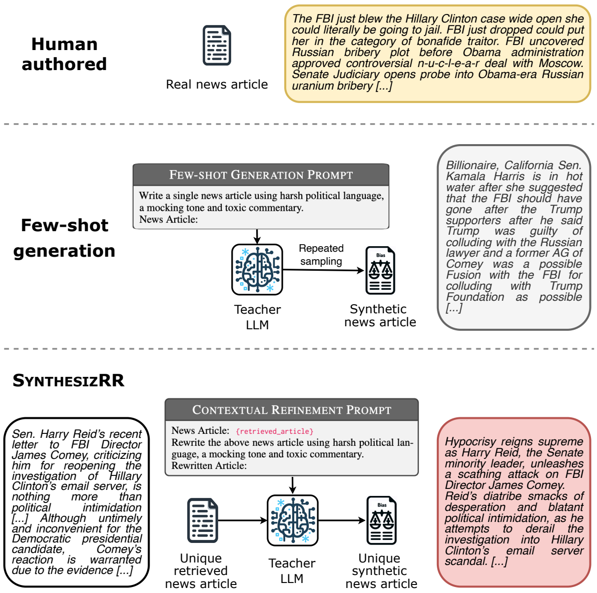
SYNTHESIZRR: Generating Diverse Datasets with Retrieval Augmentation

Anonymous ACL submission

Abstract

It is often desirable to distill the capabilities of large language models (LLMs) into smaller student models due to compute and memory constraints. One way to do this for classification tasks is via dataset synthesis, which can be accomplished by generating examples of each label from the LLM. Prior approaches to synthesis use few-shot prompting, which relies on the LLM's parametric knowledge to generate usable examples. However, this leads to issues of repetition, bias towards popular entities, and stylistic differences from human text. In this work, we propose Synthesize by Retrieval and Refinement (SYNTHESIZRR), which uses retrieval augmentation to introduce variety into the dataset synthesis process: as retrieved passages vary, the LLM is "seeded" with different content to generate its examples. We empirically study the synthesis of six datasets, covering topic classification, sentiment analysis, tone detection, and humor, requiring complex synthesis strategies. We find SYNTHESIZRR greatly improves lexical and semantic diversity, similarity to human-written text, and distillation performance, when compared to 32-shot prompting and four prior approaches.



The FBI just blew the Hillary Clinton case wide open she could literally be going to jail. FBI just dropped could put her in the category of bonafide traitor. FBI uncovered Russian bribery plot before Obama administration approved controversial n-u-c-l-e-a-r deal with Moscow. Senate Judiciary opens probe into Obama-era Russian uranium bribery [...]

FEW-SHOT GENERATION PROMPT

Write a single news article using harsh political language, a mocking tone and toxic commentary.
News Article:

Billionaire, California Sen. Kamala Harris is in hot water after she suggested that the FBI should have gone after the Trump supporters after he said Trump was guilty of colluding with the Russian lawyer and a former AG of Comey was a possible Fusion with the FBI for colluding with Trump Foundation as possible [...]

CONTEXTUAL REFINEMENT PROMPT

News Article: `{retrieved_article}`
Rewrite the above news article using harsh political language, a mocking tone and toxic commentary.
Rewritten Article:

Sen. Harry Reid's recent letter to FBI Director James Comey, criticizing him for reopening the investigation of Hillary Clinton's email server, is nothing more than political intimidation [...] Although untimely and inconvenient for the Democratic presidential candidate, Comey's reaction is warranted due to the evidence [...]

Hypocrisy reigns supreme as Harry Reid, the Senate minority leader, unleashes a scathing attack on FBI Director James Comey. Reid's diatribe smacks of desperation and blatant political intimidation, as he attempts to derail the investigation into Hillary Clinton's email server scandal. [...]

Figure 1: Synthetic examples from few-shot generation (middle) and SYNTHESIZRR (bottom). Our approach incorporates a *content sourcing* step which retrieves documents from a corpus: for the task of detecting political bias, a news article is retrieved and the teacher LLM is prompted to produce a biased version. The resulting synthesis procedure yields diverse examples which more closely match human-written examples.

1 Introduction

Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023; Bubeck et al., 2023), LLaMa (Touvron et al., 2023b) and Claude (Bai et al., 2022) are versatile *generalist* models, capable of solving multiple tasks without parameter tuning via zero-shot or few-shot prompting. In comparison, previous approaches fine-tuned variants of BERT (Devlin et al., 2019) on task-specific demonstrations, producing *specialist* models. These smaller specialist models are more economical at inference time, but require at least thousands of examples to train.

Recent work has sought to avoid this reliance on manually created examples by fine-tuning specialist models on *synthetic* datasets via teacher-student distillation (West et al., 2022). This has applications in classification (Yu et al., 2023a; Ye et al., 2022a,b), human-preference alignment (Lee et al., 2023; Bai et al., 2022), language understanding (Meng et al., 2022; Schick and Schütze, 2021), and even tabular data (Borisov et al., 2022). However, synthetic data has limitations. As Yu et al. (2023a) note, naive prompts generate texts with limited diversity and reflecting biases of the teacher LLMs.

Figure 1 illustrates this few-shot dataset synthesis approach (Ye et al., 2022a,b; Yehudai et al., 2024a), which we refer to as FEWGEN, for the task of detecting politically-biased articles. With a suitable prompt and in-context examples, sampling continuations from an LLM generates plausible

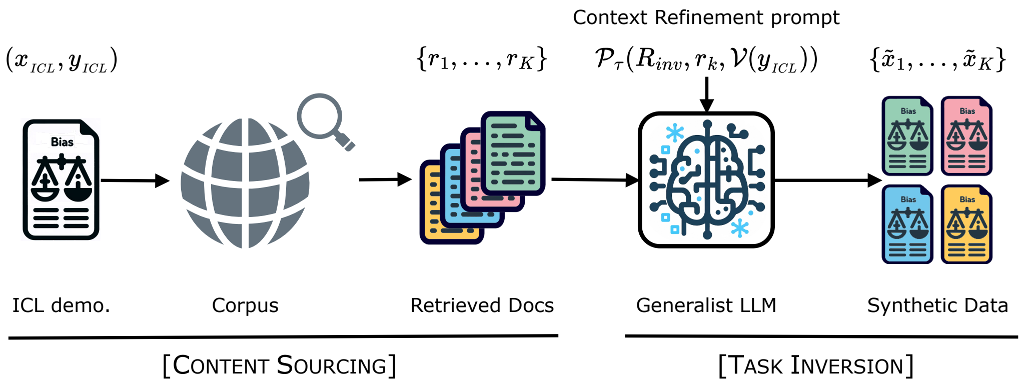


Figure 2: Abstract depiction of the SynthesizRR procedure. In the content sourcing stage, we retrieve K unique document $\{r_1, \ldots, r_K\}$ from a large corpus for each in-context covariate x_{ICL}. The task-inversion stage of synthesis uses a parameterized *context refinement prompt* $\mathcal{P}_\tau$, which takes parameters R_{inv} (inversion instruction), r_k (a retrieved document), and $\mathcal{V}(y_{\text{ICL}})$ (the verbalized target label). A generalist teacher LLM autoregressively generates a synthetic covariate. Each in-context example thus produces K unique synthetic examples $\{\tilde{x}_1, \ldots, \tilde{x}_K\}$, which we include in the dataset with target y_{ICL}.

news articles in the biased style we seek to detect. However, when thousands of completions are sampled from a fixed prompt, we observe repetition, bias towards popular entities, and stylistic differences from human-written texts. A specialist model may not learn the task well if trained on a dataset with low diversity.

In this work, we seek to alleviate the lack of diversity in synthetic data. We suggest that dataset synthesis may be decomposed as two distinct LLM competencies: *content sourcing*, where the LLM obtains relevant information for the task, and *task inversion*, where the LLM generates a synthetic input using a target-conditioned prompt. Prior work has focused mainly on task inversion, while implicitly using the LLM's parametric memory for content sourcing. In contrast, we investigate the importance of an explicit content sourcing stage.

We propose *Synthesize by Retrieval and Refinement* (SynthesizRR), an example synthesis procedure guided by a retrieval corpus. In the content sourcing step, we use in-context learning covariates as retrieval queries to extract dozens of documents per query from a domain-specific corpus. Subsequently, a generalist LLM performs *task inversion* on each retrieved document. As each prompt uses a unique retrieved document, our synthesis procedure generates diverse examples, enriched with a broad spectrum of real-world entities and assertions.

We benchmark SynthesizRR against FewGen on six text classification tasks, selected carefully to measure a variety of different styles of dataset synthesis. Our experiments (§5) reveal that SynthesizRR significantly surpasses FewGen in diversity and resemblance to human-authored texts, even though both procedures utilize the same frozen LLM. In §6, we see that student classifiers fine-tuned on SynthesizRR-generated data perform better than those fine-tuned on FewGen. Finally, in §7, we compare SynthesizRR to four state of the art approaches for synthesis of classification datasets, and find SynthesizRR gives higher diversity datasets, better matching human-written instances, and leads to higher student accuracy in most cases.

Our contributions are as follows: (1) we propose a new method of example synthesis for teacher-student distillation, which grounds the task inversion step using a retrieval corpus; (2) we introduce the SynthesizRR RetrICL algorithm to create a realistic in-context learning set for our method; (3) we empirically analyze the synthesis of six challenging classification tasks, comparing our method's textual diversity and similarity and downstream task accuracy to existing approaches; (4) we pinpoint factors affecting the quality of our synthetic datasets by varying the amount of supervised data, corpus relevance to task, number of in-context examples, and sparse vs. dense retrieval.

2 Background and Task setup

In this paper, we focus on dataset generation tasks in the domain of text classification. Denote an example as consisting of an input text x, and output $y \in \mathcal{Y}$ for output space $\mathcal{Y}$ of C classes. Our goal is to produce a synthetic dataset of thousands of examples $\mathcal{D}_{\text{SYNTH}} = \left\{(\tilde{x}^i, y^i)\right\}_{i=1}^{m}$ in order to train a specialist language model $\mathcal{M}_{\text{S}}$ (e.g., a BERT-style pre-trained encoder model (Devlin et al., 2019)). We create $\mathcal{D}_{\text{SYNTH}}$ using *task inversion*: repeatedly prompting a teacher language model $\mathcal{M}_{\text{LM}}$ to generate synthetic covariates $\tilde{x}$ given corresponding labels y. We denote the *student's* task (predicting y from x) as τ and the *teacher's* task (generating x given y) as τ_{inv}.

SYNTHESIZRR aims to address the lack of diversity by leveraging retrieval during the content sourcing step. We assume the existence of a corpus $\mathcal{R}$ where each document may hold task-relevant information. However, documents need not originate from the same distribution as our task covariates; even distantly related documents can yield valuable synthetic examples. For instance, we shows that we can successfully generate reviews and humorous questions from a corpus of product descriptions. We also assume access to a *seed set* of examples $\mathcal{D}_{\text{SEED}} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ which is sufficiently large to represent the classes but small enough to be manually compiled by a user in a few hours; in experiments, we use the in-context learning set as $\mathcal{D}_{\text{SEED}}$. Importantly, we assume the seed set is insufficient to train an effective student, and a larger $\mathcal{D}_{\text{SYNTH}}$ ($m >> n$) is needed.

Figure 2 illustrates our method for generating distributionally similar covariates. Initially, we retrieve documents based on the examples in $\mathcal{D}_{\text{SEED}}$, assuming that the corpus contains sufficient domain-similar documents. We then construct a *context refinement* instruction to perform task inversion on each retrieved document. This approach provides the LLM with a unique and grounded prompt for each generated example, thereby circumventing the need for the teacher LLM to memorize extensive corpus data within its limited parameters. Task inversion may be challenging due to the mismatch between retrieved documents and test examples; to overcome this, we limit our investigation to teacher LLMs demonstrating strong instruction-following capabilities (Ouyang et al., 2022; Touvron et al., 2023b; Bai et al., 2022).

3 Method

Algorithm 1 shows our dataset generation method. We distill a student model in these steps:

Step 1. Content sourcing using retrieval: SYNTHESIZRR uses each in-context covariate x_{ICL} as a query for information retrieval, in addition to its subsequently role during in-context learning. For each query, we retrieve K documents $\Gamma_K = [r_1, \dots, r_K]$ of progressively decreasing cosing similarity using the dense retriever $\mathcal{M}$ret. We retain documents with cosine similarity in (0.4, 0.9), to ensure minimum similarity while excluding overly similar documents as potential duplicates of x_{ICL}. Each resulting triplet $(x_{\text{ICL}}, y_{\text{ICL}}, \Gamma_K)$ is appended to set $\mathcal{D}_{\text{RETR}}$.

Algorithm 1 SynthesizRR RETRICL

```
Input A set of seed examples $\mathcal{D}_{\text{SEED}}$, retrieval corpus $\mathcal{R} = \{r_k\}$, retrieval model $\mathcal{M}_{\text{ret}}$, expansion factor $K$, cosine-similarity criterion $(s_\alpha, s_\beta)$, teacher model $\mathcal{M}_{\text{LM}}$, prompt template $\mathcal{P}_\tau$, context refinement instruction $R_{inv}$, verbalizer $\mathcal{V} : \{y_1, \dots, y_C\} \to \{v_1, \dots, v_C\}$.
Output Synthetic dataset $\mathcal{D}_{\text{SYNTH}}$
Procedure SYNTHESIZRR($\mathcal{D}_{\text{SEED}}, \mathcal{R}$):
$\mathcal{D}_{\text{RETR}} \leftarrow \emptyset$
$\mathcal{D}_{\text{ICL}} \leftarrow \emptyset$
$\mathcal{D}_{\text{SYNTH}} \leftarrow \emptyset$
▷ Content sourcing using retrieval:
for $(x, y) \in \mathcal{D}_{\text{SEED}}$ do
    $[r_1, \dots, r_K] \leftarrow \mathcal{M}_{\text{ret}}(x)$
    $\Gamma_K \leftarrow [r_1, \dots, r_K]$
    $\mathcal{D}_{\text{RETR}} \leftarrow \mathcal{D}_{\text{RETR}} \cup \{(x, y, \Gamma_K)\}$
▷ In-context learning set construction:
for $(x, y, \Gamma_K) \in \mathcal{D}_{\text{RETR}}$ do
    for $r_k \in \Gamma_K$ do
        $\mathcal{D}_{\text{ICL}} \leftarrow \mathcal{D}_{\text{ICL}} \cup \{(r_k, x)\}$ if $s_\alpha \leq \cos(x, r_k) \leq s_\beta$
▷ Task inversion:
for $(x, y, \Gamma_K) \in \mathcal{D}_{\text{RETR}}$ do
    for $r_k \in \Gamma_K$ do
        $\mathcal{D}_{\text{SHOTS}} \sim \mathcal{D}_{\text{ICL}}$
        for $j \in [1, \dots]$ until $\tilde{x}^i_j =$ <eos> do
            $\tilde{x}^i_j \sim \mathcal{M}_{\text{LM}}\left(\cdot \mid \tilde{x}^i_{<j}, \mathcal{P}_\tau(R_{inv}, r_k, \mathcal{V}(y)), \mathcal{D}_{\text{SHOTS}}\right)$
        $\mathcal{D}_{\text{SYNTH}} \leftarrow \mathcal{D}_{\text{SYNTH}} \cup \{(\tilde{x}^i, y)\}$
return $\mathcal{D}_{\text{SYNTH}}$
```

Step 2. In-context set construction: The subsequent task inversion step also benefits from in-context demonstrations, but it is challenging to construct demonstrations which effectively captures our context refinement task $r_k^i \to \tilde{x}^i$. We explored two approaches to in-context learning.

1. RETRICL: we use retrieval to construct a set of ICL examples $\mathcal{D}_{\text{ICL}}$, such that each ICL example mirrors the format of our task-inversion prompts. We select top-1 and top-2 retrieved results from the densely retrieved results, and use a cosine-similarity criterion $s_\alpha \leq \cos(x_{\text{ICL}}, r_k) \leq s_\beta$ to asses the potential match between the retrieved document r_k and x_{ICL}. Although the in-context pair may not match exactly, when used with an appropriate prompt template (Appendix H), they demonstrate the required format.

2. NON-RETRICL: a baseline method, which uses retrieval for content sourcing, but not for in-context learning. For each generation we select $N = 32$ ICL examples at random from $\mathcal{D}_{\text{SEED}}$. Each example is appended with a prefix like *"News Article:"* or *"Product details:"* but we do *not* add the context refinement instruction. After the ICL examples, we append the retrieved document r_k and context refinement instruction R_{inv} to form the final prompt. This format closely mirrors the in-context learning prompt used by FewGen, but also incorporates content-sourcing elements r_k and R_{inv}. This baseline highlights the value added by constructing $\mathcal{D}_{\text{ICL}}$ in the RetrICL approach.

Step 3. Task inversion using context refinement: The minimum elements of a task inversion prompt $\mathcal{P}_\tau$ are the context refinement instruction $\mathcal{I}_{inv}$ and target y. We use a verbalizer function $\mathcal{V}$ (Schick and Schütze, 2021; van de Kar et al., 2022) to provide a unique text representation of each label, i.e. $\mathcal{V} : \mathcal{Y} \to \{v_1, \ldots, v_C\}$. We follow prior work on classification-based task inversion (Schick and Schütze, 2021; Ye et al., 2022a,b; Yu et al., 2023b; Gao et al., 2023) and use descriptive verbalizations to induce label-separability in the final dataset.

FewGen uses the standard causal language modeling objective to induce next-token probabilities from teacher LLM, $\mathcal{M}_{\text{LM}}$. Nucleus sampling (Holtzman et al., 2019) is used to autoregressively sample next tokens until the `<eos>` token is generated. This becomes synthetic example $\tilde{x}^i$.

$$\tilde{x}^i_j \underset{p}{\sim} \mathcal{M}_{\text{LM}}\left(\cdot|\tilde{x}^i_{<j}, \mathcal{P}_\tau(I_{inv}, \mathcal{V}(y))\right) \qquad (1)$$

For each label y, we fix this prompt and sample m/C times to generate the synthetic dataset.

In SynthesizRR, we create the synthetic dataset from each triplet in $\mathcal{D}_{\text{RETR}}$. The retrieved documents $\Gamma_K = [r_1, \ldots, r_K]$ have lexical and semantic overlap with the query x_{ICL}. However, corpus documents may be distributionally dissimilar from real task covariates, due to the nature of documents or chunking process (Mialon et al., 2023). To address this, we use $\mathcal{M}_{\text{LM}}$ to perform task inversion from the content of each retrieved document, a process we refer to as *contextual refinement*. $\mathcal{P}_\tau$ is thus composed from the contextual refinement instruction $\mathcal{R}_{inv}$, each document $r_k \in \Gamma_K$, and the verbalized target for the query, i.e. $\mathcal{V}(y_{ICL})$. The LLM's context window thus sees a unique and grounded prompt when auto-regressively generating each synthetic input $\tilde{x}^i$:

$$\tilde{x}^i_j \underset{p}{\sim} \mathcal{M}_{\text{LM}}\left(\cdot|\tilde{x}^i_{<j}, \mathcal{P}_\tau(R_{inv}, r_k, \mathcal{V}(y_{ICL}))\right), \quad (2)$$

for all documents $r_k \in \Gamma_K$. We continue to use nucleus sampling to get diverse generations. Each original in-context example thus produces K unique synthetic examples $\{\tilde{x}_1, \ldots, \tilde{x}_K\}$; we call K the "expansion factor". To promote adherence to $\mathcal{R}_{inv}$, we sample pairs from $\mathcal{D}_{\text{ICL}}$ to create in-context examples following the same format. Our final dataset is constructed as:

$$\mathcal{D}_{\text{SYNTH}} = \bigcup_{(x,y,\Gamma_K)\in\mathcal{D}_{\text{RETR}}} \; \bigcup_{r_k\in\Gamma_K} \{(\tilde{x}^i, y)\}.$$

Dataset	Class	Train, Test	Corpus	Difficulty
AG News	4	115k, 7.6k	RN/Dom	Easy
ToI Headlines	10	52k, 10k	RN/Ind	Easy
Hyperpartisan	2	516, 65	RN/Dom	Medium
Polarity	2	72k*, 7.2k*	Products	Medium
Category	23	30k*, 2.4k*	Products	Medium
Humor	2	15k, 3k	Products	Hard
IMDb	2	20k, 25k	Movies	Medium
SST-2	2	54k, 872	Movies	Medium

Table 1: Dataset statistics and our estimate of task inversion difficulty. *Downsampled for convenience.

Step 4. Student distillation: The student is fine-tuned on $\mathcal{D}_{\text{SYNTH}}$ by passing the BERT `[CLS]` token embedding of $\tilde{x}$ through a feedforward layer. This produces a probability distribution over the label space C. We optimize the cross-entropy loss of the true label y. As we derive $\tilde{x}$ from a teacher LLM, this can be considered a form of symbolic knowledge distillation (West et al., 2022).

4 Experimental Setup

Tasks and their difficulty. We perform our main experiments on the first 6 datasets in Table 1, selected carefully to measure how the teacher LLM performs on task inversion tasks of varying difficulty. Previous work only benchmarked sentiment and topic classification datasets like IMDb (Maas et al., 2011) and AG News (Zhang et al., 2015). We broaden from topic classification, which primarily involves summarization during the task inversion step, which LLMs are adept at (Goyal et al., 2022). Hyperpartisan (Kiesel et al., 2019) detects bias in political news, so the task inversion step includes a more substantial rewriting of neutral retrieved articles to form biased examples. Category and Polarity are popular tasks from prior work (Yu et al., 2023a). In our setting, we generate product reviews from retrieved products, and must ensure the review belongs to the correct categorical and sentiment classes. Task inversion for Humor (Ziser et al., 2020) involves generating

Corpus	Domain	Size	Doc.	Tokens
RealNews/Dom	US/EU News	30.1M	Article	27.1B
RealNews/Reg	Regional News	2.7M	Article	2.1B
RealNews/Ind	Indian News	0.9M	Article	0.6B
Products	E-commerce	15.0M	Product	2.3B
Movie Summary	Movies	42K	Plot	0.02B

Table 2: Corpus statistics with LLaMa2 tokenizer.

humorous questions from retrieved product details, which requires additional skills from the teacher. Prompts for all tasks are in Appendix H.

Table 2 describes corpora used for retrieval. We consider five corpora in different domains, each with varying numbers of records. Three are subsets of RealNews (Zellers et al., 2019), as described in Appendix J: RealNews/Dominant (US/EU News), RealNews/Regional (Regional News), RealNews/India (Indian News). We also use Products (Amazon products metadata, (Ni et al., 2019)) and Movie Summary (movie summaries, (Bamman et al., 2013). Each task in Table 1 is associated with the corpus we consider most relevant. In §7, we compare to four prior approaches on three other tasks: IMDb (Maas et al., 2011), SST-2 (Socher et al., 2013) and AG News. These sentiment and topic tasks are less aligned with our goals and thus excluded from our main evaluation.

Models. We use Contriever (Izacard et al., 2022) for dense retrieval from each corpus. This performs a semantic match between the query and each document using cosine-similarity. In Appendix F, we also perform an ablation study using BM25 as a sparse retriever, which does lexical matching between each query-document pair.

As **teacher models**, we primarily use a frozen Llama-2 Chat 13B (Touvron et al., 2023b) for the task inversion step in SynthesizRR and FewGen. We also experiment with Claude Instant-v1 as described in Appendix K. For in-context learning (ICL) (Brown et al., 2020), we select examples randomly from the train set: 50 ICL examples/class for multi-class and 100/class for binary tasks. We believe this is a realistic number of examples that a system designer could source if they were to put some effort into building a specialist model. We explore approaches to bootstrap this seed set in limited-supervision settings Appendix D.

Specialization performance is measured on **student LMs** DeBERTa-v3-Large (435M params, He et al. (2021)) and DistilBERT (66M params, Sanh et al. (2019)).

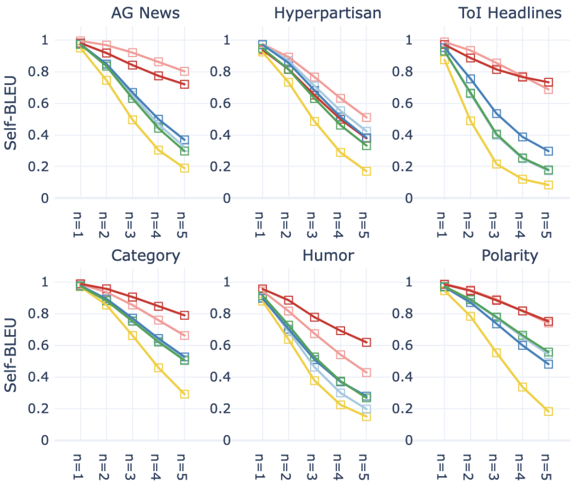


Figure 3: Self-BLEU (↓) for ngrams n=1-5. Comparison: Gold, FewGen 0-shot, FewGen 32-shot, SynthesizRR 0-shot, SynthesizRR 3-shot RetrICL, SynthesizRR 32-shot Non-RetrICL.

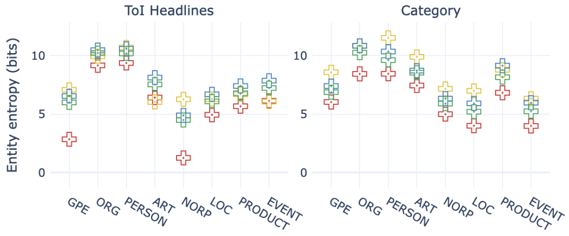


Figure 4: Entity entropy (↑) on ToI (headlines) and Category (reviews). Comparison: Gold, FewGen 32-shot, SynthesizRR 3-shot RetrICL and SynthesizRR 32-shot Non-RetrICL. Zero-shot results are similar for SynthesizRR and worse for FewGen; we omit them.

Evaluation criteria. Text generation can be challenging to evaluate objectively in multi-task scenarios (Chang et al., 2024). Therefore in §5 we evaluate synthetic text based on several criterion, to detect behaviours we observe during synthesis as in Table 3. **Self-BLEU** (Papineni et al., 2002; Zhu et al., 2018) measures lexical diversity of the dataset based on n-gram overlap between pairs of examples. **Entity entropy** measures the *diversity of entities* using the probability distribution of each of 16 entity-types, inferred using spaCy's `en_core_web_lg` (Honnibal et al., 2020). Datasets which over-represent popular entities score lower on entropy. On the other hand, **Entity recall** and **Entity KL divergence** compares the *similarity of entities* compared to Gold, and datasets which reproduce entities frequently seen in Gold data score higher. **MAUVE** (Liu et al., 2021) measures similarity to human-written text by using pretrained representations from a gpt2-xl model, indicating distributional differences in the generated text.

Method	Example
Gold	There is decent bass, but the highs are a bit soft. A quick tweak to my equalizer, and they're great. After reading several of the reviews, I was a bit worried about the sound, but now that I have them I'm very happy. They're a good price, and sooooo much better than the little ipod-like earbuds I've tried before. Those never stayed in my ear, and the bass never made me happy
FewGen	I've been a very happy customer of this company for a long time. It is fast and does everything I need it to. I would definitely recommend it to anyone looking for a good external drive. However, I do have one issue with the product. The instructions that come with it are not very clear and I had a hard time figuring out how to properly use it.
(Retrieved Product)	*Portable Laptop Microphone. Connects to 1/8" mini microphone input on laptop. Right-angle shaped. Flat-frequency response.*
SynthesizRR	The portable laptop microphone is right-angled and has a flat-frequency response, making it easy to use for online meetings and interviews. It connects to the 1/8" mini microphone input on my laptop and has worked great for the past two months, but I have noticed some distortion in the audio when I move around too much. Overall, it's a great value for the price and has made my remote work and video conferencing much more productive and efficient.

Table 3: Real and synthetic examples from "electronics" class of Category. Grey text indicates lack of specifics.

Method	Norp	Org	Person	Gpe	Recall (↑)	KL div. (↓)
			Unique Entities			
Gold	319	3943	3952	712	-	-
FewGen*	43	480	400	73	0.05	-
SynzthRR[†]	137	2718	1528	238	**0.12**	-
SynzthRR[‡]	109	1755	1012	178	0.10	-
			Total Entities			
Gold	843	7233	6096	1558	-	-
FewGen*	94	775	506	96	0.23	3.10
SynzthRR[†]	319	3991	1989	397	**0.35**	**2.35**
SynzthRR[‡]	314	2699	1464	363	0.32	2.52

Table 4: Entity similarity in Category (8K). We show the counts of unique and total entities for 4 entity-types. *Entity recall* measures the fraction of Gold entities co-occuring in the synthetic data; in the bottom half, we additionally weigh each entity by its frequency in Gold. Notation: *32-shot; [†]3-shot RetrICL; [‡]32-shot Non-RetrICL.

Method *(Dataset size)*	AG. (8K)	Hyp. (2K)	ToI (8K)	Cat. (8K)	Hum. (2K)	Pol. (4K)
			Zero shot			
FewGen	56.6	53.7	62.8	**63.2**	75.6	62.8
SynzthRR	**90.3**	**59.2**	**63.0**	61.1	**82.9**	**78.6**
			Few shot			
FewGen*	56.7	65.4	60.3	65.8	78.1	69.2
SynzthRR[†]	**92.0**	**72.8**	**87.9**	**75.2**	**87.5**	**89.9**
SynzthRR[‡]	91.8	67.9	67.2	75.1	87.0	83.2

Table 5: MAUVE similarity score (↑) using GPT2-XL embeddings. Notation: *32-shot; [†]3-shot RetrICL; [‡]32-shot Non-RetrICL.

5 Results: Intrinsic Evaluation

In this section, we focus on evaluating intrinsic properties of the generated datasets, including their diversity and entity coverage. We focus on a LLaMa-2 Chat 13B teacher LLM, retrieving from Contriever using corpora per Table 1 (we analyze changing the retrieval corpus in Appendix E). We generate datasets of size in relation to the number of Gold rows: 8K rows (AG News, ToI Headlines, Category), 4K rows (Polarity) or 2K rows (Hyperpartisan, Humor). Example generations are in Appendix I.

RQ: Does retrieval augmentation improve lexical diversity? Figure 3 shows lexical diversity within the dataset. Human-written texts (Gold) score high on lexical diversity (low Self-BLEU). FewGen texts tend to reuse the same words and phrases, leading to repeated text across generations (high Self-BLEU). SynthesizRR text has lexical diversity approaching human text for all n-gram values. We note in-context learning has an inconsistent effect; it improves the lexical diversity for news corpora but not for products.

RQ: Does SynthesizRR address entity diversity? *Popularity bias* is a phenomenon wherein LLM generations tend to over-represent popular "head" entities. This has been studied for QA tasks (Mallen et al., 2023; Kandpal et al., 2023).

In Figure 4 we see how SynthesizRR eliminates popularity bias across entity types. By sourcing from the long-tail of retrieval results ($k = 50$), the generated dataset has much higher entity entropy compared to FewGen. This positions SynthesizRR closer to Gold, which also shows high entity entropy.

RQ: How is entity similarity in synthetic data affected by grounding to an in-domain corpus? For the Category task we generate 8K product

Method *(Dataset size)*	Teacher LM	AG. (8K)	Hyper. (2K)	ToI (8K)	Categ. (8K)	Humor (2K)	Polar. (4K)	Avg
Gold		91.0	93.2	82.5	81.5	93.1	95.3	89.43
				Zero shot				
FewGen	LLaMa2	69.5	**72.6**	32.1	62.4	74.4	81.0	65.32
FewGen	ClaudeV1	75.0	57.5	23.3	47.1	49.9	87.5	56.72
SynthesizRR	LLaMa2	83.5	69.8	**74.4**	**68.9**	**82.5**	84.7	**77.32**
SynthesizRR	ClaudeV1	**83.9**	72.3	71.8	66.8	62.1	**88.7**	74.29
				Few shot				
FewGen*	LLaMa2	84.2	74.5	**73.7**	68.6	88.4	90.9	80.05
FewGen*	ClaudeV1	75.9	58.5	72.2	68.8	82.9	91.2	74.93
SynthesizRR†	LLaMa2	83.0	78.5	73.3	**72.4**	**90.2**	91.0	**81.38**
SynthesizRR‡	LLaMa2	**85.2**	**79.1**	72.8	71.9	88.8	88.2	81.00
SynthesizRR†	ClaudeV1	83.7	72.3	72.8	65.4	83.4	**91.3**	78.16
SynthesizRR‡	ClaudeV1	83.7	72.0	72.5	67.8	76.2	87.9	76.68

Table 6: Test Accuracy (↑) after distilling DeBERTa-v3-Large student from LLaMa-2 Chat 13B and Claude Instant-v1. Contriever was used as the retriever in SynthesizRR. We report the average of 5 runs and rerun in cases where std. dev. $\geq$6% (indicating one or more models failed to converge). The top half considers zero-shot synthesis and bottom half uses in-context learning, and we **bold** the best result under each paradigm. Notation: *32-shot; †3-shot RetrICL; ‡32-shot Non-RetrICL.

reviews and randomly select 8K Gold examples. In Table 4, we measure *entity recall*, and find that the occurrence of Gold entities is 100%-140% higher in SynthesizRR than FewGen. The KL divergence of each entity distribution is also lower. We finally consider the *entity coverage* (unique entities) and *entity density* (total entities). Compared to Gold, FewGen tends to produce fewer unique entities (places, events, languages, currencies, etc). Each FewGen example also has a lower density of entities, as visible in Table 3. SynthesizRR coverage and density more closely match Gold.

RQ: How distributionally similar are our generated examples and human-written examples? We see from MAUVE scores in Table 5 that zero-shot generations are quite dissimilar in both approaches compared to few-shot methods.

Surprisingly, SynthesizRR generations are much more similar to human text than FewGen, despite the fact that nothing in our content sourcing strategy explicitly guides SynthesizRR generations to match the distribution of Gold. We thus manually inspect generations and discover an interesting pattern which can be attributed to content sourcing. As shown earlier, and in Table 3, the density of entities is higher under SynthesizRR. FewGen produces generations which obey the prompt, but are very bland and do not include specifics. On the other hand, by obtaining information-rich documents, SynthesizRR is able to ground the task inversion step in details of the retrieved article/product. We hypothesise that this improves the MAUVE score towards Gold, which is similarly grounded in specifics.

6 Results: Student distillation

We have established that SynthesizRR generates more diverse datasets compared to a baseline approach. Now, we return to the application of training a specialist model based on these datasets.

Table 6 shows the results of training a DeBERTa-v3-Large student on datasets generated by SynthesizRR and FewGen. In the zero-shot setting, we find that SynthesizRR performs much better than FewGen, despite using the same frozen teacher LLM. Note that SynthesizRR uses in-context examples for retrieval here whereas FewGen does not; our method has some additional supervision here. However, in this setting, we see clear gains during the task inversion stage (↑12% for LLaMa and ↑17.6% for Claude). Thus, having access to retrieval yields a better final dataset, almost on par with 32-shot FewGen.

With ICL, 3-shot SynthesizRR using the RetrICL strategy trains better students than 32-shot FewGen (↑1.3% for LLaMa and ↑3.2% for Claude) and Non-RetrICL. We conclude that naively adding ICL examples is not an effective use of the LLM's context window. Instead, a better content sourcing strategy improves the student distillation, leading to better test performance.

Method *(Dataset)*	**LM**	MAUVE (↑) AG.	MAUVE (↑) IMDb	MAUVE (↑) SST-2	Accuracy (↑) AG.	Accuracy (↑) IMDb	Accuracy (↑) SST-2
Gold	-	-	-	-	90.8	91.3	88.2
SunGen	gpt2-xl	⧖	68.7	⧖	⧖	84.9	⧖
ReGen	BERT	68.1	⧖	⧖	82.7	⧖	⧖
S3	gpt3.5	⊗	62.0	⊗	⊗	**87.1**	⊗
AttPmt	gpt3.5-t	52.8	⧖	50.0	79.8	⧖	80.8
			Zero shot				
(Ours)	LLaMa	89.5	58.5	50.0	85.3	82.9	80.2
(Ours)	Claude	94.2	55.9	50.0	85.6	83.6	82.5
			3-shot RetrICL				
(Ours)	LLaMa	92.6	**72.6**	50.0	84.6	84.8	**83.8**
(Ours)	Claude	**95.8**	58.0	50.0	**86.0**	86.3	80.6

Table 7: MAUVE and distillation accuracy on synthetic datasets released by prior work, subsampled to 6K examples as per Appendix B. For our method, we retrieve using Contriever, and generate 6K examples using teachers LLaMa 13B Chat and Claude Instant-V1.2. Accuracy reports the average of five DistilBERT training runs using Yu et al. (2023a)'s hyperparameters (std. dev. ≤ 2.0 in all cases). Best results for each task are indicated in **bold**. Tasks not covered by prior work are marked ⊗; those evaluated without dataset release are marked ⧖.

7 Results: Comparison to prior work

We compare SynthesizRR to four competitive prior approaches: SunGen (Gao et al., 2023), ReGen (Yu et al., 2023b), S3 (Wang et al., 2023a) and AttrPrompt (Yu et al., 2023a). Table 7 evaluates overall similarity to human text and distillation accuracy, with complete details in Appendix B.

We observe SynthesizRR outperforms approaches that generate high-diversity covariates (AttrPrompt) or use content sourcing (ReGen). Even with a fixed student model, it enhances accuracy over methods that leverage student feedback (SunGen), and in Appendix C we see that student feedback can further improve the accuracy. Approaches like S3 which use iterative prompting with Chain-of-Thought reasoning (Wei et al., 2022) can provide minor accuracy improvements, but the generations are less realistic. We finally observe that ReGen, which only uses retrieval, suffers in terms of lexical diversity and student accuracy; task inversion is necessary to transform retreived contexts to match human-written covariates.

We emphasize that sentiment and topic classification are simple synthesis tasks. We include them for comparison to prior work, but believe that our experiments on more challenging tasks better represent the capacity of LLMs for dataset synthesis.

8 Related Work

Dataset synthesis using LLMs. Using LLMs to perform *task inversion* for dataset synthesis has been studied previously. Most use GPT-2XL without fine-tuning (Ye et al., 2022b,a; Gao et al., 2023; Meng et al., 2022; Schick and Schütze, 2021; Jung et al., 2023). Recent work has considered large teacher LLMs such as GPT-3 (West et al., 2022; Honovich et al., 2023; Wang et al., 2023b), PaLM-540B (Hsieh et al., 2023) and chat-tuned LLMs such as gpt-3.5-turbo (Yu et al., 2023a; Yehudai et al., 2024b; Wang et al., 2023a).

For the generation of text classification datasets, class-conditioned prompting is key. Prior approaches investigated zero-shot (Ye et al., 2022a) and iterative few-shot prompting (Ye et al., 2022b), or synthesis using seq2seq LLMs fine-tuned on a curated dataset (Lee et al., 2021). Recently, AttrPrompt (Yu et al., 2023a) established that varying prompt attributes improves diversity. Our work explores adding retrieval contexts as the source of diversity.

Retrieval-augmented generation. Our approach has many of the characteristics of in-context retrieval-augmented generation (RAG) (Lewis et al., 2020; Ram et al., 2023; Huang et al., 2023; Izacard et al., 2023). Previous studies show how RAG bypasses numerous problems associated with generating solely from parametric memory, i.e., heightened bias towards "head" entities (Mallen et al., 2023), lower lexical diversity (Holtzman et al., 2019; Jentzsch and Kersting, 2023), and hallucinated information (Zhang et al., 2023).

Using retrieval-augmented generation for synthesis of classification tasks has not been explored at the instance level. ReGen (Yu et al., 2023b) studies the retrieval-only setting for creation of topic and sentiment datasets, which are simpler than the tasks in our work. Viswanathan et al. (2023) and Gandhi et al. (2024) perform dataset-level retrieval and not instance-level retrieval.

9 Conclusion

In this work we describe how a retrieval corpus can be used to aid the synthesis of a text classification data set in specialized domains. We show that the diversity of the generated data is enhanced by including retrieved documents in a generation prompt. Compared to few-shot generation, we find that SynthesizRR produces more diverse and representative text and leads to better students.

Limitations

Most principally, our work relies on the existence of a large corpus that is close enough to the task at hand. This may be prohibitive for doing dataset generation in low-resource languages, where a large corpus of related content may not be available. It would be intriguing to explore cross-lingual transfer of content sourcing, but this would require additional experimental validation. By contrast, approaches like FEWGEN do not require this corpus.

The need for an explicit context sourcing step and increased prompt-length causes an increase in the expenses and latency, especially when using LLM APIs. Such increased expense may not be worth it in the presence of a poor quality retrieval corpus. For one, if the in-context examples are not easily reusable as queries, then SYNTHESIZRR can retrieve irrelevant documents which might not be suitable for task inversion. Furthermore, in the case of factually dubious corpus documents, the student model may end up grounding in factually incorrect information. This can be mitigated by a human-in-the-loop step to remove such documents before task inversion.

Finally, we note that the scope of our experiments is restricted to a set of classification tasks over a few English domains of text. While we believe our approach can be applied to other languages, other domains, and tasks like question answering that go beyond classification, we have not validated this in this work.

References

Anthropic. 2023. Claude v1.2 instant. `https://www.anthropic.com/news/releasing-claude-instant-1-2`.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv e-prints*, pages arXiv–2303.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets.

Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *arXiv preprint*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. 2023. Raven: In-context learning with retrieval augmented encoder-decoder language models. *ArXiv*, abs/2308.07922.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2023. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. *arXiv preprint arXiv:2305.16635*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Timothy Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *ArXiv*, abs/2102.01335.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2021. Divergence Frontiers for Generative Models: Sample Complexity, Quantization Effects, and Frontier Integrals. In *Advances in Neural Information Processing Systems*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. ArXiv:2307.03172.

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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *Transactions on Machine Learning Research*. Survey Certification.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2022. Gpt-3.5 (text-davinci-003). `https://platform.openai.com/docs/models/gpt-3-5-turbo`.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints*, pages arXiv–2307.

Mozes van de Kar, Mengzhou Xia, Danqi Chen, and Mikel Artetxe. 2022. Don't prompt, search! mining-based zero-shot learning with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7508–7520, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vijay Viswanathan, Chenyang Zhao, Amanda Bertsch, Tongshuang Wu, and Graham Neubig. 2023. Prompt2Model: Generating deployable models from natural language instructions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 413–421, Singapore. Association for Computational Linguistics.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. 2023a. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11817–11831, Singapore. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. Zerogen: Efficient zero-shot learning via dataset generation. *ArXiv*, abs/2202.07922.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024a. Achieving human parity in content-grounded datasets generation. In *International Conference on Learning Representations*.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024b. Genie: Achieving human parity in content-grounded datasets generation. *ArXiv*, abs/2401.14367.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023a. 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*.

Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. 2023b. ReGen: Zero-shot text classification via training data generation with progressive dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805, Toronto, Canada. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *SIGIR*.

Yftah Ziser, Elad Kravi, and David Carmel. 2020. Humor detection in product question answering systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 519–528, New York, NY, USA. Association for Computing Machinery.

A Risks

Although the main goal of our work is to improve text classification, our use of LLMs to generate examples does carry some conceptual risks. By generating news articles to train classifiers on, we run the risk of generating fake news and other harmful content. However, we believe this risk is mitigated by the fact that the final outcome of our system is a classifier: classification models have relatively constrained failure modes (misclassification) compared to text generation models that can mislead users. Furthermore, we do not believe our approach uniquely advances the generation of content like fake news; our advances are largely orthogonal to the technology that brings such risks.

B Detailed comparison to previous work

Here, we explore how SynthesizRR directly compares to prior work on synthesis of popular datasets. We compare against four prior approaches:

SunGen (Gao et al., 2023): uses the ZeroGen strategy to generate a large synthetic dataset (200k rows). Then, uses a custom bi-level optimization algorithm (involving the student model) to determine instance-weights of each synthetic example.

ReGen (Yu et al., 2023b): performs multi-round filtering of retrieved results using 2 BERT models; one trained for retrieval, and one classifier. Use consistency between these models to filter noisy data.

S3 (Wang et al., 2023a): Constructs a "seed dataset" (different from ours) and trains a student model. Then, extrapolate errors using an LLM and synthesizes additional data. We combine this with the seed data and repeat the process.

AttrPrompt (Yu et al., 2023a): a method focused on improving diversity and unbiasedness of generated datasets. Prompts a powerful LLM like GPT3.5-Turbo with different attributes, each along different dimensions. Attributes are extracted from a human-in-the-loop analysis of task using GPT3.5-Turbo.

Standard zero-shot and few-shot generation baselines were compared in Table 6, so we do not include them here.[1]

We benchmark three classification tasks which are popular in prior work: IMDb (Maas et al., 2011), SST-2 (Socher et al., 2013) and AG News (Zhang et al., 2015). The first two tasks are binary sentiment analysis on movie reviews, while the latter is multi-class topic classification on news.

Prior work generates much larger datasets (20k to 200k examples) and uses different student model hyperparameters. Intrinsic evaluations of dataset quality are also seldom reported. This makes it difficult to fairly compare results. Thus, we reproduce results ourselves by using the synthetic datasets released by authors.[2] Following Yu et al. (2023a), we subsample these datasets to 6,000 rows keeping a uniform distribution across classes, and generate the same number of synthetic covariates using SynthesizRR RetrICL (Algorithm 1). For the content sourcing stage of SynthesizRR, we retrieve documents from the following corpora:

- **Movies:** to generate movie reviews for IMDb and SST-2, we retrieve from the CMU Movie Summary corpus (Bamman et al., 2013), which contains 42k plot summaries.

- **RealNews/Dom:** for AG News we use RealNews/Dominant from Table 2, which contains 30M news articles from US, EU countries, UK, and Australia, which is the "dominant" portion of RealNews (see Appendix J for complete details).

DistilBERT (Sanh et al., 2019) is widely used in prior work (Yu et al., 2023a; Ye et al., 2022a; Gao et al., 2023; Wang et al., 2023a; Ye et al., 2022b), and thus we use it as the student model to measure accuracy. We use the same training hyperparams as Yu et al. (2023a), i.e. Adam optimizer (Kingma and Ba, 2015) for 5 epochs using `lr=2e-5`, `batch_size=32`, `weight_decay=1e-4` and `epsilon=1e-6`, and linear learning rate warmup for 6% of training steps.

RQ: How does SynthesizRR compare to existing approaches in terms of distilled student accuracy?

Methods like SunGen which rely on relatively weak LLM teachers like GPT2-XL (Radford et al., 2019) can perform well on topic and sentiment tasks like IMDb, but require a very high data cost (15-30x more synthetic data than SynthesizRR). In Table 8, we observe that when scaled down to 6k rows, the performance deteriorates significantly.

Approaches which use strong instruction-following LLMs like AttrPrompt, S3, and SynthesizRR can achieve similar or better performance with much smaller datasets. These methods create high-quality datasets rather than modify the student modeling process, as is done by SunGen and ProGen. SunGen performs an iterative bi-level optimization over the ZeroGen

---

[1] ZeroGen (Ye et al., 2022a) is similarly not considered.

[2] ProGen (Ye et al., 2022b) is a relevant technique but does not release datasets.

Method *(Dataset)*	**Retriever**	**Teacher LLM**	Self-BLEU-5 (↓)			Entity Entropy (↑)			Mauve (↑)			Accuracy (↑)		
			AG.	IMDB	SST-2	AG.	IMDB	SST-2	AG.	IMDB	SST-2	AG.	IMDB	SST-2
GOLD	-	-	17.1	27.9	35.5	6.6	7.5	3.2	-	-	-	90.8	91.3	88.2
SUNGEN	-	GPT2-XL	⧖	**15.4**	⧖	⧖	4.9	⧖	⧖	68.7	⧖	⧖	84.9	⧖
REGEN	BERT	-	56.5	⧖	⧖	**8.1**	⧖	⧖	68.1	⧖	⧖	82.7	⧖	⧖
S3	-	GPT3.5	⊗	62.2	⊗	⊗	5.7	⊗	⊗	62.0	⊗	⊗	**87.1**	⊗
ATTPMT	-	GPT3.5-T	39.8	⧖	71.5	6.0	⧖	3.4	52.8	⧖	50.0	79.8	⧖	80.8
ZERO SHOT														
SYNZTHRR	CONTR.	LLAMA2	29.3	66.3	41.9	7.1	5.7	4.5	89.5	58.5	50.0	85.3	82.9	80.2
SYNZTHRR	CONTR.	CLAUDEV1	31.5	51.5	45.3	6.6	5.3	4.8	94.2	55.9	50.0	85.6	83.6	82.5
SYNZTHRR	BM25	LLAMA2	28.7	62.2	36.5	7.0	5.6	5.1	90.3	60.5	50.0	84.3	74.1	**84.4**
SYNZTHRR	BM25	CLAUDEV1	30.9	50.4	36.9	6.5	5.1	**5.4**	90.8	53.2	50.0	84.2	79.1	82.6
3-SHOT RETRICL														
SYNZTHRR	CONTR.	LLAMA2	34.2	62.9	26.3	7.2	5.7	3.8	92.6	72.6	50.0	84.6	84.8	83.8
SYNZTHRR	CONTR.	CLAUDEV1	**23.7**	38.0	**24.6**	6.7	**5.9**	4.3	95.8	58.0	50.0	**86.0**	86.3	80.6
SYNZTHRR	BM25	LLAMA2	32.0	59.7	25.3	7.2	5.6	4.8	92.5	**78.7**	50.0	84.3	84.7	**84.4**
SYNZTHRR	BM25	CLAUDEV1	24.6	41.9	26.8	6.7	5.4	4.9	**96.0**	58.5	50.0	84.1	81.6	82.3

Table 8: Evaluations of synthetic datasets released by prior work. We subsample all to 6K examples (uniformly distributed across classes) before computing metrics as described in §4. Tasks not evaluated by previous authors are denoted by ⊗ while those evaluated without dataset release are marked ⧖. GPT3.5 is text-davinci-003 whereas GPT3.5-T is gpt-3.5-turbo (OpenAI, 2022), LLAMA2 is 13B Chat version (Touvron et al., 2023a), CLAUDEV1 is Instant-V1.2 version (Anthropic, 2023). Accuracy is measured on a DISTILBERT student, where we train 5 student models and report the mean accuracy (std. dev. was ≤ 2.0 in all cases). Within each dataset, we **bold** the best result. Within each dataset, we **bold** the best result.

datasets, jointly learning instance-weights and improving the student. We hypothesize these additions of the student model into the synthesis process also impact the final classification accuracy, as the dataset becomes specialized to the particular choice of student and its hyperparams. Under a standard student-distillation setup here, datasets from these approaches may not perform as well.

More complex prompting techniques like Chain-of-Thought (Wei et al., 2022) used by S3 can indeed improve the task-inversion step, though this requires much higher API costs due to longer output lengths. Chain-of-Thought prompting thus seems like a promising approach to augment SYNTHESIZRR's task-inversion step.

RQ: do we find evidence that content sourcing promotes diversity and similarity?

In Table 8, we measure the diversity of entities (Entity Entropy), lexical diversity (Self-BLEU), and similarity to GOLD texts (MAUVE) compared to prior approaches. Among prior approaches, only ATTRPROMPT (Yu et al., 2023a, Appendix E) attempts to improve diversity of the generated text, by templatizing the task inversion instruction with attributes such as `style, topic, length:min-words` and more. REGEN is the only synthesis approach which uses content sourcing (but not task inversion). We thus consider these the two most relevant baselines for SYNTHESIZRR.

We see that both REGEN and SYNTHESIZRR achieve very high entity entropy compared to ATTRPROMPT, thus underscoring the importance of a content sourcing step. Unlike SYNTHESIZRR, REGEN uses only retrieval and has no explicit task-inversion step to make the contexts similar to GOLD texts. Thus, we observe that REGEN suffers in terms of lexical diversity, MAUVE and student accuracy, compared to SYNTHESIZRR.

On the other hand, Chain-of-Thought prompting (S3), despite generating a strong classification dataset as we see in Table 8, suffers from the lack of lexical diversity and similarity to GOLD texts. This is also seen in ATTRPROMPT and previously in FEWGEN. This lends evidence to the hypothesis that synthesis without content sourcing tends to produce datasets with lower diversity, which cannot be overcome by complex prompting strategies alone.

Finally, we observe that SUNGEN shows very high diversity on IMDB, a task which involves generating movie reviews having positive or negative sentiment. However, as mentioned in Ye et al. (2022a, Section 4.6), ZEROGEN is not simply zero-shot generation; the authors first gener-

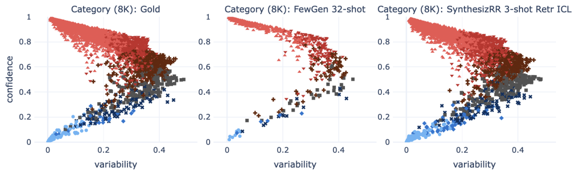


Figure 5: Data maps from a DistilBERT training run on 8K Category rows from LLaMa2. FewGen (center) is skewed towards easy-to-learn examples (top-left) while Gold (left) and SynthesizRR (right) have a higher density of ambiguous examples.

Method *(Dataset size)*	AG. (8K)	Hyp. (2K)	ToI (8K)	Cat. (8K)	Hum. (2K)	Pol. (4K)
Gold	93.8	81.6	85.2	84.8	95.5	96.6
LLaMa2 Few shot						
FewGen*	**92.4**	71.3	**85.9**	**88.1**	71.7	94.8
SynzthRR†	86.9	**78.6**	74.3	72.1	90.7	94.8
SynzthRR‡	87.6	75.5	74.9	74.5	**95.7**	**97.6**
ClaudeV1 Few shot						
FewGen*	**94.5**	63.8	**87.4**	**89.4**	85.9	99.6
SynzthRR†	87.6	**72.8**	74.8	69.4	**90.7**	99.3
SynzthRR‡	87.4	65.9	73.2	73.2	77.4	**99.7**

Table 9: Few-shot label-preservation accuracy (↑) using tuned oracle DeBERTa-v3L model. Gold row is accuracy on 20% validation split. Notation: *32-shot; †3-shot RetrICL; ‡32-shot Non-RetrICL.

ate a movie using the prompt `Movie:` and then insert the generated movie name into the prompt template `The movie review in positive sentiment for movie "<Movie>" is:` to generate an SST-2 example. SunGen which starts with ZeroGen-generated dataset and learns instance-weights (Gao et al., 2023, Section 2.2). We posit that the generated movie fulfils the same purpose as a retrieved context in SynthesizRR.

C Incorporating feedback from distilled student models

RQ: Why does SynthesizRR improve classification dataset synthesis? In this section we take a closer look at the generated classification dataset and how it affects the *training dynamics* of student models during distillation.

Aside from the final accuracy, we also consider **label preservation accuracy**, which is obtained from an "oracle" model for the task. We construct this oracle from Gold data by running a grid-search over DeBERTa-v3-Large hyperparams (Appendix K), splitting 80% of the Gold train set for fine-tuning and 20% for validation. Then, we measure the fraction of synthetic examples which the oracle classifies to belong to the prompted target class. This indicates the adherence of the generated example to the class it *should* belong to, as per the prompt.

We would expect that better label preservation means a higher-fidelity training dataset. However, Table 9 shows that FewGen datasets have very high label preservation in spite of their lower test performance. Especially on multiclass tasks (AG., ToI, Cat.), FewGen shows the highest label preservation (exceeding Gold) but this does not translate into improved student performance.

To understand this, we conduct a deeper analysis of the student training dynamics on multiclass datasets. We train a DistilBERT student for 6 epochs and plot the corresponding data-maps Swayamdipta et al. (2020). For binary tasks, the data-maps for SynthesizRR matched both FewGen and Gold, but the data maps from multi-class differed greatly. Figure 5 illustrates this difference using the Category task maps. From Figure 5 it is clear that FewGen generations tend to cluster around easy-to-learn examples (high confidence and low variability), whereas SynthesizRR contains more ambiguous examples (high variability) which Swayamdipta et al. (2020) demonstrate is essential to learning the nuances between classes.

RQ: Can we improve distillation performance by leveraging student feedback from data-maps?

Swayamdipta et al. (2020) use data-maps to filter out easy to-learn examples (top-left, red) and potentially mislabelled examples (bottom-left, blue) and obtain superior accuracy on human-generated datasets. We attempt to apply this same technique to the synthetic datasets generated by SynthesizRR and FewGen.

Concretely, we filter out the least ambiguous examples (bottom 17% variability) and retrain the DistilBERT student model on the smaller, filtered dataset. In Table 10 we find that FewGen performance degrades, whereas SynthesizRR improves (giving us new best performances on multi-class despite using only 83% of rows). We conclude that SynthesizRR generates more ambiguous examples, and this helps establish better class-separability in multi-class data sets.

Method *(Dataset size)*	AG. (6.6K)	ToI (6.6K)	Cat. (6.6K)	Avg
LLaMa2 Few shot				
FewGen*	58.0 ↓26.2	37.6 ↓36.1	48.0 ↓20.6	↓27.6
SynzthRR†	85.7 ↑2.7	76.0 ↑2.7	74.3 ↑1.9	↑2.4
SynzthRR‡	86.3 ↑1.1	75.0 ↑2.2	72.9 ↑1.0	↑1.4
ClaudeV1 Few shot				
FewGen*	71.8 ↓4.1	72.1 ↓0.1	69.3 ↑0.5	↓1.2
SynzthRR†	86.2 ↑2.5	75.3 ↑2.5	69.0 ↑3.6	↑2.9
SynzthRR‡	86.1 ↑2.4	74.6 ↑2.1	70.0 ↑2.2	↑2.2

Table 10: Test Accuracy (↑) after keeping 83% most-ambiguous examples. We report improvements compared to Table 6. Notation: *32-shot; †3-shot RetrICL; ‡32-shot Non-RetrICL.

D Bootstrapping with a synthetic seed set

A core assumption in SynthesizRR has been the existence of a small seed set of human-written (x, y) pairs for the task. This seed set is critical as it serves a dual purpose: it is used as the set of the retrieval queries, and as in-context learning examples to guide the teacher LLM's next-token distribution in the task inversion step.

In this section we consider how we can synthesize such a seed set for low-resource settings. Our core assumption is that the seed set is small (100/class for binary tasks and 50/class for multi-class tasks). Thus using FewGen with top-p = 0.9 and temperature = 0.95 and three in-context examples, we attempt to generate a diverse seed set with minimal repetitions. This bootstrapping approach makes SynthesizRR tractable when very little human data is available (just 5-15 examples per class) or no human data is available. We compare three paradigms:

1. **True zero-shot:** when we have no human data we utilize zero-shot generation to bootstrap the seed set.

2. **Low-resource:** Here, we assume we have a small number of human-written examples, e.g. 5 examples per class. This is presumed insufficient to be used as the seed set directly, but we can use it as in-context examples to guide the FewGen generator to bootstrap a realistic seed set.

3. **Sufficient:** We do not synthesize the seed set. This is the SynthesizRR paradigm we have explored in previous sections, wherein we have 50-100 Gold examples per class in our seed set.

Gold data (N)	RetrICL shots	AG. (8K)	Hyp. (2K)	ToI (8K)	Cat. (8K)	Hum. (2K)	Pol. (4K)
Gold							
All	-	91.0	93.2	82.5	81.5	93.1	95.3
True Zero-shot (0-shot FewGen seed)							
None	0-shot	66.6	68.0	60.5	60.4	76.9	76.4
None	3-shot	60.0	72.3	62.5	61.7	72.3	85.4
Low-Resource ($\binom{N}{3}$-shot FewGen seed)							
5/class	0-shot	79.9	71.7	68.1	63.4	81.3	81.3
5/class	3-shot	77.7	66.8	68.9	58.8	86.4	86.5
15/class	0-shot	78.5	72.9	69.3	65.7	77.4	84.0
15/class	3-shot	76.1	72.6	71.6	63.5	82.5	73.8
Sufficient (Gold Seed)							
Full seed	0-shot	**83.5**	69.8	**74.5**	68.9	82.5	84.7
Full seed	3-shot	83.0	**78.5**	73.3	**72.4**	**90.2**	**91.0**

Table 11: Test accuracy after distilling a DeBERTa-v3L student on a dataset generated by SynthesizRR RetrICL variant. We use the same corpus as Table 2, but vary the seed set. LLaMa2 is used as the teacher LLM. We train 5 student models and report the mean accuracy, rerunning all 5 in case of std $\geq$ 6.0. "'Full" seed implies 100 Gold examples per class for binary and 50 per class for multiclass tasks. We **bold** the best result in each dataset.

As mentioned in §4, the true zero-shot paradigm makes strong assumptions that are often unnecessarily restrictive. In practice, it is typically feasible to obtain a small amount of human-written examples (low-resource or sufficient seed), while obtaining several thousand human-written examples is still challenging.

The results of running SynthesizRR RetrICL using synthetic seed data is shown in Table 11. As a general trend, adding more human-written examples leads to better performance. Unsurprisingly, the best results are in the Sufficient paradigm, where we use 50-100 Gold examples as both retrieval queries and the the RetrICL set. True Zero-shot results (without any human input) are considerably worse. Surprisingly, however, we are able to get good distillation accuracy with just 5 examples per class rather than the full 50-100 per class, which indicates that SynthesizRR might be usable in low-resource settings where human annotated data is scarce.

In certain cases of the low-resource paradigm, we observe that the performance drops significantly from 0-shot RetrICL to 3-shot RetrICL. We attribute this to the fact that, even with 5-15 Gold in-context examples, the FewGen-generated seed

	AG News (4K)			
Corpus	DeBERTa (↑)	Mauve (↑)	Self-BLEU-5 (↓)	Entity Ent. (↑)
RN/Dom	**85.39 ± 0.8**	**92.58**	0.23	6.72
RN/Rnd	35.57 ± 6.1	83.39	**0.22**	**7.07**
RN/Reg	84.17 ± 0.7	88.88	0.26	6.72
	Hyperpartisan (2K)			
Corpus	DeBERTa (↑)	Mauve (↑)	Self-BLEU-5 (↓)	Entity Ent. (↑)
RN/Dom	**78.77 ± 2.8**	**66.94**	0.35	6.11
RN/Rnd	78.77 ± 3.5	61.45	**0.25**	**7.40**
RN/Reg	72.00 ± 2.0	65.59	0.35	6.12

Table 12: Effect of corpus-swapping for SynthesizRR 32-shot Non-RetrICL. We generate only 4k rows for AG News to reduce costs.

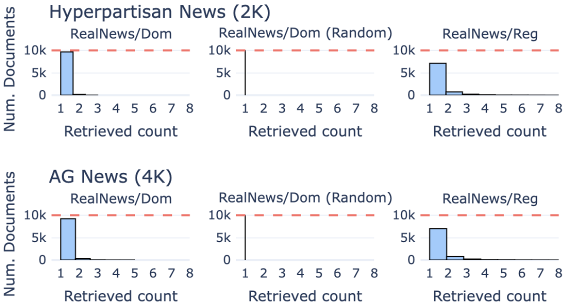


Figure 6: Retrieval counts for Hyperpartisan and AG News. The red dashed line represents the theoretical max, where all retrieved documents are unique. Note that the Random histogram plot is always 1 hence shows up as a straight line.

set might not be reflective of the true Gold examples (this behavior is reflected in the low MAUVE scores in Table 5). Thus, by conditioning on incorrect synthetic examples during RetrICL, we shift the next-token distribution away from the true distribution.

In conclusion, using FewGen to bootstrap a seed set can be a viable approach to using SynthesizRR in low-resource settings where there is not enough Gold task-data.

E Influence of retrieval corpus on domain shift

Our expectation is that SynthesizRR can flexibly specialize students to different domains by transparently changing the retrieval corpus, while keeping a frozen LLM. To quantify how changing the retrieval corpus might affect earlier metrics, we switch the news corpus for Hyperpartisan and AG News. We had assumed RealNews/Dom was the most suitable corpus (in-domain), and the others will cause domain-shift. In the following RQs, we validate the degree to which this assumption holds and the importance of information retrieval as the content sourcing mechanism in SynthesizRR.

RQ: Does modifying the corpus cause domain shift? Table 12 finds that the retrieval corpus highly influences the test performance (both student and intrinsic metrics). When grounding to a corpus with highly dissimilar entities (such as RealNews/Reg), all metrics drop significantly. Thus, we can conclude that an alternative content-source does indeed induce domain-shift. Mauve and distillation accuracy are highest for the in-domain corpus, while Self-BLEU and Entity entropy are highest for the random-retrieval results.

RQ: is retrieval essential for content sourcing? We measure the importance of retrieval by selecting top-k documents randomly from the in-domain corpus RealNews/Dom. We observe in Table 12 that retrieval using in-context learning queries plays a crucial role to the performance of AG News, as performance drops significantly in a random setting. Hyperpartisandoes not face such a drop. This matches our intuition in Table 1 that task-inversion is the more challenging step for Hyperpartisan, and a powerful LLM we can apply stylistic changes to most news articles. In both, Mauve suffers when entities no longer match Gold.

RQ: Do in-context queries retrieve redundant results? Figure 6 measures the overlap of top-50 retrieved documents from the 200 ICL queries, and finds that in most cases, the retrieved documents are unique, especially when using a large in-domain corpus. Thus, we can conclude that effective retrieval is important for the diversity of the synthetic dataset.

RQ: Can SynthesizRR work effectively with relatively small corpora? In our main results §5, we assumed the existence of a large corpus, with minimum size of 0.9M documents. As noted, this corpus need not be unlabelled examples for our task; we were able to successfully generate customer reviews and product questions for Humor, Category and Polarity tasks, while retrieving from a corpus of product information (title and description).

A potential problem with SynthesizRR is that corpuses of such massive size might be few in number. Thus, we compare the performance

Retriever *(Size)*	AG. (8K)	HYP. (2K)	TOI (8K)	CAT. (8K)	HUM. (2K)	POL. (4K)	**Avg.**
GOLD	91.0	93.2	82.5	81.5	93.1	95.3	89.43
LLAMA2 ZERO SHOT							
CONTR.	**83.5**	69.8	**74.5**	**68.9**	**82.5**	84.7	**77.32**
BM25	83.2	**74.2**	70.7	57.6	78.5	**85.4**	74.93
CLAUDEV1 ZERO SHOT							
CONTR.	**83.9**	**72.3**	**71.8**	**66.8**	62.1	88.7	**74.29**
BM25	83.2	57.2	69.8	53.7	**73.9**	**91.8**	71.60
LLAMA2 3-SHOT RETRICL							
CONTR.	**83.0**	**78.5**	**73.3**	**72.4**	**90.2**	**91.0**	**81.38**
BM25	82.1	77.9	71.9	65.4	87.5	87.4	78.69
CLAUDEV1 3-SHOT RETRICL							
CONTR.	**83.7**	72.3	**72.8**	**65.4**	**83.4**	**91.3**	**78.16**
BM25	83.0	**73.5**	70.0	52.4	82.4	90.7	75.34

Table 13: Test accuracy after distilling a DEBERTA-V3L student on a dataset generated by SYNTHESIZRR. Retrieval is done using BM25 and CONTRIEVER. We use the same seed set and corpus as Table 2. We train 5 student models and report the mean accuracy, rerunning all 5 in case of std ≥ 6.0. The top two subsections consider zero-shot synthesis and bottom two considers 3-shot RETRICL variant. We **bold** the best result in each subsection. CONTRIEVER numbers are reproduced from Table 6.

of SYNTHESIZRR on CMU MOVIE SUMMARY (Bamman et al., 2013) which is between one to three orders of magnitude smaller than other corpora in Table 6. In Table 8, we see that SYNTHESIZRR can perform suitably even with such relatively small corpora (42k movie plots). From the previous RQs, this suggests that the relevance of the corpus to the task is more important than the size of the corpus for the performance of SYNTHESIZRR.

F Dense vs sparse retrieval in SYNTHESIZRR

So far, a single dense retriever (CONTRIEVER) has been used for the content sourcing step by using a bi-encoder approach (Lee et al., 2019; Chen et al., 2017). We embed both the input in-context covariate and each corpus document, and then rank results based on cosine similarity. In §5, we retrieved $k = 500$ documents for each in-context example and after filtering, randomly sampled among these to produce a grounded set of documents on which we apply our task inversion strategy RETRICL.

In this section we explore how changing the retrieval model affects the content sourcing stage and its downstream effects. Keeping other parts of the process the same, we switch CONTRIEVER to BM25 Okapi (Robertson and Zaragoza, 2009), a popular *sparse* retrieval method. Dense retrievers like CONTRIEVER perform a semantic match between the query and document, whereas BM25 performs only a lexical match based on inverse term frequencies, with no understanding of semantics. Additionally, BM25 outputs a score which is an unbounded positive number, thus we are unable to use meaningful thresholds to bound the similarity in our RETRICL approach. Instead, we construct the RETRICL in-context set using the top-2 retrieved contexts for each ICL example and without applying the filter.

We expect that picking semantically similar information is more important to SYNTHESIZRR since we include a task inversion step, which intends to change the tone and lexical structure of the text while preserving its semantics. Thus, we want contexts which are semantically related to GOLD data, to which we can apply stylistic or formatting transformations using a task-inversion prompt to bring it closer to GOLD.

Surprisingly, in Table 8 we see that while intrinsic diversity from BM25-retrieved documents is often worse than CONTRIEVER, they both generate equally human-like text. However, comparing the DEBERTA-V3L accuracy of CONTRIEVER and BM25in Table 13, we see that a strong student model trained on a dataset obtained from the dense-retrieved document set consistently outperforms the sparse retriever BM25, which might be due to the filtering step we introduce in RETRICL. This filtering step leads to a reduction in mislabelling stemming from retrieving contexts that belong do a different class. Due to this, we conclude that dense retrieval models are potentially more suitable for SYNTHESIZRR.

G Varying number of in-context examples in RETRICL

The use of in-context examples in the RETRICL variant of SYNTHESIZRR leads to significant improvements in intrinsic and distillation metrics, as we saw in §5. Here, we do a deeper analysis on whether continually increasing the number of in-context examples yields a positive benefit.

In Figure 7 we look at the DEBERTA-V3L accuracy, entity entropy and MAUVE for our datasets with different numbers of in-context learning examples. We see that adding even a single in-context

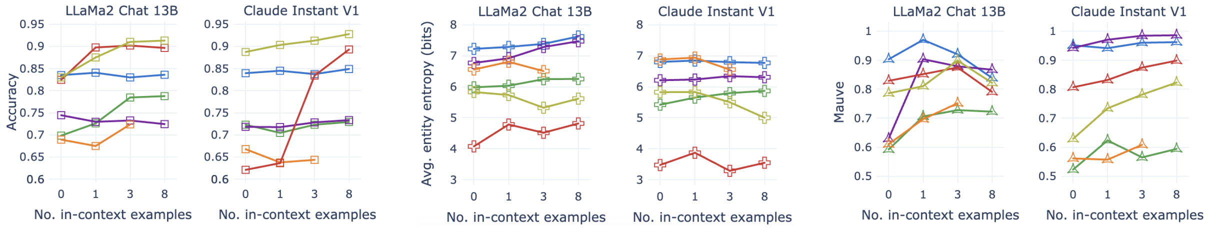


Figure 7: Left: DeBERTa-v3L test accuracy (↑), center: entity entropy (↑), right: Mauve (↑) for SynthesizRR RetrICL. We vary the number of in-context examples from 0 to 8. Teacher LLMs LLaMa-2 Chat 13B and Claude Instant-v1 are compared on 6 tasks: AG News, Hyperpartisan, ToI Headlines, Category, Humor and Polarity. We do not report Category 8-shot due to model failures.

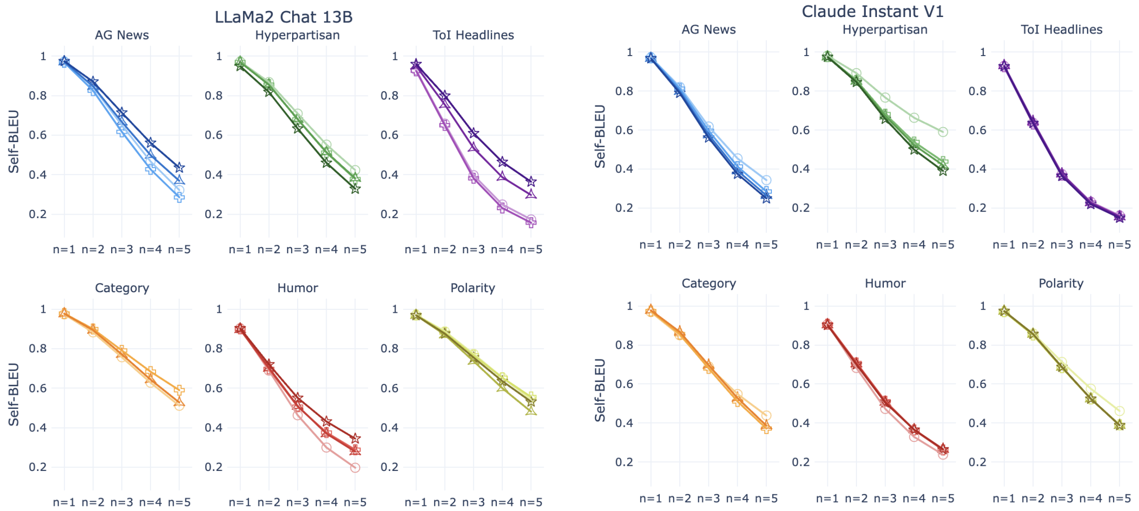


Figure 8: Lexical diversity i.e. Self-BLEU (↓) ngrams n=1-5, when varying the number of in-context examples for SynthesizRR RetrICL. We compare of teacher LLMs LLaMa-2 Chat 13B (left) and Claude Instant-v1 (right). Notation: 0-shot (●), 1-shot (+), 3-shot (△), 8-shot (★). Darker shade implies more in-context examples.

example can greatly increase the performance of all three metrics. However, no particular number of in-context examples consistently outperforms. For ClaudeV1, adding more in-context examples (up to 8) seems to always provide benefit, whereas with LLaMa2, we observe a peak and then reduction. Thus, the optimal number of in-context learning examples is a task dependent hyperparameter.

Figure 8 shows the lexical diversity i.e. Self-BLEU across datasets and number of in-context examples. As in §5 we observed that using in-context examples is neither positively nor negatively correlated with a lower Self-BLEU, despite using nucleus sampling with $p = 0.9$. This may be because for all number of shots, task inversion is performed from a single source context and thus the generation does not divert significantly from the unique n-grams of the context. Thus we conclude that to affect lexical diversity, the number of in-context learning examples has no effect and we must instead focus on changing the retrieved contexts, perhaps by using a different retrieval model.

H Task inversion prompts and label verbalizations

Here we discuss the prompt templates and verbalizations that we use for the task inversion step for both FewGen and SynthesizRR. We use descriptive verbalizations as compared to the target label.

Additionally in the prompt, we place the retrieved document near the end, as prior work indicates that intermediate placements degrade LLM recall (Liu et al., 2023).

LLMs have a fixed window-size for conditional generation, so excessively long documents are truncated (from the end) up to $r_{max} = 500$ tokens. This reserves the remaining window for in-context learning.

H.1 Hyperpartisan

Hyperpartisan is the task of detecting political bias in a news article. In transforming the retrieved news article `article_retr[k]` to one with such bias, typically there is the addition of mocking commentary and harsh political language which deeply criticizes the subject such as a person, policy or political event. On the other hand, articles in the opposite class gives a well-rounded opinion with a neutral tone. We include a length-attribute to ensure a long generation of one or two paragraphs.

Label	Verbalization
true	harsh political language, using a mocking tone and toxic commentary
false	neutral language, using a reasonable tone and politically correct commentary

Table 14: Task-inversion verbalizations for Hyperpartisan.

Prompt H.1: Hyperpartisan FewGen

In-context example:
“Write a single news article using `{label}` . The written article should be 2 to 3 paragraphs long.
News Article: `{icl[gold_text]}` ”
Prompt:
“Write a single news article using `{label}` . The written article should be 2 to 3 paragraphs long.
News Article: ”

Prompt H.2: Hyperpartisan SynthesizRR RetrICL

In-context example:
“News Article: `{icl[article_retr]}`
Rewrite the above news article using `{label}` . The rewritten article should be 2 to 3 paragraphs long.
Rewritten Article: `{icl[gold_text]}` ”
Prompt:
“News Article: `{article_retr[k]}`
Rewrite the above news article using `{label}` . The rewritten article should be 2 to 3 paragraphs long.
Rewritten Article: ”

Prompt H.3: Hyperpartisan SynthesizRR Non-RetrICL

In-context example:
“Rewritten Article: `{icl[gold_text]}` ”
Prompt:
“News Article: `{article_retr[k]}`
Rewrite the above news article using `{label}` . The rewritten article should be 2 to 3 paragraphs long.
Rewritten Article: ”

H.2 ToI Headlines

ToI Headlines is a topic classification dataset of regional news headlines in India. Here we attempt to refine the retrieved news article by summarizing it into a short headline. We use verbalizations of the content of each class, as example generation here involves summarizing the content. We add an “India” location-attribute to guide the LLM generations to include regionalization to the Indian subcontinent. A length-attribute is included to restrict the length to one sentence.

Label	Verbalization
sports	sports in India
life-style	health and lifestyle trends in India
education	Indian examinations and education
entertainment	the Indian entertainment industry
business	business-related developments in India
city	ongoing matters in any Indian city
environment	environment-related events in Indian cities
tech	technology news and the tech industry in India
elections	elections and politics in India
world	international news and events outside of India

Table 15: Task-inversion verbalizations for ToI Headlines.

Prompt H.4: ToI Headlines FewGen

In-context example:
“Write a headline for a news article about `{label}` . The headline should be a single sentence.
Headline: `{icl[gold_text]}` ”
Prompt:
“Write a headline for a news article about `{label}` . The headline should be a single sentence.
Headline: ”

Prompt H.5: ToI Headlines SynthesizRR RetrICL

In-context example:
“News Article: `{icl[article_retr]}`
Write a headline for the above news article about `{label}` . The headline should be a single sentence.
Headline: `{icl[gold_text]}` ”
Prompt:
“News Article: `{article_retr[k]}`
Write a headline for the above news article about `{label}` . The headline should be a single sentence.
Headline: ”

Prompt H.6: ToI Headlines SynthesizRR Non-RetrICL

In-context example:
“Headline: `{icl[article_retr]}` ”
Prompt:
“News Article: `{article_retr[k]}`
Write a headline for the above news article about `{label}` . The headline should be a single sentence.
Headline: ”

H.3 AG News

We consider task inversion for the AG News dataset to be generation of news summaries. We do not specify location modifiers as most Gold summaries are from US news. We add a length-attribute to restrict the output one or two sentences.

Label	Verbalization
Business	companies, industries, markets, trade, investments, entrepreneurship, economic policies, and other business-related developments
World	international news, such as politics, diplomacy, conflicts, global events, international relations, human rights issues, and significant global trends
Sci/Tech	scientific discoveries, technological advancements, innovations, research breakthroughs
Sports	professional sports leagues, major tournaments, athletes, teams, match results, player transfers, coaching changes, sports-related controversies

Table 16: Task-inversion verbalizations for AG News.

Prompt H.7: AG News FewGen

In-context example:
"Write a summary for a news article about `{label}` . The summary should be one or two short sentences.
Summary: `{icl[gold_text]}` "
Prompt:
"Write a summary for a news article about `{label}` . The summary should be one or two short sentences.
Summary: "

Prompt H.8: AG News SynthesizRR RetrICL

In-context example:
"News Article: `{icl[article_retr]}`
Write a summary for the above news article about `{label}` . The summary should be one or two short sentences.
Summary: `{icl[gold_text]}` "
Prompt:
"News Article: `{article_retr[k]}`
Write a summary for the above news article about `{label}` . The summary should be one or two short sentences.
Summary: "

Prompt H.9: AG News SynthesizRR Non-RetrICL

In-context example:
"Summary: `{icl[gold_text]}` "
Prompt:
"News Article: `{article_retr[k]}`
Write a summary for the above news article about `{label}` . The summary should be one or two short sentences.
Summary: "

H.4 Category

In the Category dataset, we determine the product category from a review written by a user. For task inversion in SynthesizRR we must retrieve a product and prompt the frozen LLM to generate a user review within the same product-category as the retrieval query. Thus, we include a style-attribute to allow minor typos in the generation and restrict to a few sentences using a length-attribute.

Label	Verbalization
magazines	magazines or periodicals covering various topics
camera_photo	photography gear including cameras, lenses, accessories, or photo editing tools
office_products	office supplies or equipment for professional and home office setups
kitchen	kitchenware, appliances, or culinary tools for cooking and dining
cell_phones_service	cell phone service accessories or service plans for communication and connectivity
computer_video_games	computers, gaming consoles, video games, or related accessories
grocery_and_gourmet_food	groceries, fruits and vegetables, gourmet treats, or specialty food items
tools_hardware	tools, hardware, or equipment for DIY projects and home repairs
automotive	auto parts, accessories, or tools for vehicle maintenance and enhancements
music_album	music albums spanning various genres and artists
health_and_personal_care	healthcare products, personal care items, or wellness essentials
electronics	electronic devices, gadgets, personal tech, or home electronics
outdoor_living	products for outdoor activities, gardening, or patio living
video	movies, TV shows, and documentaries spanning various genres and artists
apparel	clothing including casual wear, formal attire, seasonal outfits, activewear, or fashion accessories for men, women, and children
toys_games	fun or educational toys and games for kids of all ages
sports_outdoors	products for various sports and outdoor activities
books	books in various genres and formats
software	computer software for productivity or gaming covering either personal or professional needs
baby	baby essentials, gear, or toys for infants and toddlers
musical_and_instruments	musical instruments, accessories, or music production equipment
beauty	beauty products, cosmetics, or skincare essentials, makeup, hair care, fragrances, or grooming essentials
jewelry_and_watches	watches or jewelry pieces such as necklaces, bracelets, earrings, or rings, crafted in precious metals or adorned with gemstones for special occasions

Table 17: Task-inversion verbalizations for Category.

Prompt H.10: Category FewGen

In-context example:
"Write a product review about a product which is in the category of `{label}` . Include relevant product details. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: `{icl[gold_text]}` "
Prompt:
"Write a product review about a product which is in the category of `{label}` . Include relevant product details. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: "

Prompt H.11: Category SynthesizRR RetrICL

In-context example:
“Product details: `{icl[product_retr]}`
Write a product review about the above product which is in the category of `{label}` . Include relevant product details which are mentioned above. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: `{icl[gold_text]}` ”
Prompt:
“Product details: `{product_retr[k]}`
Write a product review about the above product which is in the category of `{label}` . Include relevant product details which are mentioned above. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: ”

Prompt H.12: Category SynthesizRR Non-RetrICL

In-context example:
“Review: `{icl[gold_text]}` ”
Prompt:
“Product details:
Write a product review about the above product which is in the category of `{label}` . Include relevant product details which are mentioned above. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: ”

H.5 Humor

Asking humorous product questions is a challenge of the LLM’s task inversion capabilities, as it must generate a question which is funny from the retrieved product. Not all products have obvious humorous characteristics, thus the generation requires some ingenuity. We restrict the output to only the question to prevent explanations or extraneous product generations from the LLM.

Label	Verbalization
humorous	humorous
non_humorous	solemn

Table 18: Task inversion verbalizations for Humor.

Prompt H.13: Humor FewGen

In-context example:
“Write a short `{label}` question about a product. Only include the question.
Product Question: `{icl[gold_text]}` ”
Prompt:
“Write a short `{label}` question about a product. Only include the question.
Product Question: ”

Prompt H.14: Humor SynthesizRR RetrICL

In-context example:
“Product details: `{icl[product_retr]}`
Write a short `{label}` question about the above product. Only include the question.
Product Question: `{icl[gold_text]}` ”
Prompt:
“Product details: `{product_retr[k]}`
Write a short `{label}` question about the above product. Only include the question.
Product Question: ”

Prompt H.15: Humor SynthesizRR Non-RetrICL

In-context example:
“Product Question: `{icl[gold_text]}` ”
Prompt:
“Product details: `{product_retr[k]}`
Write a short `{label}` question about the above product. Only include the question.
Product Question: ”

H.6 Polarity

Polarity is a review-sentiment classification task. In SynthesizRR, the difficulty is increased as we must generate a review from a product. For task inversion, we prompt the LLM to generate a review which can have either positive or negative sentiment and include details from the retrieved product. As with Category, we allow typos and restrict the length to a few sentences using a length-attribute in the prompt.

Label	Verbalization
positive	what the reviewer liked about the product, how the reviewer found it easy to use the product, or the reviewer’s positive experience with the product
negative	what the reviewer disliked about the product, how the reviewer found it challenging to use the product, or the reviewer’s negative experience with the product

Table 19: Task inversion verbalizations for Polarity.

Prompt H.16: Polarity FewGen

In-context example:
“Write a review about a product which discusses `{label}` . Include relevant product details. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: `{icl[gold_text]}` ”
Prompt:
“Write a review about a product which discusses `{label}` . Include relevant product details. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: ”

Prompt H.17: Polarity SynthesizRR RetrICL

In-context example:
“Product details: `{icl[product_retr]}`
Write a review about the above product which discusses `{label}` . Include relevant product details which are mentioned above. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: `{icl[gold_text]}` ”
Prompt:
“Product details: `{product_retr[k]}`
Write a review about the above product which discusses `{label}` . Include relevant product details which are mentioned above. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: ”

Prompt H.18: Polarity SynthesizRR Non-RetrICL

In-context example:
“Review: `{icl[gold_text]}` ”
Prompt:
“Product details: `{product_retr[k]}`
Write a review about the above product which discusses `{label}` . Include relevant product details which are mentioned above. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: ”

I Example generations

Here we showcase examples from the best-performing SynthesizRR approach (3-shot Non-RetrICL using LLaMa-2 Chat 13B) for each of our 6 tasks. For brevity, we do not show the ICL examples, only the retrieved article and generated text.

J Data Preprocessing

J.1 Datasets

- AG News: We use https://huggingface.co/datasets/zapsdcn/ag
- ToI Headlines: we use the data from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DPQMQH and filter headlines in following 10 topics: {sports, life-style, education, entertainment, business, city, environment, tech, elections, world}. We randomly subsample to get 5.2k rows per topic in train and 1k per topic in test.
- Humor: We use https://registry.opendata.aws/humor-detection/
- IMDb: We use https://ai.stanford.edu/~amaas/data/sentiment/
- SST-2: We use https://nlp.stanford.edu/sentiment/treebank.html

Other datasets are not pre-processed.

J.2 Corpora

- RealNews: we use the article text field and download the data from https://github.com/rowanz/grover/tree/master/realnews.
- RealNews/Regional is a subset of RealNews (Zellers et al., 2019). It includes 2.7M articles from non-US and non-EU websites. We manually check RealNews websites and identified 141 regional-news websites with headquarters in 56 non-US and non-EU countries: India, Pakistan, Nigeria, Philippines, etc.
- RealNews/India is further filtered to only include Indian news websites.
- RealNews/Dominant is the remaining 30.1M articles from 1063 news websites headquartered in 20 countries (of which over 75% are US-based).
- Products: We pull the data from https://nijianmo.github.io/amazon/index.html#complete-data and concatenate title and description

K Teacher and Student hyperparameters

K.1 Teacher LLM hyperparams

For LLaMa-2 Chat 13B, we use the implementation from HuggingFace: https://huggingface.co/TheBloke/Llama-2-13B-fp16 and run it at half-precision.

For Claude Instant-v1, we use Claude Instant v1.2: https://www.anthropic.com/news/releasing-claude-instant-1-2

We use a batch size of 1 for all generations as we have long contexts and encountered failures with higher batch sizes. We use nucleus sampling with top-p=0.9.

K.2 Student LM hyperparams

We use DeBERTa-v3-Large and DistilBERT models from HuggingFace:

Class	Example
(Retrieved document)	News Article: Gov. Andrew Cuomo Wednesdy afternoon staged a press conference to blast federal immigration officials, threatening to sue if agents don’t stop what he described as "reckless and unconstitutional enforcement" practices. Immigration and Customs Enforcement Deputy Director Thomas D. Homan responded hours later with a six-paragraph statement that accused Cuomo of "grandstanding." Cuomo cited a raid last week on a farm in Rome, N.Y. Immigration and Customs Enforcement agents entered John Collins’ property and detained a worker, Marcial DeLeon Aguilar, without presenting a warrant to Collins. That worker was later arrested and is still being detained. "The reckless and unconstitutional practices ICE is deploying in our communities violate everything we believe in New York and are an assault on our democracy," Cuomo said in a prepared statement. "I demand ICE immediately cease and desist this pattern of conduct, and if they fail to do so, I will pursue all available legal recourse and commit to doing everything in my power to protect the rights and safety of all New Yorkers." At about 7:30 p.m., Homan issued this response: "As a native New Yorker who began my law enforcement career as a police officer in the state, I was disappointed to learn about the Governor’s grandstanding today over the issue of immigration enforcement. The Governor’s comments were inaccurate and an insult to ICE’s sworn law enforcement officers who conduct their lawful mission professionally and with integrity. These brave men and woman leave the safety of their homes every day to protect this great nation and our communities.
hyperpartisan	What a joke! Governor Andrew Cuomo had the audacity to call out ICE agents for doing their jobs, and now he’s threatening to sue them? Give me a break! The nerve of this guy, talking about "reckless and unconstitutional enforcement" practices when his own state is infested with illegal immigrants who are committing crimes left and right. Cuomo is just trying to grandstand and pander to his liberal base, but the fact is, he’s failing to protect the citizens of New York. ICE agents are just doing their jobs, and they deserve respect for their hard work and dedication. Cuomo should be ashamed of himself for attacking these brave men and women. And let’s not forget, this is the same guy who is protecting child rapists and allowing them to roam free in his state. What a disgrace! Cuomo needs to focus on the real problems in his state, like the rampant crime and corruption, instead of trying to score political points by attacking ICE agents. It’s pathetic.
(Retrieved document)	News Article: This week on "Watters’ World", Jesse Watters caught up with several House Democrats who boycotted President Trump’s inauguration, including Rep. John Lewis, the Atlanta Democrat who called the Republican’s election "not legitimate". Approaching Houston Congresswoman Sheila Jackson Lee, Watters asked why she was absent from the festivities: "I have a statement on record, thank you so very much," Jackson Lee said. WATCH: Watters Tracks Down Rep. John Lewis to Ask About Trump Criticism ’Watters’ World’ EXCLUSIVE: Jesse Visits Scene of Chicago Hate Crime Attack Watters Confronts Man Who Harassed Ivanka Trump on JFK Flight As Watters introduced himself as a Fox News reporter to Rep. John Conyers of Michigan, the Detroiter kept walking toward an open elevator, "no, sir, not right now." "I didn’t boycott it, I just didn’t go to it," Conyers said. Rep. Jerrold Nadler of New York gave a similar answer when asked, "for the reasons I stated publicly; they’re on the record." Asked if he was being divisive, the Manhattan Democrat said that if anyone wanted to unite the country, they would not have voted for Trump. Lewis told Watters he was late and couldn’t speak with him as he walked with a quick gait down a hallway. An aide obstructed Watters’ cameraman as the Congressman continued toward his destination.
non-hyperpartisan	In a recent segment on "Watters’ World", Jesse Watters spoke with several members of the House of Representatives about their absence from President Trump’s inauguration. While some members provided brief statements, others declined to comment or did not respond to Watters’ inquiries. Rep. John Lewis, an Atlanta Democrat who had publicly questioned the legitimacy of Trump’s election, was approached by Watters but did not stop to speak with him. Lewis had previously stated that he would not attend the inauguration due to his disagreement with Trump’s presidency. Other members, including Rep. John Conyers of Michigan and Rep. Jerrold Nadler of New York, also declined to comment or did not respond to Watters’ questions. When asked why he did not attend the inauguration, Conyers simply stated that he "didn’t go to it". Nadler, who had publicly stated that he would not attend the inauguration due to his disagreement with Trump’s policies, reiterated that he did not support the President’s election. Watters’ segment sparked a discussion on the perceived divisiveness of the President and the responses of those who did not attend his inauguration. While some argued that the boycott was a form of protest against Trump’s policies and actions, others felt that it was a divisive move that further polarized the country. Regardless of one’s political beliefs, it is important to approach discussions and debates with respect and civility, and to strive for understanding and cooperation in the pursuit of common goals.

Table 20: Generated examples for HYPERPARTISAN task using SYNTHESIZRR.

Class	Example
(Retrieved document)	News Article: New Delhi: First reports of exit poll indicate history in West Bengal and regime change in Kerala while it’s too close to call in Tamil Nadu. An exit poll conducted by Star News television channel has projected Congress-led United Progressive Alliance (UPA)’s victory in West Bengal and Kerala. The survey projects Trinamool Congress - Congress alliance wiping out more than three decades old Left Front government in West Bengal. According to the exit polls telecast in the channel on Tuesday says the TMC and Congress will win 181 and 40 respectively while Left Front will be reduced to 62 from 227. In Kerala, the Congress-led United Democratic Front is expected to win 88 while CPM-led Left Democratic Front will win 49. In Tamil Nadu, AIADMK is projected as winning 110 and ruling DMK will have 124 seats in the 234 member assembly. In a close contest, Tarun Gogoi led Congress has a chance to create history in Assam by returning to power for a third consecutive term,post poll surveys predict. The Headlines Today poll predicts 44 of the 126 seats for Congress in Assam,with potential ally AUDF slated to win 14.
elections	Exit Polls predict regime change in Kerala, West Bengal; Tight contest in Tamil Nadu
(Retrieved document)	News Article: India’s teen sensation P V Sindhu stormed into the third round of the World Badminton Championship with a straight game victory over Olga Golovanova of Russia in women’s singles match here yesterday Copenhagen: India’s teen sensation PV Sindhu stormed into the third round of the World Badminton Championship with a straight game victory over Olga Golovanova of Russia in women’s singles match here yesterday. PV Sindhu The men’s doubles pair of Manu Attri and Sumeeth Reddy B stunned 15th seeded Japanese duo of Hirokatsu Hashimoto and Noriyasu Hirata 21-19 21-19 in 44 minutes to advance to the third round. Sindhu, seeded 11th, took 40 minutes to prevail over her Russian opponent 21-12 21-17 in the second round match at the Ballerup Super Arena here. She will next take on sixth seeded Yeon Ju Bae of Korea. Sindhu won a total of 42 points as compared to 29 by the Russian girl. The world No. 12 from Hyderabad looked a bit rusty to start with and was initially trailing in the opening game. She was playing the catching-up game till 10-10 before Sindhu managed to reel off four consecutive points and surge ahead to 14-10. There was no looking back after that, as Sindhu swiftly sealed the game in her favour with Golovanova earning just two more points. In the second game, the Russian got her act together and opened up a big lead, moving up to 11-6 at the break.
sports	15-year-old PV Sindhu creates history, enters World Badminton Championship 3rd round

Table 21: Generated examples for TOI HEADLINES task using SYNTHESIZRR.

https://huggingface.co/microsoft/deberta-v3-large, https://huggingface.co/distilbert/distilbert-base-uncased

We use the same hyperparameters for DEBERTA-V3L and DISTILBERT as (Yu et al., 2023a):

- DISTILBERT: Learning rate of 5e-5, gradient_accumulation_steps of 1, batch_size 32. We use the Adam optimizer with weight_decay of 1e-4 and epsilon of 1e-6. We use max_sequence_length of 512.

Class	Example
(Retrieved document)	News Article: SANTIAGO Nov 14 A magnitude 6.1 earthquake hit central-north Chile on Wednesday, the U.S. Geological Survey said, causing buildings to sway in capital Santiago but causing no damage. The USGS said the quake struck on the coast, 54 miles (88 km) north of Coquimbo, at 4:02 p.m. (1902 GMT). It was initially reported as a magnitude 6.0 quake. Chile's state emergency office, Onemi, said there were no reports of injuries, disruptions to basic services or infrastructure damage as a result of the tremor. The navy said the quake did not meet the conditions needed to trigger a tsunami. World No.1 copper producer Codelco said its operations were normal, and no mines in the metal-producing country reported damages. Seismically active Chile was battered in February 2010 by an 8.8 magnitude quake and ensuing tsunami, which caused billions in damage and killed hundreds.
World	A magnitude 6.1 earthquake hit central-north Chile on Wednesday, causing buildings to sway in capital Santiago but causing no damage, the USGS said.
(Retrieved document)	News Article: MONTREAL - Canadian astronaut David Saint-Jacques is bracing for some tense moments during next Monday's launch of the Soyuz rocket that will send him and two others to the International Space Station. On Oct. 11, a rocket failure forced a Soyuz capsule carrying two astronauts to abort and make an emergency landing. Russia suspended all manned space launches pending an investigation before giving the green light Nov. 1. Saint Jacques spoke to reporters today from the launch site in Kazakhstan where he's in quarantine. He said the most dangerous part of the six-month mission is the 10-minute Soyuz launch and the six hours that follow before docking. He said half of the last two-and-half years of training was dedicated to his role as Soyuz co-pilot for the trip to the space station. Saint-Jacques, 48, said once at the space station he'll be able to focus on work and life aboard the outpost. The Canadian astronaut's first space voyage, with NASA astronaut Anne McClain and Oleg Kononenkoof of the Russian space agency Roscosmos, was originally scheduled to launch Dec. 20. It was moved ahead after Russian authorities concluded their investigation into the failed launch.
Sci/Tech	commentary Saint-Jacques says the Soyuz launch is the most dangerous part of the six-month mission but he's been training for it for half of the last two-and-a-half years.

Table 22: Generated examples for AG News task using SynthesizRR.

Class	Example
(Retrieved document)	Product details: MSD ASY17296 Wiring Harness Factory Terminated MSD Ignition 6ALN (6430) Wiring Harness ASY17296 New harness with factory crimped connections a perfect item to have as an extra or replacement.
camera_photo	I have a MSD Ignition 6ALN (6430) and this wiring harness is a great addition. The factory terminated connections are perfect and make installation a breeze. The harness is well made and will definitely last. I have been using it for a few weeks now and it has performed flawlessly. Highly recommend it to anyone looking for a high quality wiring harness.
(Retrieved document)	Product details: Conair Travel Smart 3-Dial Combination Lock (6 Pack) Safely Locks Your Luggage With Tsa Recognized Locks Comes With A Red Inspection Status Feature Travel Sentry Approved
apparel	I purchased this lock for a business trip to Europe. I've used it several times already and so far, so good. The TSA inspectors have been able to open it without a hitch, which is great!

Table 23: Generated examples for Category task using SynthesizRR.

Class	Example
(Retrieved document)	Product details: Keeping Up With The Kardashians - Season 5 Region E! brings viewers back into the sensational and completely unpredictable world of one of televisions favourite families with the fourth season of the hit series Keeping Up with the Kardashians. This season, Khlo and her husband, Los Angeles Lakers star, Lamar Odom, begin their new lives as a married couple. Meanwhile, Kourtney is pregnant and back together with Scott, and Kim and her boyfriend, NFL player Reggie Bush, work on their relationship. Through all this change, the Kardashian family must find ways to evolve and keep their close-knit ties.
humorous	Does anyone else notice that this is the only show where you can hear people say things like, "OMG I am totally exhausted, I need to take a nap.? And it's not considered weird?
(Retrieved document)	Product details: Coleman Camper's Storage Tote Securely organize, store and carry camping gear 23" x 14" x 17" Lightweight 600D polyester construction Multiple pockets and compartments Convenient carry strap The Coleman Camper's Storage Tote is a great way to store, organize and carry your gear – use it in the garage, in the trunk or RV, and at the campsite. Constructed of reliable, lightweight 600D polyester, the Camper's Storage measures 23" long, 14" wide and 17" tall. It features multiple pockets and compartments for organizing all your gear, and a convenient carry strap so it's easy to bring along.
non_humorous	is this product big enough for all the camping gear for a family of four?

Table 24: Generated examples for Humor task using SynthesizRR.

Class	Example
(Retrieved document)	Product details: Birth Of Soul 1 No half measures here, but a real statement of intent, particularly with that Volume One tag. 28 cuts that helped define that thing called soul - you can argue the point, but you can't argue the music. Sublime.
positive	This album was the beginning of my love affair with soul music. These tracks are timeless and still sound amazing today.
(Retrieved document)	Product details: Beware! the Lies of Satan It's high time we pull the covers off the enemy and expose him for who he really is, the murderer, destroyer, their, deceiver and biggest liar of all time. In this book , you will begin to discover the truth about our heavenly Father and how you can stand in victory over the devil.
negative	The book does not live up to it's promise. There is no revelation of truth about our heavenly father, or any insight into Satan's lies. It is simply a polemic diatribe against Satan, with no concrete solutions to any of life's problems.

Table 25: Generated examples for Polarity task using SynthesizRR.

- DeBERTa-v3L: Learning rate of 2e-5, gradient_accumulation_steps of 8, batch_size 4. We use the Adam optimizer with weight_decay of 1e-4 and epsilon of 1e-6. We use max_sequence_length of 512.

We train all students for 6 epochs. Following (Yu et al., 2023a), we use warmup for 6% of the training steps.

K.3 Oracle model hyperparams

To train the DeBERTa-v3-Large oracle model for Label Preservation, we use a grid search over 9 combinations: 3 learning rates {2e-5, 5e-5, 1e-4} by 3 batch-sizes {1, 4, 16} (with same graident accumulation). We train on 80% of the Gold training data and use the remaining 20% as validation.

K.4 Retriever

We use Contriever from HuggingFace library: https://huggingface.co/facebook/contriever.

We pass a batch-size of 512 for embedding.

L Computational budget

We run all our models on AWS Elastic Cloud Compute[3] using 20 p3dn.24xlarge machines.

L.1 Information Retrieval

The corpora was embedded by us and the trivial was done using the Faiss library.[4] We orchestrate 80 copies of Contriever using the Ray distributed framework[5] to embed the RealNews and Products corpus in $\sim$ 3 hours each.

L.2 Dataset synthesis

For LLaMa-2 Chat 13B, we create 48 copies of the model and orchestrate it using the Ray distributed framework. Generation is done in roughly 6 hours per dataset of 8k rows.

To use Claude Instant-v1, we invoke AWS Bedrock[6] using the boto3 library[7]. Generation from Claude was done at an AWS-account level RPM of 1600 and takes roughly 4 hours per dataset on 8k rows.

L.3 Student distillation

Each DeBERTa-v3-Large student model trains for between 1 to 3 hours on a single V100 GPU on 8k rows. Each DistilBERT student model trains in 1 hour to generate the data-map.

M Licensing

We use datasets that have been released in prior work with various open licenses. Specifically:

M.1 Datasets

- AG News: custom license, described at http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
- ToI Headlines: uses Creative Commons CC0 1.0 Universal Public Domain Dedication licence as per https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DPQMQH
- Hyperpartisan: taken from SemEval 2019 Task 4, this is licensed under a Creative Commons Attribution 4.0 International License as per https://zenodo.org/records/1489920
- Humor: Community Data License Agreement – Sharing – Version 1.0 licence as per https://registry.opendata.aws/humor-detection/
- IMDb: (Maas et al., 2011) does not specify a licence but has made the data available for research at: https://ai.stanford.edu/~amaas/data/sentiment/
- SST-2: (Socher et al., 2013) does not specify a licence but has made the data available for research at: https://nlp.stanford.edu/sentiment/treebank.html

M.2 Corpora

- RealNews: custom licence as per https://docs.google.com/forms/d/1LMAUeUtHNPXO9koyAIlDpvyKsLSYlrBj3rYhC30a7Ak/viewform?edit_requested=true. The code repository is Apache Licence 2.0 as per https://github.com/rowanz/grover/blob/master/LICENSE

---

[3] https://aws.amazon.com/ec2/

[4] https://faiss.ai/index.html

[5] https://docs.ray.io/en/latest/index.html

[6] https://docs.aws.amazon.com/pdfs/bedrock/latest/APIReference/bedrock-api.pdf

[7] https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

- PRODUCTS: (Ni et al., 2019) does not specify a licence but has made the data available for research at: `https://nijianmo.github.io/amazon/index.html#complete-data`.