4. Task-Adaptive Pretraining:

Task-Adaptive Pretraining (TAPT) is **a highly effective**, **cost-efficient strategy** for specializing a pre-trained language model (LM) **for a particular job**. It acts as an intermediate learning phase between the model's initial general training and its final task-specific fine-tuning

- Second phase of pre-training in-domain leads to **gains in high** and **low resource settings**
- Adapting to the task's unlabeled data **improves performance** even after domain adaptive pretraining
- When there isn't available unlabeled data, adapting to a task corpus **augmented using simple data selection** strategies is an effective alternative (especially when resources for domain-adaptive pretraining might be unavailable

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

The CHEMPROT Example:

The Task: Chemical-Protein Relation Extraction

> The goal is to analyse abstracts of biomedical research papers and identify relationships between chemical compounds and proteins (e.g., "Compound X activates Protein Y")

The TAPT Process

1. Start with RoBERTa:
2. The Mismatch:
3. TAPT Execution:
4. Model Adaptation:
5. Result (Fine-Tuning):

Combined DAPT and TAPT:

RoBERTa ---> DAPT -----> TAPT (Domain followed by Task)

1. Main Concern is **cost.**
2. best **performance.**
3. **Outcome**: DAPT first provides a solid domain foundation, and then TAPT fine-tunes that knowledge to the specific task distribution, yielding the optimal result.

Cross-Task Transfer:

- investigates whether TAPT on **one task's data is helpful for a *different* task** in the same domain.

BioMed	RCT	ChemProt
TAPT	$87.7_{0.1}$	$82.6_{0.5}$
Transfer-TAPT	$87.1_{0.4}$ ($\downarrow$0.6)	$80.4_{0.6}$ ($\downarrow$2.2)
News	**HyperPartisan**	**AGNews**
TAPT	$89.9_{9.5}$	$94.5_{0.1}$
Transfer-TAPT	$82.2_{7.7}$ ($\downarrow$7.7)	$93.9_{0.2}$ ($\downarrow$0.6)

CS	ACL-ARC	SciERC
TAPT	$67.4_{1.8}$	$79.3_{1.5}$
Transfer-TAPT	$64.1_{2.7}$ ($\downarrow$3.3)	$79.1_{2.5}$ ($\downarrow$0.2)
Reviews	**Helpfulness**	**IMDB**
TAPT	$68.5_{1.9}$	$95.7_{0.1}$
Transfer-TAPT	$65.0_{2.6}$ ($\downarrow$3.5)	$95.0_{0.1}$ ($\downarrow$0.7)

Results:

Transfer-TAPT is consistently harmful across all domains, resulting in a performance drop and **Validation for DAPT + TAPT** .

5. Augmenting Training Data for Task-Adaptive Pretraining

- Inspired by the success of TAPT, we next investigate another setting where a larger pool of unlabeled data from the task distribution exists

1. **Using Human-Curated Corpora (§5.1)**

- The primary goal is to investigate the benefit of using **larger, readily available, human-curated unlabeled corpora** for TAPT, especially since these corpora are expected to share a similar data distribution with the final task's training set.

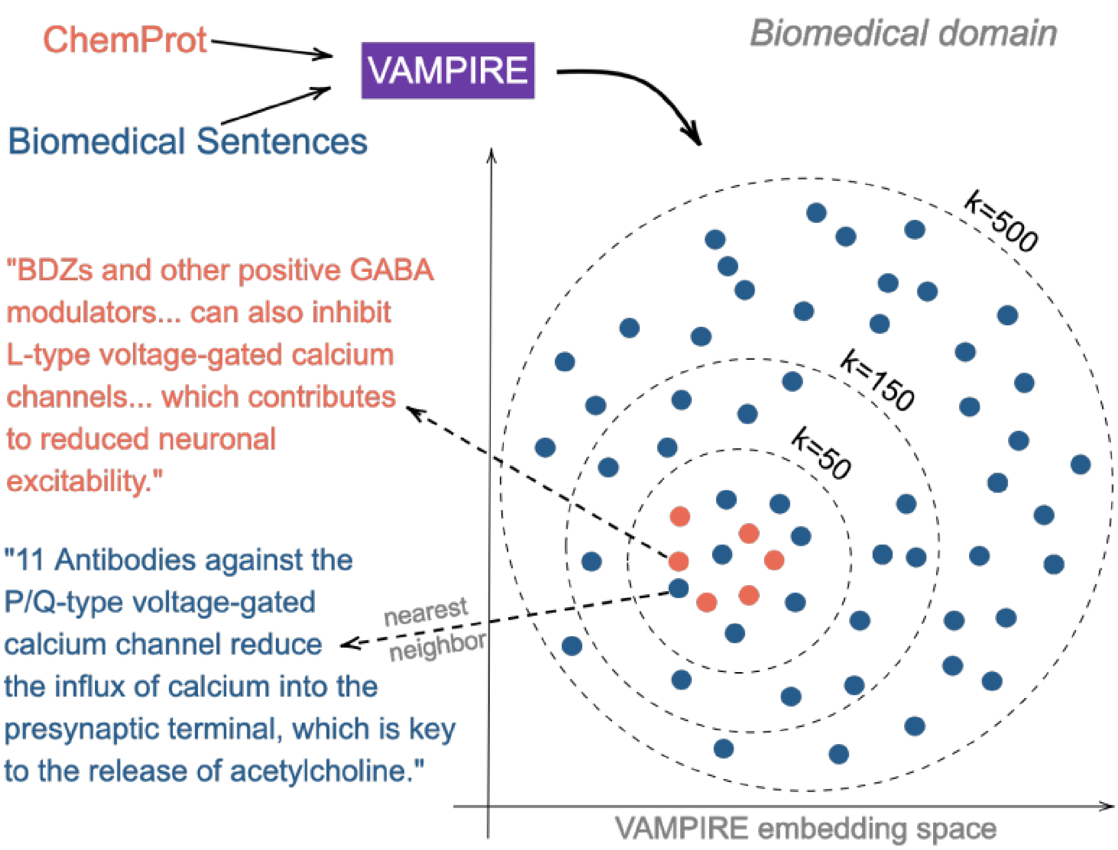
- **RCT-500: Simulated low-resource** by using only 500 labeled examples; used the **remaining 180K** examples as the adaptation data. (Data is from the **exact same distribution**).
- **HYPERPARTISAN:** Used the small low-resource documents for fine-tuning; used **5,000 documents** from the high-resource track for adaptation. (Data shares the **required political language/style**).

- **IMDB:** Used standard labeled data; used **extra, manually curated unlabeled data** collected by the original annotators for adaptation. (Data is guaranteed to be from the **same distribution**).

- **Overall Finding:** Using a larger, task-relevant unlabeled corpus for TAPT is consistently **beneficial**.

Pretraining	BIOMED RCT-500	NEWS HYP.	REVIEWS IMDB †
TAPT	$79.8_{1.4}$	$90.4_{5.2}$	$95.5_{0.1}$
DAPT + TAPT	$83.0_{0.3}$	$90.0_{6.6}$	$95.6_{0.1}$
Curated-TAPT	$83.4_{0.3}$	$89.9_{9.5}$	$95.7_{0.1}$
DAPT + Curated-TAPT	$\mathbf{83.8}_{0.5}$	$\mathbf{92.1}_{3.6}$	$\mathbf{95.8}_{0.1}$

2. **Retrieving Related Data (§5.2)**



kNN-TAPT Steps

1. **VAMPIRE Embedding:** A lightweight model is trained on a large domain corpus (e.g., 1M BIOMED sentences) to create a **shared vector space**.
2. **Mapping:** Both the small **task sentences** (e.g., CHEMPROT) and the large **domain sentences** are mapped (embedded) into this space.
3. **Query & Retrieval:** Each task sentence acts as a **query** to the domain sentences.
   - **kNN-TAPT:** The **nearest neighbors** (the most similar sentences) are identified and retrieved from the domain set.
4. **Augmentation:** The small task data is **augmented** with this highly relevant, retrieved set.
5. **Final TAPT: RoBERTa** is pre-trained using this new, larger, and highly relevant augmented corpus.

Pretraining	BioMed		CS
	ChemProt	RCT-500	ACL-ARC
RoBERTa	$81.9_{1.0}$	$79.3_{0.6}$	$63.0_{5.8}$
TAPT	$82.6_{0.4}$	$79.8_{1.4}$	$67.4_{1.8}$
RAND-TAPT	$81.9_{0.6}$	$80.6_{0.4}$	$69.7_{3.4}$
50NN-TAPT	$83.3_{0.7}$	$80.8_{0.6}$	$70.7_{2.8}$
150NN-TAPT	$83.2_{0.6}$	$81.2_{0.8}$	$73.3_{2.7}$
500NN-TAPT	$83.3_{0.7}$	$81.7_{0.4}$	$\mathbf{75.5}_{1.9}$
DAPT	$\mathbf{84.2}_{0.2}$	$\mathbf{82.5}_{0.5}$	$75.4_{2.5}$

3. **Computational Resourses:**

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

Table 9: Computational requirements for adapting to the RCT-500 task, comparing DAPT (§3) and the various TAPT modifications described in §4 and §5.

6. Related Work:

1. **Transfer Learning for Domain Adaptation (DAPT)**

- **Prior Work:** Previous studies confirmed the benefit of Domain-Adaptive Pretraining (DAPT), where LMs are continuously pretrained on a large, specific domain corpus (e.g., medical text).
- **The Paper's Contribution:** The work is more cost-effective because it focuses on continuing the pretraining of an existing powerful LM (RoBERTa), rather than training a model from scratch in the new domain, and investigates this approach across multiple domains.

2. Task-Adaptive Pretraining (TAPT)

- **Prior Work:** The benefit of **TAPT** (pretraining on the task's unlabeled data) was already recognized in related studies.
- **The Paper's Contribution:** The core novelty here is the **direct, systematic comparison** of TAPT and DAPT, and a detailed analysis of their **interplay (DAPT + TAPT)** regarding cost, relevance, and transferability.

	Training Data		
	Domain (Unlabeled)	Task (Unlabeled)	Task (Labeled)
RoBERTa			✓
DAPT	✓		✓
TAPT		✓	✓
DAPT + TAPT	✓	✓	✓
kNN-TAPT	(Subset)	✓	✓
Curated-TAPT		(Extra)	✓

Table 10: Summary of strategies for multi-phase pretraining explored in this paper.

3. Data Selection for TAPT Augmentation

- **Prior Work:** General data selection methods exist for improving transfer learning.
- **The Paper's Contribution:**
    - **Automated Selection (kNN-TAPT):** Unlike other work focused on selecting corpora to pretrain from scratch, this paper uses the lightweight **VAMPIRE** model to retrieve **highly relevant nearest neighbors** from a large domain corpus to improve TAPT performance.
    - **Curated-TAPT:** The study of using existing **human-curated data** (e.g., extra collected documents) is related to focused data collection techniques.

4. **What is "Domain"?**
    a. Broad
    b. Broadest Domain
    c. Narrow Domain
    d. Task Specific
    e. Narrowest