ABEX: Data Augmentation for Low-Resource NLU via Expanding Abstract Descriptions

Sreyan Ghosh*, **Utkarsh Tyagi***, **Sonal Kumar**, **Chandra Kiran Reddy Evuru**, **S Ramaneswaran**, **S Sakshi**, **Dinesh Manocha**
University of Maryland, College Park, USA
{sreyang,utkarsht,sonalkum,ckevuru,ramans,fsakshi,dmanocha}@umd.edu

Abstract

We present **ABEX**, a novel and effective generative data augmentation methodology for low-resource Natural Language Understanding (NLU) tasks. ABEX is based on **AB**stract-and-**EX**pand, a novel paradigm for generating diverse forms of an input document – we first convert a document into its concise, abstract description and then generate new documents based on expanding the resultant abstraction. To learn the task of expanding abstract descriptions, we first train BART on a large-scale synthetic dataset with abstract-document pairs. Next, to generate abstract descriptions for a document, we propose a simple, controllable, and training-free method based on editing AMR graphs. ABEX brings the best of both worlds: by expanding from abstract representations, it preserves the original semantic properties of the documents, like style and meaning, thereby maintaining alignment with the original label and data distribution. At the same time, the fundamental process of elaborating on abstract descriptions facilitates diverse generations. We demonstrate the effectiveness of ABEX on 4 NLU tasks spanning 12 datasets and 4 low-resource settings. ABEX outperforms all our baselines qualitatively with improvements of 0.04% - 38.8%. Qualitatively, ABEX outperforms all prior methods from literature in terms of context and length diversity [1].

Method	**Original 1:** Usually, the two of us don't agree on anything about politics. **Original 2:** The pop superstar said she was "completely inspired" by Roem's victory.
EDA (Wei and Zou)	1. The two of us dont on about politics 2. Bulge the pop superstar said she was completely inspired by roems victory
AEDA (Karimi et al.)	1. Usually, the two of us ? don't agree ; on anything ! about : politics. 2. The pop superstar ; said she was ""completely inspired"" by Roem's victory.
SSMBA (Ng et al.)	1. Usually, the two of us don't agree about anything involving politics. 2. The pop superstar said she felt was completely inspired "" by roemś victory!
AMR-DA (Shou et al.)	1. We usually don't agree on anything. 2. Pop superstars say that a complete victory for Roem and superstars will inspire them .
GENIUS (Guo et al.)	1. It about politics. It about everything. 2. The pop superstar. The singer. The songwriter.
LLaMA-2_{13B} (Touvron et al.)	1. Political disagreement is the norm between the two of us. 2. The pop star also noted that Roem's triumph had inspired her own creative process.
ZeroGen (Ye et al.)	1. The two of us may disagree on anything, but we do not agree on it. point at hand. 2. The pop icon expressed being tremendously inspired by Roem.
ABEX *(ours)*	1. President Obama has failed to reach an agreement on any political issues, including the Iran nuclear deal, and there is no consensus on the next steps. 2. Cristiano Ronaldo is inspired by Roem's victory over Manchester United, according to the Portuguese superstar.

Table 1: Comparison of augmentations generated using ABEX and our baselines on a *randomly chosen* document from HuffPost. (**1.** Politics, **2.** Entertainment). ABEX moves beyond simple text-editing or rephrasing and generates diverse augmentations by introducing a new context. Augmentations by ABEX are also more coherent and label-consistent.

1 Introduction

Improving the performance of deep learning models on downstream Natural Language Understanding (NLU) tasks requires sufficient good-quality training data. However, data annotation is an expensive, time-consuming, and noisy task (Abad and Moschitti, 2016). Data augmentation has proven to be an effective approach for overcoming the data scarcity issue in low-resource NLU tasks with limited training samples (Chen et al., 2023). The two major categories of study in data augmentation include online data augmentation by interpolation in the latent space (Guo et al., 2019; Ng et al., 2020a; Sun et al., 2020; Kumar et al., 2020; Guo, 2020; Sawhney et al., 2021) and offline data augmentation that expands an existing small-scale dataset by generating additional synthetic data (Wei and Zou, 2019; Kumar et al., 2020; Zhou et al., 2021; Kim et al., 2022; Guo et al., 2022). Owing to advancements in generative models that facilitate the creation of high-quality synthetic data, the latter is gaining traction (Yu et al., 2023).

However, generative data augmentation faces

[1] Code and data: https://github.com/Sreyan88/ABEX
* Equal Technical Contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 726–748
August 11-16, 2024

two major challenges: *diversity* in generated augmentations (Geiping et al., 2023) and *consistency* with the underlying data distribution (Chen et al., 2023). It is crucial to strike a balance between these two aspects, as overemphasizing one at the expense of the other can lead to poor downstream performance. Current augmentation methods based on text-infilling (Ghosh et al., 2023c; Guo et al., 2022; Wang et al., 2022), where the primary task is to generate a new sentence constrained with keywords, are prone to replicate biases and overfit specific linguistic patterns in the low-resource training data, thereby hurting diversity. Additionally, we show that keyword-constrained free-form generation is unable to maintain the core semantic properties of the document, like style, which proves to be critical for specific tasks (e.g., *question* style document for intent classification. See example in Table 3). Diversity also proves to be an issue with token-level editing methods (Wei and Zou, 2019; Shou et al., 2022) that rarely introduce novel entities or contexts and often randomly edits important tokens. Finally, prompt-based methods that employ Large Language Models (LLMs) require well-curated attributes selected from the data to control the distribution of the generated data (Yoo et al., 2021; Sahu et al., 2023; Yu et al., 2023).

Main Contributions. In this paper, we propose **ABEX**, a novel data augmentation methodology based on a novel paradigm - Abstract-and-Expand. We first convert an input document into a concise, abstract description of itself and then generate augmentations by expanding the resultant abstraction. The task emulates human language perception and processing: the abstraction phase mirrors how humans distill core ideas from text, focusing on essential meanings, while the expansion phase reflects human creativity in generating varied narratives from a single abstract concept, akin to human extrapolation of ideas into diverse discussions. Our proposed Abstract-and-Expand task, which differs from all tasks proposed in prior art, generates augmentations that are both more consistent and diverse. To learn the task of expanding abstract descriptions, we first synthesize a large-scale synthetic dataset by prompting LLMs and then train an Encoder-Decoder Pre-trained Language Model (BART (Lewis et al., 2019)) on the dataset. Next, we propose a simple and controllable algorithm to generate abstract descriptions for training instances in any given downstream low-resource dataset. Our proposed algorithm leverages AMR-to-Text and Text-to-AMR and generates abstract descriptions by editing Abstract Meaning Representation (AMR) graphs (Banarescu et al., 2013). Inspired by the success of mixup in data augmentation (Zhang et al., 2018), we also optionally mix AMR graphs of two sentences to boost the diversity of abstract descriptions. Finally, we synthesize diverse augmentations using the fine-tuned model and synthesized abstract descriptions. To summarize, our main contributions are:

1. We propose ABEX, a novel and effective generative data augmentation methodology for low-resource NLP. We employ a novel Abstract-and-Expand task and fine-tune an Enc-Dec PLM to learn the task. ABEX differs from all prior work in its motivation and methodology and closely mimics the human perception and processing of language.

2. We propose a simple, controllable, and training-free method for generating abstract descriptions of source documents from downstream NLU datasets. Our proposed methodology provides explicit control in the document-to-abstract generation process and overcomes the contained generation issue that LLMs face in abstract generation.

3. To evaluate the efficacy of ABEX augmentations, we experiment on 12 datasets across 4 NLU tasks under 4 low-resource settings and show that ABEX outperforms most prior works quantitatively by 0.04% - 38.8%. Additionally, generations by ABEX are superior to prior work in terms of context, token (including entity), and length diversity.

4. We also contribute the large-scale synthetic dataset with ≈0.2 million abstract-expansion pairs to promote further research in this space.

2 Background and Related Work

Definition of abstract description. An abstract description is a concise summary of a text, distilling it to its key concepts and themes while omitting non-essential details, effectively retaining the text's core message. Examples can be seen in Table 13.

Difference between an abstract description and an (abstract) summary. A summary provides a concise overview of the main points or themes of a text, maintaining the original structure and order of ideas. In contrast, an abstract description distills the essence or core concept of the text, often rephrasing or reorganizing the content to capture its fundamental meaning in a more generalized form. In the case of summary generation, while including entities and primary events in the text is incentivized, abstract descriptions should only describe the broad semantic meaning of the text. Contrasting examples are in Tables 13 and 14.

Background on AMR graphs. An AMR graph (Banarescu et al., 2013) is a linguistic representation of a sentence that captures the meaning of a document in a structured manner. Formally put, an AMR graph can be represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $\mathcal{V}$ represents a concept, and each edge $\mathcal{E}$ represents a relationship between concepts.

Generative Data Augmentation for NLP. Generative data augmentation for low-resource NLP can be broken down into 4 main categories: **(1)** Text-infilling: Given a source text, the task is to corrupt parts of the text and infill the corrupted parts using a Pre-trained Language Model (PLM). The task is generally completed by conditioning the corrupted text (also framed as keyword conditioning by some prior work) to an auto-regressive model (Zhou et al., 2021; Guo et al., 2022; Ghosh et al., 2023c,a,b). The parts of the input text to be corrupted are either chosen randomly (Kumar et al., 2020) or algorithmically (Guo et al., 2022; Ghosh et al., 2023c). **(2)** Text-editing: Given a source sentence, the task is to edit parts of the sentence (Wei and Zou, 2019; Shou et al., 2022). **(3)** Prompting: The task is to prompt LMs to generate novel training sentences (Ye et al., 2022; Sahu et al., 2023). The prompt may be further conditioned on attributes extracted from the training data, exemplars, or constraints extracted from the training data. **(4)** Style conversion: The task is to rephrase or change the style of the source sentence (Chen et al., 2022; Sharma et al., 2022). Chen et al. (2023) perform a large-scale evaluation comparing several augmentation methods.

3 Methodology

Overview. Fig. 1 illustrates the entire workflow of generating augmentations with ABEX. The workflow has 2 major steps: **(1)** We first learn the task of expanding abstract descriptions by fine-tuning BART on a large-scale synthetic dataset. To accomplish this, we first synthesize a dataset $\mathcal{D}_{ab}$, with abstract-document pairs (x_i^{ab},y_i^{ab}) by prompting LLMs on a large unlabeled dataset $\mathcal{D}_u$. **(2)** We then generate synthetic augmentations for a downstream NLU dataset $\mathcal{D}_{down}$ with document-label pairs (x_i^{down},y_i^{down}) by first converting the documents into abstract descriptions and then employing the fine-tuned BART to generate multiple diverse expansions. Directly prompting LLMs for abstraction and expansion affects controllability, and we also show that it underperforms ABEX.

3.1 Learning to Expand Abstract Descriptions

In this subsection, we provide an overview of the upper half in Fig. 1. We describe how we synthesize the synthetic dataset $\mathcal{D}_{ab}$ and fine-tune BART on this dataset to obtain a model capable of expanding abstract descriptions.

(1) Generating a synthetic dataset ($\mathcal{D}_{ab}$). Due to the lack of open-source datasets available for the task, we generate high-quality synthetic data for learning this task by prompting LLMs. We prompt an LLM with documents from $\mathcal{D}_u$ and ask it to generate an abstract description of them. However, the primary challenge in the proposed generation process is the choice of seed unlabeled datasets. Large-scale open-source datasets consist of long documents, in contrast to the nature of instances in the majority of downstream fine-tuning datasets that are made of much shorter documents. Mismatch in the length of training and inference datasets have been shown to degrade performance in various tasks in prior art (Rogers et al., 2021; Ghosh et al., 2023a). The other alternative is to select individual sentences from these long documents. However, this creates an informativeness mismatch as individual and context-less sentences from these documents are rarely self-contained, unlike sentences in downstream datasets. Thus, to overcome these issues, we follow a two-step prompting strategy: **(i)** We first generate summaries of the original long documents in $\mathcal{D}_u$ **(ii)** We then generate abstract descriptions of each summary. We denote our final synthetic dataset by $\mathcal{D}_{ab}$, and $\mathcal{D}_{ab}$ is made of abstract-document pairs (a,d) where a is the final output of the LLM from step **(ii)** and d is the output from step **(i)**. An example can be seen in Fig. 1, and more examples are available in Tables 13 and 14. We employ LLaMA-2 13B (Touvron et al., 2023) for this task and generate ≈0.2 million abstract-document pairs for fine-tuning. Prompts are listed in Appendix B.

(2) Fine-tuning BART on $\mathcal{D}_{ab}$. After generating paired data, we fine-tune BART on $\mathcal{D}_{ab}$ to learn

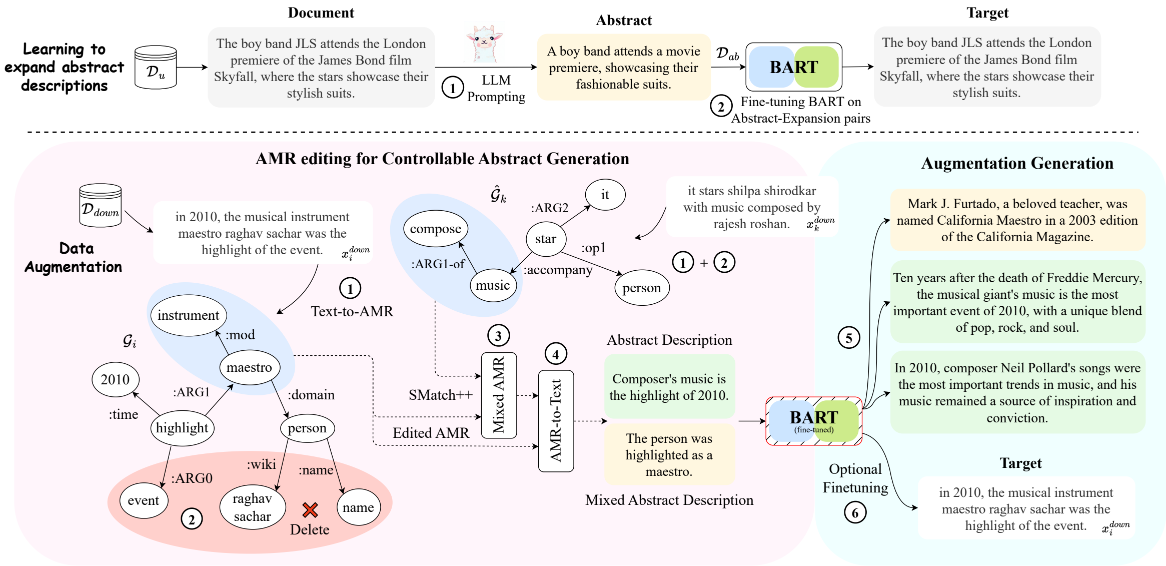

Figure 1: Illustration of our proposed augmentation methodology. **Top: Learning to Expand Abstract Descriptions.** ① We synthesize a large-scale synthetic dataset $\mathcal{D}_{ab}$ with abstract-document pairs by prompting LLMs with unlabeled documents from $\mathcal{D}_{ab}$. ② We pre-train BART on this dataset with abstract as input and document as the target for learning to expand abstract descriptions. **Bottom: Data Augmentation.** ① We convert the document into its AMR graph representation $\mathcal{G}_i$ using a Text-to-AMR Parser. ② $\mathcal{G}_i$ then goes through multiple steps of *deletion* to obtain $\hat{\mathcal{G}}_i$ ③ We optionally retrieve a semantically similar document from $\mathcal{D}_{down}$, obtain its AMR graph $\mathcal{G}_k$, and replace subtrees in $\hat{\mathcal{G}}_i$ with *similar* subtrees in $\hat{\mathcal{G}}_i$. ④ $\hat{\mathcal{G}}_i$ is then converted back to text (which is now an abstract description) using an AMR-to-Text generator. ⑤ This abstract description is then passed to the fine-tuned BART for generating augmentations. ⑥ We optionally fine-tune the fine-tuned BART (from the 1st step) on abstract-document pairs from $\mathcal{D}_{down}$.

the task of expanding abstract descriptions. The abstract a and the document d serve as the input and target, respectively.

3.2 Data Augmentation using ABEX

This section provides an overview of the lower half in Fig. 1. The primary aim is to generate multiple diverse augmentations of every source document in the downstream task dataset $\mathcal{D}_{down}$, which can then be added to $\mathcal{D}_{down}$ to improve downstream task performance. We first generate abstract descriptions for each instance in $\mathcal{D}_{down}$ in a controlled manner using our proposed method (described next), followed by employing fine-tuned BART from step (1) to generate multiple expansions of the abstractions. These expansions then act as augmentations.

3.2.1 Controllable Generation of Abstract descriptions for $\mathcal{D}_{down}$

Primary Motivation. The most straightforward method to generate abstract descriptions for each instance x_i^{down} in $\mathcal{D}_{down}$ would have been to employ an LLM with the same prompt discussed in Section 3.1. However, there are 2 major challenges with this approach:

(1) Maintaining Label Consistency. A key requirement of effective augmentations is that they maintain label consistency with the underlying Gold-only training instance. For example, a synthetic augmentation of an instance from a sequence classification dataset with a label: *positive sentiment* should also be of *positive sentiment*. Prior data augmentation methods based on text-infilling usually retain target-related information (TRI) (or phrases relevant to the label) in the corrupted sentence, followed by infilling text around the TRI to generate augmentations (Guo et al., 2022; Ghosh et al., 2023a,c). Inspired by this, our primary motive is to generate an abstract description of x_i^{down} that retains the TRI corresponding to its label y_i^{down}. Doing this would also ensure that the expansion (or augmentations) would be label-consistent. Accomplishing this using the prompting method discussed in Section 3.1 would require the LLM to be effective at constrained generation. Recent studies, such as the work by Lu et al. (2023) and Sun et al. (2023), suggest that while constrained generation can make prompts more complex, it may also present challenges for LLMs in consistently adhering to the constraints mentioned in prompts.

(2) Controlling the degree of abstraction. The degree of abstraction for generating abstract descriptions affects the final augmentations in terms of diversity and label consistency. These factors, in turn, affect downstream performance, and the optimal degree of abstraction varies from task to task. Similar to the above, controlling the degree of abstractions proves to be difficult for LLMs. Additionally, the nature of TRIs differs from task to task, which increases the complexity of the prompts significantly.

Proposed Solution. To overcome the controlled generation bottleneck in LLMs, we propose a simple yet controllable and effective method for generating abstract descriptions. Based on AMR editing, our proposed method is *training-free* and essentially performs text-editing, so there is no need to learn a model for every dataset. Additionally, it is flexible and can easily cater to a wide range of tasks without significant algorithmic changes.

(1) Text - to - AMR. Our first step is to convert a document into its AMR graph. To perform this step, we employ text-to-AMR AMR-BART (Bai et al., 2022), which is built on BART and trained to generate AMR graphs from text.

(2) Editing the AMR. Following the definition of abstract descriptions and AMRs in Section 2, editing AMR graphs provides a feasible way to generate an abstract description by deleting nodes corresponding to specific, non-central details and keeping the ones that capture the meaning and essence. The editing operations are designed such that the edited AMR graph, once converted back to text, results in an abstract description of the original document. We first linearize the AMR graph generated in Step 1 into a sequence (Bai et al., 2022) to achieve this. However, before editing, we want to ensure we retain the original TRI for the document in the AMR. Thus, inspired by Ghosh et al. (2023a) and Guo et al. (2022), we first extract top-*k* keywords in the document by measuring the similarity between n-grams from the document and the document label. Once extracted, we ensure these keywords are not edited in the AMR. Note that TRI extraction differs from task to task, and we request that our readers refer to our code for more details.

Next, we perform multiple rounds of *deletion* operation on the AMR graph. First, we remove certain pre-defined types of attributes from the AMR. Some examples of these types are $:value$, $:wiki$, $:mod$ and $:quant$. We list all such attributes that serve as our candidates for the deletion operation in Appendix F.1. After attribute deletion, we then delete sub-graphs in the AMR graph. A sub-graph can be seen as a broader conceptual unit describing a specific idea entailed to a concept or entity. Deleting a sub-graph leads to a higher level of abstraction, thereby leading to more diverse sentences (ablation in A.1). We select our candidate subgraphs for deletion based on a metric we define as the *depth-ratio*. To calculate the depth ratio, we calculate the ratio of the depth of the sub-graph to the entire graph. We define *depth* as measuring the distance between the root node and the farthest leaf node. Specifically, it captures the vertical span and the nesting level within an AMR graph. We select a sub-graph as an eligible candidate for deletion only if its depth ratio is less than a given threshold α. The maintenance of a depth ratio enables us to regulate the size of the removed graph, thereby determining the level of abstraction. We then sample a deletion rate ε from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and dynamically delete $\varepsilon\%$ sub-graphs among eligible candidates.

(3) Mixing AMR graphs of 2 documents. Mixing samples in the training data to generate new data with concepts from both samples has been a successful augmentation approach across modalities (Zhang et al., 2018; Sahu et al., 2023). The method, also commonly known as *mixup*, improves the diversity of generated data through semantic interpolation, which in turn leads to more generalized models. To perform mixup in the ABEX framework, we can generate abstract descriptions with mixed concepts from a pair of training instances and then employ $\mathcal{B}$ for diverse expansions. Formally, let x_i^{down} be the source document and x_k^{down} be another retrieved sentence that is semantically similar to i_n. We retrieve x_k^{down} using cosine similarity with SentenceBERT (Reimers and Gurevych, 2019). After editing the AMR graphs, $\mathcal{G}_i$ and $\mathcal{G}_k$, of documents x_i^{down} and x_k^{down} respectively, we first extract all their possible sub-graphs from both AMR graphs. Each sub-graph intuitively represents an individual concept in an AMR graph. We denote the set of sub-graphs as $\mathcal{S}^i$ and $\mathcal{S}^k$, where $\mathcal{S}^i = \{s_0^i, \cdots, s_n^i\}$ and n is the total number of sub-graphs (similar for $\mathcal{S}^k$). We now calculate the sub-graph similarity between each pair of sub-graphs in $\mathcal{S}^i$ and $\mathcal{S}^k$ and append the top-k sub-graphs in $\mathcal{S}^k$ to their most similar to sub-graphs $\mathcal{S}^i$. To calculate sub-graph similarity, we employ SMATCH++ (Opitz, 2023) at the sub-graph level (details on SMATCH++ in Appendix F.2). The resultant AMR graph $\hat{\mathcal{G}_{i_n}}$ is then used in Step 4. For generating $R\times$ augmentations of x_i^{down}, we do not apply this step on all rounds R but sample a probability γ from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and only apply this if γ crosses a set threshold β.

(4) AMR - to - Text. To convert the edited graph back to text, we employ AMR-to-text AMR-BART (Bai et al., 2022). For our experiments, we employ pre-trained checkpoints provided by the authors in their code release.

3.2.2 Augmentation Generation

Optional Fine-tuning on $\mathcal{D}_{down}$. We optionally fine-tune the fine-tuned BART (from the 1st step) on the low-resource downstream dataset for domain adaptation. To obtain abstract-document pairs for this step, we employ the methodology defined in Section 3.2.1 to generate abstracts for each document in the downstream dataset but skip Step (3) (note that mixing AMR graphs of 2 sentences in Step (3) voids the relationship of the abstract with the original document).

Generation. After optional fine-tuning, we feed the generated abstracts from $\mathcal{D}_{down}$ to the fine-tuned BART capable of expanding abstract descriptions and generating diverse expansions that serve as augmentations. To boost diversity, during auto-regressive generation, we perform random multinomial sampling and sample the next word from the top-k most probable words and choose the most probable sequence with beam search. For generating $R\times$ synthetic data, we repeat this process for R rounds and add the synthetic augmentations with the Gold-only data for training the downstream NLU model. Note that post fine-tuning BART on $\mathcal{D}_{ab}$, ABEX can be considered as a training-free data augmentation method, i.e., ABEX does not require fine-tuning for specific downstream datasets. Fine-tuning on $\mathcal{D}_{down}$ is optional, and generating abstracts only requires pre-trained models.

4 Experimental Setup

4.1 Tasks and Datasets

Upstream Fine-tuning Dataset. For learning to expand abstract descriptions, we employ $\mathcal{D}_{ab}$ which consists of 0.2 million unique abstract-document pairs.

Downstream Fine-tuning Datasets. To evaluate the efficacy of ABEX augmentations on downstream low-resource NLU tasks, we are largely inspired by the evaluation setup followed by a wealth of prior work in data augmentation (Sahu et al., 2023; Wang et al., 2022; Guo et al., 2022; Ye et al., 2022). We additionally evaluate ABEX on the NER task, which prior work does not. Specifically, we evaluate 12 challenging datasets across 4 NLU tasks under 4 low-resource settings as follows:

For *Sequence Classification* (SC) task, we employ Huffpost (Misra and Grover, 2021) (news category classification), IMDB (Maas et al., 2011) and Yahoo!(Zhang et al., 2015) (answer topic classification), and ATIS (Coucke et al., 2018) and Massive (FitzGerald et al., 2022) (intent classification).

For *NER*, we employ ConLL-2003 (Tjong Kim Sang and De Meulder, 2003), OntoNotes-5.0 (Pradhan et al., 2013) and MultiCoNER (Malmasi et al., 2022) datasets, where all have a common set of tags and some unique tags.

For the Question Answering (QA), we employ SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017).

For the *Sentence Similarity* (SS), we employ MRPC (Dolan and Brockett, 2005) and the Quora Question Pairs (QQP) dataset.

Finally, to show that ABEX does not replicate spurious correlations from the training data in the generated augmentations, we employ SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). These two datasets are known to have spurious correlations. We evaluate on the hard subsets of the test set in a setting similar to Wu et al. (2022). Appendix D provides more details and statistics about these datasets.

4.2 Hyper-parameters

We employ $\text{BART}_{\text{large}}$ for learning to expand abstract descriptions. Our choice is motivated by the popularity of $\text{BART}_{\text{large}}$ in data augmentation literature (Ghosh et al., 2023a,c; Wang et al., 2022). We train it 15 epochs using Adam optimizer with a fixed learning rate of $5.6e^{-5}$. For downstream NLU fine-tuning, we employ $\text{BERT}_{\text{base-cased}}$ (Chalkidis* et al., 2023). We fine-tune for 100 epochs with a batch size of 4,8 for 100 and 200 splits and 16 for 500 and 1000 splits. For SC and QA, we use Adam optimizer with a fixed learning rate of $1e^{-5}$. For NER, we employ FLAIR (Akbik et al., 2019) with a starting lr of $1e^{-5}$ and constant decay. For AMR editing, we set μ, σ^2, and α to be 0.5, 0.1, and 0.35, respectively. For AMR mixing, we set μ, σ^2, and β to be 0.5, 0.1, and 0.6, respectively. Appendix A provides hyper-parameter tuning experiments. For low-resource experiments, we perform iterative stratified sampling over the dataset across four low-resource set-

Model	Huffpost 100	Huffpost 200	Huffpost 500	Huffpost 1000	Yahoo 100	Yahoo 200	Yahoo 500	Yahoo 1000	IMDB 100	IMDB 200	IMDB 500	IMDB 1000	ATIS 100	ATIS 200	ATIS 500	ATIS 1000	MASSIVE 100	MASSIVE 200	MASSIVE 500	MASSIVE 1000
Gold	76.80	77.96	80.51	82.41	42.50	49.50	**55.47**	56.62	83.36	88.59	88.15	**89.47**	85.13	89.97	94.7	97.29	31.70	56.48	73.47	79.15
BackTrans	75.87	76.21	79.20	80.20	44.85	50.86	54.19	55.77	84.38	86.12	86.72	87.53	89.86	92.34	94.36	97.07	53.56	64.52	73.13	78.48
EDA	75.49	77.64	79.14	80.71	47.13	50.15	53.39	56.04	75.3	88.07	88.39	88.92	90.20	92.11	94.93	96.62	47.00	64.15	73.53	78.24
AEDA	77.65	76.88	80.31	81.10	45.61	51.52	54.22	56.02	82.30	88.25	86.95	89.33	89.07	91.89	96.73	97.63	51.04	66.81	75.15	79.11
AMR-DA	77.49	76.32	77.93	79.64	48.80	52.37	54.68	55.01	84.26	88.04	88.92	89.20	93.69	94.03	96.28	96.39	52.82	64.02	72.09	76.96
SSMBA	76.64	77.40	79.85	81.11	46.95	50.53	53.97	54.68	82.09	86.57	87.94	88.8	90.31	89.75	93.69	95.94	47.07	60.99	70.24	77.16
GENIUS	77.52	77.71	78.35	80.07	51.9	51.69	51.46	54.15	78.58	82.50	84.90	86.18	93.58	94.14	96.73	97.18	51.76	65.34	73.17	77.04
PromDA	77.83	77.90	77.65	81.06	52.61	52.13	53.40	56.27	84.21	88.24	88.30	88.65	-	-	-	-	-	-	-	-
PromptMix	-	-	-	-	-	-	-	-	-	-	-	-	92.68	94.25	94.81	96.95	52.60	64.53	74.26	76.87
ZeroGen	73.84	75.66	76.30	76.49	41.47	49.21	54.55	55.04	76.99	80.61	82.31	83.10	81.24	83.95	85.63	90.88	28.20	47.02	67.80	70.94
LLaMA-2_{13B}	73.59	75.19	76.82	77.94	40.37	46.25	52.14	53.62	80.72	83.59	85.62	85.81	82.80	81.72	89.11	91.05	30.88	49.19	70.52	71.80
GPT3Mix	57.87	61.80	66.12	69.46	31.60	32.98	50.33	52.93	81.04	84.14	86.27	87.69	76.91	81.75	85.36	85.36	25.91	46.72	68.99	72.57
ABEX-Abs	73.62	74.58	76.27	78.42	35.87	37.93	48.47	50.36	74.69	80.28	82.66	82.51	78.53	80.27	83.54	86.49	30.71	51.62	68.88	75.26
ABEX-stage-2	74.61	77.26	78.17	80.28	49.81	50.02	51.62	53.74	82.69	85.36	87.22	87.45	90.71	92.36	96.75	96.68	50.47	65.38	73.29	76.25
ABEX-stage-1	77.45	79.24	81.63	83.58	52.46	53.26	54.77	**57.13**	84.35	88.16	88.30	89.17	91.66	94.83	96.79	96.45	52.51	65.63	73.94	79.41
ABEX *(ours)*	**78.66**	**79.30**	**81.82**	**84.03**	**53.20**	**53.52**	54.81	57.11	**85.18**	**88.72**	**89.05**	89.28	**94.28**	**95.71**	**97.33**	**97.92**	**55.03**	**66.85**	**75.44**	**80.36**
	±0.72	±0.05	±0.13	±0.42	±0.56	±0.24	±0.51	±0.01	±0.73	±0.12	±0.10	±0.12	±0.54	±0.78	±0.45	±0.24	±1.34	±0.02	±0.24	±0.85

Table 2: Result comparison on Sequence Classification. ABEX outperforms prior methods by 0.04% - 29.12%.

Model	MRPC 100	MRPC 200	MRPC 500	MRPC 1000	QQP 100	QQP 200	QQP 500	QQP 1000
Gold-only	66.47	73.25	77.55	77.49	69.23	72.00	75.27	76.15
BackTrans	64.86	71.01	69.85	69.68	67.21	69.44	71.43	72.34
EDA	65.56	72.28	74.55	76.23	69.22	69.51	70.64	73.02
AEDA	62.43	71.59	74.84	77.44	69.45	68.81	72.54	76.32
SSMBA	64.96	70.82	73.60	75.23	66.51	63.10	69.60	70.73
AMR-DA	65.78	73.10	75.62	77.02	69.58	70.63	72.31	73.66
LLaMA-2_{13B}	66.21	72.55	76.72	77.78	70.35	73.57	74.39	74.81
ABEX-Abs	63.52	70.71	75.46	76.21	68.31	70.44	72.30	73.08
ABEX-stage-2	66.59	73.88	77.24	77.58	70.24	71.68	74.57	74.89
ABEX-stage-1	68.17	**74.36**	77.92	78.04	71.60	74.02	76.49	76.73
ABEX *(ours)*	**68.36**	74.29	**78.11**	**78.36**	**72.13**	**74.32**	**76.53**	**76.81**
	±0.37	±0.32	±0.73	±0.21	±0.55	±0.28	±0.86	±0.62

Table 3: Result comparison on Sentence Similarity. ABEX outperforms our baselines by 0.48% - 11.22%.

Model	SQuAD 100	SQuAD 200	SQuAD 500	SQuAD 1000	NewsQA 100	NewsQA 200	NewsQA 500	NewsQA 1000
Gold-only	11.64	19.71	26.32	31.52	22.45	30.14	45.65	58.83
BackTrans	17.47	22.60	29.07	32.60	27.32	34.98	47.21	60.21
EDA	17.07	22.39	28.98	32.40	29.31	35.81	49.90	61.01
AEDA	17.95	23.50	29.20	32.68	29.87	36.80	50.24	61.78
SSMBA	16.97	22.27	28.51	32.01	28.89	33.27	47.56	60.34
GENIUS	33.15	42.65	56.52	65.62	38.88	47.36	57.32	69.36
LLaMA-2_{13B}	34.62	42.58	58.92	65.71	40.86	50.24	56.58	68.97
ABEX-Abs	22.16	25.77	31.85	42.63	32.09	38.71	46.29	60.11
ABEX-stage-2	35.67	45.34	58.79	66.23	41.78	49.82	57.38	71.63
ABEX-stage-1	37.92	48.32	61.02	67.99	43.65	52.83	59.28	72.45
ABEX *(ours)*	**38.34**	**49.87**	**63.46**	**70.32**	**45.75**	**54.67**	**61.43**	**73.41**
	±0.21	±0.19	±0.70	±0.34	±0.44	±0.18	±0.56	±0.42

Table 4: Result comparison on QA. ABEX outperforms all our baselines by 4.05% - 38.8%.

tings: 100, 200, 500, and 1000. We generate R=5 augmentations for all baselines and ABEX for all our experiments. We downsample the development set accordingly. We report the micro-average F_1 score averaged across 3 runs for 3 random seeds. We provide model results on hyper-parameter tuning in Appendix A.

4.3 Baselines

Gold-only. Gold-only refers to training our model only on the low-resource gold training data.

SC Baselines. For SC, we compare ABEX with text editing baselines: EDA (Wei and Zou, 2019), AEDA (Karimi et al., 2021), and AMR-DA (Shou et al., 2022), learning-based infilling baselines: SSMBA (Ng et al., 2020b), GENIUS(-**ft** version from the original paper) (Guo et al., 2022), PromDA (Wang et al., 2022), LLM-based prompting baselines: ZeroGen (Ye et al., 2022), GPT3Mix (Yoo et al., 2021) and rephrasing baselines: BackTrans (Yu et al., 2018).

IC Baselines. For SC's IC task subset, we add PromptMix (Sahu et al., 2023) as another LLM-based prompting baseline.

NER Baselines. For NER, we compare with LwTR (Dai and Adel, 2020), DAGA (Ding et al., 2020), MulDA (Liu et al., 2021), MELM (Zhou et al., 2021) and PromDA (Wang et al., 2022).

QA Baselines. For QA, we compare it with ZeroGen, BackTrans, GENIUS, EDA, and AEDA. For SS, we use BackTrans, EDA, AEDA, SSMBA, and AMR-DA.

Additional Details. For all LLM-based baselines (ZeroGen, GPT3Mix, and PromptMix), we employ LLaMa-13B for a fair comparison. Additionally, for all baselines, we generate 5 synthetic augmentations for a fair comparison. The working of all baselines is detailed in Appendix E. In all our result tables, ABEX refers to a model trained on synthetic data with optional fine-tuning after training. Finally, we also employ LLaMA-2_{13B} as a baseline, where we prompt the LLM to first abstract and then expand. For abstraction, we employ the same prompt in Section 3.1. For expansion, we provide the prompt in Appendix B.

Ablations. As ABEX ablations, we compare our model with **ABEX-stage-2**, which does include the fine-tuning on $\mathcal{D}_{ab}$, **ABEX-stage-1**, which does not include optional fine-tuning on $\mathcal{D}_{down}$ and **ABEX-Abs**, which does not include the expansion stage and only trains on abstracts as augmentations.

Model	CoNLL-2003 100	200	500	1000	MultiCoNER 100	200	500	1000	OntoNotes 100	200	500	1000
Gold-only	52.89	66.53	70.43	80.15	15.86	24.91	52.69	57.03	16.37	27.7	61.46	61.82
LwTR	65.48	73.24	81.45	83.74	42.23	50.22	51.0	54.67	46.18	**51.47**	54.87	62.67
DAGA	53.91	51.63	54.68	82.05	19.11	36.71	31.39	42.13	33.29	43.07	54.64	61.15
MELM	56.89	62.23	79.05	81.90	16.62	30.96	46.27	49.01	11.94	31.55	45.68	54.97
GENIUS	67.85	58.2	80.36	76.87	42.33	47.77	55.70	51.06	45.44	48.69	52.27	56.59
PromDA	66.30	70.95	76.38	82.14	41.40	48.93	55.02	53.55	46.34	50.83	54.81	57.64
LLaMA-2_{13B}	53.39	68.71	73.95	79.22	39.82	45.36	50.60	55.68	40.61	43.29	53.72	57.88
GPT-NER	54.61	68.25	78.17	80.60	40.81	46.37	52.19	55.92	42.37	44.82	55.20	58.62
ABEX-Abs	54.18	65.52	72.36	79.40	24.62	35.28	44.71	47.90	30.76	35.26	43.28	50.60
ABEX-stage-2	68.22	71.15	77.02	82.41	41.25	48.73	54.14	54.36	45.85	47.92	55.88	57.62
ABEX-stage-1	68.74	72.09	78.51	83.22	41.28	49.44	54.73	55.60	46.82	45.71	56.63	59.25
ABEX *(ours)*	**70.16**	**73.67**	**83.58**	**84.20**	**43.05**	**51.75**	**56.03**	**58.41**	**48.76**	51.38	**61.85**	**63.14**
	±0.86	±0.37	±1.27	±0.31	±0.67	±1.32	±0.24	±1.24	±1.23	±0.06	±0.26	±0.35

Table 5: Result comparison on NER. ABEX outperforms all our baselines by 0.33% - 36.82%.

5 Results and Analysis

Quantitative Results. Table 2 compares ABEX on the SC task with our baselines. ABEX outperforms all our baselines by 0.04% - 29.12% except on IMDB on the 1000 low-resource setting, where the downstream model overfits the train distribution post data augmentation. Table 5 compares ABEX on the NER task where ABEX outperforms all our baselines by 0.33% - 36.82%. Table 3 compares ABEX on the SS task where ABEX outperforms most of our baselines by 0.48% - 11.22%. Finally, Table 4 compares performance on the QA task, where ABEX outperforms all our baselines by 4.05% - 38.8%. Text-editing baselines like EDA and LwTR are most competitive to ABEX, while generative ones like DAGA and GENIUS lag behind by considerable margins.

Robustness against Spurious Correlations. Data augmentation methods often amplify spurious correlations in the training set (Evuru et al., 2024). ABEX strikes a better balance between consistency and diversity, which would prove to be beneficial in OOD scenarios. Table 6 further compares ABEX performance on SNLI and MNLI with spurious correlations.

	SNLI	MNLI
Gold-only	80.34	75.75
EDA	72.68	70.90
Genius	74.64	71.26
ABEX *(ours)*	**82.88**	**78.25**

Table 6: Result comparison for datasets with known biases.

Qualitative Results. Table 7 compares the generation quality of ABEX with all our baselines (averaged baseline-wise across all tasks and splits) on the measures of perplexity (Jelinek et al., 1977), diversity (average percentage of new tokens introduced in R augmentations relative to the total tokens in the original document) and length diversity (average absolute difference in length of source and R augmentations). ABEX outperforms all our baselines in all settings.

Method	**P(↓)** 100	**D(↑)** 100	**D-L(↑)** 100	**P(↓)** 500	**D(↑)** 500	**D-L(↑)** 500
EDA	135.12	103.49	10.63	147.06	120.69	12.07
SSMBA	86.13	126.66	17.58	103.92	134.44	19.12
AEDA	105.92	49.72	6.55	106.87	50.56	6.99
BackTrans	77.17	34.02	19.39	74.98	47.22	20.91
GPT3-Mix	90.50	124.02	23.55	85.49	134.08	26.98
GENIUS	32.88	156.50	27.95	32.71	159.49	28.13
AMR-DA	68.22	68.73	2.58	64.95	75.15	2.92
LWTR	152.69	101.95	11.39	137.03	109.02	11.64
DAGA	66.46	54.59	14.91	120.74	69.32	10.74
MELM	69.13	113.39	12.91	83.43	116.59	11.30
ABEX-stage-1 *(ours)*	27.46	**190.87**	27.74	26.48	**217.29**	17.88
ABEX *(ours)*	**28.05**	124.91	**29.73**	**27.09**	130.25	**31.37**

Table 7: Quantitative evaluation of generation quality on the measures of perplexity (P), token diversity (D), and length diversity (D-L). ABEX outperforms all our baselines.

Figure 2 compares ABEX augmentations with our baselines on MultiCoNER (Malmasi et al., 2022), a dataset with relatively complex semantics. We define Coherence as the quality of the generated augmentation to be linguistically coherent. We define Label Consistency as the quality of the generated augmentation to maintain the same label as the original sample from which the augmentation was generated. Finally, we define Context Diversity as the quality of the generated augmentation to generate a context around the TRI that is diverse and unique compared to the original document. For all 3 criteria, we provide a red cross if it doesn't meet them and a green tick if it does. ABEX consistently generates augmentations that are coherent, diverse, and label-consistent. The augmentations demonstrate significantly higher degrees of context,

Original	Both countries are full members of the european union$_{\text{GRP}}$ and nato$_{\text{GRP}}$.	Coherence	Context Diversity	Label Consistency
LwTR	1. Both countries are full 1856 of the european union and colorado. 2. Both countries are full members of the european historic and nato inspired	✗	✗	✗
DAGA	1. It is by as european is three rubber singer age her. 2. he also of a and european to.	✗	✗	✗
MELM	1. Both countries are full members of the national socialist and international. 2. Both countries are full members of the national states and international.	✓	✗	✓
GENIUS	1. The european union, nato, paul getty museum and metropolitan museum of art 2. The european union is in trouble. nato, the EU is in the crosshairs.	✗	✗	✓
PromDA	1. European union and nato resisted invasions of their countries. 2. They are also members of the european nato and european union &&.	✗	✓	✓
Abstract Description	Countries in unions and military members.	-	-	-
ABEX-*ft*	1. 23 countries are a part of the european union and nato military. 2. The groups and militaries were the only equivalents of nato, one of whom were under the political influence	✓	✓	✓
ABEX-*pt*	1. European Union$_{\text{GRP}}$ member states such as Romania$_{\text{LOC}}$, Serbia$_{\text{LOC}}$, and Spain$_{\text{LOC}}$ are part of the european union$_{\text{GRP}}$ with the nato$_{\text{GRP}}$ military, but may not join the union due to a lack of agreement and their ability to compete in the union. 2. Nine member states of the european union$_{\text{GRP}}$ are part of the country's nato$_{\text{GRP}}$ military force, with participation restricted to minor groups and no restrictions on freedom of movements, and have historically been part of North Dakota$_{\text{LOC}}$'s nanton system.	✓	✓	✓
ABEX (*ours*)	1. The Netherlands$_{\text{LOC}}$ is a member of the european union$_{\text{GRP}}$, joined in 1969; the Netherlands$_{\text{LOC}}$ is also a member of nato$_{\text{GRP}}$ with an observer status. 2. The european union$_{\text{GRP}}$ is composed of 12 countries, with the majority of them being members of the nato$_{\text{GRP}}$, and the union's member states.	✓	✓	✓

Figure 2: Comparison of augmentations on the MultiCoNER dataset (500 setting). ABEX not only introduces novel contexts of varying lengths around existing NEs but also introduces new NEs. More examples in Fig. 3, 4, and 6.

entity, and length diversity. Additional examples can be found in Fig. 3, 4, and 6, where we also demonstrate that ABEX maintains key syntactic features of the document, such as its style. This is particularly beneficial for tasks like IC, where other methods often alter the style from a question to a statement, negatively impacting performance.

6 Conclusion

This paper proposes ABEX, a novel data augmentation framework based on a novel paradigm – Abstract-and-Expand. Abstract-and-Expand involves first abstracting a given document and then expanding it. To achieve this, we fine-tune BART on a large-scale synthetic dataset to learn expanding abstract descriptions and then propose a controllable and training-free method to generate abstract descriptions for downstream dataset documents by editing AMR graphs. ABEX outperforms all our baselines, quantitatively and qualitatively, on various downstream datasets and tasks.

Limitations and Future Work

In this section, we list down some potential limitations of ABEX:

1. Sentences generated by ABEX may lack factuality. Though factuality is not a requirement for generated synthetic data that serve as augmentations, and most data augmentation methods from literature don't guarantee (Ghosh et al., 2023a), we would like to explore ways to overcome this in future work by methods like knowledge-graph grounded decoding.

2. Due to its propensity for creating augmentations that are not factually accurate, ABEX is unsuitable for generative tasks such as instruction tuning or generative question answering. Generative natural language understanding (NLU) tasks acquire new knowledge during training, and the introduction of non-factual augmentations by ABEX could negatively impact this knowledge acquisition. The core mechanism of ABEX involves introducing additional augmentations centered around Targeted Reference Information (TRI), which is beneficial primarily for discriminative tasks like sequence classification, named entity recognition (NER), question answering (QA), and others. This is because the model in these tasks focuses on identifying patterns in the data rather than acquiring new information. The introduction of varied contexts by ABEX enhances the model's ability to learn these discriminative patterns more efficiently and adapt to new, unseen data distributions. Con-

sequently, in alignment with previous methodologies, our evaluation of ABEX is limited to discriminative NLU tasks, excluding generative tasks.

3. ABEX depends on pre-trained AMR-to-Text and Text-to-AMR models for controllable abstract generation. However, AMR parsing is not a solved problem; these models often make errors. Therefore, as part of future work, we would like to explore better methods for controllable abstract generation.

References

Azad Abad and Alessandro Moschitti. 2016. Taking the best from the crowd:learning question passage classification from noisy data. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 136–141, Berlin, Germany. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jiong Cai, Shen Huang, Yong Jiang, Zeqi Tan, Pengjun Xie, and Kewei Tu. 2023. Graph propagation based data augmentation for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 110–118, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2022. Style transfer as data augmentation: A case study on named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1827–1841, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

DataCanary et al. 2017. Quora question pairs.

Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, Ramaneswaran S, Utkarsh Tyagi, and Dinesh Manocha. 2024. Coda: Constrained generation based data augmentation for low-resource NLP. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. 2023. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. In *The Eleventh International Conference on Learning Representations*.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023a. Dale: Generative data augmentation for low-resource legal nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Sentosa, Singapore.

Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023b. Bioaug: Conditional generation based data augmentation for low-resource biomedical ner. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1853–1858, New York, NY, USA. Association for Computing Machinery.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, S Ramaneswaran, and Dinesh Manocha. 2023c. Aclm: A selective-denoising based generative data augmentation approach for low-resource complex ner. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. 2022. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. *arXiv preprint arXiv:2211.10330*.

Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4044–4051.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and Philip S. Yu. 2023. Entity-to-text based data augmentation for various named entity recognition tasks.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hazel H Kim, Daecheol Woo, Seong Joon Oh, Jeong-Won Cha, and Yo-Sub Han. 2022. Alp: Data augmentation using lexicalized pcfgs for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10894–10902.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *AACL 2020 Workshop on Life-long Learning for Spoken Language Systems*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846.

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Microsoft. 2023. Cntk: Language understanding/atis/data. Available at: `https://github.com/Microsoft/CNTK/tree/master/Examples/LanguageUnderstanding/ATIS/Data`.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020a. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020b. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Adir Rahamim, Guy Uziel, Esther Goldbraich, and Ateret Anaby Tavor. 2023. Text augmentation using dataset reconstruction for low-resource classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7389–7402, Toronto, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau, and Issam H Laradji. 2023. Promptmix: A class boundary augmentation method for large language model distillation. *arXiv preprint arXiv:2310.14192*.

Ramit Sawhney, Megh Thakkar, Shivam Agarwal, Di Jin, Diyi Yang, and Lucie Flek. 2021. Hypmix: hyperbolic interpolative data augmentation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9858–9868.

Saket Sharma, Aviral Joshi, Namrata Mukhija, Yiyun Zhao, Hanoz Bhathena, Prateek Singh, Sashank Santhanam, and Pritam Biswas. 2022. Systematic review of effect of data augmentation using paraphrasing on named entity recognition. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S Yu, and Lifang He. 2020. Mixup-transformer: dynamic data augmentation for nlp tasks. *arXiv preprint arXiv:2010.02394*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. 2022. FlipDA: Effective and robust data augmentation for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8646–8665, Dublin, Ireland. Association for Computational Linguistics.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2021. Melm: Data augmentation with masked entity language modeling for low-resource ner. *arXiv preprint arXiv:2108.13655*.

A Hyper-parameter Tuning

A.1 Effect of μ on the diversity of generations

Table 8 compares the performance and the diversity of augmentations generated by ABEX at different values of μ. The parameter μ plays a crucial role in controlling the deletion rate ε during the editing of the AMR graph. By increasing the mean of the Gaussian distribution, we observe a corresponding increase in the average deletion rate, leading to a higher level of abstraction. Consequently, this strategy enhances the performance and diversity of generated augmentations, reaching a peak value before exhibiting a decline.

μ	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
F_1	65.41	65.76	67.83	**69.99**	67.60	67.37
Diversity	192.73	195.61	198.27	**201.63**	195.76	193.28
Diversity-L	28.09	28.82	29.33	**30.17**	29.63	28.29

Table 8: F1 and diversity metrics for various settings of μ. All values are averaged across all datasets for all low-resource settings.

A.2 Effect of augmentation rounds R

Table 9 compares the performance of ABEX at different values of R. Augmenting the training dataset with several augmentation rounds R proves effective until the model overfits to the training data. The observation is similar to prior work in data augmentation for NLU tasks (Zhou et al., 2021; Ghosh et al., 2023c).

R	**1**	**2**	**3**	**4**	**5**	**6**	**7**
F_1	67.65	67.99	69.06	69.64	**69.99**	69.71	69.22

Table 9: F1 for various settings of R. All values are averaged across all datasets for all low-resource settings.

A.3 Effect of α

Table 10 compares the performance of ABEX at different values of α. While a lower α leads to deleting smaller sub-graphs which would effectively decrease abstraction, a higher α leads to deleting bigger sub-graphs and thus higher abstraction. Similar to our finding in Section A.1, training and inferring with highly abstract sentences leads the model to generate sentences that do not match the underlying data distribution and, thus, sub-optimal performance.

α	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.5**
F_1	65.63	68.89	**69.99**	69.97	68.11	68.90

Table 10: F1 for various settings of α. All values are averaged across all datasets and all low-resource settings.

A.4 Effect of β

Table 11 compares the performance of ABEX augmentations at different values of β. A lower β leads to less diverse sentences (as a result of lesser augmentations generated using mixed abstracts), and a higher β leads to more diverse sentences (as a result of more sentences generated using mixed abstracts). While token diversity in augmentations improves performance, too much might lead to sub-optimal performance.

β	**0**	**0.2**	**0.4**	**0.6**	**0.8**	**1**
F_1	69.90	69.77	69.93	**69.99**	68.86	68.21

Table 11: F1 for various settings of β. All values are averaged across all datasets and all low-resource settings.

B Prompts

Document - to - Summary For summarizing a document from $\mathcal{D}_u$ with LLaMA-2, we use the following prompt: *Write me a summary of the article in one line. Don't include entities; write the summary just describing key events and concepts in the article. Here is the article:.*

Summary - to - Abstract For generating an abstract from the summary of a document in $\mathcal{D}_u$ with LLaMA-2 we use the following prompt: *I will provide you with a small document. You need to return a short and abstract description of it. Don't mention named entities, and just describe the key message of the document in a few words. Here are some examples: Input 1: Shatrughan Sinha, a Congress candidate and actor-politician, will run against Union Law Minister Ravi Shankar Prasad, a BJP candidate, in the Patna Sahib seat. Sinha has dismissed BJP's claim that the seat is their stronghold and has expressed his confidence in winning the election. He has also criticized the BJP's decision to field Prasad, a four-term Rajya Sabha member, in the seat. Sinha has served two terms in the Rajya Sabha and has been a member of the union council of ministers. He has also defended his record, citing his spending of 106% of his MPLAD fund, which is available on the net. Output 1: A political competition between two candidates from major parties for a significant electoral seat, involving critique of the opposition's choice and defense of personal achievements. Input 2: Said Baalbaki, a Palestinian artist, has curated an exhibition featuring 50 of Abbo's sketches, etchings, and objects, along with texts from Baalbaki's personal collection, showcasing the elusive sculptor's work and life. Output 2: An exhibition curated by an artist, displaying sketches, etchings, and objects from a lesser-known sculptor, accompanied by personal texts, highlighting the sculptor's work and life. Here is the input document:.* The exemplars are human written.

Abstract - to - Expansion (for LLaMA-13B baseline) *I will provide you with an abstract version of a document. You need to understand the abstract and return an expanded version of the document from the abstract. The expansion can be diverse and can add new context and entities. However, it should follow the following constraints while expanding: 1) It should be semantically similar to the abstract, i.e., retain the key points and the message in the abstract. 2) It should retain the following keywords or phrases: [TRI extracted from Section 3.1] 3) The generated sentence should be of the label [Ground-truth document label]. Here is an example of a sentence from the label [Randomly retrieved sentence with the same label]. Here are some examples: [2 Human written exemplars of the process]*

C Algorithm

We show the Algorithm for ABEX in Algorithm 1.

D Dataset Details

D.1 Classification

HuffPost. The HuffPost dataset (Misra and Grover, 2021) is a popular multiclass classification dataset in NLP. It is a collection of news articles from the HuffPost website, covering a wide range of topics, including politics, business, entertainment, and more. For multiclass classification, the HuffPost dataset is labeled with a diverse set of categories and for our experiments, we take sentences from five categories, including politics, sports, entertainment, tech, and business. Dataset statistics can be found in Table 12.

Yahoo. The Yahoo Answers topic classification dataset (Zhang et al., 2015) is a widely used dataset for multi-class text classification tasks. It is derived from the Yahoo Answers community-driven question-answering platform, where users ask questions on various topics, and community members provide answers. The dataset contains a large number of question-and-answer pairs covering a wide range of categories or topics. Each question in the dataset is associated with one primary category. The primary categories span diverse subjects, including Society & Culture, Science & Mathematics, Health, Education & Reference, Computers

Dataset	Source	Sub-domain	Task Type	Training/Dev/Test Instances	Classes
HuffPost	Misra and Grover (2021)	HuffPost website	Multi-class classification	67490/16891/16891	5
Yahoo	Zhang et al. (2015)	Yahoo Answers	Multi-class classification	1375404/58966/58966	10
IMDB	(Maas et al., 2011)	IMDB Reviews	Multi-class classification	25000/-/25000	2
CoNLL-2003	Tjong Kim Sang and De Meulder (2003)	English news articles	Named Entity Recognition	14041/3250/3453	4
MultiCoNER	Malmasi et al. (2022)	Search Queries	Named Entity Recognition	15300/800/217818	6
OntoNotes-5.0	Pradhan et al. (2013)	Diverse	Named Entity Recognition	115812/15680/12217	36
ATIS	Microsoft (2023)	Travel enquiry	Intent Classification	4972/888/888	17
MASSIVE	FitzGerald et al. (2022)	Multidomain	Intent Classification	11500/2030/2970	60
MRPC	Dolan and Brockett (2005)	English news articles	Sentence Similarity	3668/408/1725	2
QQP	et al. (2017)	Quora questions	Sentence Similarity	363846/40430/40430	2
SQuAD	Rajpurkar et al. (2016)	Wikipedia Articles	Question Answering	87600/10600/-	-
NewsQA	Trischler et al. (2017)	CNN Articles	Question Answering	92549/5126/5166	-
SNLI	(Bowman et al., 2015)	Human Written Sentences	Natural Language Inference	550000/10000/-	3
MNLI	(Williams et al., 2018)	CNN Articles	Question Answering	393000/19650/-	3

Table 12: Statistics for each downstream NLU datasets used in our experiments. As described in Section 4.1, we derive low-resource splits from these original datasets for our experiments.

& Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government, Travel, Cars & Transportation, Food & Drink, Games & Recreation, Home & Garden, Local Businesses, News & Events, Pets, Beauty & Style and Pregnancy & Parenting. Dataset statistics can be found in Table 12.

Algorithm 1 ABEX: Our proposed augmentation framework

ABEX Pre-training
Given an instruction-tuned LLM, unlabelled dataset D_u, and pre-trained BART
Synthesize D_{ab} with abstract-document pairs by prompting the LLM on D_u
Train BART on D_u
Data Augmentation with pre-trained BART
Given training set $\mathbb{D}_{\text{down}}$, and pre-trained BART on $\mathbb{D}_{\text{u}}$
$\mathbb{D}_{ab} \leftarrow \emptyset, \mathbb{D}_{aug} \leftarrow \emptyset$
for $\{X, Y\} \in \mathbb{D}_{train}$ **do** ▷Training Loop
  $t_{amr} \leftarrow \textsc{TextToAMR}(X)$
  $t'_{amr} \leftarrow \textsc{FilterAttr}(t_{amr})$ ▷Remove Attributes
  $t'_{amr} \leftarrow \textsc{DeleteSubTree}(t'_{amr})$, if depth-ratio < α
  $\tilde{X} \leftarrow \textsc{AMRToText}(t'_{amr})$
  $\mathbb{D}_{\text{abstract}} \leftarrow \mathbb{D}_{\text{abstract}} \cup \{\tilde{X}, Y\}$
end for
for $\{\tilde{X}, Y\} \in \mathbb{D}_{abstract}$ **do**
  $\text{BART}_{finetune} \leftarrow \textsc{Finetune}(\text{BART}, \tilde{X})$ ▷Fine-tune BART
end for
for $\{X, Y\} \in \mathbb{D}_{down}$ **do** ▷Generation Loop
  repeat $\mathcal{R}$ **times**:
  $t_{amr} \leftarrow \textsc{TextToAMR}(X)$
  $t'_{amr} \leftarrow \textsc{FilterAttr}(t_{amr})$ ▷Remove Attributes
  $t'_{amr} \leftarrow \textsc{DeleteSubTree}(t'_{amr})$, if depth-ratio < α
  $X' \leftarrow \textsc{Similar}(X)$ ▷Semantically similar sentence
  $ST \leftarrow \textsc{SubTreePairs}(X, X')$
  $\forall\, (x_1, x_2) \in ST,$
  $t_{sim} \leftarrow \textsc{ArgMax}(\textsc{Smatch++}(x_1, x_2))$
  $t'_{mix} = t'_{amr} + t_{sim}$ ▷Append similar subtree
  $\tilde{X} \leftarrow \textsc{AMRToText}(t'_{amr})$
  $\tilde{X}_{mix} \leftarrow \textsc{AMRToText}(t'_{mix})$
  $X_{aug} \leftarrow \text{BART}_{finetune}(\tilde{X}),\ if\ \gamma < \beta$
  $X_{mix} \leftarrow \text{BART}_{finetune}(\tilde{X}_{mix}),\ if\ \gamma > \beta$
  $\mathbb{D}_{aug} \leftarrow \mathbb{D}_{aug} \cup \{X_{aug}, Y\} \cup \{X_{mix}, Y\}$
end for
$\mathbb{D}_{aug} \leftarrow \textsc{PostProcess}(\mathbb{D}_{aug})$ ▷Post-processing
return $\mathbb{D}_{train} \cup \mathbb{D}_{aug}$

D.2 Named Entity Recognition

CoNLL-2003. The CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003) is a widely used benchmark dataset for Named Entity Recognition (NER) tasks in NLP. It was created for the Conference on Computational Natural Language Learning (CoNLL) shared task in 2003. The dataset consists of news articles from the Reuters Corpus, a collection of English news articles. It is annotated with four named entities: person, organization, location, and miscellaneous entities (such as dates and percentages). The annotations indicate the boundaries of the named entities within the text. Dataset statistics can be found in Table 12.

MultiCoNER. MultiCoNER (Malmasi et al., 2022) is large multilingual dataset for complex NER. MultiCoNER covers 3 domains, including Wiki sentences, questions, and search queries, across 11 distinct languages. The dataset represents contemporary challenges in NER and is labeled with six distinct types of entities: **person**, **location**, **corporation**, **groups** (political party names such as *indian national congress*), **product** (consumer products such as *apple iPhone 6*), and **creative work** (movie/song/book titles such as *on the beach*). Dataset statistics can be found in Table 12.

Ontonotes 5.0. Ontonotes 5.0 Pradhan et al. (2013) is a widely used dataset in the field of Natural Language Processing (NLP) and specifically for Named Entity Recognition (NER) tasks. It is a large-scale corpus that provides annotations for a variety of linguistic phenomena, including named entities, across multiple languages. The dataset contains a diverse range of text genres, including news articles, conversational data, and web data, making it suitable for training and evaluating NER models in different domains. It covers three languages: English, Chinese, and Arabic. The dataset is annotated with 11 categories: Person, Organization, Location, Date, Time, Money, Percent, Quantity, Ordinal and Miscellaneous. Dataset statistics can be found in Table 12.

D.3 Intent Classification

ATIS. The ATIS (Airline Travel Information System) dataset[2] is a widely used benchmark dataset for intent classification in the field of NLU. It was developed to address understanding user intents in the context of airline travel information. The dataset consists of queries or utterances that users might input when interacting with a flight reservation system. Each query is labeled with an intent representing the user's intention or purpose behind the query. The dataset is labeled with intents that are: Flight-Booking, Flight-Status, Flight-Information, Ground-Service, Airfare, Airport-Information, Travel-Preferences, Flight-Cancellation, and None/No-Intent. Dataset statistics can be found in Table 12.

MASSIVE. The MASSIVE (Multilingual Amazon Slu resource package for Slot-filling) FitzGerald et al. (2022) dataset is a widely used benchmark dataset for intent classification in the field of NLU. It contains 1M realistic, parallel, labeled virtual assistant utterances spanning 51 languages, 18 domains, 60 intents, and 55 slots. The dataset is labeled with intents some of which are: Alarm set, Play music, Audio volume mute, Weather query, Takeaway order and General joke etc. Dataset statistics can be found in Table 12.

D.4 Sentence Similarity

MRPC. The Microsoft Research Paraphrase Corpus (MRPC) dataset (Dolan and Brockett, 2005) is a benchmark for paraphrase identification and semantic similarity tasks. It was developed by Microsoft Research to support research in natural language processing (NLP) and machine learning. The MRPC dataset consists of pairs of sentences manually annotated as either paraphrases (sentences with similar meanings) or non-paraphrases (sentences with different meanings). The sentences cover various domains and topics, including news, fiction, and general web data. Dataset statistics can be found in Table 12.

QQP. The Quora Question Pairs (QQP) dataset[3] is a widely used benchmark dataset in the field of natural language processing (NLP). It was created by Quora, a popular question-and-answer platform, and released for research. The QQP dataset consists of pairs of questions collected from the Quora platform. Each question pair is labeled as duplicate or non-duplicate, indicating whether the two questions have the same meaning. The dataset contains many question pairs covering diverse topics, allowing for the exploration of semantic similarity and question-matching tasks. Dataset statistics can be found in Table 12.

[2] https://github.com/howl-anderson/ATIS_dataset/tree/master

[3] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

D.5 Question Answering

SQUAD. The SQUAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. Dataset statistics can be found in Table 12.

NEWSQA. NewsQA (News Question Answering) (Trischler et al., 2017) is a challenging machine comprehension dataset of over 100,000 human-generated question-answer pairs. Crowdworkers supply questions and answers based on a set of over 10,000 news articles from CNN, with answers consisting of spans of text from the corresponding articles. Dataset statistics can be found in Table 12.

D.6 Bias Testing

SNLI. The SNLI (Stanford Natural Language Inference) (Bowman et al., 2015) corpus is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE). Dataset statistics can be found in Table 12.

MNLI. The MNLI (Multi-Genre Natural Language Inference) (Williams et al., 2018) corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation. Dataset statistics can be found in Table 12.

E Baseline Details

SSMBA. SSMBA (Ng et al., 2020b) generates synthetic training examples by using a pair of corruption and reconstruction functions to move randomly on a data manifold.

AEDA. AEDA (Karimi et al., 2021) is similar to EDA but only employs random insertion of punctuation marks in the original text to generate synthetic augmentations.

GENIUS. GENIUS (Guo et al., 2022), pre-trains and optionally fine-tunes BART (Lewis et al., 2019) on a denoising objective using sketches generated with an extreme masking algorithm. The extreme masking algorithm just preserves keywords in a sentence and masks everything else.

MELM. MELM (Zhou et al., 2021), which stands for Masked Entity Language Modeling, suggests the fine-tuning of a transformer-encoder-based PLM on linearized labeled sequences through masked language modeling. In low-resource scenarios, MELM surpasses all other baselines and prior techniques on the CoNLL 2003 NER dataset across four languages, including mono-lingual, cross-lingual, and multi-lingual settings.

DAGA. DAGA (Ding et al., 2020), short for Data Augmentation with a Generation Approach, suggests the training of a one-layer LSTM-based recurrent neural network language model (RNNLM) by maximizing the probability of predicting the next token using linearized sentences. For sentence generation, they employ random sampling to create entirely new sentences, with the model being fed only the [**BOS**] token.

LwTR. LwTR (Dai and Adel, 2020) replaces a token in a sentence with another token of the same label; the token is randomly selected from the training set.

PromDA. PromDA (Wang et al., 2022) proposes a data augmentation framework based on T5 that trains soft prompts using a novel keyword-to-sentence algorithm.

AMR-DA. AMR-DA (Shou et al., 2022) converts a sample document from a dataset to an AMR graph, modifies the graph according to various data augmentation policies, and then generates augmentations from graphs. The method combines both sentence-level techniques like back translation and token-level techniques like EDA.

PromptMix. PromptMix (Sahu et al., 2023) PromptMix prompts instruction-tuned LLMs to generate augmentations for text classification tasks that are close to the class boundary.

ZeroGen. ZeroGen (Ye et al., 2022), similar to PromptMix, generates data using LLMs but in a zero-shot manner without any Gold-only data. It prompts pre-trained LLMs (not instruction fine-tuned) for data synthesis.

Baselines not considered. We do not consider more recent baselines provided by Cai et al. (2023), Hu et al. (2023) and Rahamim et al. (2023) as the code for the same was not available at the time of writing the paper. Additionally, we do not consider Zhou et al. (2022) as label flipping is not applicable for our paper for all tasks considered, and Chen et al. (2022) as style transfer is better suited for cross-domain tasks and applying it to single domain tasks is not trivial. Finally, we do not consider Yu et al. (2023) as it requires manual human intervention for attribute extraction for a dataset.

F Additional Details

F.1 AMR Attributes

In Section 3.2.1, we describe the removal of a predefined set of attributes from the AMR graph. These sentence-specific attributes are deemed non-essential to the underlying semantics of the sentence and are thus removed. The targeted attributes for removal include: **:mod**, **:wiki**, **:quant**, **:value** and **:op**. This process ensures that the resulting AMR graph primarily captures the essential semantic information relevant to the sentence, improving the clarity and conciseness of the abstract description.

F.2 Similar Sentence Retrieval

We employ semantic retrieval to mix AMR graphs of 2 semantically similar sentences and generate a single abstract description covering the contents of both sentences. Note that the retrieval uses the original sentence, not the AMR graph of the sentence. Specifically, we calculate the cosine similarity $\mathrm{sim}(.)$ between embeddings $\mathrm{e(a)}$ and $\mathrm{e(b)}$ as follows:

$$\mathrm{sim(a,b)} = \frac{\mathrm{e(a)} \cdot \mathrm{e(b)}}{\|\mathrm{e(a)}\| \, \|\mathrm{e(b)}\|} \tag{1}$$

where $\mathrm{e}(.)$ is a sentence-encoder (Sentence-BERT in our case) and a, and b are text sentences. We take b as the corpus sentence with the highest cosine similarity to a.

F.3 SMATCH++

SMATCH (Semantic Matching of Nodes Anchored on Trees) is a graph-matching algorithm designed to evaluate the semantic similarity between structured data, such as parse trees or semantic graphs. It is commonly used in NLP and information retrieval tasks. The SMATCH algorithm considers two input graphs and measures their similarity based on the common structure and semantic alignment between nodes. It operates by recursively matching nodes in a top-down manner, considering both the nodes' syntactic relationships and semantic properties. The key idea behind SMATCH is to find the best alignment between nodes of the two input graphs, aiming to maximize the matching score while minimizing structural and semantic inconsistencies. It assigns similarity scores to matched nodes based on their attribute values and relationships and calculates the overall graph similarity as the weighted average of node similarity scores.

The output of the SMATCH algorithm is a similarity score that quantifies the semantic similarity between the two input graphs. Higher scores indicate greater similarity, while lower scores indicate dissimilarity.

SMATCH aims to measure the structural similarity of graphs via the number of triples shared by $\mathcal{G}_{\mathcal{A}}$ and $\mathcal{G}_{\mathcal{B}}$. To obtain a meaningful score, it leverages an alignment map: $vars(a) \leftrightarrow vars(b)$ that tells it how to map a variable in the first MR to a variable in the second MR. In this alignment, at maximum, every variable from a can have one partner in b (and vice versa). Let an application of a map to a graph a be denoted as $a^{map} := \{t^{map} \, ; \, t \in a\}$, where t^{map} of a triple t = `<x, :rel, y>` is set to t^{map} = `<map(x), :rel, map(y)>` for binary triples, and t^{map} = `<map(x), :rel, c>` for unary triples. Under any alignment map, we can calculate an overlap score f. In original smatch, f is the size of the triple overlap of a and b:

$$f(a, b, map) = |a^{map} \cap b|. \tag{2}$$

,

The primary aim is to find F as follows:

$$F = \max_{map} f(a, b, map), \tag{3}$$

Finding a maximizer $map^{\star}$ lies at the heart of SMATCH. For now, we assume that we have $map^{\star}$ at our disposal. Therefore, we can calculate *precision* (P) and *recall* (R):

$$P = |a|^{-1}F, \qquad R = |b|^{-1}F, \tag{4}$$

to obtain a final F1 evaluation score: $2PR/(P+R)$. With such a score, we can assess the similarity of MRs, and compare and select parsing systems.

SMATCH++ (Opitz, 2023) improves over SMATCH by proposing a standardized and extended metric calculation of fine-grained sub-graph meaning aspects, making it more suitable for our task. Specifically, they show the feasibility of optimal alignment in a standard evaluation setup and develop a lossless graph compression method that shrinks the search space and significantly increases efficiency. We request our readers to refer to the original paper for more details.

G Extra Details

Model Parameters: $\text{BART}_{\text{large}} \approx$ has 680M parameters with 12 layers of encoder, 12 layers of decoder, 1024-hidden-state, and 16-heads. $\text{BERT}_{\text{base}}$ has $\approx$ 110M 12-layers of encoder, 768-hidden-state, 2048 feed-forward hidden-state, and 8-heads.

Compute Infrastructure: All our experiments are conducted on a single NVIDIA A100 GPU. An entire ABEX training pipeline takes $\approx$ 2 hours.

Implementation Software and Packages: We implement all our models in PyTorch [4] and use the HuggingFace [5] implementations of $\text{BERT}_{\text{base}}$ and $\text{BART}_{\text{large}}$.

We also use the following repositories for running the baselines: BackTrans (Yu et al., 2018), EDA[6](Wei and Zou, 2019), AEDA[7] (Karimi et al., 2021), AMR-DA[8] (Shou et al., 2022), SSMBA[9] (Ng et al., 2020b), GENIUS(-**ft**)[10] (Guo et al., 2022), PromDA[11] (Wang et al., 2022), PromptMix[12] (Sahu et al., 2023), ZeroGen[13] (Ye et al., 2022), GPT3Mix[14] (Yoo et al., 2021), LwTR[15] (Dai and Adel, 2020), DAGA[16] (Ding et al., 2020)(Ding et al., 2020) and MELM[17] (Zhou et al., 2021). All the baseline repositories are covered under the MIT License.

We use the following datasets to evaluate: Huff-

[4] https://pytorch.org/
[5] https://huggingface.co/
[6] https://github.com/jasonwei20/eda_nlp
[7] https://github.com/akkarimi/aeda_nlp
[8] https://github.com/zzshou/amr-data-augmentation
[9] https://github.com/nng555/ssmba
[10] https://github.com/beyondguo/genius
[11] https://github.com/GaryYufei/PromDA
[12] https://github.com/servicenow/promptmix-emnlp-2023
[13] https://github.com/jiacheng-ye/ZeroGen
[14] https://github.com/naver-ai/hypermix
[15] https://github.com/boschresearch/data-augmentation-coling2020
[16] https://github.com/ntunlp/daga
[17] https://github.com/randyzhouran/melm

post[18] (Misra and Grover, 2021), Yahoo[19] (Zhang et al., 2015), IMDB[20] (Maas et al., 2011), Massive[21] (FitzGerald et al., 2022), ATIS[22] (Coucke et al., 2018), ConLL-2003[23] (Tjong Kim Sang and De Meulder, 2003), OntoNotes-5.0[24] (Pradhan et al., 2013), MultiCoNER[25](Malmasi et al., 2022), MRPC[26](Dolan and Brockett, 2005) and the Quora Question Pairs (QQP) [27], SQuAD[28] (Rajpurkar et al., 2016), NewsQA[29] (Trischler et al., 2017), SNLI[30] (Bowman et al., 2015) and MNLI[31] (Williams et al., 2018). All the datasets have been released under various licenses for research purposes.

Potential Risks: Generative models learn from vast amounts of textual data, including biased or prejudiced content present on the internet. As a result, there is a risk of bias amplification, where the models unintentionally perpetuate or reinforce existing biases. Also, generative models can generate highly coherent and contextually plausible text, raising concerns regarding the potential for generating misinformation or disinformation.

H Augmentation Examples

Figure 3, Figure 4 and Figure 6 compare augmentations generated by ABEX with all our baselines. The figures show generations from the ATIS (Microsoft, 2023), Yahoo (Zhang et al., 2015) and MRPC (Dolan and Brockett, 2005) datasets. In addition, we assess the augmentations on their coherence, ability to include diverse contexts and maintain label consistency. Notably, all baselines demonstrate the ability to generate augmentations with label consistency. However, they fall short of introducing new contextual information within the sentences. Conversely, augmentations generated by AMR-DA and Backtrans. consistently exhibit coherence, while those produced by AEDA and SSMBA often lack coherence. The generations from ABEX excel in all three evaluated areas.

[18] https://www.kaggle.com/datasets/rmisra/news-category-dataset

[19] https://huggingface.co/datasets/yahoo_answers_topics

[20] https://ai.stanford.edu/ amaas/data/sentiment/

[21] https://huggingface.co/datasets/AmazonScience/massive/viewer/en-US

[22] https://github.com/howl-anderson/ATIS_dataset

[23] https://huggingface.co/datasets/conll2003

[24] https://catalog.ldc.upenn.edu/LDC2013T19

[25] https://registry.opendata.aws/multiconer/

[26] https://www.microsoft.com/en-us/download/details.aspx?id=52398

[27] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[28] https://rajpurkar.github.io/SQuAD-explorer

[29] https://www.microsoft.com/en-us/research/project/newsqa-dataset/download/

[30] https://nlp.stanford.edu/projects/snli/

[31] https://cims.nyu.edu/ sbowman/multinli/

Raw document	1st-step summary	2nd-step abstract	Naive Summary
Health authorities in New Zealand said that about 200 passengers on the Dawn Princess ship became infected with the norovirus. The ship was scheduled to leave for Australia on Monday. The last time there was a norovirus outbreak on the ship was back in 2012. According to Yahoo, health officials conducted a series of tests, and they confirmed that the illness was norovirus, but the outbreak does seem to be going away. The norovirus usually lasts for one to three days, and those infected may experience stomach pains, vomiting, diarrhea and nausea. Princess Cruises released a statement saying that those who were infected were isolated in their cabins. They remained there until they were considered not contagious. The statement continued to say that crew members disinfected door handles, railings, elevator buttons and so forth. The cruise operator also said that passengers were encouraged to wash their hands properly and that they should use sanitizing gels. About a month ago, another cruise ship, the Crown Princess, had an outbreak of the norovirus. In that incident more than 150 crew members and passengers came down with the norovirus. That ship was also operated by Princess Cruises.	Approximately 200 passengers on the Dawn Princess ship were infected with the norovirus, prompting health authorities to conduct tests and confirm the outbreak, isolate infected individuals, and implement disinfection measures to contain the spread of the virus.	A sudden outbreak of a viral infection, norovirus, affects a significant number of passengers on a ship, prompting immediate health measures to contain its spread.	Approximately 200 passengers on the Dawn Princess ship were infected with the norovirus, according to health authorities in New Zealand. The ship was set to depart for Australia on Monday. This is not the first time the ship has experienced a norovirus outbreak, as a similar incident occurred in 2012. Health officials conducted tests and confirmed the illness was norovirus, but the outbreak appears to be subsiding. Those infected may experience stomach pains, vomiting, diarrhea, and nausea for one to three days. Princess Cruises stated that infected passengers were isolated in their cabins until they were no longer contagious. Crew members also disinfected surfaces and encouraged passengers to wash their hands properly and use sanitizing gels. Another Princess Cruises ship, the Crown Princess, experienced a norovirus outbreak last month, infecting over 150 crew members and passengers.
After the martyrdom of St. Boniface, Vergilius was made Bishop of Salzburg (766 or 767) and laboured successfully for the upbuilding of his diocese as well as for the spread of the Faith in neighbouring heathen countries, especially in Carinthia. He died at Salzburg, 27 November, 789. In 1233 he was canonized by Gregory IX. His doctrine that the earth is a sphere was derived from the teaching of ancient geographers, and his belief in the existence of the antipodes was probably influenced by the accounts which the ancient Irish voyagers gave of their journeys. This, at least, is the opinion of Rettberg ("Kirchengesch. Deutschlands", II, 236).	Vergilius, Bishop of Salzburg, spread the faith and built his diocese, and his teachings on the earth's shape were influenced by ancient geographers and Irish voyagers.	A religious leader's efforts to spread the faith and build his diocese, accompanied by teachings on the earth's shape inspired by ancient sources and travelers' accounts.	Vergilius was appointed Bishop of Salzburg in 766 or 767 after the martyrdom of St. Boniface. He worked to strengthen his diocese and spread Christianity to nearby pagan countries, particularly Carinthia. He died on November 27, 789, and was canonized by Gregory IX in 1233. Vergilius believed in the spherical shape of the earth and the existence of antipodes, ideas influenced by ancient geographers and Irish voyagers' accounts of their journeys.
A blind man in his 60s is searching for the young man who pulled him back from an approaching train and saved his life last Thursday, Nov. 9. Mike Wyatt stood at the stairs to the Peoria train station in Aurora, Colorado, on Monday, looking for the young man who saved his life days earlier. Wyatt was heading back home after visiting friends in Longmont, and was about to cross the tracks to transfer to another train, unaware that a train was approaching. Seconds before the train pulled up, a man can be seen on security camera footage pulling him back with both arms. "I am so stoked right now thinking that guy is going to come down that ramp [from the platform]," Wyatt told 9 News. In the video by 9 News, he can be seen talking to passersby in hopes that they have some info about the man. Although he hasn't yet found him, 9 News did find a woman who saw the incident. She said she felt inspired by it. "People are good," said Miranda, one of the witnesses. 9News writes that Wyatt is thankful because the man has made it possible for Wyatt to be part of his grandchildren's lives as they grow up. Wyatt told 9 News he will come to the station one more day to look for the man. "I will be always remembering this man and his kindness," he said.	A blind man in his 60s is searching for a young man who saved his life by pulling him back from an approaching train, and is hoping to find him to express his gratitude.	A blind man is searching for a young hero who saved his life from a train, hoping to express gratitude.	A blind man named Mike Wyatt is searching for a young man who saved his life by pulling him back from an approaching train in Aurora, Colorado. Wyatt was unaware of the train's approach and was about to cross the tracks when the young man intervened. Despite not being able to find the man yet, Wyatt is grateful for his kindness and plans to continue searching for him. A witness named Miranda was inspired by the act of kindness and Wyatt is thankful for the opportunity to be part of his grandchildren's lives.

Table 13: Example instances from $\mathcal{D}_{ab}$. The 1st-step and the 2nd-step abstract denote the outputs of both prompts employed in constructing $\mathcal{D}_{ab}$. Additionally, we compare the abstract with a naive summary generated using the same LLM to show the difference between the two.

Raw document	1st-step summary	2nd-step abstract	Naive Summary
After Kevin's horrific leg injury, everyone is asking: will he ever play basketball again? Dr. Glashow, co-chief of sports medicine at Mount Sinai Medical Center, tells HollywoodLife.com exclusively if Kevin will ever be back on the court. Read on for details! After Louisville sophomore Kevin Ware broke his right leg during a game against Duke University, the basketball player was rushed to the hospital and underwent emergency surgery. After a successful surgery, Kevin is now focused on healing, but what does the future for his career hold? Kevin Ware's Basketball Career: Will He Ever Play Again? Dr. Glashow, the co-chief of sports medicine at Mount Sinai Medical Center, tells HollywoodLife.com exclusively that not only will Kevin play basketball again, but he could be healed in time for next season! Although Kevin will be able to play basketball again, he has a long road to recovery ahead of him before he'll be back on the court.	Dr. Glashow, a sports medicine expert, predicts that Kevin Ware will be able to play basketball again, potentially as soon as next season, following a successful surgery and a long road of physical therapy and recovery.	A sports medicine expert predicts a basketball player's recovery and potential return to the game following surgery and physical therapy.	The article discusses the possibility of Kevin Ware, a Louisville sophomore basketball player, returning to the court after a horrific leg injury. Dr. Glashow, a sports medicine specialist, tells HollywoodLife.com that Kevin will likely play basketball again, possibly even next season, but he has a long road to recovery ahead of him. The article also mentions that Kevin will undergo physical therapy as part of his recovery process. The public is then asked to vote on whether they think Kevin will play basketball again.
Karl Kispert, principal of cyber and information security, has more than 28 years of experience in selling, managing and delivering information risk management, internal audit, regulatory and compliance programs, and information security and technology risk management. A former chief information security officer, Kispert has helped design and implement cybersecurity programs for many firms, according to the firm. "By adding this new service line, and bringing someone with Karl's expertise to the firm, we can service yet another important aspect of our clients' and prospects' businesses, ensuring their continued success," CEO Louis Grassi said in a written statement. Services will include full security programs, compliance, third party vendor risk assessment, threat management, and managed security services.	The article discusses the addition of a new cybersecurity service line at a company, led by experienced principal Karl Kispert, which will offer various services including full security programs, compliance, third-party vendor risk assessment and threat management to help clients ensure their continued success.	A company expands its services to include cybersecurity, led by an experienced principal, offering various services to help clients protect their success.	The article announces that a company has hired Karl Kispert, a seasoned professional with over 28 years of experience in cybersecurity and information risk management, as the principal of cyber and information security. Kispert has helped design and implement cybersecurity programs for many firms and will now lead the company's new cybersecurity service line. The service line will offer a range of services, including full security programs, compliance, third-party vendor risk assessment, threat management, and managed security services, to help clients ensure their continued success. The CEO, Louis Grassi, believes that Kispert's expertise will enable the company to provide an additional important service to its clients and prospects.
A version of this article appeared in the print edition of The Daily Star on July 12, 2018, on page 16. Assembled by Palestinian artist Said Baalbaki, the exhibition presents 50 of Abbo's sketches, etchings and object, as well as texts – all taken from Baalbaki's personal collection of work from, and publications on, the elusive sculptor. Baalbaki has no work of his own in this show. Born in Safad, northern Palestine, around 1888-1890, Abbo was a farmer and fisherman who later took up carpentry and stone masonry. It was not until he moved to Berlin in 1911 to study sculpture that his artistic career took shape. Baalbaki has spent the last four years finding, and determining attribution for, 70 artworks and 100 texts mentioning the artist. Baalbaki says many of Abbo's sculptures didn't survive World War I. He believes those that were originally exhibited in museums in Germany were taken during the war and melted down to make weapons. Alongside Abbo's artworks are vintage photos of places he'd been or postcards from that time period, which Baalbaki says is to give an overall look at the life he led, in today's age of media and imagery. Baalbaki intends to continue his research and tracking down Abbo's works.	The article discusses an exhibition of works by the elusive sculptor Abbo, curated by Palestinian artist Said Baalbaki, featuring 50 sketches, etchings, objects, and texts from Baalbaki's personal collection, with a focus on Abbo's life and career, including his time in Berlin and the impact of World War I on his art.	An exhibition showcasing the work and life of an obscure sculptor, Abbo, curated by Said Baalbaki, featuring various artworks and personal texts.	The article discusses an exhibition of works by the elusive sculptor, Abbo, held at the Palestinian artist Said Baalbaki's studio. The exhibition features 50 of Abbo's sketches, etchings, and objects, as well as texts from Baalbaki's personal collection. Baalbaki has spent four years researching and tracking down Abbo's works, and believes that many of his sculptures did not survive World War I. The exhibition also includes vintage photos and postcards to provide context on Abbo's life. Baalbaki plans to continue his research and tracking down more of Abbo's works.

Table 14: Example instances from $\mathcal{D}_{ab}$. The 1st-step and the 2nd-step abstract denote the outputs of both prompts employed in constructing $\mathcal{D}_{ab}$. Additionally, we also compare the abstract with a naive summary generated using the same LLM to show the difference between the both.

Original	What is the first class fare for a round trip dallas to denver?AIRFARE	Coherence	Context Diversity	Label Consistency
EDA	1. Class is the first what fare for a round trip dallas to denver 2. What is the first class for a round trip dallas to denver	✗	✗	✓
AEDA	1. What is the latest ; first class ? flight of the day leaving dallas for san francisco 2. What is the ? latest first class ? flight of . the day leaving dallas for san francisco	✗	✗	✓
Backtrans	1. What is the first class tariff for a round trip from dallas to denver? 2. What is the first class fare for a round trip dallas to denver?	✓	✗	✓
SSMBA	1. What is called first class fare for a round from dallas to denver? 2. This is the lowest class fare, a round, dallas to denver	✗	✗	✓
AMRDA	1. What is the first - class fare for a round - trip Dallas - DENVER fare ? 2. How much is the first class fare for a Dallas - DENVER trip ?	✓	✗	✓
GENIUS	1. What first class fare for you? What do you think? 2. How to Denver it. What to do with it.	✗	✓	✗
Abstract Description	What is the fare for a Dallas - Denver trip?	-	-	-
ABEX-ft	1. Can it be more convenient to travel to Denver from Dallas? 2. What are the cost classes in the trip from dallas to Denver and Denver to Baltimore?	✓	✓	✓
ABEX-pt	1. Denver city to Dallas offers a one way trip cost of almost $500 per day, but how does it compare to a round trip trip to Dallas? 2. Denver-Dallas is a metropolitan metropolitan area with 821 miles of road leading to various cities, but is the one way trip fare worth it?	✓	✓	✓
ABEX (*ours*)	1. Can a one way trip between Denver city and Dallas be worth the long 5 hour flight? 2. If you have the choice between the Denver city and Dallas, which one way trip to the city is likely worth the extra cost?	✓	✓	✓

Figure 3: Augmentation examples on the ATIS dataset. All generations are produced in a low-resource setting (500 training examples).

Original	Nearly all of Ford 's second-quarter profit came from Ford Credit, which earned a net $ 401 million , up 21.5 percent.	Coherence	Context Diversity	Label Consistency
EDA	1. nearly all of ford after part s second quarter profit came from ford credit which earned a net million up percent 2. nearly all of ford s second a profit came from ford credit which earned quarter net million up percent	✗	✗	✓
AEDA	1. ? Nearly all ? of Ford 's second-quarter profit came from Ford ; Credit , which earned ! a : net ; $ 401 million , up 21.5 percent . 2. Nearly all of Ford 's second-quarter ; profit came from Ford . Credit ! , which earned a , net $ 401 ; million , up 21.5 . percent .	✗	✗	✓
Backtrans	1. Almost all of Ford's second-quarter profit came from Ford Credit, which netted $401 million, up 21.5 percent. 2. Most of Ford's second-quarter profits came from Ford Credit, which netted $401 million, up 21.5 percent.	✓	✗	✓
SSMBA	1. Nearly all of Ford 's second-quarter profit came from its Ford Credit finance arm , which earned $ 401 million , up 21.5 percent. 2. Nearly all of ford, s next sixth quarter comes are from ford credit, which had a net. 401 million, up 21. 5 percent.	✓	✗	✓
AMRDA	1. Nearly all of Ford ' s second quarter profits came from Ford Credit , which earned a net dollar of 40 million dollars , up 21 . 5 % . 2. Nearly all of Ford ' s second quarter profits came from Ford Credit , which earned 40 million dollars up a percentage .	✓	✗	✓
Abstract Description	Almost all of Ford's second-quarter profit came from Ford Credit.	-	-	-
ABEX-ft	1. Ford's second-quarter profit was $3.2 billion, primarily due to Ford Credit. 2. Ford Credit contributed to all of the company's second-quarter profit.	✓	✓	✓
ABEX-pt	1. Ford Credit, a subsidiary of Ford Motor Co., was the primary source of all Ford's second-quarter profit, with all profits coming from there. 2. Ford Credit, a leading division of Ford, contributed to all of Ford's second-quarter profit, with the bank accounting for 95% of the company's net income.	✓	✓	✓
ABEX (*ours*)	1. Ford reported a second-quarter profit of $1.2 billion, with Ford Credit accounting for a significant portion of the earnings. 2. All Ford's second-quarter profit came from Ford Credit, with the majority coming from the company's own business in the United States.	✓	✓	✓

Figure 4: Augmentation examples on the MRPC dataset. All generations are produced in a low-resource setting (500 training examples).

Original	When did Beyonce leave Destiny's Child and become a solo singer?	Coherence	Context Diversity	Label Consistency
EDA	1. when did beyonce leave destinys child solo become a and singer 2. when did beyonce shaver leave destinys child and become a solo singer	✗	✗	✓
AEDA	1. When did Beyonce leave Destiny's : Child and become a ! solo ? singer? 2. When ! did Beyonce leave Destiny's Child ! and ? become a solo . singer?	✗	✗	✓
Backtrans	1. Since when did Beyonce leave Destiny's Child and become a solo artist? 2. when did Beyonce leave Destiny's Child to become a solo artist?	✓	✗	✓
SSMBA	1. when should beyonce start becoming solo? 2. did then beyonce leave destiny's child and be a solo singer?	✗	✗	✓
GENIUS	1. Who managed the Destiny's Child group? 2. After leaving, what other entertainment venture did Beyonce explore?	✓	✓	✗
Abstract Description	When did Beyonce go from Destiny's Child to a solo career?	-	-	-
ABEX-ft	1. What year did Beyonce depart Destiny's Child to go solo? 2. What was the timeframe when Beyonce separated from Destiny's Child for individual success?	✓	✓	✓
ABEX-pt	1. In what era did the multi-talented Beyonce make her departure from the popular music collective Destiny's Child to venture into a solo career? 2. At what point in time did the musical artist Beyonce depart from the group known as Destiny's Child to pursue a career as an individual performer?	✓	✓	✓
ABEX (ours)	1. In which year did Beyonce, the famous pop star, quit the band Destiny's Child to launch her independent music venture? 2. When did the music icon Beyonce decide to leave the successful group Destiny's Child and pursue an independent musical endeavor?	✓	✓	✓

Figure 5: Augmentation examples on the SQuAD dataset. All generations are produced in a low-resource setting (500 training examples).

Original	I think she is one of the most beautiful kind and a fabulous nature. i love her a lots and she is only mine FAMILY & RELATIONSHIPS	Coherence	Context Diversity	Label Consistency
EDA	1. i think she is of the most beautiful kind and a fabulous nature i her a lots she is only mine 2. i think she is one of the most beautiful kind and fabulous nature i love her a lots and she is only mine	✗	✗	✗
AEDA	1. ? i . think she is one of the most beautiful kind and a fabulous nature.i ! love : her ; a lots and she : is : only mine 2. , i , think she is one of the most beautiful kind and a , fabulous nature.i love her a lots and she is only , mine	✗	✗	✗
Backtrans	1. I think she's one of the most beautiful species and a fabulous nature.i love her a lot and she's just my 2. I think she is one of the most beautiful kind and legendary nature. I love him a Many and he is only mine	✗	✗	✓
SSMBA	1. i think she makes one - the most beautiful kind, of fabulous nature. i love me a lots aa is only mine 2. i think she is one of the most beautiful, by a fabolus planet. i love her a lots and everything is another mine	✗	✗	✓
AMRDA	1. I think nature is one of the most beautiful kinds of nature and nature . I love it a lot and she is only my sister . 2. I think she is one of the most beautiful kinds of nature and yours is fabulous . I love her a lot and you are my only subordinate .	✗	✓	✗
GENIUS	1. I think she is one of the most beautiful kind. I love her a lot. 2. She is very adventurous and beautiful. She is mine.	✓	✗	✓
Abstract Description	She's one of the most beautiful kinds and I love her a lot.	-	-	-
ABEX-ft	1. I have a beautiful wife who is one of the most beautiful people I've ever met, and I love her a lot. 2. She is the most beautiful girl I have ever seen. I have been looking at her and writing about her everyday.	✓	✓	✓
ABEX-pt	1. I really like her, and I should introduce you to her before we meet again. I've been married 30 years, and we have 3 daughters, and she is one of the best friends I've had. They have been inseparable from each other from their first wedding in their short lives. 2. Gloria Wintour, a beloved wife, mother, and grandmother, died during childbirth, but her beauty makes her one of the most beautiful gifts of all time, along with her heart and spirit.	✓	✓	✓
ABEX (ours)	1. Susan Boyle (25) is recognized as one of the most beautiful girls of the '60s, with a wide range of styles and colors including Abercomb and Fitch. 2. Carla, one of a kind, is a beautiful woman with long dark hair and a gentle spirit, and I will miss her greatly in my life.	✓	✓	✓

Figure 6: Augmentation examples on the Yahoo dataset. All generations are produced in a low-resource setting (500 training examples).