settings

Anonymous authors
Paper under double-blind review

Abstract

Deep learning models for text classification typically achieve strong performance on in-distribution (ID) data but often fail to generalize to out-of-distribution (OOD) inputs. This degradation frequently arises because models rely on spurious background cues (e.g., specific syntax or register) learned during training, which become unreliable when the domain changes. While recent Test-Time Augmentation (TTA) approaches have enabled robustness in black-box settings, they often rely on unconstrained rewriting strategies. For instance, standard In-Context Rewriting (ICR) instructs Large Language Models (LLMs) to modify input details to match ID exemplars, creating a high risk of semantic drift and label flipping, particularly when using smaller, resource-constrained LLMs. In this work, we propose a Background-Aware TTA (BA-TTA) framework that strictly disentangles style from semantics. Unlike prior methods that encourage broad paraphrasing, we utilize a semantic-constrained alignment strategy that enables small, efficient LLMs to transform specific background attributes, such as tone and sentence structure, to match in-distribution priors while explicitly enforcing the preservation of original meaning. This approach mitigates OOD degradation by neutralizing spurious background shifts, allowing frozen black-box models to process inputs in their native distribution without risking semantic corruption. Empirical evaluations across multiple text classification benchmarks demonstrate that our targeted alignment strategy outperforms unconstrained augmentation baselines. By generating higher-fidelity augmentations, our method achieves superior OOD robustness with reduced computational overhead, establishing a viable path for deploying robust in resource-limited black-box environments.

1 Introduction

Deep learning models for text classification are often expected to perform reliably not only on in-distribution (ID) inputs drawn from the training distribution, but also on unseen out-of-distribution (OOD) data encountered in deployment Hupkes et al. (2023). Achieving such robustness is essential for building safe and trustworthy NLP systems Song et al. (2026), especially in application domains where errors can be costly, such as spam detection, toxicity monitoring, and healthcare diagnostics. However, OOD generalization in NLP remains a persistent challenge due to the inherently diverse and dynamic nature of language. Real-world data is subject to continuous distributional shifts, where changes in topics, tones, dialects, or writing styles can substantially alter the input distribution. For instance, a sentiment classifier trained on formal movie reviews may fail catastrophically when applied to casual tweets, even if the underlying sentiment is identical. When these background characteristics change at test time, predictions can become unreliable because the model often relies on spurious correlations Sagawa et al. (2020) (e.g., associating specific syntax with a label) rather than robust semantic features.

While a rich body of prior work aims to improve OOD robustness, most existing methods depend on white-box access to model parameters. Techniques such as fine-tuning Bommasani et al. (2021), gradient-based adversarial training, domain-invariant representation learning Nguyen et al. (2021) or uncertainty calibration

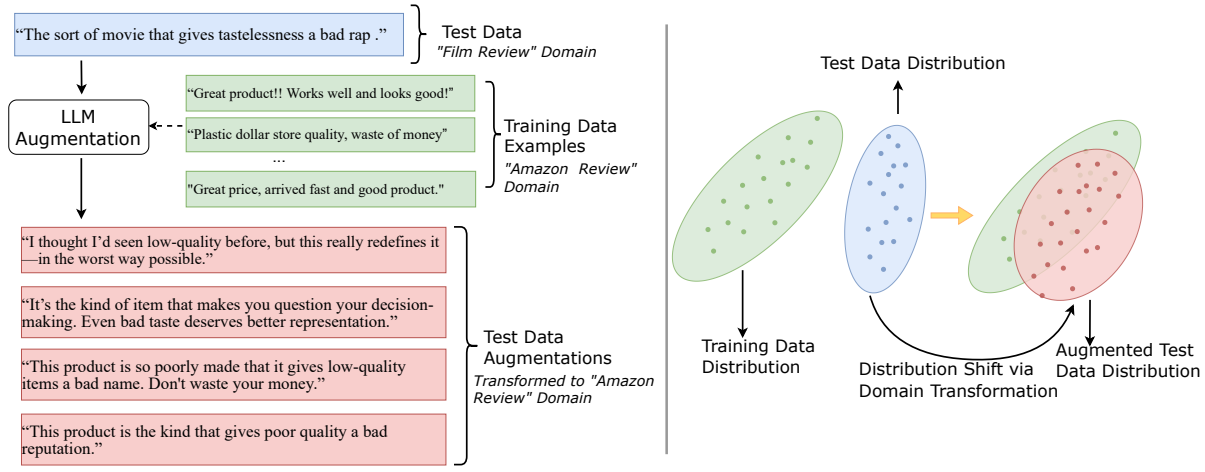

Figure 1: **Test-Time Augmentation with Domain Transformation:** LLM-based test-time augmentation via domain transformation. OOD input and ID examples are fed into LLM with background shift prompts. We show that this process can shift OOD distribution closer to ID distribution.

require modifying the model weights or accessing internal gradients Yuan et al. (2023a). These requirements are often unrealistic in many deployment scenarios. In many practical applications, models are deployed as immutable components on edge devices, or accessed via query-only interfaces (e.g., proprietary large language model APIs) where weights are hidden. Under these "black-box" constraints, parameter-centric robustness techniques are fundamentally inapplicable. Consequently, recent work has explored Test-Time Augmentation (TTA) as a practical alternative O'Brien et al. (2024); Cohen & Giryes (2024). By generating multiple variants of a test input and aggregating their predictions, TTA attempts to bridge the domain gap at inference time. However, conventional TTA methods relying on simple word- or phrase-level perturbations (e.g., synonym replacement) often fail to produce sufficiently rich transformations to correct complex stylistic shifts.

To address these limitations, recent approaches like LLM-TTA have leveraged Large Language Models (LLMs) to perform In-Context Rewriting (ICR). These methods prompt an LLM to rewrite OOD inputs to statistically resemble a set of in-distribution exemplars. While effective at masking distribution shifts, these methods typically rely on unconstrained rewriting strategies. For example, standard ICR prompts often instruct the LLM to "modify details if necessary" to match the training style. This introduces two critical limitations. First, unconstrained rewriting creates a high risk of semantic drift, where the LLM, prioritizing style alignment over content fidelity, accidentally alters the class label (e.g., hallucinating positive sentiment to match a positive ID exemplar). Second, to mitigate the noise introduced by these hallucinations, these methods must rely on the statistical power of large ensembles (aggregating predictions over many augmentations), which strains the computational budget in resource-constrained deployment scenarios.

A primary goal of test-time augmentation is to provide an immediate, "plug-and-play" solution for robust inference. However, for this to be viable in real-world applications, such as on-premise servers or latency-sensitive edge devices, the augmentation process itself must be computationally efficient. While Large Language Models provide the necessary flexibility for domain transformation, the shift toward resource-constrained deployment necessitates strategies that can function effectively with smaller, open-weights models (e.g., 7B parameters). We demonstrate that by introducing explicit semantic constraints, we can focus the model's limited capacity purely on background alignment. This allows a 7B-parameter model to produce high-fidelity, label-preserving transformations that traditionally would be expected only from much larger architectures.

In this work, we propose a BA-TTA framework (see Figure 1) grounded in the observation that domain shift can often be decomposed into two distinct axes: semantic shift (changes in meaning) and background shift (changes in style/syntax) Arora et al. (2021); Yuan et al. (2023a). We identify that OOD performance

		In-Distribution	Out-of-Distribution (OOD)		
Model	**Training Source**	**Amazon (Test)**	**SST-5**	**Dynasent**	**SemEval**
T5	Amazon	90.11%	76.12%	47.73%	50.07%
BERT	Amazon	90.38%	68.47%	42.71%	44.97%
Llama-2	Amazon	97.15%	74.16%	42.73%	39.21%

Table 1: Model performance on Sentiment Analysis, comparing In-Distribution (ID) baselines with Out-of-Distribution (OOD) robustness.

		In-Distribution	Out-of-Distribution (OOD)	
Model	**Training Source**	**Civil Comments (Test)**	**ToxiGen**	**Adv. Civil**
T5	Civil Comments	90.57%	65.78%	46.97%
BERT	Civil Comments	88.46%	66.74%	30.46%
Llama-2	Civil Comments	97.71%	65.57%	27.67%

Table 2: Model performance on Toxicity Detection, comparing In-Distribution (ID) baselines with Out-of-Distribution (OOD) robustness results.

degradation is frequently driven by models' reliance on spurious background cues, such as domain-specific phrasing or sentence length, hich become unreliable at test time even when the semantic signal remains unchanged. To mitigate this, we move beyond the "blind" rewriting of prior work and introduce a semantic-constrained alignment strategy. Instead of asking an LLM to broadly "paraphrase" the input, we prompt small, efficient LLMs to precisely align only the specific background characteristics (e.g., tone, register) of OOD inputs to the ID distribution, while strictly enforcing the preservation of semantic meaning.

We validate our approach across seven open-source datasets and one new synthetic benchmark, spanning sentiment analysis, toxicity detection, and topic classification. Our experiments demonstrate that by explicitly correcting background shift, we can achieve superior robustness compared to generic LLM-based rewriting. Crucially, our method demonstrates that strict semantic constraints allow smaller, less capable LLMs to perform effective augmentation, achieving state-of-the-art robustness even when using identical model sizes and inference budgets as baseline methods.

Our contributions are threefold:

1. **Decomposition of Domain Shift**: We empirically demonstrate that significant performance degradation stems from spurious background features (style, syntax) rather than semantic content, identifying "background alignment" as the optimal intervention for OOD robustness.

2. **Semantic-Constrained Rewriting**: We introduce a black-box augmentation framework that uses targeted constraints to rewrite OOD inputs. Unlike standard ICR which risks label flipping, our method strictly preserves task-relevant semantics, enabling the safe use of small LLMs.

3. **Practicality in Resource-Constrained Settings**: We demonstrate the effectiveness of our framework using a 7B-parameter open-weights model (Stable Beluga 2). Our results show that targeted background alignment is a highly efficient path to robustness, allowing practitioners to achieve significant OOD gains on consumer-grade hardware without the need for massive, proprietary APIs or high-latency ensembles.

2 Related Work

Test-Time Adaptation Test-time adaptation refers to techniques that adjust model parameters or intermediate representations on-the-fly during inference, specifically in response to OOD data Liang et al. (2025); Yuan et al. (2023b). Unlike traditional adaptation procedures, test-time adaptation does not assume the

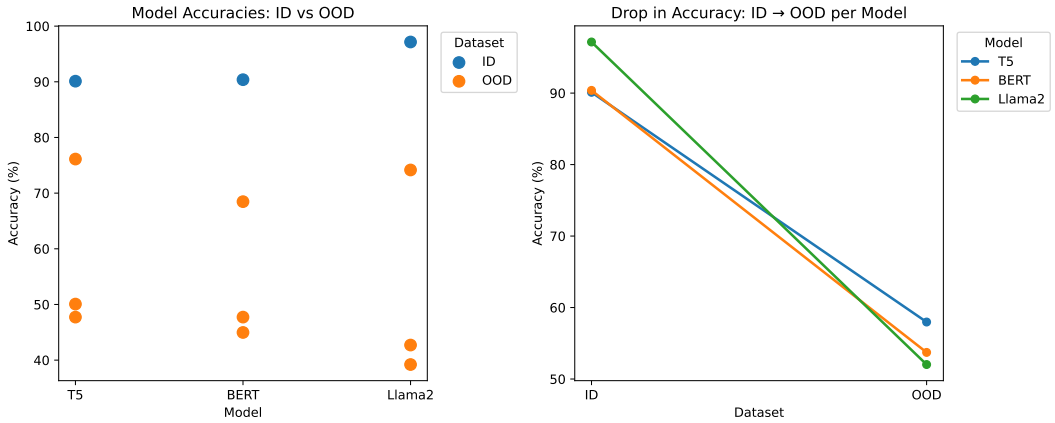

Figure 2: **ID → OOD Performance Gap:** While fine-tuned LLMs achieve high accuracy on ID data, they experience substantial drops when evaluated on OOD datasets, highlighting a critical generalization dilemma.

availability of labelled samples from the target distribution, and it often operates under stringent latency constraints.

Early works on test-time adaptation focus on adapting batch normalization statistics to align activations with the new distribution. For instance, Sun et al. (2020) proposed a "test-time training" method (TTT) in which a self-supervised auxiliary task is optimized at inference to refine feature representations. This method reduces the mismatch between training and test distributions by continuously updating selected parameters. Similarly, Wang et al. (2020) introduced TENT, which fine-tunes only the batch normalization parameters via entropy minimization on unlabelled test data. These approaches have demonstrated that lightweight adaptation can improve performance when encountering domain shift.

Subsequent research explores alternative adaptation strategies, such as feature alignment Wang et al. (2023); Jung et al. (2023) and domain-specific normalization layers Wu et al. (2024); Zhang et al. (2023). Recent advances include meta-learning schemes that prepare the model for rapid parameter updates at test time Liang & Chen (2025). While promising, these methods usually require careful hyperparameter tuning and may struggle to preserve the model's generalization capabilities when facing highly disparate OOD samples.

TTA extends the conventional notion of data augmentation, which is commonly employed at training phase, into the inference phase. By creating multiple augmented variants of an incoming sample and aggregating the predictions, TTA can mitigate the impact of distributional shifts on the final output Zhang et al. (2022); O'Brien et al. (2024). Typical augmentations involve random cropping, flipping, color jittering, or geometric transformations in vision tasks Shorten & Khoshgoftaar (2019), back translation Sennrich et al. (2015) or paraphrasing for text tasks Wieting et al. (2017), aiming to enforce consistency of model predictions across perturbed inputs.

One line of work integrates TTA with model uncertainty estimation, which is also called calibration Kumar et al. (2019), to weight predictions from each augmented view. This procedure offers robustness by smoothing decision boundaries around ambiguous regions in the input space. However, it can become computationally expensive, particularly in large-scale deployments, as multiple forward passes are required. Addressing efficiency concerns, methods such as adversarial test-time augmentation have emerged, balancing the diversity of augmentations with feasible inference time.

Despite performance gains, TTA alone may not be sufficient to fully handle severe domain shifts—especially if the augmentations do not closely mimic real-world variations. As a result, combining TTA with test-time adaptation schemes can further improve robustness by enabling the model to update its parameters while also smoothing predictions over various augmented examples.

No.	Sample
1	we should build a wall on the border to keep them from coming in illegally.
2	having a disability means you can get everything for free
3	they are taking all of our jobs
4	the return of people of color to the spaces where they were once enslaved
5	iran is not the same thing as isis

Table 3: Samples from ToxiGen dataset

No.	Sample
1	What a stuppid thing to say. TTake a course on trading. By your logic markets would stay at the same price in perpetuity. I can assure you that at this poinyt in the cycle it will take surprisingly little to go from a buyers market to a sellers market. When the sellers overpower buyers you'll see prices tank. Te fact that I have to even explani this elementary concept reminds me thaft there is no fix for stupi.d The former alignment of Route 29 in the northern part
2	I misspoke the way I should have said it" Political speech for, "I'm an idilt and I got caught in my lie to one part of the country by another part of the country." I was just happy to see Calgary booing our selfie seeking PM. Some where in Quebec there is a village that ks missing it's idiiot and we need to get him back tto them before he completely ruins this country.
3	Anoter cause of the sewage spills rather than "human error" could be better attributed to stuppidity, Incompetence, negligence. I suppose it is kind of late in the game since Capt Kirk well into his second and llast term as Mayor but maybe not too late to fire the environmental services director who seems to be a prime example oof the Peter principle, you know rising to the level of your incompetence. A further example of Starr 's allegedly anti

Table 4: Samples from Adv.Civil dataset

Augmentation for Text Text-based models also face OOD challenges, including shifts in vocabulary, style, or topic. Text augmentation at training time, such as synonym replacement, back-translation, or paraphrasing Sennrich et al. (2015); Wieting et al. (2017), is widely used to improve model generalization. However, fewer works explore augmenting text samples directly at test time.

Existing text augmentation strategies often rely on transformations that preserve semantic meaning but vary the lexical or syntactic form Shorten et al. (2021). For instance, synonyms or short paraphrases can be generated using language models O'Brien et al. (2024), then passed through the model to produce an ensemble of predictions. In the presence of domain shift (e.g., changes in terminologies or introduction of new expressions), these augmentations can expose the model to equivalent meanings expressed in different forms, offering additional robustness.

Effectively applying text augmentations at test time poses unique challenges. Minimal changes in word choice might not suffice if the domain shift significantly alters context or linguistic style. Furthermore, real-time text augmentation can introduce latency overhead and risk generating unnatural phrases that degrade model confidence. Recent studies attempt to mitigate these issues by controlling augmentation quality and relevance through learned policies. Overall, test-time text augmentation remains an area for further exploration, especially in conjunction with adaptive methods designed for OOD scenarios.

LLMs as Augmentation Engines Recent advances in large language models (LLMs) have opened new opportunities for test-time adaptation. By conditioning on natural language prompts, LLMs can perform high-quality paraphrasing, style transfer, and contextual rewriting without labeled supervision. These capabilities allow LLMs to serve as powerful black-box augmentation engines, capable of reshaping test-time inputs to better resemble training data. However, existing applications of LLM-based rewriting typically

LLM-TTA: In-Context Rewriting (ICR)

```
### Instructions ###
The assistant is to paraphrase the input text as if it was one
of the examples. Change the details of the text if necessary.

### Style Examples ###
<style_transfer_exemplars>

### Input Text ###
Now paraphrase ```"{input_sentence}"``` as if it was one
of the examples. Change the details of the text if necessary.

Return the text in the format: ```Paraphrased Text```

### Paraphrased Text ###
Paraphrased Text:
```

BA-TTA: Domain Transformation

```
### Instructions ###
Rewrite the input sentence into the <ID domain>, adjusting only background
features such as writing structure, sentence flow, and discourse style.

CRITICALLY:
- Preserve all toxic, abusive, insulting, or hateful expressions EXACTLY.
- Do NOT soften, weaken, sanitize, or remove any toxic or profane words.
- Do NOT replace slurs or insults with milder alternatives.
- Do NOT change the target of the toxicity.
- Do NOT add meta-language about toxicity (e.g., "this comment is offensive").
- Only modify non-semantic stylistic aspects (punctuation, connective words,
flow, formatting).

###Example Sentences and their domain:###
{examples}

###Input Sentence:
{input_sentence};

### Trnasformed Text ###
Rewrite the sentence in <ID domain> while keeping all toxic expressions
verbatim and maintaining the same level of toxicity. Only shift background
writing style.
```

Figure 3: **LLM-TTA Prompt Templates.** Left: In-Context Rewriting (ICR) prompt, which instructs the model to paraphrase the input sentence using provided style exemplars and format the output as "Paraphrased Text". Right: Domain Transformation prompt, which directs the model to convert an input sentence into an equivalent expression in a specified target domain, matching style, tone, vocabulary, and structure, while preserving its original meaning.

treat it as a generic paraphrasing tool, without a principled framework for targeting the underlying structure of distributional shifts.

Decomposing Distribution Shifts in Text: Background vs. Semantic Features Emerging evidence suggests that distributional shifts in NLP are governed by two distinct and often orthogonal components Arora et al. (2021):

• Background Shifts: These involve changes in surface-level or contextual features, such as domain (e.g., news vs. social media), style (formal vs. informal), register, or vocabulary. Such shifts often preserve the core semantic content but alter the distributional characteristics that models rely on during training.

• Semantic Shifts: These occur when the task-relevant meaning of inputs changes, such as the introduction of new classes in classification tasks (e.g., previously unseen intents or topics). These shifts are more fundamental, as they affect the decision boundary learned by the model.

3 Methodology

3.1 Preliminary & Problem Definition

Consider a text classification task where each example consists of an input $x \in X$ and an output label $y \in Y$. The input space $X \subset \mathbb{R}^n$ and output space Y contain K classes. We assume access to a training dataset D_{train} consisting of pairs (x, y) sampled from the training data distribution $p_{\text{train}}(x, y)$. A model f is trained on this dataset to learn a mapping $f : X \to Y$ that generalizes to unseen data.

At test time, however, the model may encounter an input $x' \in X$ drawn from an unknown distribution $p_{\text{OOD}}(x, y)$, which may differ from the original training distribution. The goal is to ensure that the model remains robust and performs well on such out-of-distribution (OOD) data despite potential distributional shifts.

Decomposition of Distribution Shifts in Text. As proposed in prior work Arora et al. (2021), any representation of an input x can be decomposed into two independent and disjoint components: background features: $\phi_b(x) \in \mathbb{R}^m$, semantic features: $\phi_s(x) \in \mathbb{R}^n$.

The overall probability distribution of x can be expressed as $p(x) = p(\phi_s(x))p(\phi_b(x))$. Ideally, background features $\phi_b(x)$ should be independent of the label, while semantic features $\phi_s(x)$ should be label-dependent. Formally, for any class label $y \in Y$: $p(\phi_b(x)|y) = p(\phi_b(x))$, $p(\phi_b(x)|y) \neq p(\phi_s(x))$. This distinction allows us to categorize textual distribution shifts into two major types:

- Background Shift: Occurs when the domain, style, or context of text changes, even if the underlying semantics remain the same. For example, transitioning from Amazon product reviews to film reviews.
- Semantic Shift: Occurs when the meaning or class distribution of text changes, such as the emergence of new or unseen classes during inference.

Our proposed approach aims to mitigate the impact of background shifts while preserving semantic integrity, thereby enhancing OOD robustness.

3.2 BA-TTA: Robustness through Test-Time Domain Transformation

Transformation & Augmentation To bridge the distribution gap between OOD and ID data, we define an augmentation function a_z, which transforms an input x into an augmented version x'. Our objective is to decouple the distribution shift into two components:

- Background Alignment: Shift background features to match the ID distribution: $p(\phi_b(x')) \approx p_{ID}(\phi_b(x))$.
- Semantic Preservation: Strictly maintain the semantic features of the original input: $p(\phi_s(x')) \approx p(\phi_s(x))$.

To implement this separation, we employ a semantic-constrained prompting strategy (see Figure 3, Right). Unlike standard In-Context Rewriting (ICR), which approximates the target distribution by broadly mimicking exemplars and instructing the model to "change details if necessary", our prompts explicitly instruct the model to transform specific background attributes (e.g., tone, sentence structure, vocabularies) while strictly enforcing the preservation of equivalent semantic meaning.

This explicit decomposition serves two critical functions. First, it prevents semantic drift, ensuring that the augmentation process does not accidentally alter the class label. Second, by simplifying the instruction from "open-ended rewriting" to "targeted alignment," we enable the effective use of smaller, resource-efficient LLMs to parametrize a_z. This allows us to generate high-fidelity augmentations without requiring the massive parameter counts typically needed for safe unconstrained rewriting.

Inference and Aggregation At inference time, multiple augmented versions of an OOD input x' are generated using the augmentation functions. These augmented inputs are then passed through the model f, and their predictions are aggregated to obtain a more robust final prediction.

Let $A(x)$ represent the set of augmented samples generated from x. The final prediction $\hat{y}$ is computed as $\hat{y} = \arg\max_y \sum_{x' \in A(x)} p(y|f(x'))$.

This aggregation strategy helps smooth out distributional variances, ensuring that the model predictions remain stable and resilient against OOD variations.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our proposed test-time augmentation (TTA) framework, we utilize datasets across three text classification tasks: sentiment analysis, toxicity detection, and news topic classification.

These tasks reflect real-world scenarios where models are required to generalize beyond their training distribution. Our selection of in-distribution (ID) and out-of-distribution (OOD) datasets is designed to capture both background shifts (changes in contextual framing) and semantic shifts (variations in meaning), which are the two primary factors contributing to distribution gaps. Please see Table 3 and Table 4 samples from OOD datasets.

Sentiment Classification For sentiment classification, we use a benchmark that includes a three-way classification task with labels: positive, neutral, and negative. The ID dataset consists of Amazon reviews McAuley & Leskovec (2013), representing a consumer-driven review domain. The OOD datasets are DynaSent Potts et al. (2020), SST-5 Socher et al. (2013), and SemEval Nakov et al. (2019); Yuan et al. (2023a), introducing variations in linguistic style, domain, and sentiment annotation schemes. These shifts capture both background changes (e.g., different sources of user-generated content) and potential semantic variations (e.g., differences in sentiment annotation granularity).

Toxicity Detection The toxicity detection task is framed as a binary classification problem, distinguishing between toxic (positive) and non-toxic (negative) language. The ID dataset is Civil Comments Borkan et al. (2019), a collection of user comments from a moderated online platform. The OOD datasets include an adversarially augmented version of Civil Comments - AdvCivil Borkan et al. (2019), as well as two additional datasets. Here, the OOD evaluation specifically targets adversarial shifts and diverse online environments, challenging the model's ability to decouple toxic intent from benign community slang.

News Topic Classification For the topic classification task, we focus on a four-way categorization problem distinguishing between World, Sports, Business, and Science/Technology topics. The ID dataset is AG News, representing a standard distribution of formal journalistic reporting. The OOD dataset is Twitter-Topic. These benchmarks introduce significant background shifts, moving from structured news articles to informal social media posts, effectively testing the model's resilience to changes in register and length while the underlying topical semantics remain consistent.

4.2 Baseline Methods

Word-Level Augmentations As a representative of heuristic-based TTA, we include word-level perturbation methods widely used in NLP robustness research. Following the implementation by Lu et al. (2022), we utilize the nlpaug library to perform stochastic word insertion and substitution. Specifically, each word in the input text has a 30% probability of being augmented, up to a maximum of 10 words per input. Substitutions are generated using BERT-based contextual prediction to maintain local fluency. However, unlike our method, these perturbations are local and structure-agnostic; they introduce noise to smooth predictions but fail to systematically address the broader stylistic or domain shifts (e.g., syntax, register) that characterize true distribution mismatch.

Back-Translation We also compare against Back-Translation, a standard "whole-text" augmentation strategy often used for paraphrasing. We employ an English $\leftrightarrow$ German translation loop, where the input is translated to German and then back to English to generate a paraphrase. While this approach captures some global semantic structure, it functions as a blind paraphraser: it alters the text without any explicit guidance on which background features (e.g., tone, formality) should be aligned to the ID distribution. Consequently, it may either fail to correct the domain shift or accidentally introduce semantic errors during the translation process.

Standard In-Context Rewriting (ICR) We select the LLM-based TTA method proposed by O'Brien et al. (2024) as our primary baseline, as it represents the current state-of-the-art for black-box robustness. This approach utilizes In-Context Learning to adapt OOD inputs: the LLM is provided with 16 randomly selected in-distribution (ID) exemplars and instructed to rewrite the test input to resemble them.

Crucially, this baseline employs an unconstrained prompting strategy. As noted in their implementation, the model is instructed to "change the details of the text if necessary" to achieve stylistic alignment. While this generates diverse augmentations that capture the broad "vibe" of the ID data, it prioritizes stylistic mimicry over semantic fidelity. We follow their experimental setup by generating N stochastic rewrites per input

and aggregating predictions, providing a direct comparison between their diversity-driven rewriting and our semantic-constrained alignment.

4.3 Implementation Details

Task Models Task Models We investigate black-box robustness across diverse model architectures to ensure our findings generalize beyond a single paradigm. We select representative models for the three primary architectural families: Encoder-Only: BERT Devlin et al. (2019), using the fine-tuned checkpoints from O'Brien et al. (2024). Encoder-Decoder: T5-Large Raffel et al. (2020) , a generative model optimized for sequence-to-sequence tasks. Decoder-Only: Llama-2 (Touvron et al., 2023), representing the family of causal large language models. This diverse selection allows us to verify that our Background-Aware TTA is agnostic to the underlying classifier architecture. Table 1, Table 2 and Figure 2 shows the ID-OOD performance gap.

LLM-based Domain Transformation To demonstrate the effectiveness of our framework in resource-constrained settings, we utilize Stable Beluga 2-7B (SB2) (Mahan et al., 2023) as our primary augmentation backbone. SB2 is a fine-tuned version of Llama-2-7B, optimized for instruction following. Unlike prior works that rely on massive proprietary models (e.g., GPT-4), we show that a small-scale, open-weights model is sufficient for high-fidelity augmentation when guided by our semantic-constrained prompts (Figure 4).

For each test input, we generate $N = 4$ stochastic augmentations using SB2. We provide 16 randomly selected ID exemplars in the context window to guide the background alignment. The use of a 7B-parameter model significantly reduces the computational overhead compared to 70B+ or API-based baselines, making our approach viable for local deployment.

Inference and Aggregation Strategy To mitigate the variance inherent in stochastic generation, we adopt an ensemble aggregation strategy ($N_{total} = 5$, including the original input):

For Discriminative Models (BERT): We average the calibrated softmax probability distributions across all five inputs. The final prediction is determined by the class with the highest average probability: $\hat{y} = \arg\max_c \frac{1}{N} \sum_{i=1}^{N} P(y_c|x_i)$. This leverages the model's confidence scores to smooth out noise.

For Generative Models (T5 & Llama-2): Since these models output discrete text labels, we employ Majority Voting. The final class label is determined by the most frequent prediction among the augmented set. This consensus mechanism ensures that the final decision is robust to outliers or hallucinations in any single augmentation.

4.4 Results

The out-of-distribution (OOD) robustness performance for three sentiment benchmarks (SST-5, Dynasent, SemEval), three toxicity detection benchmarks (ToxiGen, Adv. Civil, ImplicitHate), and a news topic dataset (Tweets) is reported across fine-tuned T5, BERT, and Llama-2 models. Performance is evaluated under three primary regimes: no augmentation, standard augmentation baselines (insert, substitute, back-translation), and two LLM-based test-time augmentation (TTA) approaches: in-context rewriting and our proposed domain transformation approach.

Across almost every task model and OOD dataset, the domain transformation approach yields the best robustness. On average, this method achieves the highest accuracy for T5, BERT, and Llama-2 across most task categories. Specifically, it reaches an averaged performance of 59.95% in sentiment, 68.59% in toxicity, and 91.11% in news topic classification for T5, outperforming all other baselines. For BERT, it achieves 66.08% in toxicity and 90.33% for news topic classification, and for Llama-2, it reaches 63.54% in toxicity and 91.53% in news topic classification (refer to Table 5). At the individual dataset level, domain transformation outperform baseline methods nearly on all benchmark datasets (see Table 6).

The substantial improvements over the non-augmentation baseline demonstrate that task models, when unassisted, can struggle significantly under distribution shifts. By aligning background representations while maintaining semantic integrity, our framework effectively transforms test inputs into the ID domain for more robust prediction. In sentiment analysis, this approach delivers notable gains on challenging OOD datasets. For instance, on a fine-tuned BERT task model, our method outperforms the "none" augmentation approach

	Sentiment			Toxicity			News → Tweets		
Augmentation	T5	BERT	Llama	T5	BERT	Llama	T5	BERT	Llama
None	57.97%	52.05%	52.04%	58.89%	53.91%	53.65%	89.01%	88.57%	86.42%
Word insertLu et al. (2022)	56.69%	51.38%	48.28%	57.62%	53.36%	51.85%	89.96%	88.87%	91.36%
Word substituteLu et al. (2022)	54.57%	49.93%	44.25%	57.00%	52.96%	50.37%	89.18%	88.51%	90.43%
Back translationSennrich et al. (2016)	54.16%	50.05%	48.45%	57.70%	57.13%	55.80%	88.41%	88.41%	90.37%
In-context rewriting	59.66%	**56.20%**	50.50%	63.76%	60.84%	54.88%	90.38%	89.51%	90.05%
Domain transformation	**59.95%**	55.94%	**52.20%**	**68.59%**	**66.08%**	**63.54%**	**91.11%**	**90.33%**	**91.53%**

Table 5: **OOD TTA Performance Summary.** Results are averaged across the three OOD shifts for Sentiment and Toxicity for each model architecture.

	Sentiment			**Toxicity**			**News**
Augmentation	**SST-5**	**Dynasent**	**SemEval**	**ToxiGen**	**Adv. Civil**	**ImplicitHate**	**Tweets**
None	72.92%	44.40%	44.75%	66.00%	35.07%	65.38%	88.00%
Word insertLu et al. (2022)	69.93%	42.65%	43.77%	66.70%	30.78%	65.35%	90.06%
Word substituteLu et al. (2022)	64.40%	41.65%	42.70%	64.19%	30.87%	65.27%	89.37%
Back translationSennrich et al. (2016)	67.10%	41.91%	43.65%	64.76%	41.34%	64.53%	89.06%
In-context rewriting	72.82%	**47.71%**	45.83%	63.35%	49.88%	66.26%	89.98%
Domain transformation	**74.75%**	46.67%	**46.67%**	**66.63%**	**65.94%**	**65.63%**	**90.99%**

Table 6: **Detailed OOD Performance per Dataset.** We compare various augmentation methods across individual datasets for the primary task model.

by +5.22 percentage points and surpasses the in-context rewriting baseline by +1.68 percentage points on the SST-5 OOD dataset (see Table 7).

The divergent impact of test-time domain transformation is particularly striking in toxicity benchmarks. On the Adv. Civil dataset, which features long, context-rich comments drawn from real-world discussions, our method yields substantial robustness gains comparing with ICR baseline method, such as over +14.7 percentage points with T5, +14.4 percentage points with BERT, and +19.06 percentage points with Llama-2 (see Table 7). This indicates that augmenting nuanced background features helps the model generalize to complex, paragraph-level toxicity. By contrast, on ToxiGen, where examples are deliberately concise probes of implicit bias, the same transformation leads to performance degradation. We attribute this drop to the lack of broader contextual cues in ToxiGen, as short, targeted utterances offer little auxiliary information for the domain transformation to exploit. Overall, BA-TTA consistently outperforms both the unaugmented setting and in-context rewriting across the majority of OOD evaluations (Figure 5).

5 Limitations

While our proposed test-time domain transformation framework demonstrates promising improvements in OOD robustness, it is not without limitations. First, the reliance on LLMs for in-context augmentation introduces addition computational and financial costs, particularly when processing large-scale datasets or operating in real-time applications. Second, as observed in our toxicity experiments, domain transformation yields substantial gains on Adv. Civil, where inputs are longer, context-rich comments, but degrades performance on ToxiGen's concise, sentence-level probes. This suggests that when target samples lack broader contextual cues, the transformation may introduce noise rather than informative variation, limiting its applicability to tasks or datasets composed of very short, self-contained utterances. Additionally, the augmented representations may not always align perfectly with the intended in-distribution (ID) characteristics, especially for highly divergent domains, which could limit the generalisability of the approach. Finally, while our experiments focused on sentiment analysis and toxicity detection, further evaluation across a broader set of NLP tasks is needed to validate the robustness and scalability of the framework.

Task Model	Augmentation Method	Sentiment Analysis			Toxicity Detection			News Topic	Summary
		SST-5	Dynasent	SemEval	ToxiGen	Adv. Civil	ImplicitHate	Tweets	Avg.*
T5	Without augmentation	76.12%	47.73%	50.07%	65.78%	46.97%	63.93%	89.01%	62.80%
	Word insert	74.35%	46.23%	49.49%	66.52%	42.11%	64.23%	89.96%	61.84%
	Word substitute	70.34%	44.72%	48.65%	64.51%	41.75%	64.73%	89.18%	60.55%
	Back-Translation	69.12%	45.44%	47.93%	65.57%	43.93%	63.61%	88.41%	60.57%
	In-Context Rewriting	76.12%	**52.11%**	50.74%	64.72%	60.07%	**66.50%**	90.38%	65.81%
	Domain Transformation (Ours)	**77.99%**	50.55%	**51.30%**	**66.63%**	**74.76%**	64.37%	**91.11%**	**68.10%**
BERT	Without augmentation	68.47%	42.71%	44.97%	66.74%	30.46%	64.53%	88.57%	58.06%
	Word insert	68.47%	41.50%	44.18%	**68.11%**	27.06%	64.92%	88.87%	57.59%
	Word substitute	64.93%	41.62%	43.25%	65.36%	28.16%	65.36%	88.51%	56.74%
	Back-Translation	66.60%	39.81%	43.73%	64.51%	43.20%	63.67%	88.41%	58.56%
	In-Context Rewriting	72.01%	**48.63%**	47.96%	63.24%	52.79%	**66.49%**	89.51%	62.95%
	Domain Transformation (Ours)	**73.69%**	44.91%	**49.21%**	66.42%	**67.23%**	64.59%	**90.33%**	**65.20%**
Llama-2	Without augmentation	**74.16%**	42.75%	39.20%	65.47%	27.79%	67.69%	86.42%	57.64%
	Word insert	66.98%	40.21%	37.64%	65.47%	23.18%	66.89%	91.36%	55.96%
	Word substitute	57.93%	38.61%	36.20%	62.71%	22.69%	65.72%	90.43%	53.47%
	Back-Translation	65.58%	40.49%	39.29%	64.20%	36.89%	66.31%	90.37%	57.59%
	In-Context Rewriting	70.34%	42.38%	38.79%	62.08%	36.77%	65.80%	90.05%	58.03%
	Domain Transformation (Ours)	72.57%	**44.55%**	**39.49%**	**66.84%**	**55.83%**	**67.94%**	**91.53%**	**62.68%**

Table 7: **Out-of-distribution robustness performance** across various task models. We compare our **Background-Aware Domain Transformation** with several baselines. Bold values indicate the best performance for each model per dataset. (*Avg. is calculated across the seven listed datasets.)

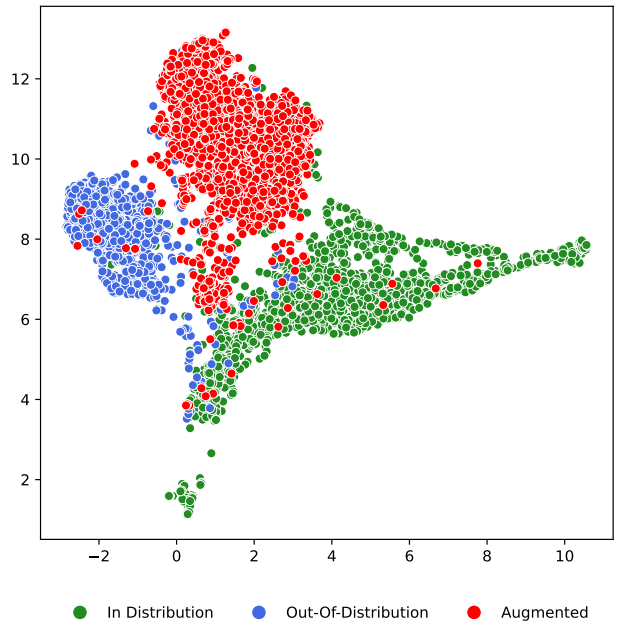

Figure 4: **2D UMAP Projection of Text Embeddings:** visualization of distribution shift and test-time augmentation. Green points represent in-distribution samples, blue points are out-of-distribution samples, and red points are the augmented examples generated at test time. Augmented samples bridge the gap between in- and out-of-distribution regions, illustrating how TTA expands the model's support to improve robustness.

6 Conclusions

In this paper, we introduced BA-TTA, a novel test-time domain transformation framework designed to enhance the robustness of NLP tasks under out-of-distribution (OOD) conditions. By leveraging Large Language Models (LLMs) for background-aware domain transformation, our approach generates augmented representations of test inputs whose background information is aligned with in-distribution (ID) data. Our empirical results across sentiment analysis, toxicity detection, and news topic classification benchmarks

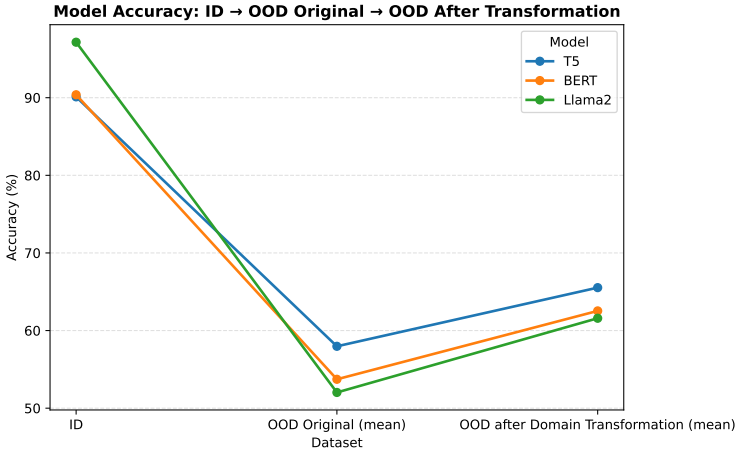

Figure 5: **OOD Performance Increase:** Comparison of T5, BERT and Llama 2 accuracies on in-distribution (ID) data, the mean out-of-distribution (OOD) splits before domain transformation, and the mean OOD splits after applying domain transformation. All three models experience a substantial performance drop when evaluated on the original OOD data, with T5 falling from 90.1% to 57.97%, BERT from 90.4% to 52.05%, and Llama 2 from 97.2% to 52.04%.

demonstrate that this transformation yields superior robustness, outperforming baseline methods across almost every OOD dataset. Specifically, the proposed method achieved the highest averaged performance for T5 (68.10%), BERT (65.20%), and Llama-2 (62.68%). The impact was particularly striking on context-rich benchmarks like Adv. Civil, where our method yielded gains of over +27.8 percentage points with T5, +36.8 percentage points with BERT, and +28.0 percentage points with Llama-2 (see Table 7). These results suggest that aligning background representations significantly improves classification accuracy without requiring modifications to model weights.

Future work may explore more sophisticated aggregation methods, the interplay between multiple augmentation techniques, and a deeper analysis of which components of the domain transformation contribute most to the observed robustness. Extending this framework to additional NLP tasks, as well as investigating its potential in combination with adversarial training methods, could further broaden its applicability and effectiveness.

References

Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Gilad Cohen and Raja Giryes. Simple post-training robustness using test time augmentations and random forest. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3996–4006, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American*

chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186, 2019.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023.

Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, and Jaegul Choo. Cafa: Class-aware feature alignment for test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19060–19071, 2023.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Yunsheng Liang and Kai Chen. Automatic test-time adaptation for heterogeneous contexts in meta-learning. *Neural Computing and Applications*, pp. 1–22, 2025.

Helen Lu, Divya Shanmugam, Harini Suresh, and John Guttag. Improved text classification via test-time augmentation. *arXiv preprint arXiv:2206.13607*, 2022.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172, 2013.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.01973*, 2019.

A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems*, 34:5264–5275, 2021.

Kyle O'Brien, Nathan Ng, Isha Puri, Jorge Mendez, Hamid Palangi, Yoon Kim, Marzyeh Ghassemi, and Thomas Hartvigsen. Improving black-box robustness with in-context rewriting. *arXiv preprint arXiv:2402.08225*, 2024.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. Dynasent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*, 2020.

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

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 86–96, 2016.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Ping Song, Adegboyega Ojo, and Edward Curry. Trustworthy requirements for foundation models—a comprehensive survey and roadmap. *Engineering Applications of Artificial Intelligence*, 163:113111, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20050–20060, 2023.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*, 2017.

Yanan Wu, Zhixiang Chi, Yang Wang, Konstantinos N Plataniotis, and Songhe Feng. Test-time domain adaptation by learning domain-aware batch normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15961–15969, 2024.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023a.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023b.

Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18971–18981, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

A Appendix

You may include other additional sections here.