# Domain Adaptation via Prompt Learning for Alzheimer's Detection

**Shahla Farzana***
Institute for Population and Precision Health
University of Chicago
sfarza3@uic.edu

**Natalie Parde**
Department of Computer Science
University of Illinois Chicago
parde@uic.edu

## Abstract

Spoken language presents a compelling medium for non-invasive Alzheimer's disease (AD) screening, and prior work has examined the use of fine-tuned pretrained language models (PLMs) for this purpose. However, PLMs are often optimized on tasks that are inconsistent with AD classification. Spoken language corpora for AD detection are also small and disparate, making generalizability difficult. This paper investigates the use of domain-adaptive prompt fine-tuning for AD detection, using AD classification loss as the training objective and leveraging spoken language corpora from a variety of language tasks. Extensive experiments using voting-based combinations of different prompting paradigms show an impressive mean detection $F_1$=0.8952 (with std=0.01 and best $F_1$=0.9130) for the highest-performing approach when using BERT as the base PLM.

## 1 Introduction

Alzheimer's disease (AD) results in gradual impairment of memory, executive function, and language abilities (Alzheimer's Association, 2023). It is not treatable, but its effects can be slowed through medical or lifestyle interventions (Kivipelto et al., 2017; Sharma, 2019). Spoken language assessment offers a non-invasive, inexpensive, and scalable way to automate AD screening (de la Fuente Garcia et al., 2020; Khojaste-Sarakhsi et al., 2022), and the natural language processing community has recently launched popular challenges to accelerate progress on AD detection (Luz et al., 2020, 2021a). Many approaches have used manual features (Martinc and Pollak, 2020; Rohanian et al., 2021; Farzana and Parde, 2023), although recently the use of pretrained language models (PLMs) has also grown prominent (Balagopalan et al., 2020; Yuan et al., 2020; Li et al., 2021; Ye et al., 2021; Syed et al., 2021; Farzana and Parde, 2022; Wang et al., 2022). Currently, the most common pipelined architecture for AD detection feeds text embeddings produced by PLMs into back-end classification layers optimized for the task (Ye et al., 2021; Wang et al., 2022), but the discrepency between AD classification and the PLM feature extractor's loss function remains unaddressed. Furthermore, since AD detection datasets are small and involve varied language tasks, cross-domain[1] application of fine-tuned PLMs remains underexplored.

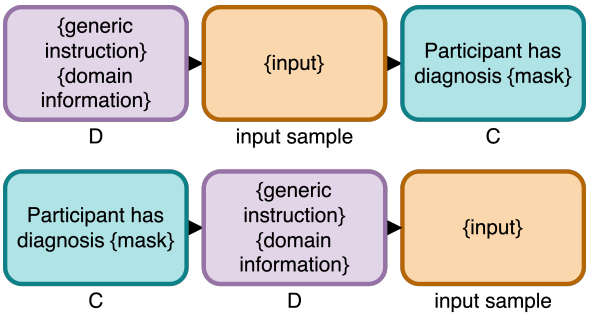



Figure 1: High-level template overview for domain adaptation via prompt-based fine-tuning, showing the domain-specific template (D) and class-specific template (C) in two sample formats.

We address these limitations by framing cross-domain AD detection as a supervised domain adaptation problem, and experiment with cloze-style prompt-based learning (Sivarajkumar and Wang, 2023; Wang et al., 2023a) for this purpose. There has been limited research regarding the effectiveness of prompt-based learning for medical applications (Taylor et al., 2023), particularly in the cognitive health domain (Wang et al., 2023b). We introduce a novel prompt learning approach for cross-domain adaptation, **D**omain **A**daptation via **P**rompt-based **F**ine-tuning (DAPF, partially summarized in Figure 1), that optimizes PLMs along

*Work completed in the Department of Computer Science at the University of Illinois Chicago.

[1]We define *domains* in this context as different spoken language tasks with reference labels pertaining to AD status.



*Findings of the Association for Computational Linguistics: EACL 2024*, pages 15963–15976

with domain-adaptive prompt parameters while training the model with source and target domain data. This dynamically adapts the classifier to each domain. We show that DAPF outperforms both p-tuning (Li and Liang, 2021) and prompt-based fine-tuning (Wang et al., 2023b) that lacks domain information. Our contributions include:

- We design and adapt prompt learning techniques for supervised domain adaptation across spoken language AD tasks.
- We introduce a new prompt learning paradigm (DAPF) for supervised domain adaptation.
- We empirically validate DAPF's cross-domain generalizability for AD detection.

In contrast, prior relevant prompting research focused on a particular AD corpus (Wang et al., 2023b) or clinical tasks with different characteristics (Sivarajkumar and Wang, 2023; Taylor et al., 2023). Results from our experiments show that DAPF outperforms contemporary alternatives. Additionally, domain-adaptive manual prompt fine-tuning outperforms domain-adaptive soft prompting with frozen PLM weights for AD detection.

## 2 Related Work

### 2.1 Language Model Prompting

Cloze-style prompt-based learning (Brown et al., 2020; Ben-David et al., 2022) with PLM backbones has performed competitively across many tasks, especially in few-shot learning scenarios. However, limited prior research has studied its effectiveness for medical applications (Sivarajkumar and Wang, 2023; Taylor et al., 2023). The prompts themselves can be constructed in numerous ways. *Manual prompting* uses carefully constructed templates with discrete keywords or tokens to elicit the desired response from the language model. These prompts are highly interpretable, although the human prompt designers may fail to identify the most optimal prompts, leading to underperformance in some tasks (Shin et al., 2020; Gao et al., 2021).

Recent studies have also experimented with optimizing *soft prompts* in continuous embedding space. A popular soft prompting technique is *prefix tuning*, which prepends a prefix of soft tokens (tunable embeddings) to the input layer and each layer in the encoder stack and then optimizes those tokens (Li and Liang, 2021). Other popular versions of soft prompt tuning prepend trainable continuous embeddings to the original sequence of input word embeddings (Lester et al., 2021) or use a long short-term memory (LSTM) encoder to capture the sequential representations of soft prompts (Liu et al., 2022a). Soft prompting is generally parameter efficient, yields high performance in low-resource settings (Li and Liang, 2021), and is fairly robust with respect to domain transfer (Lester et al., 2021). However, soft prompts have poor interpretability. There is also limited evidence that the language model understands the prompt's meaning (Webson and Pavlick, 2022), although some research suggests that embeddings arising from soft prompting techniques are close when they are clustered semantically or task-specifically (Lester et al., 2021; Zhang et al., 2021; Su et al., 2021).

### 2.2 Domain Adaptation via Prompt Learning

Most PLMs are trained on large, general-domain datasets. Domain-specific PLMs must be pre-trained from scratch on domain-specific datasets (Alsentzer et al., 2019; Lee et al., 2019) or adapted to target domain data using fine-tuning approaches (Gururangan et al., 2020; Hedderich et al., 2021). In low-resource settings, a popular approach for doing this is to employ zero- or few-shot cross-domain learning, coupled with unsupervised or semi-supervised domain adaptation techniques (Hedderich et al., 2021).

Some existing domain adaptation algorithms use domain-invariant features to minimize the discrepancy between domains (Long et al., 2015, 2017), or align samples from the source and target domain via linear projection (Sun and Saenko, 2016). A family of zero- or few-shot domain adaptation techniques, DANN (Ganin and Lempitsky, 2015) and CDAN (Long et al., 2018), distinguish source and target samples using a domain discriminator and then extract domain-invariant features. Although aligning domains using these approaches works well in some cases, it can also distort feature representations in settings with complex underlying data distributions (Cai et al., 2019). Bridging the gap between semantic and domain representation, domain-adaptive prompt learning methods for unsupervised domain adaptation can be an effective and efficient solution for unimodal and multimodal classification (Hu et al., 2022; Ge et al., 2023).

Recent studies suggest that prompt tuning holds promise in low-resource settings across diverse language tasks (Ben-David et al., 2022; Zhao et al., 2023; Goswami et al., 2023). The dynamic

prompting approach *SwitchPrompt* (Goswami et al., 2023) prompts PLMs in low-resource domains with domain-specific keywords and soft prompt vectors, showing performance superior to baseline methods in few-shot and all-data settings. Since AD detection is characterized by varied low-resource datasets that involve different spoken language tasks, it is an opportune test bed for investigating this technique as a domain adaptation strategy.

### 2.3 Prompt Learning for Healthcare Tasks

Prompt-based learning can also help circumvent the scarcity of high-quality data for healthcare tasks. For example, HealthPrompt (Sivarajkumar and Wang, 2022) shows promise at classifying new diseases and phenotypes from clinical texts using prompt learning. Position-based prompting (Abaho et al., 2022) automatically adjusts prompt templates, eliminating the need to prepare handcrafted prompts when probing PLMs for rare domain knowledge. It results in improved performance in recalling biomedical entities encountered during training and generalizing across unseen prompts for health outcome generation. Zhang and Guo (2024) present MDSD-T5 and MDSD-BERT to extract comprehensive semantic features using prompt learning followed by a post-fusion strategy, achieving promising performance in fine-grained depression detection. Although all of these prior approaches also probed the PLM's knowledge base in pursuit of healthcare outcomes, none focused on cognitive health outcomes specifically.

## 3 Methodology

### 3.1 Task Definition

Given a set of labeled source domain data $D_s$, $N_s$ is the number of samples in the source domain. $D_s$ is thus defined as the set of all paired samples $x_i$ and labels $y_i$ for the source domain: $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$. Correspondingly, given a set of low-resource labeled target data $D_t$ and letting $N_t$ be the number of samples in the target domain, then $D_t$ is the set of all paired samples $x_i$ and labels $y_i$ for the target domain: $D_t = \{(x_i, y_i)\}_{i=1}^{N_t}$.

We domain adaptively train a model jointly on data from the source domain and the low-resource target domain, with the goal of transferring knowledge from source to target. This allows us to empirically validate whether (a) data from different distributions, and (b) domain-adaptive joint training, improve performance in the target classification task. We share the same class labels (in our case, AD and CONTROL) across all domains.

| Generic Instruction | Domain Information |
|---|---|
| (a) `Participant narration on`<br>(b) `This is participant's`<br>(c) `This narrative response was collected from the participant when asked to` | ***DB and ADReSS:***<br>(1) `picture description`<br>(2) `description of a picture`<br>(3) `cookie theft picture description`<br>***CCC:***<br>(1) `health`<br>(2) `experience with health and wellbeing`<br>(3) `health and wellbeing` |

Table 1: Domain-specific prompt (D) template for different domains, composed from both the generic instruction and specific domain information. For example, for the DB and ADReSS domains, one domain-specific prompt may be: `Participant narration on cookie theft picture description`.

### 3.2 Domain Adaptation via Prompt-based Fine-tuning (DAPF)

To implement DAPF, we frame the classification problem as a probabilistic determination of which label token should fill a masked template position. We design a template (recall Figure 1) with three segments: a placeholder for the input text (*T*), a domain-specific prompt (*D*), and a class-specific prompt (*C*). The domain-specific prompt is the novel component, fostering capture of specialized domain information. Domain-specific prompts have two segments: a *generic instruction*, which is transportable across domains, and specific *domain information*, which changes for each domain. Table 1 shows examples of domain-specific templates for several AD detection domains, described in §4.1. The class-specific template is tailored for the target task and consistent across all domains; we use:

- **(x):** `Patient has diagnosis <MASK>.`
- **(y):** `Participant has diagnosis <MASK>.`
- **(z):** `Diagnosis is <MASK>.`

The input texts are concatenated with the domain-specific and class-specific prompts as defined in Figure 1. The positions of *D*, *T*, and *C* with respect to one another may vary. We probabilistically fill `<MASK>` in *C* with tokens from the PLM's vocabulary-bounded answer space, and these are then mapped to our label space (AD or CONTROL). We train the DAPF model jointly on samples from

source domains and the target domain. The model predicts logits representing the answer probabilities to fill <MASK>. These logits are normalized with a softmax layer to compute binary cross-entropy AD detection loss over the logits for the source ($\mathcal{L}_s$) and target ($\mathcal{L}_t$) data label tokens as follows:

$$\mathcal{L}_s = -\frac{1}{N_s}\sum_{i=1}^{N_s} \log P(\hat{y_i^s}, y_i^s) \quad (1)$$

$$\mathcal{L}_t = -\frac{1}{N_t}\sum_{i=1}^{N_t} \log P(\hat{y_i^t}, y_i^t) \quad (2)$$

The model can then be trained in an end-to-end manner with a total loss:

$$\mathcal{L} = \mathcal{L}(D^s) + \mathcal{L}(D^t) \quad (3)$$

Existing domain adaptation methods train the classifier on the source domain to learn a conditional probability distribution $P(y|x^s)$. By aligning the marginal distribution of $P(f(x^s))$ and $P(f(x^t))$ (where $f(.)$ is the text encoding function), they can directly make use of the conditional probability for inference on the target domain. However, when the conditional probability distribution varies ($P(y|x^s) \neq P(y|x^t)$), these methods risk performance degradation (Wang et al., 2020). DAPF differs from these methods because it does not align marginal distributions; instead, it learns two conditional probability distributions $P(y|x^s)$ and $P(y|x^t)$ by learning domain-specific prompts coupled with input from the respective source and target domains. Hence, DAPF can handle both conditional and marginal distribution shift.

**DAPF Ensemble.** We experimented with DAPF models using different domain- and class-specific prompt locations and lengths and combined the three best-performing models via late fusion using majority voting to create a DAPF ensemble model. We also experimented with ensembling different base PLMs, since this has been shown to boost AD detection performance by exploiting the constituent PLMs' complementarity (Wang et al., 2022).

### 3.3 Alternative Techniques

**SwitchPrompt.** To broaden our investigation of prompt learning techniques for domain adaptation, we also adapted SwitchPrompt (Goswami et al., 2023) for AD detection. Unlike DAPF, SwitchPrompt dynamically generates domain-specific keywords based on the input example. It generates soft embedding vectors for those keywords using the PLM, and concatenates the discrete keyword embedding vectors with the soft prompt vectors to retrieve domain-specific knowledge from the PLM. It dynamically switches between general and domain-specific soft prompts using a gating mechanism based on the input instance, while keeping the underlying PLM frozen. This dynamic switching facilitates retrieval of different kinds of input-relevant knowledge from the PLM. SwitchPrompt is jointly trained on the source and target domain data similarly to DAPF, but unlike DAPF, manual prompt design is not required.

**P-Tuning (Baseline).** To compare our proposed DAPF model's performance with a strong baseline, we leveraged P-tuning v2 which employs deep prompt tuning (Li and Liang, 2021; Qin and Eisner, 2021). In contrast to the original P-tuning approach (Lester et al., 2021; Liu et al., 2023), where continuous prompts are only inserted into the input embedding space, P-tuning v2 inserts prompts in different layers of the pretrained language model as prefix tokens. Therefore, P-tuning v2 offers more per-task capacity and more tunable task-specific parameters (increasing from 0.01% to 0.1%-3%) while presenting a parameter-efficient alternative to fine-tuning. Unlike other prompt tuning (Liu et al., 2023) approaches that use a language modeling head to predict masked tokens (Schick and Schütze, 2021b), P-tuning v2 applies a randomly-initialized BERT-like (Devlin et al., 2019) classification head on top of the tokens. Since P-tuning v2 has matched fine-tuning performance in a variety of smaller-scale natural language understanding tasks (Liu et al., 2022b), we included it as a strong baseline for our proposed DAPF model. We also included an ensemble condition, similarly to our DAPF ensemble, for P-tuning.

## 4 Evaluation

### 4.1 Data

We evaluate DAPF using three publicly available datasets, representing separate domains. These datasets are the most widely used datasets for AD detection research in the NLP community, and the only for which public access is available.[2] Charac-

[2]Researchers are still required to obtain permission from the dataset creators prior to using each of these datasets, via established processes that range from email request (Becker et al., 1994) to full review and approval by local and external Institutional Review Boards (Pope and Davis, 2011).

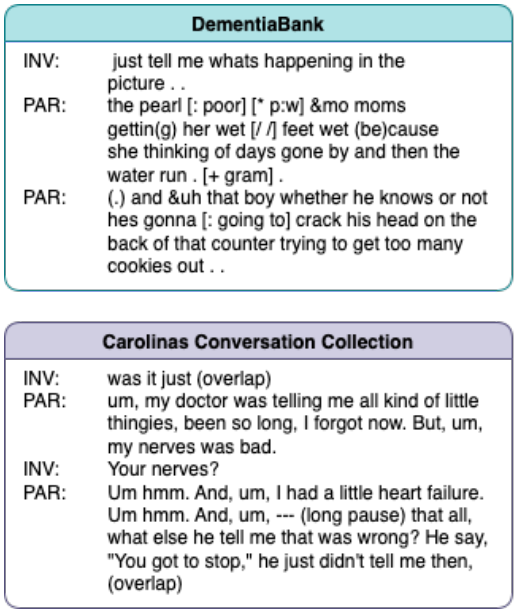



Figure 2: Characteristic language samples from DB (Becker et al., 1994) and CCC (Davis et al., 2017).

| Dataset | | # P | # T | L | SD |
|---|---|---|---|---|---|
| $ADReSS_d$ | tr | 54 | 54 | 125.5 | 81.8 |
| | te | 24 | 24 | 95.0 | 47.0 |
| $ADReSS_c$ | tr | 54 | 54 | 134.7 | 59.4 |
| | te | 24 | 24 | 120.0 | 72.0 |
| $DB_d$ | | 162 | 243 | 124.8 | 67.9 |
| $DB_c$ | | 99 | 303 | 133.9 | 67.4 |
| $CCC_d$ | | 46 | 97 | 1320.7 | 1059.1 |
| $CCC_c$ | | 36 | 192 | 776.9 | 469.7 |

Table 2: Descriptive dataset characteristics. The subscripts *d* and *c* refer to *dementia* and *control*, respectively. *Length* (L) is provided as average number of words per transcript. DB and CCC have differing *# Participants* (P) and *# Transcripts* (T) because some participants in those datasets had multiple recorded interviews. ADReSS is subdivided into standardized *(tr)ain* and *(te)st* partitions established by the dataset's creators.

teristics of these datasets are provided in Table 2. In Figure 2, we provide samples from two of these datasets, quoted directly from Becker et al. (1994) and Davis et al. (2017), to illustrate language differences between task-oriented and conversational Alzheimer's disease detection domains.

**DementiaBank (DB).** DB (Becker et al., 1994) contains audio recordings and manual transcriptions of neuropsychological tests administered to participants with and without AD. The neuropsychological tests include a picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1972), often referred to as the "Cookie Theft Picture Description Task." In this task, participants are presented with a picture stimulus depicting numerous events, including a boy stealing a cookie from a jar. They are asked to describe everything they see occurring in the picture. We use an English-language subset of 546 transcripts from the Cookie Theft Picture Description Task, of which 243 were collected from 162 subjects diagnosed with probable AD and 303 were collected from 99 healthy control subjects.

**Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS).** ADReSS (Luz et al., 2021b) is a subset of DB created for a series of shared tasks on AD detection. Control and AD subjects were age-, gender-, and diagnosis-matched, resulting in a balanced set of 156 samples (78 with AD and 78 control subjects). The goal in developing ADReSS was to eliminate possible biases that may arise due to label and demographic imbalance in the original DB, at the expense of dataset size. It presents an interesting opportunity for comparison of balanced and unbalanced versions of the same source data.[3]

**Carolinas Conversation Collection (CCC).** CCC (Pope and Davis, 2011) contains recorded English conversational interviews of individuals with and without AD, collected by researchers studying language and healthcare across numerous institutions. Members of the control cohort have one interview with a clinical professional and one with a demographically-similar community peer, whereas members of the AD cohort have 1-10 interviews with researchers and student visitors. The goal of these interviews is to elicit autobiographical narrative pertaining to health and wellness, but conversation topics vary considerably. It is less commonly used in the NLP community than DB or ADReSS, although has recently appeared in some studies pertaining to interaction patterns and dementia status (Nasreen et al., 2021) and dementia-related linguistic anomalies in human language (Li et al., 2022). We used the transcribed subset of this corpus, with 97 transcripts from 46 people with AD and 192 transcripts from 36 control subjects.

**Data Preprocessing.** All datasets are in interview format, with interviews including utterances

[3]Since these datasets are drawn from the same source, we do not adapt DB to ADReSS or vice versa in our experiments.

from both a participant and an interviewer. We preprocessed the data to retain only the participant's utterances from the input text. In doing so, we concatenated all participant utterances for a given transcript into a single block rather than feeding each utterance into the our models individually, thereby allowing the model to consider the entire interview context in a manner similar to a real diagnosis scenario. Neither DB nor ADReSS transcripts included explicit disclosures of dementia status; we stripped any explicit disclosures of dementia status in transcripts from the CCC dataset before feeding them to the model to prevent undue influence to downstream classification results. While concatenating the domain-specific and class-specific prompts with the input text from transcripts, we kept the prompt template consistent and truncated additional input tokens beyond the chosen input sequence length. We applied this strategy uniformly across all class labels and domains.

### 4.2 Experimental Settings

By comparing DAPF with P-tuning and SwitchPrompt, we were able to assess performance relative to both continuous prompt tuning and a domain adaptation approach found to be effective in other settings (Goswami et al., 2023). We considered two *source* → *target* adaptations: DB → CCC and CCC → ADReSS. This enabled performance comparisons both when transferring from task-oriented picture description interviews to conversational discourse and vice versa. The second setting had the added benefit of testing on a popular benchmark, allowing direct comparison with external AD detection results. We evaluated each model defined previously (see §3) under each adaptation setting.

We used BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as our base PLMs, both from the HuggingFace library.[4] Given our AD dataset sizes, we focused on BERT and RoBERTa since prompt-based fine-tuning is known to support performance improvements in low-resource settings with smaller PLMs (Schick and Schütze, 2021a). We used the OpenPrompt framework (Ding et al., 2022) to implement and evaluate DAPF. We trained our models with a batch size=16 and maximum sequence length=512 for both BERT and RoBERTa. We used the AdamW optimizer with a ConstantLRwithWarmup[5] learning rate scheduler, a prompt learning rate of $0.5$, and a PLM learning rate of $1e-05$ when fine-tuning the PLM. We performed prompt-based fine-tuning for 10 epochs to update all parameters of the PLMs. For SwitchPrompt, we used a total prompt length of 16 with soft prompt length=6, and 10 dynamically chosen keywords from the target domain. For the P-Tuning baseline, we experimented with different prompt lengths ($\{16, 32, 40, 64\}$) and batch sizes ($\{8, 16, 32\}$). Our reported results were achieved using the best-performing hyperparameter combination identified via grid search.

Experiments with DAPF models were performed on a $V100$ GPU. Each reported result is the average performance across three runs with different random seeds, with each run trained for 10 epochs. When training DAPF in the CCC → ADReSS setting, each epoch took $20.72$ seconds. When cross-validating DAPF in the DB → CCC setting, training each fold took approximately $39.38$ seconds.

## 5 Results

We present our results in Tables 3 and 4. Since we experimented with different prompt lengths and positions for DAPF, we report results for the three top-performing DAPF models (these were also the models included in DAPF Ensemble). We indicate the prompt form in the *Template* column of Tables 3 and 4. Alphanumeric characters in parentheses in the prompt form specify components defined in Table 1. Overall, we observed performance boosts for the BERT and RoBERTa DAPF models over P-tuning in both the DB → CCC and CCC → ADReSS settings. We also observed large performance improvements over SwitchPrompt in BERT-based DAPF; given the extent of the performance difference, we did not study this condition further.

### 5.1 CCC → ADReSS Results

When adapting from CCC → ADReSS (Table 3), we observe that the top-performing BERT-based DAPF model (using a template of the form T+D(a)(3)+C(y), or `[INPUT] + [Participant narration on cookie theft picture description.] + Participant has diagnosis <MASK>.`) outperforms P-Tuning$_L$ (P-tuning with BERT-large-uncased)[6] by $33.20\%$ in accuracy and $34.64\%$ in $F_1$. BERT-based DAPF

[4] https://huggingface.co/models

[5] https://huggingface.co/transformers/v2.9.1/main_classes/optimizer_schedules.html#transformers.get_constant_schedule

[6] P-Tuning$_B$ and P-Tuning$_L$ refer to the base and large versions of the respective PLM used for the P-Tuning model (e.g., BERT base and BERT large).

| PLM | Fine-Tuning | Template | Acc. | $F_1$ |
|---|---|---|---|---|
| BERT | P-Tuning$_B$ | — | 55.21 (0.05) | 45.78 (0.12) |
| | P-Tuning$_L$ | — | 59.38 (0.03) | 58.04 (0.02) |
| | DAPF | T + D(a)(3) + C(y) | 88.89 (0.01) | 88.80 (0.01) |
| | DAPF | D(b)(2) +T + C(y) | 87.50 (0.02) | 87.46 (0.02) |
| | DAPF | D(a)(1) + C(x) + T | 86.11 (0.01) | 86.02 (0.01) |
| | **DAPF Ensemble** | — | **90.28** (0.01) | **89.52** (0.01) |
| | Switch-Prompt | — | 59.03 (0.11) | 48.65 (0.19) |
| RoBERTa | P-Tuning$_B$ | — | 70.83 (0.00) | 70.50 (0.00) |
| | P-Tuning$_L$ | — | 54.17 (0.04) | 45.80 (0.12) |
| | DAPF | T + D(b)(3) + C(y) | **87.50** (0.00) | **87.50** (0.00) |
| | DAPF | T + D(b)(3) + C(x) | **87.50** (0.03) | 87.43 (0.03) |
| | DAPF | T + D(a)(3) + C(x) | 81.94 (0.06) | 81.80 (0.06) |
| | DAPF Ensemble | — | **87.50** (0.03) | **87.50** (0.03) |
| BERT, RoBERTa | P-Tuning | — | 66.67 (0.06) | 61.50 (0.10) |
| | DAPF | — | **89.58** (0.02) | **88.39** (0.02) |

Table 3: Accuracy and $F_1$ (on the transcript level) on the ADReSS test set. CCC and ADReSS (train) are used as source and target data, respectively, when training jointly with 73.46% source data (33.78% AD) and 26.54% target data (50% AD) under the supervised P-tuning and DAPF paradigms. For the DAPF *Template* structure, *D*: domain-specific prompt (content inside the parantheses indicates the generic prompt and domain information from Table 1), *T*: transcript text, and *C*: class-specific prompt.

| PLM | Fine-Tuning | Template | Acc. | $F_1$ |
|---|---|---|---|---|
| BERT | P-Tuning$_B$ | — | 50.00 (0.00) | 34.20 (0.07) |
| | P-Tuning$_L$ | — | 50.63 (0.01) | 35.35 (0.09) |
| | DAPF | D(b)(2) + T + C(y) | 83.19 (0.14) | **81.11** (0.14) |
| | DAPF | D(b)(1) + T + C(y) | 82.45 (0.15) | 79.87 (0.14) |
| | DAPF | D(b)(3) + T + C(y) | 82.06 (0.14) | 80.85 (0.14) |
| | **DAPF Ensemble** | — | **85.28** (0.01) | 78.81 (0.01) |
| | Switch-Prompt | — | 52.01 (0.12) | 51.33 (0.09) |
| RoBERTa | P-Tuning$_B$ | — | 72.57 (0.09) | 68.89 (0.11) |
| | P-Tuning$_L$ | — | 50.56 (0.01) | 35.73 (0.08) |
| | DAPF | T + D(a)(3) + C(x) | 81.18 (0.14) | **79.33** (0.14) |
| | DAPF | T +D(a)(3) + C(z) | 78.94 (0.14) | 76.28 (0.16) |
| | DAPF | D(b)(1) + T + C(y) | 73.28 (0.16) | 66.90 (0.22) |
| | DAPF Ensemble | — | **82.05** (0.02) | 72.55 (0.03) |
| BERT, RoBERTa | P-Tuning | — | 53.18 (0.06) | 34.72 (0.10) |
| | DAPF | — | **85.28** (0.01) | **78.81** (0.01) |

Table 4: Transcript level results when DB is the source dataset and CCC is the target (standard deviation reported in parentheses). Five-fold cross-validation is used in all cases with each fold having 70.3% source data (55.5% AD) and 29.7% target data (33.6% AD). *Template* shows the template structure for DAPF, with *D*: domain-specific prompt, *T*: transcript text, and *C*: class-specific prompt.

achieves an overall accuracy of 88.89% $\pm$ 0.01 and $F_1$=88.80 $\pm$ 0.01. DAPF Ensemble outperforms P-Tuning$_L$ in accuracy and $F_1$ by an even larger margin of 34.23% in accuracy and 35.17% in $F_1$, for an overall accuracy of 90.28% $\pm$ 0.01 (best accuracy=0.9167) and $F_1$=89.52 $\pm$ 0.01 (best $F_1$=0.9130). This approaches the performance of the current state-of-the-art model on the ADReSS test set, which achieves an accuracy of 93.75% by first automatically transcribing input recordings and then passing them to a pipelined classification architecture with a masked language modeling fine-tuning objective (Wang et al., 2022).

Experimental results using RoBERTa-base models also outperformed the P-tuning baselines. In this case, the RoBERTa DAPF model (using a template of the form T+D(b)(3)+C(y), or `[INPUT] + [This is participant's cookie theft picture description.] + Participant has diagnosis <MASK>.`) achieved the highest performance, outperforming RoBERTa-base P-tuning (P-Tuning$_B$) by 19.05% and 19.03% in accuracy and $F_1$, respectively. We reiterate that in general, the adaptation of the SwitchPrompt model for AD detection did not prove to work well;[7] we see that it was outperformed by P-Tuning$_L$ and all DAPF models by a great margin. The BERT + RoBERTa DAPF Ensemble outperformed the respective P-Tuning (BERT-large P-Tuning$_L$ + RoBERTa-base

[7]We speculate that it may be because it does not involve fine-tuning; perhaps the model struggles since its weights are not updated to attend to respective segments of the input.

P-Tuning$_B$) ensemble, and underperformed the BERT-only DAPF Ensemble, likely due to the performance discrepancies between the BERT- and RoBERTa-based DAPF models.

### 5.2 DB → CCC Results

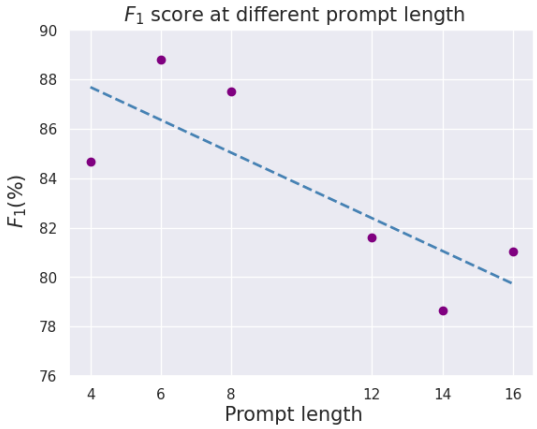
When adapting from DB → CCC (Table 4), the top-performing BERT-based DAPF model (using template D(b)(2)+T+C(y), or `[This is participant's experience with health and wellbeing] + [INPUT] + [Participant has diagnosis <MASK>]`) outperforms P-Tuning$_L$ by 39.14% in accuracy and 56.42% in F$_1$.[8] Our BERT-based DAPF Ensemble outperforms P-Tuning$_L$ in terms of accuracy by even larger margins of 40.63% and F$_1$ by 55.15%. In line with observations of the BERT-based DAPF model, the RoBERTa-based DAPF Ensemble also outperformed RoBERTa-base P-Tuning$_B$ in terms of both accuracy and F$_1$ by 11.55% and 5.04%, respectively. The adaptation of SwitchPrompt for AD detection again did not prove to work well in this setting, but it did outperform the P-tuning models in accuracy and F$_1$ by a very small margin. We find that the BERT + RoBERTa DAPF Ensemble outperforms the the respective P-tuning (BERT-large P-Tuning$_L$ + RoBERTa-base P-Tuning$_B$) ensemble by a large margin.

### 5.3 Follow-Up Analyses

**Word Choice in Prompts.** We used the terms "diagnosis" and "patient" with careful consideration in our class-specific prompt templates. Our use of "diagnosis" was motivated by an interest in nudging the model to consider the health angle of the given context, since our underlying goal was to classify the cognitive health status of the subject. Our use of "patient," applied consistently across both AD and control classes for the source and target datasets, was driven by an interest in learning how probing with more specific to general (patient versus participant) terms impacted the final classification. Since we experimented with different combinations of the class-specific prompt in different positions, we were able to observe the performance of many more prompt templates than could fit in Tables 3 and 4 (we included the top-performing prompt templates for the DAPF models in those tables). Overall, we found that models using "patient" in the class-specific prompt (c(x)) performed favorably compared to other DAPF models without significant differences.

[8]There is currently no standardized state-of-the-art benchmark for the CCC dataset.


(a) Few-Shot Learning




(b) Prompt Length

Figure 3: Analyzing performance with varied few-shot learning and prompt length conditions for the $CCC \rightarrow ADReSS$ setting.

**Few Shot Learning.** In addition to our primary experiments, we investigated how many target domain data points were needed during training for the model to generalize with reasonable accuracy. We studied performance with $n$-shot (with $n \in \{2, 4, 8, 16, 32\}$) target domain data from each class while training with the source domain data for CCC → ADReSS, and present the results in Figure 3a. We observe that although in extremely low-resource few-shot settings ($n \in \{2, 4, 8\}$), the BERT-based DAPF model's output fluctuates, beyond that point with $n \in \{16, 32\}$ the accuracy is quite comparable to that observed in the full-data training settings reported in Table 3.

**Prompt Length.** We also experimented with the length of the domain-specific prompts (Table 1), since prior research has suggested that prompt

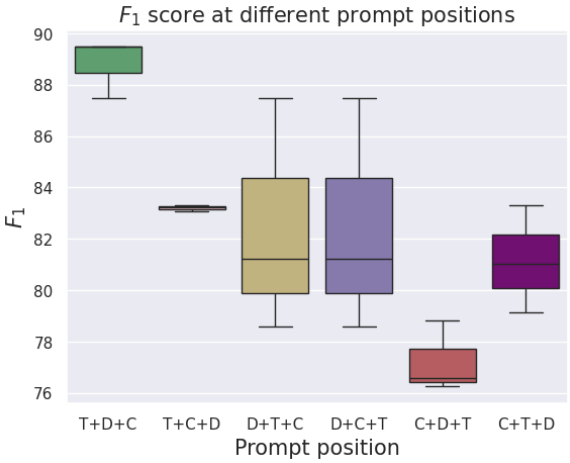



Figure 4: Analysis on varied prompt position for the $CCC \rightarrow ADReSS$ setting.

length may affect the performance of prompt tuning approaches (Li and Liang, 2021). Specifically, we varied domain-specific prompt length for the high-performing prompt order *T+D+C* for the CCC → ADReSS setting. We observed that $F_1$ generally declines when the domain-specific prompt length extends beyond length=8 (see Figure 3b).

**Prompt Position.** Finally, prior research has suggested that prompt phrase location may influence what information is captured, potentially affecting task performance (Wang et al., 2023b). We experimented with different combinations of prompt phrase locations (recall Figure 1 for two examples) for domain- and class-specific prompts with respect to the input text. We conducted this study for the CCC → ADReSS setting with our top performing BERT-based DAPF model (Table 3) and and its five other prompt position variations. Each model was run with three different random seeds. We show the $F_1$ across all six model variations in Figure 4.

We observe that our top-performing model has the class-specific context (*C*) in the last position. The lowest-performing models with prompt order *C+D+T* and *C+T+D* are significantly outperformed by the top-performing model (*T+D+C*). To further support this finding, in Tables 3 and 4 we also observe that our top-performing DAPF models have the class-specific context (*C*) next to the domain-specific prompt (*D*) or input text (*T*) with both the BERT and RoBERTa settings. It is intuitive and logical to have the domain-specific prompt and the input text before *C*, as it places the context in sequence such that it is available prior to the masked token that must be predicted.

## 6 Conclusion

In this paper we systematically and comprehensively investigated the use of domain adaptation via prompt learning for AD detection. We proposed a novel prompt learning paradigm, **D**omain **A**daptation via **P**rompt-based **F**ine-tuning (DAPF), for this purpose. We showed that DAPF yields similar performance to the current state-of-the-art (Wang et al., 2022) on the ADReSS test set when using a BERT base PLM and ensembling across three top-performing discrete prompt templates of different forms. We compare DAPF across multiple domain adaptation settings, base PLMs, and prompt template structures relative to strong baselines including P-tuning and SwitchPrompt, a competitive contemporary prompt-based domain adaptation approach. In doing so, we also study the interplay between prompt length, prompt template order, and overall performance.

In follow-up analyses, we find that DAPF also performs competitively in few-shot settings, achieving performance comparable to full-data training settings when using only 16 target domain training samples. Given the widespread data limitations in healthcare tasks such as AD detection, this further supports the utility and anticipated appeal of DAPF as a novel training paradigm. We make all source code and models publicly available[9] to encourage further experimentation by others.

## 7 Limitations

Our work is limited by several factors. First, we conduct our work primarily using popular, publicly available AD detection datasets, all of which are in English. Thus, it is unclear whether our findings generalize to other languages, especially since there are also fewer PLMs available for non-English use. Second, we only experiment with two backbone PLMs (BERT and RoBERTa). These models are well-suited to cloze-style prompting and are popular across a broad range of classification tasks, including AD detection; however, it may be the case that other PLMs yield different results. Finally, as is always the case with manual prompt design it is possible that our constructed prompt templates are suboptimal. We experimented with many subtle prompt variations during our initial design process; nonetheless, we recognize that we may have missed better-performing alternatives.

[9] https://github.com/treena908/DAPF

Collectively, these limitations present intriguing avenues for follow-up work.

## 8 Ethical Considerations

This research was guided by a broad range of ethical considerations, taking into account factors associated with fairness, privacy, and intended use. Although many of these are described throughout the paper, we summarize those that we consider most critical in this section.

**Data Privacy and Fairness.** This research was approved by the Institutional Review Board at our institution. Access was granted for all datasets used in this research, and our use is governed by approved protocols unique to each dataset. DementiaBank, ADReSS, and the Carolina Conversations Collection are all publicly available following access request protocols specified by their governing organizations. We refer readers to the citations throughout this work if they are interested in obtaining access to this data. We are unable to share it directly, although we can share our processing scripts and other code to facilitate reproducibility of our work by others.

**Intended Use.** Automated models for AD detection from spoken language present potential benefits in real-world scenarios: they offer opportunity to expand healthcare access, minimize cost of care, and reduce caregiver burden. However, they may also pose risks if used in unintended ways. We consider intended use of the work reported here to extend to the following:

- People may use the technology developed in this work to study language differences between individuals with and without AD, as a way of building further understanding of the condition.
- People may use the technology developed in this work to further their own research into low-resource NLP tasks, including those associated with this and other healthcare problems.
- People may use the technology developed in this work to build early warning systems to flag individuals about potential AD symptoms, *provided that the technology is not misconstrued as an alternative to human care* in any way.

Any use outside of those listed above is considered an unintended use. To safeguard against unintended use of our work, we remind readers that dataset access must be granted through the approved channels by the creators of the respective datasets used in this work. This may include processes ranging from email request to full review and approval by local and external Institutional Review Boards. We reiterate our caution against using any findings from this paper to build systems that function as intended or perceived replacements for human medical care.

## Acknowledgements

We thank the anonymous reviewers for their helpful feedback, which was incorporated in the final version of this manuscript. This work was partially supported by the National Science Foundation under Grant No. 2125411. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2022. Position-based prompting for health outcome generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 26–36, Dublin, Ireland. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alzheimer's Association. 2023. 2023 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection. In *Proc. Interspeech 2020*, pages 2167–2171.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the*

*Association for Computational Linguistics*, 10:414–433.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 2060–2066. AAAI Press.

BH Davis, C Pope, K Van Ravenstein, and W Dou. 2017. Three approaches to understanding verbal cues from older adults with diabetes. *The Internet Journal of Advanced Nursing Practice*, 16(1).

Sofia de la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

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.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, Dublin, Ireland. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2022. Are interaction patterns helpful for task-agnostic dementia detection? an empirical exploration. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–182, Edinburgh, UK. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2023. Towards domain-agnostic and domain-adaptive dementia detection from spoken language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11965–11978, Toronto, Canada. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. 2023. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11.

Harold Goodglass and Edith Kaplan. 1972. *The assessment of aphasia and related disorders*. Lea & Febiger.

Koustava Goswami, Lukas Lange, Jun Araki, and Heike Adel. 2023. SwitchPrompt: Learning domain-specific gated soft prompts for classification in low-resource domains. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2689–2695, Dubrovnik, Croatia. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

M Khojaste-Sarakhsi, Seyedhamidreza Shahabi Haghighi, SMT Fatemi Ghomi, and Elena Marchiori. 2022. Deep learning for alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, 130:102332.

Miia Kivipelto, Francesca Mangialasche, and Tiia Ngandu. 2017. Can lifestyle changes prevent cognitive impairment? *The Lancet. Neurology*, 16.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. 2021. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6423–6427.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022a. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. *ArXiv*, abs/1502.02791.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. 2018. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1647–1657, Red Hook, NY, USA. Curran Associates Inc.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2208–2217. JMLR.org.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021a. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Proc. Interspeech 2021*, pages 3780–3784.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021b. Detecting cognitive decline using speech only: The adresso challenge. *medRxiv*.

Matej Martinc and Senja Pollak. 2020. Tackling the adress challenge: A multimodal approach to the automated recognition of alzheimer's dementia. In *Interspeech*.

Shamila Nasreen, Julian Hough, and Matthew Purver. 2021. Rare-class dialogue act tagging for Alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–300, Singapore and Online. Association for Computational Linguistics.

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Morteza Rohanian, Julian Hough, and Matt Purver. 2021. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *ArXiv*, abs/2106.15684.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Kamlesh Sharma. 2019. Cholinesterase inhibitors as alzheimer's therapeutics (review). *Molecular Medicine Reports*, 20.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Sonish Sivarajkumar and Yanshan Wang. 2022. Healthprompt: a zero-shot learning paradigm for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2022, page 972. American Medical Informatics Association.

Sonish Sivarajkumar and Yanshan Wang. 2023. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2022:972–981.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2021. On transferability of prompt tuning for natural language processing. In *North American Chapter of the Association for Computational Linguistics*.

Baochen Sun and Kate Saenko. 2016. *Deep CORAL: Correlation Alignment for Deep Domain Adaptation*, pages 443–450.

Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. 2021. Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling. *IEEE Access*, 9:88377–88390.

Niall Taylor, Yi Zhang, Dan W. Joyce, Ziming Gao, Andrey Kormilitzin, and Alejo Nevado-Holgado. 2023. Clinical prompt learning with frozen language models. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11.

Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, Chenxi Yue, Haiyang Zhang, Yi-Hsueh Liu, Xiang Li, Bao Ge, Dajiang Zhu, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. 2023a. Prompt engineering for healthcare: Methodologies and applications. *ArXiv*, abs/2304.14670.

Jindong Wang, Yiqiang Chen, Wenjie Feng, Han Yu, Meiyu Huang, and Qiang Yang. 2020. Transfer learning with dynamic distribution adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(1).

Yi Wang, Jiajun Deng, Tianzi Wang, Bo Zheng, Shoukang Hu, Xunying Liu, and Helen Meng. 2023b. Exploiting prompt learning with pre-trained language models for alzheimer's disease detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi Wang, Tianzi Wang, Zi Ye, Lingwei Meng, Shoukang Hu, Xixin Wu, Xunying Liu, and Helen M. Meng. 2022. Exploring linguistic feature and model combination for speech recognition based automatic ad detection. In *Interspeech*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, and Helen Meng. 2021. Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6433–6437.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimerâ€™s disease. pages 2162–2166.

Jun Zhang and Yanrong Guo. 2024. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recognition Letters*, 178:167–173.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *ArXiv*, abs/2108.13161.

Wenbo Zhao, Arpit Gupta, Tagyoung Chung, and Jing Huang. 2023. SPC: Soft prompt construction for cross domain generalization. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 118–130, Toronto, Canada. Association for Computational Linguistics.